Zeile wird in eine eigene Lochkarte übertragen. Diese Fragekarten und eine auf Magnetband oder -platte verzeichnete Liste von Formelnummern, die die Vorselektion passiert haben, werden der Maschine mit dem Rechercheprogramm eingegeben und von ihr mit dem topologischen Speicherband verglichen. Dieses enthält – um Einlesezeit zu sparen – die Hauptmatrizen in stark gekürzter Form: Zwar sind die redundanten Angaben noch explizit vorhanden, aber die weitaus meisten Nullelemente der Matrizen sind durch einfache Kunstgriffe unterdrückt; so wird nur wenig Rechenzeit für das Wiederaufblähen im Kernspeicher benötigt.

Als Antworten druckt der Computer entweder die Literaturhinweise oder die Nummern von Referaten aus, die wir gegebenenfalls kopieren und dem Fragesteller zuschicken. Auf Wunsch kann aus der Hauptmatrix maschinell auch die Strukturformel rekonstruiert und ausgedruckt werden. Das geschieht beispielsweise schon während der Einspeicherung zu Kontrollzwecken; allerdings liefern die gebräuchlichen Schnelldrucker kein sehr schönes Formelbild.

4. Schlußbemerkung

Die Computerprogramme für die hier beschriebene Recherche sind für das IBM-System 360 (OS) geschrieben und ausgetestet[12] – mit Ausnahme der erwähnten vier Fragebedingungen. Auch diese werden noch einprogrammiert: Mit „Komplex" sollen koordinative Bindungen und auch wichtige Wasserstoffbrücken abgefragt werden; durch „Reaktion" wird im Rahmen einer Reaktionenfolge ein Atom angerufen, das reagiert hat oder reagieren wird; „Stereo" soll nach einem von *Petrarca, Lynch* und *Rush*[13] vorgeschlagenen System prüfen, ob ein Atom die gewünschte Chiralität oder *cis-/trans*-Stellung aufweist. (Voraussetzung für die praktische Anwendung dieses Verfahrens ist freilich eine wirtschaftliche Einspeicherungsmethode; Programme, die diese mit Hilfe unserer Formellesemaschine erreichen sollten, wurden für die IBM 7090 schon geschrieben, konnten aber noch nicht auf das System 360 umgestellt werden.) Mit „Isotop" sollen schließlich Verbindungsklassen auffindbar gemacht werden, die an bestimmter Stelle markiert sind. Wir hoffen, mit diesem Dokumentationssystem besonders dem präparativ arbeitenden Chemiker ein Werkzeug in die Hand gegeben zu haben, mit dem er nicht nur seine Fragen an den schon erarbeiteten chemischen Wissensschatz schneller und zielsicherer als bisher beantwortet bekommt; wir glauben vielmehr, daß gerade die Möglichkeit zu einer großen Zahl neuartiger Fragestellungen die chemische Forschung befruchten und anregen kann, wenn der Chemiker diese Möglichkeiten erkennt und nützt. Dies könnte seinen Arbeitsstil noch rationeller gestalten, und obendrein würde die Gefahr, längst Bekanntes unfreiwillig nachzuarbeiten, sicher gemindert.

Eingegangen am 11. September 1969 [A 766]

[12] Die topologischen Einspeicherungs- und Rechenprogramme (IBM System /360) schrieb Herr *Peter Schilling*; er steuerte für die Details viele interessante Ideen bei.

[13] *A. E. Petrarca, M. F. Lynch* u. *J. E. Rush*, J. chem. Documentation 7, 154 (1967).

Tosar – ein topologisches Verfahren zur Wiedergabe von synthetischen und analytischen Relationen von Begriffen

Von Robert Fugmann, Herbert Nickelsen, Ingeborg Nickelsen und Jakob H. Winter[*]

Bei mechanisierten Suchsystemen in Literaturspeichern ist ein ständig wachsendes Bedürfnis zu verzeichnen, nicht nur die Fachbegriffe selbst als Suchbedingung formulieren zu können, sondern auch die charakteristischen Verknüpfungen, unter denen diese Fachbegriffe in der Fragestellung erscheinen. Auf diese Weise läßt sich die Treffsicherheit der mechanisierten Literatursuche erheblich steigern. Das System Tosar wurde entwickelt, um die Literatursuche speziell mit elektronischen Rechenanlagen in dieser Hinsicht zu vervollkommnen.

1. Einführung

Mit dem ständigen Anwachsen der chemischen Fachliteratur gewinnen alle Verfahren an Bedeutung, welche es ermöglichen, einschlägige Publikationen zu einer wissenschaftlichen oder technischen Fragestellung wiederzufinden. Die herkömmlichen Registerwerke sind hierfür nur so lange eine gute Hilfe, wie die Fragestellung auf nur einen einzigen Begriff abzielt und nur sofern für dieses Thema ein Schlagwort im Register existiert. Je stärker aber die Spezialisierung in Wissenschaft und Technik sich entwickelt, desto weniger läßt sich ein Fragethema mit nur einem einzigen Begriff, z. B. mit einer Strukturformel, beschreiben. Eine Fragestellung nach Terephthalsäurederivaten oder nach Propylen allein würde heute derartig viel Literatur als Antwort ermitteln, daß sie von niemandem mehr überblickt werden könnte, es sei denn,

[*] Dr. R. Fugmann, Dr. H. Nickelsen, Dr. I. Nickelsen und Dr. J. H. Winter
Farbwerke Hoechst AG
623 Frankfurt/Main-Höchst und
IDC-Internationale Dokumentationsgesellschaft für Chemie mbH
6 Frankfurt/Main

er beabsichtigt, hierüber eine Monographie zu schreiben und hierfür mehrere Monate Arbeitszeit aufzuwenden, einschließlich der Literaturstudien.

Was den Fragesteller im Regelfall interessieren dürfte, sind vielleicht Verfahren zur Oligomerisation von Propylen zu Hexenen und höheren Olefinen, mit einem bestimmten Organoaluminium-Katalysator, unter Wiederverwendung des nicht umgesetzten Propylens zum kontinuierlichen Austreiben von laufend entstehenden Verunreinigungen aus dem Katalysator, wobei dieser fortgesetzt reaktiviert wird. Das Thema der Fragestellung ist also in Wirklichkeit viel spezifischer als es diesem einen Begriff „Propylen" allein entspricht. Dieser Frage-Begriff ist mit einer ganzen Reihe anderer Begriffe „koordiniert", und nur diejenigen Publikationen sind für den Fragesteller von Interesse, bei welchen auch die anderen Begriffe der Fragestellung vorkommen, und zwar im gewünschten Zusammenhang.

Wollte man nun auch für derartig hochspezifische Fragestellungen, von denen vorstehend ein Beispiel genannt wurde, treffende Schlagwörter im Register vorsehen, um die einschlägige Literatur schärfer lokalisieren zu können, so müßten diese Schlagwörter die Form ganzer Sätze annehmen. Hierfür würden viele Milliarden derartiger hochgradig „zusammengesetzter" Schlagwörter benötigt. Aber niemand wäre als Dokumentar in der Lage, ein Schlagwortvokabular von dieser Größe bei der Aufbereitung der Literatur verläßlich zu memorieren, so daß er jeder Publikation das jeweils treffende Schlagwort zuteilen könnte oder deren mehrere. Die verwandtschaftlichen Beziehungen zwischen diesen Schlagwörtern wären nämlich nicht mehr zu überblicken. Gerade dies aber ist eine unabdingbare Voraussetzung für die Wahl des treffendsten Schlagwortes aus einer Gruppe von annähernd zutreffenden. Außerdem wären die hohen Kosten für die Herstellung solcher Register prohibitiv, ebenso wie die unüberwindlichen Schwierigkeiten bei ihrer technischen Handhabung. Schon bei einem viel kleineren Vokabular von derartigen vorgefertigten „präkoordinierten" Schlagwörtern, wie sie etwa bei den herkömmlichen Sachregistern benutzt werden, läßt die Vollständigkeit der Literaturzitate unter einem Schlagwort sehr zu wünschen übrig. Dies ist die zwangsläufige Folge mangelnden Überblickes über das Vokabular beim Einspeichern und der begrenzten Möglichkeiten in den herkömmlichen, streng eindimensional wirksamen Registerwerken. Von einer weiteren Vergrößerung solcher Schlagwortvokabularien, um stärker spezifischen Fragestellungen gewachsen zu sein, kann man sich also keinen Vorteil versprechen.

Die Arbeit mit einem derartig hochgradig präkoordinierten Vokabular von Schlagwörtern würde aber auch noch aus einem anderen Grunde scheitern: Häufig sind nämlich die Aussagen eines Autors derartig allgemein und schließen derartig viele Einzelfälle ein, daß die ausführliche Wiedergabe eines jeden Einzelfalles mit einem Einzelschlagwort einen untragbar großen Arbeitsaufwand bedeuten würde. Beispielsweise kann in einem Oligomerisationsprozeß als Ausgangsstoff entweder Äthylen oder Propylen oder 1-Buten oder 2-Buten oder Isobutylen oder 1,3-Butadien oder ein bestimmtes Penten genannt sein. Für all diese Prozesse müßten getrennte Schlagwörter gewählt werden, um zu verhindern, daß die gleichzeitige Teilnahme mehrerer dieser Stoffe an der Reaktion vorgetäuscht wird. Wenn in einem Copolymerisationsprozeß als monomere Ausgangsstoffe A und (B oder C oder (D und E)) und F und/oder G genannt sind, so muß man jede Hoffnung aufgeben, derartig komplizierte logische Begriffsverknüpfungen mit einem einzigen präkoordinierten Schlagwort darstellen zu können.

Diese Schwierigkeiten lassen sich nun nicht dadurch überwinden — wie zuweilen angenommen wird — daß man den vollen Originaltext einer Publikation in den Speicher einer Rechenanlage eingibt und gleichzeitig die Möglichkeit schafft, jedes Wort in einem solchen Text oder überhaupt jede beliebige Zeichenfolge wiederzufinden, z. B. „Propylen" oder „Oligomerisation". Zwar mag es bei oberflächlicher Betrachtung erscheinen, als ob hierdurch alle Voraussetzungen für eine verläßliche Recherche gegeben sind, denn scheinbar ist alle gebotene Information in abrufbarer Form in den Speicher eingegangen. Aber die Maschine braucht für den Suchvorgang einen genauen Auftrag, welche besonderen Zeichenfolgen und Wörter für den Fragesteller in den eingespeicherten Texten gesucht werden sollen, und der Trugschluß liegt darin zu glauben, daß es leicht sei, ja sogar vom eigenen Literaturstudium her geläufig sei, einen solchen Suchauftrag zu formulieren. Es ist aber etwas Grundverschiedenes, ob man aus einem vorgelegten Text die gemeinte Bedeutung erkennen will, wie es bei der sachverständigen Lektüre durch den Menschen stets geschieht — oder ob man umgekehrt (!) vorauszusehen gezwungen ist, in welchen Wörtern ein Autor das interessierende Thema beschrieben haben könnte. Im ersten Falle handelt es sich gleichsam um ein aposteriorisches, im zweiten Falle aber um ein apriorisches Urteil zur gleichen Sachlage.

Derartige Voraussagen sind überhaupt nur dann verläßlich möglich, wenn eine Gesetzmäßigkeit bekannt ist, nach welcher sich das betreffende Ereignis einstellt. Es ist aber bis heute noch keine Gesetzmäßigkeit bekannt geworden, nach welcher ein Autor in jedem Einzelfall seine Wörter zur Beschreibung eines bestimmten Themas wählt. So muß man sich aufs Raten und Probieren verlegen, wenn man durch die maschinelle Recherche im Originaltext etwa nach Propylen suchen will („C_3H_6"?, „Propen"?, „C_2- bis C_4-Olefine"?, „olefinische Crackgase im Siedebereich von 50—60 °C bei 20—22 atm"?, „... die gasförmigen Reaktionsprodukte aus der Isopropanol-Dehydratisierung"?, usw.). Und wie soll dem Umstand Rechnung getragen werden, daß möglicherweise überhaupt nur die Strukturformel gezeichnet ist? Sicherlich kann man auch auf diesem Wege gelegentlich zu glücklichen Funden kommen. Von einem verläßlichen Verfahren kann indessen keine Rede sein.

Ein erster Schritt zur Lösung dieses Problems besteht darin, daß man zunächst bei der Literaturauswertung die vom Autor genannten oder angedeuteten Begriffe in eine Form bringt, die für die spätere Recherche *voraussehbar* ist. Dies geschieht dadurch, daß man entweder bei der Einspeicherung laufend Buch führt, welche Bezeichnungen schon für einen bestimmten Begriff vorgekommen sind, oder daß man gleich alle wesentlichen Begriffe in ein festgelegtes, genormtes Vokabular übersetzt. Auch verwendet man nicht ein hochgradig zusammengesetztes präkoordiniertes Schlagwort für die angetroffenen Begriffe und Begriffsverknüpfungen (etwa „kontinuierliche hexen- und nonenbildende Propylenoligomerisation mit fortlau-

fender Trialkylaluminiumkatalysator-Regenerierung" usw.), sondern stattdessen eine Anzahl von einfacheren Schlagwörtern, die gleichsam nur als begriffliche Bausteine dienen. Einer Patentschrift werden also beispielsweise die Schlagwörter „Cracken", „Dimerisierung", „fraktionierende Destillation", „Hexen", „Isopren", „Katalysator", „kontinuierliches Verfahren", „Propylen", „Nonen", „Regenerierung", „Trialkylaluminium", „Trimerisierung", „Wiedergewinnung" zugeteilt. Bei einer Fragestellung der oben genannten Art wird das hochgradig zusammengesetzte Schlagwort, unter dem man die einschlägige Literatur zusammengestellt wünscht, immer erst zum Zeitpunkt der Fragestellung gebildet, d. h. durch „Postkoordination". Eine besondere, vieldimensionale Suchtechnik muß es dann leisten, daß in einem Literaturspeicher nur diejenigen Publikationen gefunden werden, in denen sämtliche Schlagwörter der Fragestellung in Koordination vorkommen. Ein Vokabular dieser einfachen Schlagwörter kann viel kleiner und übersichtlicher gehalten werden als ein entsprechendes Vokabular für alle vorweggenommenen, präkoordinierten Kombinationen, die sich aus diesen Bausteinen aufbauen lassen. Nach diesem Prinzip arbeiten heute alle mechanisierten Dokumentationsmethoden wie Handlochkarten- und Maschinenlochkarten-Systeme, Magnetbandspeicher usw.

Führt man nun nach diesem einfachen Prinzip mechanisierte Literaturrecherchen aus, so kann das Ergebnis trotzdem noch enttäuschend sein. Zwar erfüllen sämtliche lokalisierten Literaturstellen mit Sicherheit die Bedingung, daß in ihnen alle Einfachbegriffe der Fragestellung genannt sind, aber sehr oft werden sie dort in gänzlich anderen Zusammenhängen vorkommen, als es dem Thema des Fragestellers entspricht. Beispielsweise kann sich in einer Publikation die Oligomerisation auf das Hexen beziehen, die Crackreaktion hat zum Propylen geführt, und die Wiedergewinnung hat sich auf überschüssiges Hexen bezogen, und nicht etwa auf Propylen. Je größer eine Literatursammlung auf einem Gebiet ist, desto störender muß sich zwangsläufig dieser Ballast auswirken. Wenn beispielsweise unter tausend maschinell ermittelten Literaturstellen sich nur zehn bis zwanzig zutreffende befinden, welche in der Menge der unzutreffenden erst mühsam herausgesucht werden müssen, so ist die Brauchbarkeit eines solchen Dokumentationssystems in Frage gestellt.

Es hat also auf längere Sicht nicht sein Bewenden damit, daß man hochgradig zusammengesetzte Begriffe aus gutem Grund in ihre begrifflichen Bestandteile zerlegt und diese lediglich aufzählt. Es muß vielmehr auch dargestellt werden, nach welchem Bauplan diese Einfachbegriffe im Text der Originalpublikation zusammengefügt sind. Erst hierdurch wird der Sinn der vom Autor gemachten Aussage treffend wiedergegeben, und erst hierdurch läßt sich ein solcher Sinnzusammenhang als Suchbedingung für die maschinelle Literaturrecherche formulieren.

Die Dokumentationsliteratur ist reich an Vorschlägen und an Erfahrungsberichten zur Wiedergabe von Begriffsverknüpfungen in Einspeicherung und Anfrage. Einen Überblick über den Stand der Forschung auf diesem Gebiet vermittelte ein Symposium, das speziell diesem Thema gewidmet war [1]. Relativ weite Verbreitung hat beispielsweise ein Verfahren gefunden, bei welchem man einen jeden Begriff mit einem entsprechenden syntaktischen Merkmal kennzeichnet (Role-Indikator, Funktor, Relator). Hierdurch wird mit mehr oder minder großer Deutlichkeit wiedergegeben, welche Funktion ein Begriff im Bezug zu den begleitenden Begriffen in einer Publikation ausübt. „Ausgangsmaterial", „Reaktionsprodukt", „isolierter, von Verunreinigungen abgetrennter Stoff", „Hilfsstoff bei Synthesen" usw. sind Beispiele dafür, auf welche Art häufig die Funktion eines Stoffbegriffs angedeutet wird. Daß hierbei nicht zur Geltung kommt, für *welche* Synthese (von mehreren beschriebenen) der Stoff als Ausgangs- oder Hilfsstoff dient, von *welchen* Verunreinigungen ein anderer Stoff befreit worden ist usw., sind Ungenauigkeiten, die sich bei der mechanisierten Recherche zwangsläufig in einem entsprechenden Restanteil von Ballast auswirken.

Bei einem anderen Verfahren betrachtet man diejenigen Begriffe als besonders eng zusammengehörig, welche im gleichen Satz eines Textes oder im gleichen Abschnitt vorkommen. Alle Begriffe eines solchen formalen Bereiches bezeichnet man mit einem gemeinsamen Zeichen („link"), durch welches sie unterscheidbar werden von den Begriffen eines anderen Bereiches. Da sich aber die Sinnzusammenhänge in einem Text häufig nicht mit diesen formalen Grenzen decken, kommt es auch hier zu Ungenauigkeiten und Verfälschungen.

Auch graphentheoretische Methoden sind schon angewendet worden, um Begriffsverknüpfungen maschinell zu handhaben [2]. Jedoch sind diese Graphen stets rein maschinenintern aufgebaut und nicht etwa intellektuell durch einen Fachmann bei Anblick einer zu speichernden Publikation konstruiert worden.

Wir beschreiben im folgenden ein neues graphisches Modell zur Darstellung der erörterten Begriffsverknüpfungen, für das wir die Kurzbezeichnung „Tosar" gewählt haben [*]. Bei diesem Verfahren wird vom Dokumentar ein Graph gezeichnet, in welchem die Begriffsverknüpfungen bei einzuspeichernden Dokumenten oder bei Fragestellungen wiedergegeben sind. Dadurch, daß beim Aufbau eines solchen Graphen gewisse Prinzipien eingehalten werden müssen, nehmen die Begriffsverknüpfungen eine für die spätere Recherche voraussehbare Form an. Alle Begriffe und Begriffsverknüpfungen in diesem Graphen werden codiert und in den Speicher einer programmgesteuerten elektronischen Rechenanlage übernommen. Hiernach sind mit Hilfe entsprechender Maschinenprogramme diese Graphen unter allen einschlägigen Fragestellungen wieder auffindbar.

---

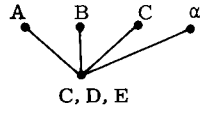
[1] *J. M. Perreault*, Proceedings of the International Symposium on Relational Factors in Classification. Information Storage and Retrieval *3*, 177 (1967).

[2] *G. Salton*, Commun. Amer. Assoc. Computing Machinery *1962*, 103.

[*] *T*opologische Wiedergabe von *s*ynthetischen und *a*nalytischen *R*elationen von Begriffen.

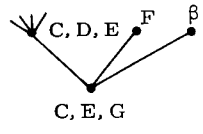
2. Prinzipien des graphischen Modells

Ein Graph ist ein abstraktes System aus Strecken und Punkten, dessen mathematische Gesetzmäßigkeiten durch die Graphentheorie beschrieben werden, ein Teilgebiet der Topologie. Den Punkten ordnen wir Begriffe zu, deren Verknüpfungen mit anderen Begriffen dargestellt werden sollen. Die Begriffsverknüpfungen selbst treten in üblicher Weise als Verbindungsstrecken zwischen den Punkten in Erscheinung. Jeder experimentell vollzogene (oder vollzogen gedachte) Vorgang wird durch eine Stufenfolge von Niveaus dargestellt. Die wesentlichen Begriffe *vor* Ablauf des betreffenden Vorganges (α, β, γ, δ usw.) werden in einem eigenen Niveau angeordnet. Die Ergebnisbegriffe *nach* Ablauf des Vorganges stehen an einem Punkt vereint auf einem tiefer gelegenen Niveau. Alle Eingangsbegriffs-Punkte sind mit dem Ergebnis-Punkt durch eingezeichnete Strecken verbunden:

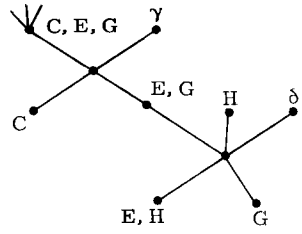


„Die Stoffe A, B und C werden der Reaktion α unterworfen, wobei die Stoffe D und E neu entstehen und der Stoff C (etwa ein Lösungsmittel) noch vorhanden ist". – Daß es sich bei den Begriffen A, B, C um Stoffbegriffe, bei α um einen Reaktionenbegriff handelt, wird in einer hier nicht näher zu erörternden Weise ebenfalls wiedergegeben.

Das Stoffgemisch C, D, E kann einer weiteren Reaktion β unter Zusatz des Stoffes F unterworfen worden sein, wobei aus D und F der Stoff G entsteht. C und E sind noch unverändert im Gemisch enthalten:

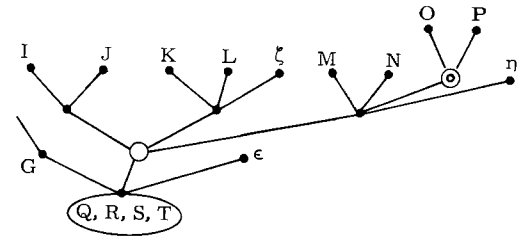


Vom Stoffgemisch C, E, G wird hiernach das Lösungsmittel C durch Destillation γ entfernt, wonach der Rückstand, bestehend aus E und G, mit einem Lösungsmittel H getrennt (δ) wird, in welchem nur E löslich ist.



Ebenso wie die $\vee$-Figur für das Vereinigen von Stoffen charakteristisch ist, so kennzeichnet die umgekehrte Figur $\wedge$ jede Trennung eines Stoffgemisches, oder auch eine Teilung eines Stoffsystems. Entstehen hierbei mehrere Fraktionen, so nimmt die topologische Figur zum Beispiel die folgende Gestalt an:

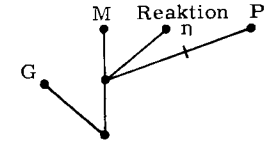
Der Stoff G wird nun entweder mit dem Stoffgemisch I und J oder aber mit dem Produkt der Reaktion ζ aus K und L umgesetzt in der Reaktion ε. Es kann stattdessen für die Reaktion ε mit G aber auch ein Produkt verwendet werden, das aus M und N durch die Reaktion η erhalten worden ist, wobei in dieser Reaktion O und/oder P zugegen gewesen sind.



Es entsteht die Stoffklasse Q, z. B. die Einzelverbindungen R oder S oder T, je nach den Reaktionspartnern von G.

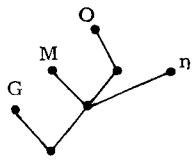
Schließen sich zwei Begriffe gegenseitig aus, wie das Stoffgemisch I und J einerseits und das Reaktionsprodukt aus K und L andererseits, so werden sie an einen als Kreis gezeichneten Punkt gebunden. Besteht diese Ausschließlichkeit unmittelbar an einem Ergebnispunkt, so werden sie *in* eine Kreisfigur eingetragen, wie die Begriffe Q, R, S und T. (Die Ausschließlichkeit der Begriffe Q bis T ist insofern gegeben, als an dieser Stelle immer nur jeweils einer dieser Begriffe gefunden zu werden braucht, nicht aber eine Kombination dieser Begriffe.) Besteht hingegen diese Ausschließlichkeit nur potentiell, so wird statt des Kreises ein Doppelkreis gezeichnet, wie es für die Begriffe O und P dargestellt worden ist. Wie man sieht, kann sich die Logik bei dieser Darstellungsweise über beliebig viele Stufen erstrecken, wobei die Übersichtlichkeit der Zusammenhänge voll erhalten bleibt.

Für unsere Dokumentationszwecke kam es aber besonders darauf an, diese Darstellungsweise der Begriffsverknüpfungen einzuspeichern und mit einer Fragestellung maschinell konfrontieren zu können. Eine Fragestellung mit irgendeiner Kombination der Begriffe α bis T soll dann und nur dann einen gespeicherten Graphen als Antwort ermitteln, wenn die gesuchten Begriffe genau in der gewünschten Verknüpfungsart angetroffen werden. Beispielsweise wird die obenstehende Figur als einschlägig auf eine Fragestellung erkannt, welche folgendermaßen darzustellen wäre:



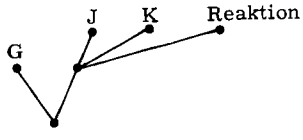
„G soll mit einem Reaktionsprodukt von M umgesetzt worden sein. Bei der Reaktion η von M dürfen beliebige andere Stoffe zugegen gewesen sein, nur nicht P". (Die durchgestrichene Verbindungsstrecke zu P im Fragegraphen bedeutet, daß P im gezeichneten Zusammenhang negiert ist.) Der Speichergraph enthält

nämlich die Teilaussage: „G wird mit dem Reaktionsprodukt von M umgesetzt, das in Gegenwart von O (und nicht unbedingt von P!) entsteht":



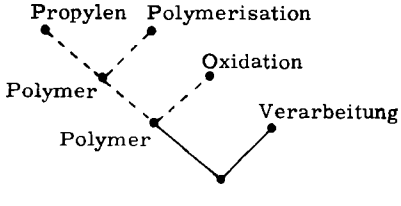
Offenkundig darf bei einer mechanisierten Literaturrecherche das Verbot eines bestimmten Begriffes nur dann ausgesprochen werden, wenn die Wirksamkeit dieses Verbotes auch genau auf den gemeinten Zusammenhang beschränkt werden kann. Aus dem bloßen Auftreten eines negierten Begriffes in einer Publikation allein dürfen noch keine Rückschlüsse gezogen werden. Denn der negierte Begriff kann in einem gänzlich peripheren, „unschädlichen" Zusammenhang genannt sein, etwa als Substanz für einen Eigenschaftsvergleich, als Hilfsstoff für die Synthese für den Fragesteller uninteressanter Folgeprodukte usw. Wenn aber doch mit Negationen in der Fragestellung gearbeitet wird, ohne daß die geschilderten Voraussetzungen erfüllt sind, wie es zuweilen in der modernen mechanisierten Dokumentation geschieht, so führt dies unausweichlich zum Verlust relevanter Information.

Die gespeicherte Figur ist hingegen nicht einschlägig auf eine Fragestellung: „Umsetzung von G mit einem Reaktionsprodukt aus J und K":



Es spielt in diesem Modell keine Rolle, ob die Struktur eines Stoffes bekannt ist oder nicht und ob es überhaupt eine treffende Bezeichnung für den Stoff gibt. Seine Eigenart, etwa als Reaktionsprodukt aus bestimmten Ausgangsstoffen, als Extrakt aus bestimmten Organen oder Pflanzenteilen usw., ist durch die topologisch definierte Lage in einem solchen Graphen eindeutig festgelegt. Auf diese Weise ist der Stoffbegriff im Graphen sicher wieder auffindbar.

Vorgänge, deren experimentelle Durchführung in der zu speichernden Publikation oder Patentschrift nicht ausdrücklich beschrieben worden ist, sondern nur angedeutet ist, werden mit unterbrochenen Linien dargestellt. Wenn etwa die Verarbeitung eines partiell oxidierten Polypropylens beschrieben worden ist, ohne jede detaillierte Angabe über die Herstellung des Polymeren selbst, so ergibt sich die folgende Figur:



Bei der Fragestellung nach einem bestimmten Prozeß hat man dann die Wahl, ob man als Antwort nur Publikationen wünscht, in denen der betreffende Vorgang *ausdrücklich beschrieben* ist, oder ob auch die *angedeuteten* Vorgänge von Interesse sind.

Bei näherer Untersuchung erweisen sich überraschend viele Eigenschafts- und Verfahrens-Begriffe als hochgradig zusammengesetzt. In der Dokumentation würde man sie besser in der Form handhaben, daß man sie einer begrifflichen Analyse unterwirft und hierbei die in ihnen enthaltenen Einfachbegriffe feststellt. Diese Einfachbegriffe *und* die zwischen ihnen herrschenden Begriffsverknüpfungen wären die idealen Stellvertreter für die zusammengesetzten Begriffe. „Fungistatische Thiocarbaminsäureester", „bienenunschädliche Insektizide", „Zonenschmelzverfahren", „mit Glycerin oberhalb 60 °C nur begrenzt mischbare halogenierte Kohlenwasserstoffe" usw. sind Beispiele für derartige zusammengesetzte Begriffe. Einer solchen analytischen Betrachtungsweise stand bisher immer entgegen, daß zwar die konstituierenden Einfachbegriffe genügend exakt darstellbar waren, nicht aber deren charakteristische wechselseitigen Verknüpfungen. Deswegen wäre bei der analytischen Handhabung dieser Begriffe immer ein wesentlicher Teil der Information verloren gegangen, der im zusammengesetzten Begriff noch enthalten ist. Dies hätte zwangsläufig zu den schon erörterten Ungenauigkeiten bei der machinellen Recherche geführt. Ein Verfahren zur exakten und übersichtlichen Darstellung von Begriffsverknüpfungen ebnet den Weg zu einer konsequent analytischen Betrachtungsweise auch von nichtstrukturchemischen Begriffen.

2.1. Beispiel für die Anwendung des graphischen Modells

Abschließend seien an einem Beispiel aus der chemischenVerfahrenstechnik (Referat zum französischen Patent 1356225) einige der vielen Arten von Fragestellungen gezeigt, welche mit diesem Modell beim gegenwärtigen Stande der Programmierung heute bereits beantwortbar sind:

Bei der Oligomerisation von Olefinen, z. B. Propylen, zu ungesättigten Oligomeren, z. B. Hexenen, Nonenen, verliert der Katalysator dadurch zunehmend an Wirksamkeit, daß er sich mit den gebildeten Olefin-Oligomeren belädt. Aus dem Reaktionsgemisch der Oligomerisation wird ein wiederverwendbarer Katalysator folgendermaßen regeneriert:

Man destilliert in einer Fraktionierkolonne A unter Druck das Propylen und die Oligomerisationsprodukte vom Katalysator ab. In einer nachgeschalteten zweiten Kolonne B wird Propylen vom Gemisch der Oligomerisate, hauptsächlich Hexen und Nonen, durch fraktionierende Destillation getrennt. Das Propylen geht teilweise in den Oligomerisations-Reaktor zurück, teilweise wird es in den unteren Bereich der Kolonne A eingeleitet, um den sich dort ansammelnden Katalysator von den letzten Beimengungen an Oligomerisaten zu befreien. Dies geschieht bei einer Temperatur, die unter dem Zersetzungsbereich des Katalysators und auch unter dem Siedepunkt der verunreinigenden Oligomerisate liegt.

Der gereinigte Trialkylaluminium-Katalysator wird am Boden der Kolonne A kontinuierlich entnommen und dem Oligomerisationsreaktor wieder zugeführt. — Die gebildete Hexenfraktion kann zur Herstellung von Isopren durch Cracken verwendet werden.

Graphisch sind diese Vorgänge etwa wie in Abbildung 1 darzustellen.

Es sei erwähnt, daß vielfach ein Referat zu einer Patentschrift überhaupt erst dann abgefaßt werden kann, *nachdem* man sich die Zusammenhänge in irgendeiner Weise zeichnerisch vergegenwärtigt hat.

Die Teilung oder Trennung eines Stoffsystems und die nachträgliche Wiedervereinigung der Stoffströme wird in einem Graphen für diesen Prozeß dadurch evident, daß ein Cyclus entsteht.

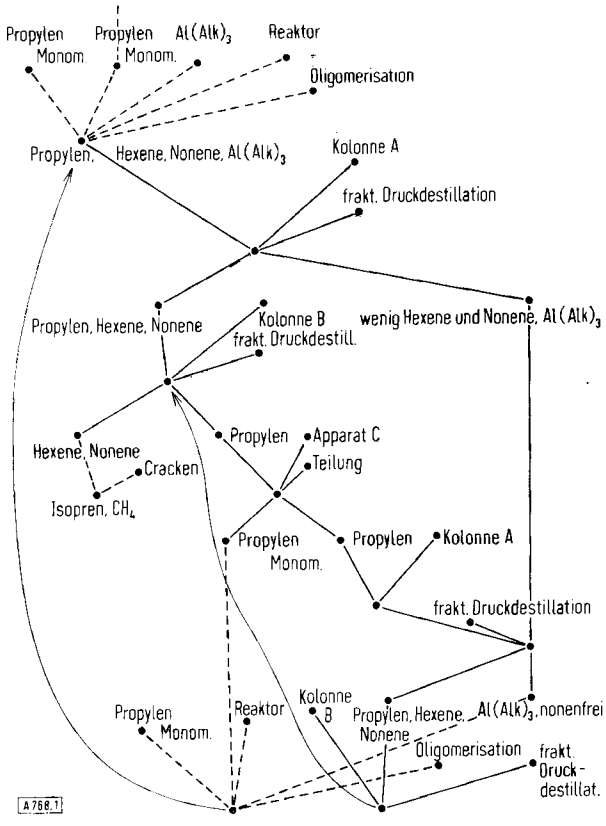


Abb. 1. Verknüpfung der Begriffe im Franz. Pat. 1356225.

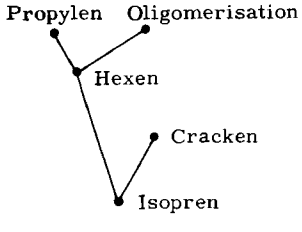
Wenn ein Stoff oder ein Stoffgemisch in einem Reaktionskreislauf geführt wird, so äußert sich auch dies im Graphen in charakteristischer Weise. Wenn man den rein chronologischen Laufweg eines solchen Stoffes verfolgt, so tritt dieser Stoff zweimal nacheinander in derselben Apparatur auf, nachdem er sie zwischendurch einmal verlassen hat, und in dieser Apparatur ist er beide Male mit genau den gleichen anderen Begriffen stofflicher und reaktionschemischer Art vergesellschaftet. Die Punkte, an denen, zeitlich verschoben, die gleiche Begriffsvergesellschaftung anzutreffen ist, sind im obenstehenden Graphen durch nicht-geradlinig geführte Verbindungspfeile zusammengeführt.

Es ist auch ersichtlich, ob ein Stoff einer Reaktion neu zugeführt ist oder ob er aus den Reaktionsprodukten zurückgewonnen und in die Reaktion zurückgeführt worden ist. Beispielsweise wird nur ein Teil des Propylens, welches in die Oligomerisation eingesetzt wird, frisch zugeführt. Ein anderer Teil ist aus dem Prozeß zurückgewonnen worden. Der Punkt, welcher frisch zugeführtes Propylen repräsentiert, hat keinen Verbindungsstrich nach oben. — Diese Kennzeichnung wird gegenwärtig allerdings noch nicht von der Codierung erfaßt und ist noch nicht maschinell recherchierbar.

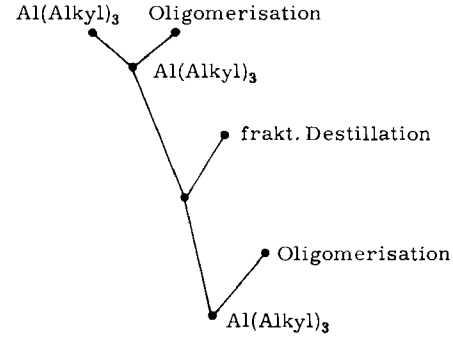
Wir betrachten nun eine Reihe von einschlägigen Fragestellungen aus diesem Gebiet und beschreiben ihre Formulierung gemäß unserem Modell sowie das Ergebnis, welches der maschinelle Vergleich der Fragestellungen mit dem Speichergraphen erbracht hat:

1. „Herstellung von Isopren durch Cracken von Hexen, das durch Oligomerisation von Propylen gewonnen wurde".

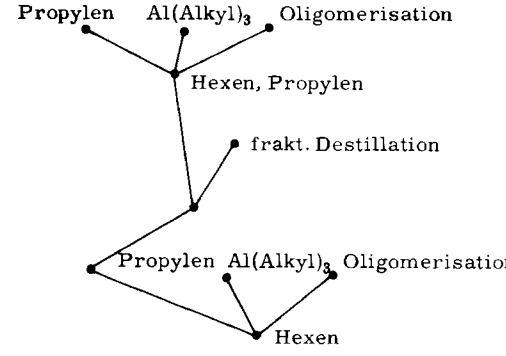
Die Frage läßt sich wie folgt graphisch formulieren:



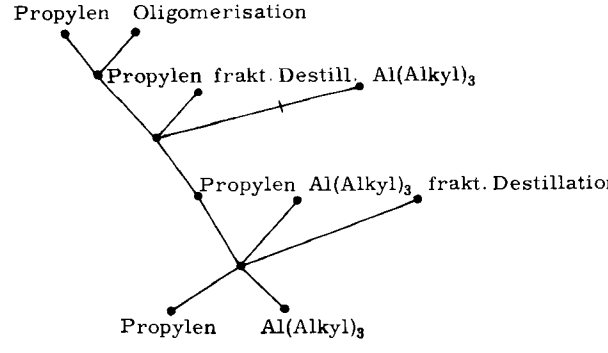
2. „Verwendung eines $Al(Alkyl)_3$-Katalysators zu einer Oligomerisation und seine Wiederverwendung für den gleichen Prozeß nach Reinigung durch fraktionierende Destillation".



3. „Oligomerisierung von Propylen mit Hilfe von $Al(Alkyl)_3$ zu Hexen sowie Abtrennung und Reinigung des überschüssigen Propylens durch fraktionierende Destillation und Rückführung des Propylens".

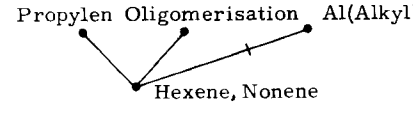


4. „Fraktionierende Destillation von $Al(Alkyl)_3$ mit Hilfe von zugeführtem, von $Al(Alkyl)_3$ freiem Propylen, das zuvor fraktionierend destilliert und aus einen Oligomerisationsprozeß wiedergewonnen wurde".



Beim topologischen Vergleich der Fragegraphen mit dem Speichergraphen wird in allen Fällen Übereinstimmung erzielt, d. h. die bibliographischen Daten des Speichergraphen werden als Antwort auf diese Fragestellung ermittelt.

Hingegen würde diese Patentschrift zurecht nicht als Antwort auf eine Fragestellung der Art: „Oligomerisation von Propylen mit anderen Katalysatoren als mit $Al(Alkyl)_3$, unter Bildung von Hexenen und Nonenen" gefunden werden:



3. Prinzipien des Vergleiches von Fragegraph und Speichergraph

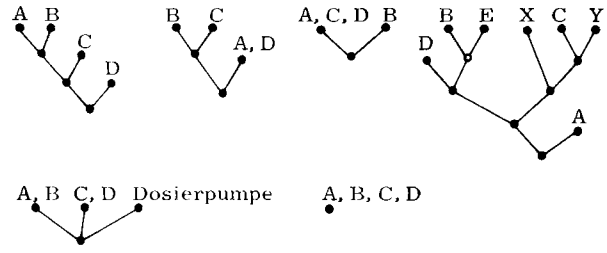
Wie aus den Beispielen ersichtlich ist, kann es sich bei der mechanisierten Prüfung, ob ein Speichergraph für einen Fragegraphen einschlägig ist, nicht um einen einfachen Prozeß handeln. Keinesfalls könnte z. B. ein Verfahren befriedigen, bei welchem lediglich nach solchen Speichergraphen oder Teilen eines Speichergraphen gesucht wird, auf welche sich die Struktur des Fragegraphen sozusagen kongruent abbilden läßt. Exakt ausgedrückt, hat es nicht sein Bewenden damit, daß man den Fragegraphen homöomorph und simplizial in den Speichergraphen abzubilden versucht und diejenigen Speichergraphen als einschlägig für die Fragestellung betrachtet, bei denen dies gelingt.

Ein solches einfaches Verfahren würde schon beim ersten Fragebeispiel versagen. Denn beim einschlägigen Speichergraphen zur Aufarbeitung der Propylen-Oligomerisationsprodukte existiert keineswegs eine direkte Verknüpfung zwischen dem gebildeten Hexen und dem hieraus entstandenen Isopren, wie im Fragegraphen gezeichnet. Vielmehr macht im Speichergraphen das Hexen eine Reihe von Trenn- und Reinigungsschritten durch, bevor es dem Crackprozeß unterworfen wird. Dies jedoch sind Vorgänge, deren Stattfinden oder Nichtstattfinden für den Fragesteller gänzlich peripher sein können und bezüglich derer er sich bewußt nicht festlegen will. Auch könnte das Hexen zwischendurch, d. h. bevor es weiterreagiert, mit anderen Stoffen zusammengebracht worden sein, ohne daß hierdurch eine Publikation an Interesse für den Fragesteller verliert. Das Hexen kann zur Abtrennung von Verunreinigungen Einschlußverbindungen gebildet haben, in Zweiphasensystemen verteilt worden sein usw. Wäre der Fragesteller gezwungen, sich stets auch bezüglich solcher Details festzulegen, die für ihn uninteressant oder unwesentlich sind, und deren Auftreten in den ihn interessierenden Publikationen unvorhersehbar ist, so wäre ihm mit einem solchen Dokumentationssystem nicht gedient. Mit solchen einschränkenden Festlegungen würde er sich selbst wertvolle Informationsquellen vorenthalten. Er wäre etwa in der gleichen Lage wie beim Summenformelregister, wenn er Literatur zu einer Stoff*gruppe* sucht. Auch hier muß er sich in höchst unerwünschter Weise bezüglich der Art und Anzahl von – für ihn! – belanglosen Substituenten am ihn allein interessierenden Grundtyp festlegen.

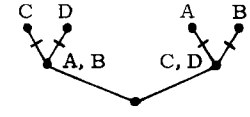
Es muß aus dem gleichen Grund für den Fragesteller auch die Möglichkeit bestehen, etwa bei der Frage nach einem bestimmten Mehrkomponenten-System offen zu lassen, in welcher Reihenfolge die Komponenten zusammengebracht worden sind, ob weiterhin zwei von ihnen etwa vorgemischt gewesen sind, bevor sie mit einer weiteren Komponente (oder mit einer Mischung weiterer Komponenten) vereinigt worden sind usw. Ein Fragegraph der Struktur

A, B, C, D

muß daher, wenn diese Anforderungen erfüllt sein sollen, beispielsweise von den folgenden Speichergraphen erfüllt werden:



Kommt es hingegen darauf an, daß etwa ein Gemisch von A und B mit einer ebenfalls vorgebildeten Mischung von C und D vereinigt worden ist, so soll der entsprechende Fragegraph



nur von den Speichergraphen der Art

erfüllt werden.

Für den Vergleich zweier Graphen im oben beschriebenen System bieten ausgebaute mathematische Theorien keine Hilfe. Zwar beschreibt die Graphentheorie – also die eindimensionale kombinatorische Topologie – ausführlich die allgemeinen Eigenschaften von Graphen, und die hier beschriebenen Frage- und Speichergraphen nehmen in dieser Hinsicht keinerlei Ausnahmestellungen ein. Aber für das eigenartige Prinzip, nach welchem der Vergleich eines Fragegraphen mit einem Speichergraphen abgewickelt werden muß, damit die Anforderungen der Chemie-Dokumentation erfüllt werden, bietet die Graphentheorie keinen Lösungsansatz.

Was weiterhin die Logistik anbetrifft, so lehrt sie zwar, wie man logische Ausdrücke in der verschiedenartigsten Weise umformen kann. Die oben beschriebene Logik, nach welcher die Relevanz eines Speichergraphen für einen Fragegraphen beurteilt werden muß, behandelt sie indessen nicht.

Diese mathematischen Disziplinen legen lediglich gewisse maschinell verarbeitbare Darstellungsformen für die Publikation im Speicher einerseits und für den Inhalt der Fragestellung andererseits nahe. Das Prinzip hingegen, nach welchem die beiden Graphen bei der praktischen Literaturrecherche miteinander konfrontiert werden müssen, mußte eigens von uns für diesen Zweck konstruiert werden, und zwar aufgrund zahlreicher pragmatischer Einzelüberlegungen. Sie haben ihren Niederschlag in einem recht umfangreichen Computerprogramm gefunden. Die Vervollkommnung dieses Programms, um noch mehr Komfort für die Einspeicherung und noch mehr Treffsicherheit und Flexibilität für die Anfrage zu erreichen, ist noch im Gange.

Damit ein solches Maschinenprogramm in der mechanisierten Chemie-Dokumentation nutzbringend eingesetzt werden kann, muß es mit einem anderen Programm verbunden werden, welches das Auffinden der *gesuchten Begriffe selbst* zum Ziele hat, unabhängig von ihren wechselseitigen Verknüpfungen. Solange nicht sichergestellt ist, daß in einem gespeicherten Dokument die Begriffe der Fragestellung auch tatsächlich vorkommen, hat es nämlich keinen Sinn, daß man die Maschine auf die Suche nach der gewünschten Verknüpfung dieser Begriffe schickt. Gegenwärtig sind wir damit beschäftigt, den hintereinandergeschalteten Ablauf zweier solcher Maschinenprogramme mit unterschiedlicher Aufgabenstellung zu verwirklichen. Zur Vorselektion nach den Begriffen selbst dient uns das in der IDC[3] bereits eingeführte Verfahren zum Recherchieren nach chemischen Fachbegriffen.

4. Ausblick

Betrachtet man die vertraute Strukturformel aus dieser Perspektive, so kann man auch sie als einen Graphen auffassen, wenn auch mit besonderen Eigenschaften. Nicht ein Stoffbegriff besetzt hier einen Punkt, sondern ein Atom. Die Bindungen zwischen den Atomen fungieren als Strecken in diesem Graphen. Tatsächlich ist die Strukturformel auch als ein immer noch hochgradig zusammengesetzter Begriff aufzufassen, aufgebaut aus vielen „Einfachbegriffen", nämlich den Atomen selbst. Nun ist es eine mehr als hundert Jahre alte Erfahrung, daß die analytische Betrachtungsweise und Darstellungsweise für den Strukturformelbegriff ungleich anschaulicher und übersichtlicher ist, als die komplexe Nomenklatur oder gar eine willkürlich festgelegte Trivialbezeichnung für einen Stoff. Darüber hinaus ist diese Betrachtungsweise wissenschaftlich auch ungleich fruchtbarer, weil sie verwandtschaftliche Beziehungen zwischen den Stoffbegriffen unmittelbar ins Auge springen läßt.

Aus dieser Perspektive stellen die beschriebenen Graphen nichts anderes dar als die Übertragung der Strukturformelidee auf stärker ausgeweitete Zusammenhänge, nämlich auf die Zusammenhänge innerhalb einer Publikation. Der Gewinn an Übersichtlichkeit wird deutlich, wenn man sich die Mühe macht, den oben gezeichneten Graphen mit dem Originaltext der verfahrenstechnischen Patentschrift zu vergleichen. Wir verfolgen daher sehr aufmerksam, welchen Anklang solche Graphen bei unseren Kollegen in der Forschung und im Betrieb finden, als reines Mittel zur Schnellinformation, eventuell zusätzlich zu einem Referat oder zum Originaltext.

Diese Arbeit wurde durch Mittel des Bundesministeriums für Bildung und Wissenschaft gefördert. Für anregende Diskussionen danken wir dem Arbeitskreis für Makromolekularchemie-Dokumentation.

Eingegangen am 7. Januar 1970 [A 768]

[3] *S. Rössler* u. *A. Kolb*, J. Chem. Doc. 10, Nr. 2, S. 128 (1970); *E. Meyer*, Angew. Chem. *82*, 605 (1970); Angew. Chem. internat. Edit. *9*, Heft 8 (1970).

Der Dokumentationsring der chemisch-pharmazeutischen Industrie; Ziele und Methoden

Von Wolfgang Nübling und Walter Steidle [*]

Die im Dokumentationsring zusammenarbeitenden Firmen wählten für die Verschlüsselung der pharmazeutisch-chemischen Literatur das Lochbildverfahren. Chemische, biochemische und medizinische Fakten einer Publikation werden dabei in Grundbegriffe zerlegt, denen auf der Lochkarte jeweils eine Position zugeteilt ist. Dieses Verfahren hat sich seit zwölf Jahren bewährt.

1. Einleitung

Der 1958 gegründete Dokumentationsring[1] der chemisch-pharmazeutischen Industrie hat sich die Aufgabe gestellt, chemische und verwandte Fachliteratur durch Verbesserung vorhandener und Entwicklung neuer Methoden intensiver zu erfassen und damit besser nutzbar zu machen. Der Dokumentationsring versetzt seine Teilnehmer in die Lage, bei angemessenem eigenen Aufwand über das große Gesamtmaterial der Gemeinschaft verfügen zu können. Jedes Mitglied bearbeitet einen gleich großen Anteil der Quellen (Zeitschriftenliteratur, Patentschriften, andere Literatur). Diese Arbeitsteilung gestattet es auch kleineren

[*] Dr. W. Nübling
E. Merck
61 Darmstadt 2, Postfach 4119
Dr. W. Steidle
Knoll AG
67 Ludwigshafen

[1] Gründungsmitglieder: Farbenfabriken Bayer AG, Leverkusen; CIBA AG, Basel (Schweiz); Knoll AG, Ludwigshafen; E. Merck, Darmstadt; Dr. Karl Thomae GmbH, Biberach.