Recherchen mit dem IDC-System

Von Martin A. Lobeck[*]

Das Dokumentations-System der Internationalen Dokumentationsgesellschaft für Chemie mbH (IDC) arbeitet mit Mitteln der elektronischen Datenverarbeitung. Es können Strukturen, Reaktionen und Sachverhalte gespeichert und recherchiert werden. Strukturen und Reaktionen werden als Terms einer Facettenklassifikation codiert, Sachverhalte im Klartext gespeichert. Zur Zeit sind 900000 Formeln, darunter zahlreiche Markush-Formeln, aus Patent- und Zeitschriftenveröffentlichungen auf Magnetband gespeichert. Die Recherche in diesem Bestand geschieht mit IDC-Programmen, die pro Frage durchschnittlich weniger als 1 Minute Computerzeit (CPU-Zeit) benötigen. Die Strukturterms können maschinell aus topologischen Einspeicherungen erzeugt werden. Erfaßt wird die niedermolekulare organische Chemie; eine Erweiterung auf das Makromolekulargebiet ist möglich und vorgesehen.

1. Einleitung

Die IDC (Internationale Dokumentationsgesellschaft für Chemie mbH, Frankfurt) hat es sich zum Ziel gesetzt, chemische Veröffentlichungen – Literatur und Patente – zu codieren und so ihren Inhalt einer Suche mit Methoden der elektronischen Datenverarbeitung zugänglich zu machen. Das erste Gebiet, das in Angriff genommen wurde, war das der niedermolekularen organischen Chemie, weil es hier am schwierigsten ist, Fragen nach Partialstrukturen zu beantworten, die in der Praxis häufig gestellt werden. Bei der IDC werden deshalb Strukturformeln organischer Substanzen so eingespeichert, daß auch nach Teilstrukturen gesucht werden kann, ferner nach Reaktionen organischer Substanzen mit organischen und anorganischen Reaktionspartnern und nach Sachverhalten.

2. Aufbau des IDC-Systems

2.1. Strukturverschlüsselung

Die Entwicklung begann in den fünfziger Jahren bei den Farbwerken Hoechst. *Fugmann* und seine Mitarbeiter fanden die damals existierenden Methoden der Strukturverschlüsselung unbefriedigend, besonders wenn große Bestände zu durchsuchen waren. Das in Hoechst konzipierte und auch benutzte System wurde bald auch von den Farbenfabriken Bayer und der Badischen Anilin- und Soda-Fabrik übernommen, die auch Verbesserungen und neue Ideen verwirklichten. Erst 1967 wurde die IDC gegründet, die inzwischen 13 Chemiefirmen als Mitglieder hat.

Bei der Gründung lagen bereits über 400000 Speichersätze auf Magnetband vor, eine Zahl, die sich inzwischen mehr als verdoppelt hat. Jedem Speichersatz entsprechen dabei eine oder mehrere verwandte Strukturformeln, die nach dem GREMAS-Verfahren [1] codiert sind.

[*] Dr. M. A. Lobeck
Henkel und Cie GmbH
4 Düsseldorf 1, Postfach 1100

[1] Genealogisches *RE*cherchieren mit *MA*gnetband-*S*peichern.

2.1.1. Dreierterms

Das GREMAS-System beruht auf einer Facettenklassifikation. Jedes Kohlenstoffatom einer organischen Verbindung erhält mindestens einen Term, der aus drei Buchstaben besteht und kurz Dreierterm genannt wird.

Der erste Buchstabe des Dreierterms, Genusbuchstabe genannt, sagt etwas darüber aus, wieviele Heterobindungen das betrachtete C-Atom betätigt, meist auch, zu welchen Atomen. Als Heterobindung wird dabei jede Bindung betrachtet, die nicht zu Wasserstoffatomen oder zu anderen C-Atomen führt.

Genus B:	Amine	z.B.
BA.	primär	CH_3-NH_2
BB.	sekundär	$(CH_3)_2NH$
BC.	tertiär	$(CH_3)_3N$
BD.	quaternär	$(CH_3)_4N^{\oplus}$
Genus H:	**Halogenverbindungen**	**z.B.**
HA.	Fluor	C_6H_5-F
HB.	Chlor	C_6H_5-Cl
HC.	Brom	C_6H_5-Br
HD.	Jod	C_6H_5-J
Genus R:	**unsubstituierte Ketten-C-Atome**	**z.B.**
RA.	gesättigt	CH_3-CH_3
RB.	olefinisch	$CH_2=CH_2$
RC.	acetylenisch	$CH\equiv CH$

Abb. 1. Beispiele für Genus und Spezies.

Abbildung 1 zeigt, daß z. B. für Amine der Genusbuchstabe B gilt, für Halogenverbindungen Genus H, und daß R als erster Buchstabe des Dreierterms gesetzt wird, wenn das C-Atom an kein Heteroatom geknüpft ist.

Der zweite Buchstabe des Terms, die Spezies, präzisiert den Inhalt des ersten. So bedeutet bei Genus B ein A als Zweitbuchstabe ein primäres Amin, ein B ein sekundäres, ein C ein tertiäres usw. Im Genus H haben die Zweitbuchstaben A, B, C usw. die Bedeutung Fluor, Chlor, Brom usw., im Genus R geben sie an, ob das C-Atom Doppel- oder Dreifachbindungen betätigt oder nicht.

Der letzte Buchstabe der Dreierterms, die Subspezies, sagt aus, in welcher Umgebung sich das betrachtete C-Atom befindet. Sitzt es in oder an einem Heterocyclus, einem Aromaten, einer Kette, mit oder ohne Doppelbindungen (Abb.2)?

Durch die Kombination von Buchstaben sind weit mehr als 3000 Dreierterms definiert; man erhält also ein recht feinmaschiges Netz, das die Struktur schon recht gut wiedergeben müßte. Tatsächlich hat man sich zu Anfang auch mit diesen Terms begnügt, mußte aber bald feststellen, daß der Ballastanteil bei vielen Fragestellungen doch ganz erheblich war.

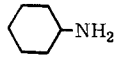
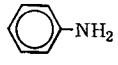
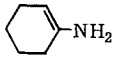
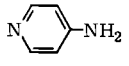
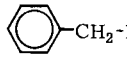
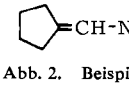
BA.	primäre Amine
$CH_3-CH_2-NH_2$	BAA
$CH_2=CH-NH_2$	BAF
$CH\equiv C-NH_2$	BAM
	BAQ
	BAR
	BAT
	BAS
	BAD
	BAH

Abb. 2. Beispiele für Subspezies.

2.1.2. Bezirksterms

Es wurden daher Bezirksterms variabler Länge (Y-Terms) geschaffen, die Aussagen darüber machen, womit z. B. ein Ring oder eine Kette substituiert sind, ob und wieviele Doppel- oder Dreifachbindungen vorhanden sind, bei Aromaten auch, ob *ortho*-, *meta*- oder *para*-Substitution vorliegt.

Bei den Aromaten ist man bei mehr als zweifacher Substitution sogar so weit gegangen, die Besetzung aller sechs Positionen zu spezifizieren (Z-Terms). Bei allen Bezirksterms hat man sich damit begnügt, den Substituenten nur mit dem Genusbuchstaben, also nicht übermäßig genau, zu kennzeichnen.

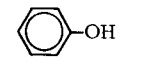
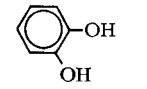
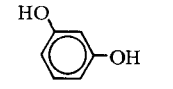
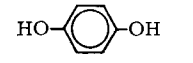
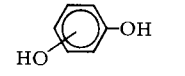
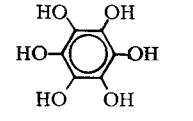
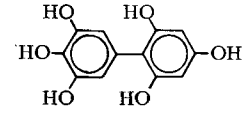
	Dreierterms	Bezirksterms
	SAF EAR	YT E– – –
	SAF EAR	YT EE2–
	SAF EAR	YT EE3–
	SAF EAR	YT EE4–
	SAF EAR	YT EE1–
	SAF EAR	ZT EEEEEE– – – –
	SAF EAR	ZT EEEØTØ– – – – und ZT ETEØEØ– – – –

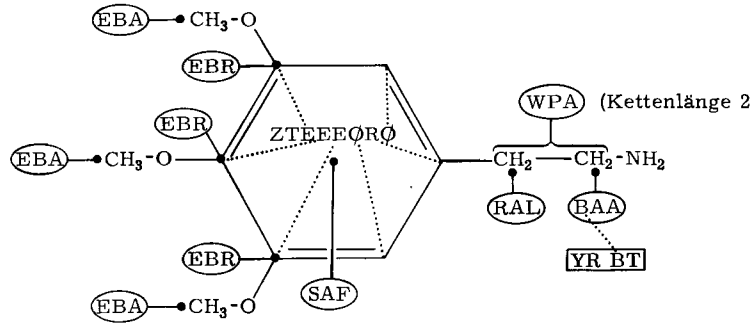
Abb. 3. Bezirksterms differenzieren Strukturen mit gleichen Dreierterms.

Wie man in Abbildung 3 sieht, fallen bei völlig gleicher Codierung der Verbindungen mit Dreierterms die Bezirksterms ganz verschieden aus, die Strukturen werden also unterscheidbar.

2.2. Beispiel einer Strukturverschlüsselung

Das Beispiel der vollständigen Codierung einer Verbindung soll anschließend auch zur Illustration einer Recherche dienen.

Man erkennt, daß in Abbildung 4 jedem C-Atom ein Dreierterm zugeordnet ist. In der endgültigen Zusammenstellung („Auflistung") werden mehrfach



BAA EBA EBR RAL SAF WPA
YRBT--
ZTEEEØRØ----

Abb. 4. Vollständige Codierung einer Strukturformel.

vorkommende Terms nur einmal berücksichtigt. Die Ätherbindung wird einmal von der Alkylseite her betrachtet, einmal von der Phenylseite, und führt zu Terms EBA und EBR; SAF steht für den Benzolring, oder genauer, für jedes seiner C-Atome; RAL für das nicht mit Heteroatomen verknüpfte, aromatisch substituierte CH_2; BAA für das NH_2-substituierte CH_2. Hinzu kommt ein Term WPA für eine Kettenlänge von zwei C-Atomen, der gleichzeitig dartun soll, daß es zahlreiche Dreierterms gibt, die hier nicht besprochen sind, u.a. viele Terms für Substitutions- und Kondensationsangaben bei Hetero- oder Carbocyclen, zahlreiche Terms, die Aussagen über pharmakologische oder mikrobiologische Wirksamkeit der behandelten Struktur machen und andere, die Angaben über spektroskopische oder analytische Eigenschaften enthalten.

2.3. Reaktionsverschlüsselung

Zur Reaktionsverschlüsselung (Abb. 5) werden die Dreierterms jedes reagierenden Kohlenstoffatoms herangezogen. In einem zwölfstelligen Term wird in den ersten drei Stellen durch @DR festgelegt, daß eine Reaktionscodierung folgt, dann folgt der Dreierterm des C-Atoms *vor*, anschließend der *nach* der Reaktion, außerdem werden an den Stellen 10 bis 11 besondere Merkmale codiert, z.B. Ringschluß oder -sprengung oder Kettenverlängerung (ZL). An Stelle 12 kann das Kohlenstoffatom numeriert werden, damit es wiedererkannt wird, wenn weitere Reaktionsschritte in der betreffenden Arbeit beschrieben sind. Bei der Recherche kann dann über mehrere Schritte hinweg gesucht werden.

Da die Reaktionsterms mit besonderen Programmen befragt werden müssen, werden stets auch Reaktionskurzterms, also eine einfachere Codierung der Reaktion, erzeugt, die bei der Verschlüsselung des jeweiligen Endprodukts mitgespeichert wird und hier 2RB, 2RC und 3ZL bzw. 2II und 2RA lautet. Diese Reaktionskurzterms können dann wie alle Dreierterms recherchiert werden. In vielen Fällen reichen sie aus; das eigentliche Reaktionsrechercheprogramm braucht dann nicht nachgeschaltet zu werden.

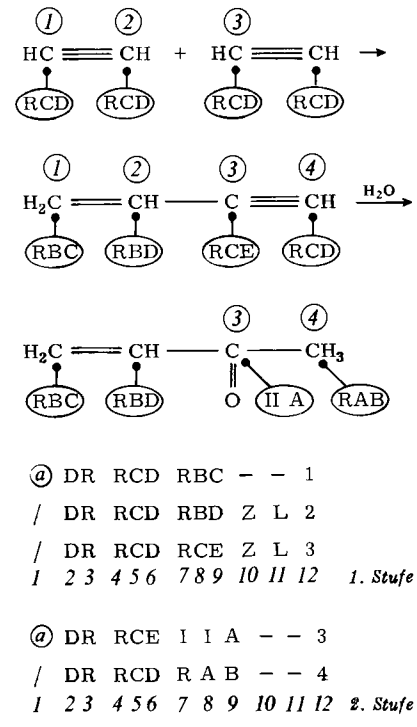


Abb. 5. Codierung einer Reaktion mit Einleitungszeichen (1—3), Anfangsterm (4—6), Endterm (7—9), Reaktionsbedingungen (10—11) und Numerierung der C-Atome (12).

2.4. Sachverhalte

Häufig vorkommende Sachverhalte erhalten, wie bereits erwähnt, eigene Dreierterms, die übrigen werden im Klartext eingespeichert, ebenso auch anorganische Reaktionspartner. Bei den Sachverhalten ist dem Verschlüsseler völlig freie Wortwahl zugestanden worden, um in späteren Stadien früher unbekannte Begriffe nicht in ein vorgesehenes Sprachkorsett zwängen zu müssen. Das bedingt andererseits eine sehr intensive Nachbearbeitung, denn jeder neu auftauchende

a)

THESAURUS-ALPHABET-REGISTER, BAND-BR. 09194

RT 001699 B PHOTOCHLORIERUNG
EO 000428 B PHOTOCHROMIE
EO 000428 E PHOTOCHROMISM
RT 002341 B PHOTODIMERISIERUNG
SF 000221 E PHOTOGRAPHIC DEVELOPER
SF 001203 E PHOTOGRAPHIC EMULSION
ZZ 000220 B PHOTOGRAPHIE
SF 001203 B PHOTOGRAPHISCHE EMULSION
SF 000221 C PHOTOGRAPHISCHE ENTWICKLER
SF 000221 B PHOTOGRAPHISCHER ENTWICKLER
ZZ 001062 B PHOTOGRAPHISCHER FILM
RB 001701 B PHOTOIONISATION
RT 001701 B PHOTOIONISATION
RE 001701 B PHOTOIONISIERUNG
RT 001701 B PHOTOIONISIERUNG
RT 001698 B PHOTOISOMERISATION
RT 001698 B PHOTOISOMERISIERUNG
UM 000520 B PHOTOKOLORIMETRIE
EO 001562 E PHOTOLUMINESCENCE
EO 001562 B PHOTOLUMINESZENZ
RB 001694 B PHOTOLYSE
RB 001147 B PHOTOLYSE IN FLUESSIGPHASE
RB 001694 E PHOTOLYSIS
UM 000519 B PHOTOMETRIE
UM 000519 E PHOTOMETRY
RT 001700 B PHOTONITROSIERUNG

A 765 6a

b)

GENERIC CHART OF THESAURUS NO. 0066 STATE 27.10.69 (D) (UTS)

1 2 3 4 5 6 7 8 9 A B C D E F G H I J K L M N P Q R S T U V W X Y Z

000430 LUMINESZENZ / KALTES LEUCHTEN / LUMINESCENCE
V 001579 LUMINESZENZANALYSE
V 001580 LUMINESZENZ-SPEKTRALANALYSE
001561 U SENSIBILISIERTE LUMINESZENZ
001562 U PHOTOLUMINESZENZ / PHOTOLUMINESCENCE
000095 U FLUORESZENZ / FLUORESZENZSTRAHLUNG / FLUORESZENZFAEHIGKEIT / FLUORESZENZVERMOEGEN / FLUORESCENCE
V 001259 OPTISCHER AUFHELLER
V 001584 FLUORESZENZANALYSE
V 001585 FLUORESZENZTITRATION
V 001586 FLUORESZENZINDIKATOR
V 001587 FLUORESZENZMIKROSKOPIE
000432 U SENSIBILISIERTE FLUORESZENZ
001582 U RESONANZFLUORESZENZ
001583 U STOKES-FLUORESZENZ
000431 U PHOSPHORESZENZ / PHOSPHORESZIEREND / PHOSPHORESCENCE
000433 U MOLEKUEL-PHOSPHORESZENZ
001588 U KRISTALL-PHOSPHORESZENZ
001589 Z MOLEKUELPHOSPHORE
001590 Z KRISTALLPHOSPHORE
001591 Z NACHLEUCHTDAUER
001563 U CHEMILUMINESZENZ / CHEMOLUMINESZENZ / CHEMILUMINESCENCE
000434 U SENSIBILISIERTE CHEMILUMINESZENZ
001592 Z CHEMILUMINESZENZ-INDIKATOR
001564 U BIOLUMINESZENZ
001565 U KRISTALLOLUMINESZENZ

A 765 6b

Abb. 6. Thesaurus-Ausdruck. Beispiel: *Photolumineszenz* a) im alphabetischen Register und b) im hierarchischen Begriffsfeld.

Begriff wird geprüft, Beziehungen zu anderen Begriffen werden festgestellt, etwa: Zugehörigkeit, Relation Teil-Ganzes, hierarchische Relationen (Ober-, Unterbegriffe), Synonymität. Es wird eine Definition oder eine Begriffserläuterung formuliert und schließlich die Begriffs-Sammlung periodisch als Thesaurus ausgedruckt, in dem von einem alphabetischen Register aller Begriffe (Abb. 6a) auf die zugehörigen Begriffsfelder verwiesen wird, die alle Relationen erkennen lassen (Abb. 6b).

2.5. Einspeicherung

Bei der IDC werden alle zu speichernden Formeln codiert und die Terms ebenso wie die Sachverhalte auf Lochbelegen niedergeschrieben. Dabei erhalten Sachverhalte und anorganische Reaktionspartner ebenso wie die Reaktionsterms verschiedene Einleitungszeichen, die bei der späteren Recherche geprüft werden und verhindern, daß man an Speicherstellen sucht, die mit Sicherheit keine Information der gewünschten Art enthalten. Die Strukturen werden dagegen nicht alle intellektuell verschlüsselt, sondern nur der Anteil, für dessen Bewältigung die Kapazität der Formellesemaschinen nicht ausreicht, mit denen die Formeln maschinell und topologisch aufgenommen werden. Die Strukturterms werden dann mit einem Computerprogramm aus dem topologischen Speichersatz erzeugt [2].

Die mit Lochkarten oder Lochstreifen aufgenommenen Daten werden nach mehrfacher Prüfung auf Magnetband gespeichert; aus technischen Gründen benutzt man für Reaktionen und Sachverhalte ein vom „Strukturband" getrenntes Ergänzungsband. Jede Struktur erhält einen besonderen Speichersatz, in dem aber auch alle notwendigen bibliographischen Daten enthalten sind, z.B. eine Dokument-Nummer der ausgewerteten Veröffentlichung. Über Speichersatz- und Dokument-Nummer kann man dann auch getrennte Recherchen im Struktur- und Ergänzungsband anschließend korrelieren.

Für jedes Dokument werden vom Verschlüsseler auch noch eine oder mehrere Gebietsangaben mitgeliefert, durch die die Veröffentlichung grob klassifiziert wird; z.B. bedeutet B: Organische Chemie allgemein, D: Pharmazeutische Chemie, E: Schädlingsbekämpfung. Diese Gebietskennzeichnung kann bei der Recherche mitverlangt werden und so für häufig vorkommende Substanzen die Zahl der Antworten einengen.

Diese Angaben über die Einspeicherung sind nicht vollständig, reichen aber zum Verständnis der folgenden Rechercheformulierungen wohl aus.

3. Recherche

3.1. Organisatorisches

Jeder Chemiker erhält gleich nach seinem Eintritt in unsere Firma Recherchebögen, wie sie Abbildung 7 zeigt. In einer Kurzanleitung wird ihm das Notwendigste über die Recherchemöglichkeiten gesagt. Die Bögen enthalten Abteilung und Namen des Chemikers, ferner eine dreistellige Frager-Nummer. Der Chemiker zeichnet die gewünschte Struktur und eventuelle Nebenbedingungen auf und übergibt den Bogen der Dokumentationsabteilung.

Hinter der dreistelligen Frager-Nummer werden die Fragen dieses Chemikers von der Dokumentation beim Eingang zweistellig fortlaufend gekennzeichnet. Das führt dazu, daß alle Anfragen, Computerlisten usw. für einen Chemiker in der Ablage stets an einem Platz zu finden sind.

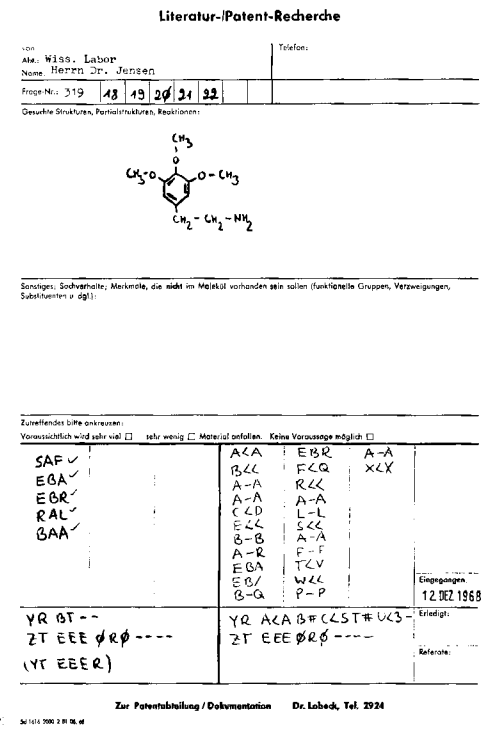


Abb. 7. Recherchebogen.

3.2. Recherchenverschlüsselung

Auf dem Recherchebogen wird bei jeder Frage zunächst die angegebene Struktur in der oben skizzierten Weise verschlüsselt, ebenso weitere codierbare Kriterien, z.B. bestimmte Herstellungsverfahren. Es liegt nahe, die gleichen Terms, die man bei der Speicherung der Verbindung benutzt, auch als Suchbedingungen zu verwenden. Eine solche Formulierung ist tatsächlich möglich und liefert auch alle Speichersätze der gesuchten Formel – leider aber mit einem meist sehr hohen Ballastanteil. Das Suchprinzip entspricht genau dem in Sichtlochkarten angewendeten Verfahren. Es leuchtet ein, daß damit nur Minimalforderungen an einen Speichersatz gestellt werden können: Die geforderten Terms *müssen* darin enthalten sein, es können aber durchaus noch weitere vorkommen. Man erhält deshalb nicht nur die gewünschte, sondern auch kompliziertere Strukturen. Abhilfe läßt sich nur schaffen, wenn alle unerwünschten Terms verbietbar sind. So selbstverständlich das klingt, so selten ist eine solche Recherchenmöglichkeit in Dokumentationssystemen vorgesehen, soweit sie sich nicht elektronischer Datenverarbeitungsanlagen bedienen.

Die große Zahl der Dreierterms läßt es nicht zu, jeden zu verbietenden Term einzeln zu benennen. Die praxisnahe Art der Formulierung im IDC-System soll an-

[2] *E. Meyer*, Angew. Chem. *82*, 605 (1970); Angew. Chem. internat. Edit. *9*, Nr. 8 (1970).

hand einer Modellrecherche nach der in Abbildung 4 codierten Struktur gezeigt werden.

Gehen wir von den Strukturterms aus, wie sie im Recherchebogen (Abb. 7) links unten eingetragen sind, so müssen diese bei der Frageformulierung in eine strikt alphabetische Reihenfolge gebracht werden. Diese Forderung nach alphanumerischer Sequenz ließe sich mit einem einfachen Sortierprogramm leicht umgehen. Bei der Einspeicherung von Strukturen geschieht das auch, nicht aber bei der Frage. Hier treten nämlich die Verbote hinzu, und es ist sinnvoll, die möglichen Genera, also die Erstbuchstaben der Terms, von Anfang an durchzugehen und sich zu fragen, ob in diesem Genus überhaupt ein Term gefordert ist oder nicht. Ist das z. B. für Genus A nicht der Fall, so kann es mit der einfachen Formulierung A < A verboten werden. Das Programm läßt Erweiterungen auf ganze Bereiche zu, H < L bedeutet, daß im Speichersatz kein Dreierterm mit Genus H, I, J, K, L vorkommen darf.

Die Forderung, daß ein Term im Speichersatz vorkommen soll, formuliert man durch Angabe des Terms selbst, z. B. BAS für eine primäre Aminogruppe an einem Heterocyclus. Ist die Forderung allgemeiner, so kann ein Bereich gefordert werden, aus dem mindestens ein Term gefunden werden muß. Die Formulierung erfolgt dann mit dem „Beliebig"-Zeichen *, für das im Speichersatz jedes Zeichen stehen kann.

Auf die Forderung H * *, die besagt, daß irgendwo im Molekül mindestens einmal Halogen vorkommen muß, antworten also alle Speicherterms, die mit H beginnen. Das Beliebig-Zeichen kann erläutert, d. h. eingeengt werden: in H * * B—C R—S erläutert der zweite Term B—C den ersten Stern, R—S den zweiten, die Formulierung fordert Halogen (H), und zwar entweder Chlor (B) oder Brom (C), aromatisch (R) oder heterocyclisch (S) gebunden; andere halogenhaltige Verbindungen, z. B. Jodbenzol (HDR) oder Chloräthan (HBA), fallen nicht in den geforderten Bereich und werden nicht ausgegeben. Fluorchlorbenzole (HAR HBR) werden dagegen mitgeliefert, weil ja HBR verlangt ist.

Will man das verhindern, so muß man innerhalb eines Genus sowohl Forderungen als auch Verbote aufstellen. Es resultiert eine „Nur"-Bedingung: H << B—C R—S bedeutet die gleiche Forderung wie vorhin, zusätzlich aber sind alle H-Terms außerhalb des definierten Bereichs jetzt ausdrücklich verboten. Ferner gibt es reine Verbotsterms. H / / B—C R—S ist die genaue Negation des entsprechenden Terms mit * *; A < C bedeutet, daß alle Terms verboten sind, die mit A, B oder C beginnen. Nehmen wir nun das Beispiel aus Abbildung 4, so ist festzustellen, daß Genus A nicht vorkommen soll, B aber in BAA (und nur dort) verlangt ist; Formulierung deshalb: A < A B << A—A A—A. Ganz ähnlich sind die weiteren Formulierungen zu verstehen. Die Recherche selbst ist zum besseren Verständnis in verschiedenen Versionen in einem Teil des IDC-Speichers durchgeführt worden (Abb. 8).

ANFRAGE-KENNZCHNG.	ANZAHL DER ANTWORTEN	----------- STRUKTUR-FRAGETERMS -----------
31918	432	BAA EBA EBR RAL SAF WPA
31919	168	A□A B□□ A-A A-A C□D E□□ B-B A-R EBA EB/ B-Q EBR F□Q P□□ A-A L-L S□□ A-A F-F T□V W□□ P-P A-A X□X
31920	46	YRA□AB *C□ST* U□3--- ZTEEE□ R□----
31921	28	A□A B□□ A-A A-A C□D E□□ B-B A-R EBA EB/ B-Q EBR F□Q R□□ A-A L-L S□□ A-A F-F T□V W□□ P-P A-A X□X YRA□AB *C□ST* U□3--- YTA□DF EF*F□Q R*S□4-
31922	22	A□A B□□ A-A A-A C□D E□□ B-B A-R EBA EB/ B-Q EBR F□Q P□□ A-A L-L S□□ A-A F-F T□V W□□ P-P A-A X□X YRA□AB *C□ST* U□3--- ZTEFF□ F□----
31923	35	BAA EBA EBR RAL SAF WPA YRA□AB *C□ST* U□3--- ZTEEE* F*----

[A 765.8]

Abb. 8. Zahl der Antworten bei verschiedener Formulierung einer Frage. □ entspricht <.

Verwendet man lediglich die durch die Struktur geforderten Dreierterms, so führt die Suche zu über 400 Antworten. Nimmt man die Verbote aller übrigen Dreierterms hinzu, so sinkt die Zahl der Antworten auf weniger als die Hälfte. Mit Y- und Z-Terms allein, ohne Dreierterms, bekommt man 46 Antworten. Die übliche Formulierung ist in den Fragen 31921 und 31922 angegeben. Hier sind die Y- bzw. Z-Terms zusätzlich zu den Dreierterms verlangt, und man erhält 28 bzw. 22 Antworten.

Wir sind nun bisher davon ausgegangen, daß die zu suchende Struktur vollständig angegeben wird. Das ist keineswegs die häufigste Art von Anfrage, denn eine solche Verbindung ist ja auch über Register gut aufzufinden. Anders sieht es aus, wenn der Anfragende Partialstrukturen nennt. Ist in unserem Modellfall beispielsweise eine *beliebige* weitere Substitution des Benzolrings zugelassen, so können natürlich keine Terms mehr verboten werden; es bleiben nur die zu fordernden Dreierterms und die entsprechend modifizierten Bezirksterms. Für diesen Fall erhält man 34 Antworten.

Zwischen diesen Extremen – vollständige Angaben einerseits, Partialstruktur mit beliebiger Substitution andererseits – gibt es zahlreiche Zwischenformen, die bei der Formulierung oft in verschiedener Weise berücksichtigt werden können. Da das System von Chemikern konzipiert wurde, haben chemisch ähnliche Gruppen auch ähnliche Dreierterms, die leicht in verkürzter Schreibweise zusammen gefordert oder auch verboten werden können.

Bei der Codierung von Patentansprüchen stößt man oft auf Markush-Formeln, d. h. Strukturen, bei denen ein Grundkörper an einer oder mehreren Stellen mit verschiedenen, oft zahlreichen Substituenten alternativ besetzt ist. Bei der Speicherung solcher Formeln werden die alternativen Teilstücke auf besonders markierte Plätze gebracht und sind auf diese Weise einem Verbot in der Fragestellung entzogen. Hat man etwa eine an einer Position wahlweise mit Chlor oder Hydroxyl substituierte Verbindung eingespeichert, so soll diese auch gefunden werden, wenn die chlorierte Verbindung gesucht, OH aber verboten ist. Analog werden auch im Fragesatz alternative Gruppen mit besonderen Zeichen markiert, wobei wieder mehrere Alternativzentren möglich sind. Die Möglichkeit zu bequemen Alternativfragen ist für eine erfolgreiche Recherche sicher nicht ausschlaggebend. Eine Verbindung, die ein Alternativzentrum mit drei Substituenten besitzt, kann auch durch drei Einzelfragen leicht ermittelt werden. Hat man es aber mit drei Zentren und je sechs Substituenten zu tun, so ist man doch erleichtert, statt $6^3 = 216$ Anfragen nur eine einzige formulieren zu müssen. Bei zahlreichen anderen Systemen ist dagegen die Suche nach Markush-Formeln eine recht unerfreuliche Angelegenheit.

Bei der Recherche nach Reaktionen können in ähnlicher Weise sehr spezielle, aber auch allgemeine Fragen gestellt werden; meist wird man gleichzeitig Strukturbedingungen kennen und verlangen. Die Recherchenprogramme zur Struktur- und Reaktionssuche laufen zwar getrennt ab, die Ergebnisse können aber nachträglich korreliert werden.

Bibliographische Daten der verschlüsselten Veröffentlichung selbst sind nicht direkt recherchierbar, weil sie nicht mitgespeichert werden. Es kann also z. B. nicht nach Autoren gesucht werden. Auf dem Patentgebiet existieren aber Konkordanzbänder, die zu jedem Referat die zugehörige Patentnummer liefern. Mit Hilfe dieser Bänder und weiterer Dateien, die nähere Angaben zu den Patent-Nummern enthalten, kann dann z. B. verhindert werden, daß eigene Patente im Recherchenergebnis enthalten sind.

Die auf den Recherchebögen codierten Fragen werden auf Lochkarten abgelocht. Zur Recherche dient eine Datenverarbeitungsanlage IBM/360-50.

3.3. Fehlerprüfung

Alle Anfragen werden zunächst in einem Diagnoseprogramm einer Prüfung auf formale Fehler unterzogen. Die Anfragen werden mit ihrer Nummer und den Codierungen aufgelistet, bei gefundenen Unstimmigkeiten, z. B. Nichteinhaltung der alphabetischen Reihenfolge, wird ein entsprechender Fehlerkommentar ausgedruckt.

Lochkarten mit einer Fragenummer, zu der ein Fehlerkommentar gegeben wurde, werden vom Operator aus dem Fragekartenpaket entfernt und die restlichen Karten erneut der Diagnose unterzogen. Die nunmehr fehlerfreie Auflistung dieser Fragen dient der Dokumentation als Beleg für diesen Maschinenlauf. Das Diagnoseprogramm benötigt etwa 2 bis 5 Sekunden Rechenzeit.

3.4. Sequentielle Recherche

Sofort im Anschluß an das Diagnoseprogramm beginnt die eigentliche Recherche. Der gesamte Speicher wird sequentiell abgesucht, ein Verfahren, das auf den ersten Blick unrationell anmutet. Um so überraschender ist es zu hören, daß der Zeitbedarf für eine einzeln recherchierte Frage bei nur etwa 90 Sekunden liegt, aber bei gleichzeitiger Recherche von mehreren Fragen meist auf etwa die Hälfte, manchmal bis auf 20 Sekunden zurückgeht. Mit Zeitbedarf ist die CPU-Zeit gemeint, der Zeitbedarf der zentralen Recheneinheit, der Grundlage für die Kostenabrechnung ist.

3.5. Screening

Die hohe Suchgeschwindigkeit wäre allerdings nicht möglich, wenn wirklich jeder Term jedes Speichersatzes überprüft würde; sie wird nur durch das „Screening" erzielt, einen Kunstgriff, den *Ernst Meyer* einführte und durch den die vorher sehr zeitaufwendige und damit kostspielige maschinelle Recherche zu einer relativ billigen Routine geworden ist.

Beim Screening wird im Speicher von vornherein jedem Speichersatz ein Maschinenwort mit 32 bits vorangestellt. In diesem Wort ist für jeden möglichen Genusbuchstaben von A bis X ein Bit reserviert; kommt jetzt etwa Genus C im Speichersatz vor, so wird durch ein besonderes Programm schon bei der Einspeicherung in diesem Screen-Wort das 3. Bit, das für C reserviert ist, belegt. Bei der Fragenaufbereitung wird in einem Fragescreenwort (ebenfalls maschinell) für alle *verbotenen* Genera ein Bit belegt. Bei der Recherche werden dann zunächst in einer sehr schnellen Maschinenoperation das Frage- und Speicherscreenwort miteinander verglichen (Abb. 9).

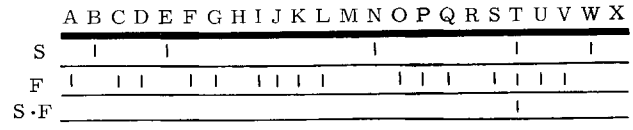


Abb. 9. Screening. Im ersten Screenwort werden die Genusbuchstaben des Speichersatzes S (hier also B, E, N, T, W) mit den im Fragesatz F *verbotenen* Genera (hier also A, C, D, F, G, I—L, O—Q, S, T, U, V) verglichen. Das logische Produkt S · F der beiden Sätze muß Null sein. Im Beispiel ist das nicht der Fall: T ist im Speichersatz vorhanden, im Fragesatz verboten.

Sind beide irgendwo an der gleichen Stelle belegt, so bedeutet das ja, daß der Speichersatz ein Genus enthält, das in der Frage ausdrücklich verboten wurde. Es lohnt sich also nicht, diesen Speichersatz näher zu prüfen; der Computer überspringt ihn und wendet sich dem Screenwort des nächsten Satzes zu. Neben diesem Screenwort gibt es vier weitere, die geforderte Kombinationen von Erst- und Zweitbuchstaben oder von Zweit- und Drittbuchstaben der Dreierterms überprüfen. Schon das erste Screenwort erlaubt es aber, im Mittel 99% aller Speichersätze ohne nähere Prüfung zu verwerfen.

3.6. Antwortenausgabe

Die vom Computer gefundenen Antworten werden auf einem Magnetband zwischengespeichert, nach Frage-Nummern sortiert und dann in einer Liste ausgedruckt. In der Dokumentation suchen Hilfskräfte die entsprechenden Referate heraus und kopieren sie oder vergrößern sie vom Mikrofilm. Die kopierten Referate werden in der Dokumentation gesichtet und nach Zutreffendem oder Ballast sortiert. Der Anfragende erhält beides, vor allem weil auch unzutreffende Verbindungen strukturell dicht beeinander und bei der gesuchten Verbindung angesiedelt sind und deshalb immerhin interessant sein *könnten*. Die Computer-Liste verbleibt bei der Dokumentation und wird abgelegt, ebenso die Lochkarten der Frage.

Übersteigt die Zahl der Antworten die Zahl 200, so wird man gewöhnlich zurückfragen, ob der Frager wirklich so viele Antworten will oder ob er die Frage enger fassen kann. Eine durchschnittliche Zahl der Antworten anzugeben, ist zwar möglich, aber wenig sinnvoll. Das leuchtet ein, wenn man bedenkt, daß zahlreiche Fragen nach Strukturen vorkommen, die als neu zum Patent angemeldet werden sollen. In solchen Fällen wird also die Antwort „Keine Unterlagen vorhanden" ersehnt und auch meist erhalten. Sehr hohe Antwortzahlen versucht andererseits die Dokumentation zu verhindern, so daß die Hauptmenge der Anfragen 0—100 Referate liefert.

3.7. Neueinspeicherungen, Abonnementsfragen

Etwa alle drei Wochen liefert die IDC auf Magnetband Neueinspeicherungen. Bevor diese dem Hauptbestand zugefügt werden, was wegen der sequentiellen Anordnung durch einfaches Anhängen an den Bestand möglich ist, wird eine Recherche des Neuzugangs mit allen vorliegenden Abonnementsfragen durchgeführt.

Diese Art von Daueraufträgen ist durchaus als eine Form der oft propagierten Selective Dissemination of Information, also als ein SDI-Dienst, anzusehen. In vielen Fällen wird eine Abonnementsfrage breiter gehalten als die ursprüngliche Frage an das ältere Material. Bei uns laufen stets etwa 60–70 Fragen im Abonnement.

4. Kritische Daten für ein Dokumentationssystem

Neben der Kenntnis von Codierung sowie Recherchenablauf und -möglichkeiten in einem System sind für jeden Dokumentar andere Punkte von hohem Interesse:

1. Erfassungsbereich (welche Veröffentlichungen werden codiert?)
2. Analysentiefe (wie genau wird gearbeitet? Wird jede genannte Verbindung auch verschlüsselt?)
3. Ballastanteil und, eng damit zusammenhängend, die Gefahr, relevante Dokumente nicht wiederzufinden
4. Fehlerquote
5. Zeitbedarf: a) für die Einspeicherung, b) für die Fragenverschlüsselung, c) für die maschinelle Recherche
6. Kosten für: a) die Einspeicherung, b) die Recherche.

Diese Punkte sollen im folgenden kurz diskutiert werden.

4.1. Erfassungsbereich und Analysentiefe

Den Vorwurf, zu wenig zutreffendes Material zu liefern, muß jede Dokumentation hinnehmen. Die IDC hat kaum Material vor 1959 codiert; die zur Codierung benutzten Quellen erfassen andererseits nicht vollständig die interessierende Literatur, während die Patente schon besser abgedeckt sind.

Als Quellen dienen zur Zeit:

a) der Fortschrittsbericht der Farbenfabriken Bayer, der etwa 600 Zeitschriften auswertet,

b) der Chemische Informationsdienst (vor 1970 die Schnellreferate des Chemischen Zentralblatts) und

c) der Patentschnellbericht, der von der Badischen Anilin- und Soda-Fabrik, den Farbwerken Hoechst und den Farbenfabriken Bayer herausgegeben wird. Er enthält Referate der Patentdokumentationsgruppe (PDG) von chemischen Patenten aus den Ländern Belgien, Frankreich, England, den Niederlanden, Österreich, der Bundesrepublik Deutschland, der DDR, der Sowjetunion, den USA und Japan.

Beim Chemischen Informationsdienst wird nicht das Referat, sondern die ursprüngliche Veröffentlichung für die Codierung zugrundegelegt. Bei allen Quellen werden stets alle genannten Verbindungen verschlüsselt (mit einigen Ausnahmen, u.a. für besonders häufige Lösungsmittel, die man nicht gern als Rechercheantwort erhalten möchte).

4.2. Ballast

Die Ballastmengen („false drops") im IDC-System hängen stark von der gesuchten Verbindung ab. Man kann nur einige halbwegs brauchbare Regeln dafür aufstellen:

1. je komplizierter die Struktur ist, desto geringer wird der absolute Ballastanteil,
2. je allgemeiner die Struktur formuliert wird, desto höher wird der Ballast, weil die Zahl der verbietbaren Terms abnimmt.

Nicht vermeidbar ist systembedingter Ballast. Bei diesem lassen sich z.B. Isomere oft nicht trennen. 3-Octanol unterscheidet sich nicht von 1- oder 2- oder 4-Octanol. Auch bei Angaben von Kettenlängen läßt sich nicht aussagen, welche von mehreren Ketten gemeint ist. Bei manchen Heterocyclen kann der Sättigungsgrad nicht genau beschrieben werden. Eine Abhilfe ist sofort möglich, wenn topologisch gesucht werden kann.

Da die IDC in Zukunft nahezu alle Formeln in Form einer topologischen Verschlüsselung speichern wird, aus der die Dreier- und Bezirksterms maschinell erzeugt werden, steht für solche Fälle die topologische Suche als Feinselektionsmittel zur Verfügung. Diese gegenseitige Ergänzung beider Methoden ist für uns ein entscheidender Faktor für zukünftige Planungen in der Dokumentation, denn sie bewirkt, daß die älteren und zukünftigen Einspeicherungen mit den gleichen Methoden befragt werden können. Die topologischen Suchmöglichkeiten machen den Speicher zukunftssicher: Auch bei den zu erwartenden riesigen Datenbeständen muß die Dokumentation weder kapitulieren noch von vorn beginnen.

4.3. Verluste und Fehlerquote

Für die Inversion des Ballastproblems, nämlich den Verlust dessen, was früher gespeichert worden ist, legt fast jede Steilkartei, die nach sachlichen Gesichtspunkten geordnet ist, Zeugnis ab. Die Verlustquote wächst gewöhnlich mit der Größe der Kartei und fällt dann wieder auf Null, weil niemand mehr die Kartei befragt. Im IDC-System ist diese Gefahr gering. Verluste treten praktisch nur durch Codier- und Ablochfehler auf. Die Ablochfehler werden durch Vierfach-Ablochung weitgehend und rechtzeitig erkannt. Viele Schreib- und Codierfehler lassen sich durch gezielte Fehleranfragen (z.B. nach nichtexistierenden Terms oder Termkombinationen) ermitteln. Der verbleibende Fehler-Rest ist nicht genau bekannt, liegt aber wohl unter 1% der Terms. Auch das bedeutet aber nicht, daß diese Verbindungen dann auch grundsätzlich nicht gefunden werden können.

Bei der Recherche können ebenfalls Codier- und Ablochfehler auftreten; überdies muß der Dokumentar alle Formulierungen kennen, die in früheren Stadien des Systems benutzt wurden. Einige Verschlüsselungs-

möglichkeiten existierten in den ersten Jahren nicht; manche Fragen müssen deshalb an verschiedene Bereiche des Speichers verschieden gestellt werden.

4.4. Zeitbedarf

Verzögerungen bei der *Einspeicherung* waren früher zuweilen recht spürbar. Der Rückstand ist aber fast völlig aufgeholt. Die IDC rechnet für 1970 mit einem Zeitraum von 8–10 Wochen zwischen Eingehen der Referate und Auslieferung des Magnetbandes mit der Codierung an die IDC-Gesellschafter.

Der Zeitbedarf für die *Codierung einer Frage* ist sehr unterschiedlich. Er kann Minuten betragen, manchmal aber auch eine halbe Stunde ausmachen, wenn die Frage mit der Facettenklassifikation schlecht vereinbar ist. Zeitraubend sind fast immer Fragen mit pauschalen Substitutionsangaben, z.B. „0 bis 2 OH-Gruppen und 0 bis 3 Halogenatome im Molekül".

Die *Recherchenzeit* pro Frage wurde bereits mit 20 bis 90 Sekunden angegeben. Die in Assembler geschriebenen Programme haben einen Kernspeicherbedarf von 54 K bytes. Da hiervon nur 12 K auf die eigentlichen Recherchenprogramme entfallen, kann man auch mit kleineren Kernspeichern arbeiten, wenn auch nicht ganz so schnell.

Die Zeit für die Nachbearbeitung der Recherchen hängt naturgemäß von der Zahl der Antworten ab. In aller Regel erhalten die Chemiker aber die Referatkopien binnen 24 Stunden nach Stellung der Frage.

Um die hochgezüchtete Recherchentechnik genau zu kennen, bedarf es einer gründlichen *Einarbeitung*. Da das System aber logisch und gut durchdacht ist, kann es nach einer Anlaufzeit von 4 bis 8 Wochen auch von begabten Chemotechnikern beherrscht werden.

4.5. Kosten

Die Recherchekosten hängen vom jeweils benutzten Computer und dem Abrechnungsmodus des Rechenzentrums ab. Die angegebenen CPU-Zeiten geben jedem Interessenten genügende Anhaltspunkte; die Kosten für die Recherche fallen aber ohnehin gegenüber dem Mitgliedsbeitrag für die IDC wenig ins Gewicht. Hier sind zur Zeit von jedem Gesellschafter je Chemiker rund 1000 DM/Jahr zu zahlen, wobei aber eine Mindestzahl von 50 Chemikern zugrunde gelegt wird. Dieser Beitrag verringert sich, wenn ein Gesellschafter nur Literatur oder nur Patente recherchieren will, er sinkt natürlich auch, wenn sich die Gesellschafterzahl der IDC erhöht. Gesellschafter ohne eigenen Computer können bei der IDC recherchieren lassen.

5. Ausblick

Abschließend kann festgestellt werden, daß uns im IDC-System eine nicht gerade billige, aber zuverlässige und zukunftssichere Recherchenmethode zur Verfügung steht. Anfragen können recht bequem und flexibel verschlüsselt werden. In naher Zukunft soll auch die makromolekulare Chemie in den Erfassungsbereich einbezogen werden, ein weiterer Schritt zum noch weit entfernten Endziel der vollständigen und schnellen Dokumentation der chemischen Literatur der Welt.

Eingegangen am 13. Januar 1970 [A 765]

Vielseitige maschinelle Suchmöglichkeiten nach Strukturformeln, Teilstrukturen und Stoffklassen[1]

Von Ernst Meyer[*]

Das Prinzip der topologischen Formelcodierung und maschinellen Recherche wird kurz erläutert. Durch den Ausbau dieser Methode ist es jetzt möglich geworden, nicht nur nach beliebigen Teilen von Strukturformeln maschinell und ballastfrei zu suchen, sondern dabei auch Fragebedingungen an die Strukturformel zu stellen, die zwar wohldefiniert sind und vom Chemiker gern benutzt werden, sich aber nicht ausschließlich durch Elementsymbole und Bindungsstriche darstellen lassen.

1. Einleitung

Eine der wichtigsten Aufgaben der Dokumentation für die organische Chemie ist das Auffindbarmachen von Strukturformeln und Stoffklassen, die durch gemeinsame Partialstrukturen gekennzeichnet sind. Das geschah bisher meist mit „Fragmentcodes"; mit ihnen verschlüsselt man eine Auswahl von Strukturmerkmalen durch bestimmte Codesymbole, die dann einzeln oder in Kombination miteinander abgerufen werden können. Der Vorteil dieser Codierung liegt vor allem in der leichten Abfragbarkeit. Nachteile sind der intellektuelle Aufwand beim Verschlüsseln und vor allem die Beschränkung auf eine begrenzte Zahl von im Schlüssel vorgesehenen Merkmalen: Sucht man Stoffklassen mit Partialstrukturen, für die es kein eigenes Code-Symbol gibt, so muß man die Frage verfälschen und bei den Antworten Ballast

[*] Dr. E. Meyer
Ammoniaklaboratorium C 6
der Badischen Anilin- & Soda-Fabrik AG
67 Ludwigshafen

[1] Die Arbeiten wurden zum Teil von der IDC Internationale Dokumentationsgesellschaft für Chemie mbH, Frankfurt, unterstützt.