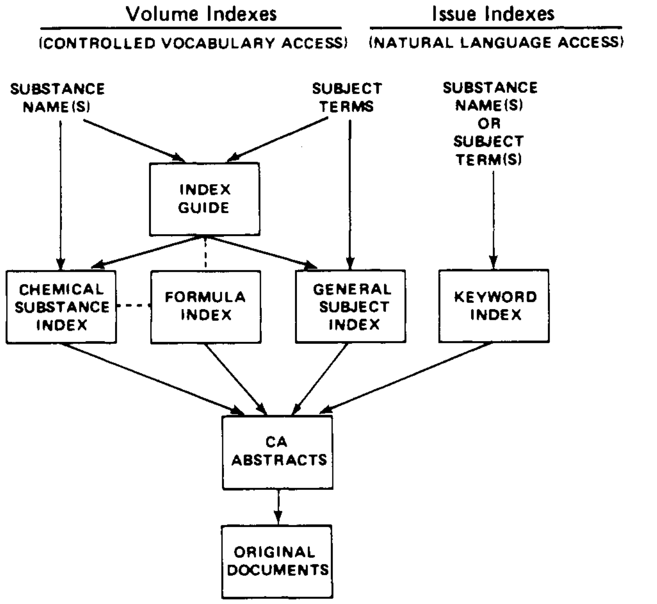


Figure 14. Access routes to documents through *Chemical Abstracts*.

the desired class of substances.

The hierarchies of General Subject Headings are found in Appendix I of the Index Guide. For each of 66 subject areas, CA index headings are listed in order of increasing specificity. An alphabetic index identifies the various hierarchies in which a given index heading may be found. Thus, a searcher who has a single topic in mind, e.g., synthetic fibers, may turn to it and find related headings, both more and less specific such as Fibers and Acetate Fibers, respectively (Figure 13). These will provide additional access points to the General Subject Index where information of interest may be found.

SUMMARY

Figure 14 summarizes access to the chemical literature using printed CA issue and volume indexes. Access is available weekly in the natural-language Keyword Index using the terminology current in a given field of chemistry for both substances and general subjects. Access with a greater depth of indexing is available in the semiannual, controlled-vocabulary Chemical Substance, Formula, and General Subject Indexes. The existence of controlled access points in these indexes requires use of the Index Guide for effective and efficient use of the indexes. All of these access points lead back to the CA abstract which identifies a document and summarizes its technical content, thus allowing printed CA to fulfill its role as "Key to the World's Chemical Literature".

The title of the original document, the name(s) of the document's author(s), all the bibliographic information necessary to identify the document, the Keyword Index entries, and all of the in-depth volume index entries are also available in *CA Search*, a new biweekly computer-readable service offered by CAS. Use of this file for computer-assisted access to the polymer literature will be the topic of a subsequent paper.

REFERENCES AND NOTES

(1) O'Dette, R. E. "The CAS Data Base", *Pure Appl. Chem.* **1977**, *49*, 1781-1792.

(2) O'Dette, R. E. "The CAS Data Base Concept", *J. Chem. Inf. Comput. Sci.*, **1975**, *15*, 165-169.

(3) Vander Stouw, G. G. "Computer Programs for Editing and Validation of Chemical Names", *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 232-236.

(4) Nelson, R. D.; Hensel, W. E.; Baron, D. N.; and Beach, A. J. "Computer Editing of General Subject Heading Data for *Chemical Abstracts* Volume Indexes", *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 85-94.

(5) Loening, K. L.; Metanomski, W.; and Powell, W. H. "Indexing of Polymers in *Chemical Abstracts*", *J. Chem. Doc.* **1969**, *9*, 248-251.

(6) American Chemical Society, "Subject Coverage and Arrangement of Abstracts by Section in *Chemical Abstracts*", 1975.

POLIDCASYR: The Polymer Documentation System of IDC[†]

ROBERT FUGMANN

Hoechst A.-G., 6230 Frankfurt/M, Postfach 80 03 20, Federal Republic of Germany

Received January 8, 1979

The IDC indexing system was adapted to the requirements of the polymer field. Polymer structures and the important syntactical relations between structures, including monomers, can be encoded precisely and searched as well. For nonstructural concepts (properties, etc.), a novel kind of controlled vocabulary, which can be supplemented continually by free terms, was developed.

INTRODUCTION

In the indexing and retrieval of polymer literature we encounter some problems that play only a minor part in the nonpolymer literature.

First, the *structure* of a polymer is less clearly defined than that of a nonpolymer compound. An indexing language for polymers must therefore cope with various degrees of structural uncertainty and vagueness and must not be limited to precisely defined structures (cf. refs. 1-3, 9-13).

Secondly, polymers most often are not defined structurally, but rather by their route of preparation. In such cases, one should not be obliged to resort to assumptions regarding the structure of the polymer. Rather, it should be possible to represent this specific route of preparation as precisely as possible. This includes indexing the monomers involved as well as the particularities of their *logical and syntactical relations* just as they are recorded in a document.

Thirdly, the indexing of properties, processes, and uses deserves particular attention in the polymer literature, for the essence of a document frequently consists precisely of these *nonstructural concepts*, and much less of statements regarding structures and syntheses.

Consequently, if one has at his disposal an indexing system developed specifically for the nonpolymer literature, he will encounter serious difficulties in using that system in the polymer field. This was the plight of IDC (International Documentation in Chemistry, Frankfurt/M., Hamburger Allee 26, Federal Republic of Germany) and, in particular, of our company as a member firm of IDC 11 years ago. At that time, it was also realized, that no indexing system was available that satisfied the relatively high requirements of the IDC firms with respect to the effectiveness of machine searches, in particular

[†] Presented before the Division of Chemical Information, Symposium on "Retrieval of Polymer Information", 176th National Meeting of the American Chemical Society, Miami Beach, Fla., Sept 13, 1978.

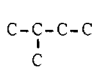
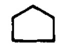
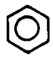
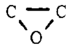
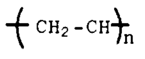
		Example
Chains of directly connected carbon atoms (nonpolymer)	"YR..."	
Aliphatic rings	"YS..."	
Aromatic rings	"YT..."	
Heterocyclic rings	"YU..."	
Chains of directly connected carbon atoms (polymer)	"Y6..."	

Figure 1. GREMAS terms for molecular regions.

with respect to the ratios of precision and recall. It was decided, therefore, that research and development work should be undertaken to close this methodological gap and also the gap in the literature coverage of IDC existing up to that point (cf. ref. 1).

In view of the expected complexity of the problem, three different committees of experts were established. These committees took up work in the three different areas that I have mentioned: structures, syntactical and logical relations, and nonstructural concepts. All the member firms of IDC supported this joint enterprise, at least through financial contributions. It became obvious very early that part of the results to be expected would be of general interest and applicable to other fields of scientific documentation as well. This part of the work was supported by additional governmental funds.

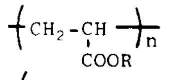
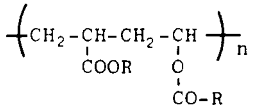
This paper presents an overview of the IDC polymer documentation system, for which we have chosen the acronym POLIDCASYR, which stands for Polymer Documentation System of IDC with Inclusion of Analytical and Synthetic Concept Relations".

POLYMER STRUCTURES

As far as the indexing of chemical structures and reactions is concerned, the GREMAS system[4] is well established among the IDC firms. In order to extend the indexing capabilities of the GREMAS system to handle the requirements of the polymer field, only a few enhancements had to be introduced into the system.

One of these supplements was the introduction of a new kind of term for molecular regions in addition to the terms already existing. In the GREMAS system we differentiate between four kinds of molecular regions, and thus have separate region terms for aliphatic chains of carbon atoms, for alicycles, for aromatic rings, and for heterocycles (Figure 1). What was missing in this set of regions were the very large chains of carbon atoms of statistically uncertain length, such as those formed in the polymerization of olefins. The introduction of a term for this particular molecular region constituted one of the most important enhancements to the GREMAS system. The numeral "6" was chosen to characterize these new terms. The main purpose of all these terms is to record the kind and number of substituents occurring in the corresponding region. It is common to these terms that they begin with the character "Y". They are differentiated by their second character, which is "R" for the chain of carbon atoms, "S" for alicyclic rings, "T" for aromatic rings, and "U" for heterocyclic rings, as shown in Figure 1. The kind of substituent attached to such a molecular region is recorded in the corresponding region terms together with the number of occurrences of such a substituent.

The length of these terms is variable, so that as many substituents can be recorded as are present in a region. Let the character "E" represent a hydroxy group or its derivatives. Likewise, "N" may represent a carboxylic group or carboxy derivative. Then we arrive at the region terms for our examples as shown in Figure 2. How often a particular region occurs in a molecular structure is also recorded. This is expressed simply by adding the corresponding numeral or range of numerals to the region term. This is demonstrated for the example of the dimer of ethylene glycol.

$HO-CH_2-CH_2-OH$	"YREE"
$HO-CH_2-CH_2-O-CH_2-CH_2-OH$	"YREE·02"
$HO-CH_2-CH_2-\underset{OH}{CH}-COOH$	"YREEN"
	"Y6N"
	"Y6EN"

"E" indicates a hydroxy (derivative) group

"N" indicates a carboxylic (derivative) group

Figure 2. Examples of terms for molecular regions.

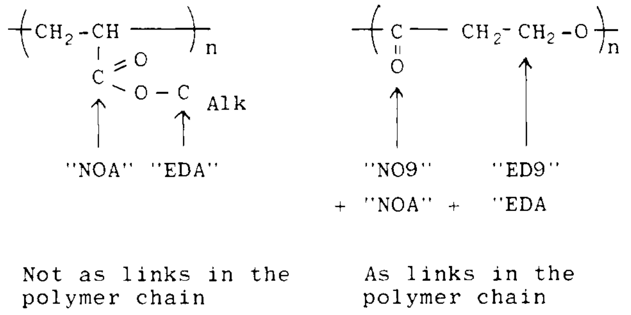

"NOA"
"EDA"
Alk
"NO9"
"ED9"
+ "NOA" +
"EDA
Not as links in the polymer chain
As links in the polymer chain

Figure 3. GREMAS terms for carbon atoms.

It is obviously impossible to make *quantitative* statements for the substitution of a chain of carbon atoms that has resulted from olefin polymerization. Instead, the substituents are recorded merely *qualitatively* in the term for such a molecular region. For example, acrylic homopolymers are encoded as "Y6N", and copolymers of acrylic acid and vinyl acetate are described as "Y6NE".

Another enhancement to the GREMAS system was the following. In phrasing a search for a polymer structure it is necessary to differentiate carbon atoms that are members of the "backbone" from those that are located in side chains. This was achieved by the introduction of a new kind of GREMAS term for carbon atoms.

In the GREMAS system, an individual carbon atom is represented by a term that consists of three characters. For example, the terms "NOA" and "EDA" are assigned to the carboxylic ester group in polyacrylic esters (Figure 3). The third position of such a three-character term for the carbon atom always indicates whether a carbon atom is a member of a chain or of an aliphatic ring, whether it is saturated or unsaturated, etc. The character "9" was introduced to denote that a carbon atom is a member of the polymer backbone. These terms are merely additive to the ones that were already in use for the nonpolymer field.

In addition to the structure codes for polymers, keywords of a general kind are used for classes of polymers. If a polymer is described only in very general terms, e.g., as an epoxy resin, aminoplast, phenoplast, etc., then this polymer is indexed solely under these keywords. As far as the nonpolymer structures are concerned, their connection tables can largely be me-

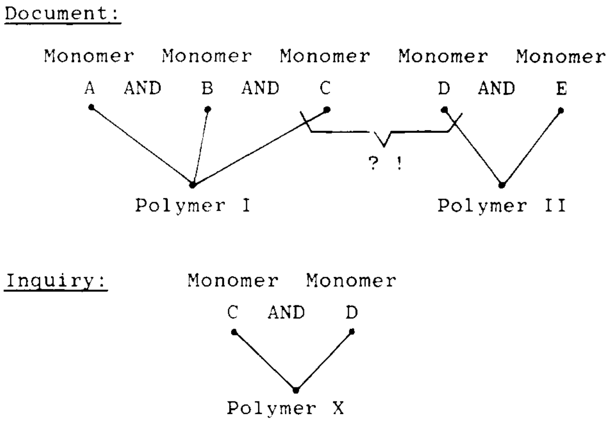


Figure 4. Syntactical relations between compounds.

chanically generated from the connection tables of Chemical Abstracts Service (cf. ref 4, pp 118 and 201).

SYNTACTICAL AND LOGICAL CONCEPT RELATIONS

As already mentioned, polymer structures are encoded only when they are explicitly presented in a document. This leads us to our second problem, namely, the topic of representing the *syntactical and logical relations* between chemical compounds and the processes that may be described in a document. If our aim is to retrieve polymer literature from a computerized file with an adequate degree of precision, then it is obviously not sufficient merely to enumerate all the polymers and the monomers presented in a document. For instance, two different polymers may be described in a document, each resulting from the polymerization of a typical set of monomers. This situation is depicted in the graphs of Figure 4 for polymers I and II. Polymer I is synthesized from the monomers A and B and C. Polymer II requires a different set of monomers consisting of D and E.

Now, imagine a search for the polymer X, to be synthesized from the monomers C and D. Obviously, the model document is *not* a relevant answer to the question. It is true, that the use of the monomers C and D for polymer synthesis is described within the document, but the monomers occur in the wrong syntactical relation in this document. They belong to different polymers.

In the long run, one cannot solve this problem by establishing a vocabulary of polymer names, each name comprising a specific combination of two, three, four, or more monomers. An example is the descriptor "Butadiene–Acrylonitrile–Styrene Copolymers", for the number of conceivable copolymers is almost unlimited. A vocabulary of names for all these polymers would have to be very large from the outset and would continually and rapidly increase. The size of such a vocabulary and its growth rate would be serious obstacles to reliable handling. The retrieval of polymers and especially of polymer classes would be correspondingly unreliable. We shall discuss this general problem in some detail in our third topic, which deals with a vocabulary of nonstructural concepts.

Let us consider still another model document that is common in the polymer field (Figure 5). Polymer III is formed by the copolymerization of monomer A with either monomer C or monomer D. In this case, it is crucial to represent not only the purely syntactical relation between the monomers, but also their important logical relation. Let us assume, that a search is made for a polymer Y composed of monomer C *and* monomer D. In this case, the model document is *not* a relevant answer, since either C or D, but always exclusively only one of them, was used for the polymerization. This mutual exclusiveness is expressed in the graph by the small circle to which the concept nodes are attached. The bold dot is used as a connective node between two concepts to denote the logical AND. These are conventions agreed upon in the TOSAR system[5] which is the graphical system specifically developed for representing logical and syntactical relations between concepts. Of course, all these relations can also be used as search parameters in a similarly organized search graph.

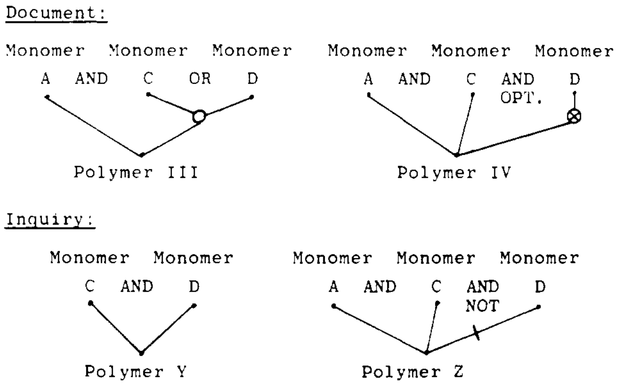


Figure 5. Logical relations between compounds.

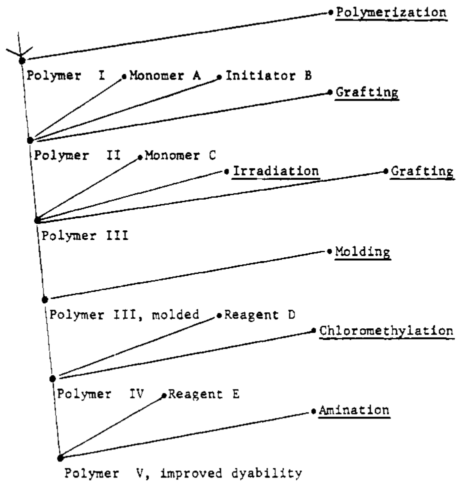


Figure 6. Succession of processes.

Quite common to the polymer field are searches in which the *absence* of a certain monomer is desirable in a particular polymer of interest. One may be interested, e.g., in copolymers of butadiene, acrylonitrile, and a third monomer *other* than styrene. For such a query our model document would constitute a perfectly relevant answer. This is true, even though the undesired monomer is mentioned explicitly in the document. The point is that it is recorded only as one possible monomer, not as an essential one. This can be recognized and taken into account in a TOSAR search.

Let us finally consider another essential feature of a document, which is again of a markedly syntactical nature, and which cannot be represented by any kind of indexing language vocabulary, without that vocabulary being severely overtaxed. I refer here to the *sequence in time* in which the various processes and reactions may take place during the preparation of a polymer. Again, the very graphical display technique of the TOSAR system is useful for representing these essentials of a model document (Figure 6). After a period of familiarization, one can recognize clearly from such a graph that, following an initial polymerization, two different grafting processes take place. The first one is carried out with monomer A in the presence of initiator B. The second grafting which is later in time is initiated by irradiation, and a *different* monomer is used. After the polymer formed in this way is molded, it is chloromethylated on the surface and then aminated. The product has several favorable properties, one of which is recorded in the diagram. In the natural language text of an author, these properties may be scattered throughout the entire text, including title and footnotes. For the IDC indexer, it is obligatory to collect all the properties at the place of the corresponding substance. This improves considerably the conceptual transparency and lucidity of the entire representation. That such a graphical representation is conducive to conceptual transparency will not come as a surprise to the chemist. It is our common experience that the structural formula, which is likewise of a graphical nature, is often much more lucid and easier to comprehend than the text of the corresponding nomenclature. This suggests the analogy that the TOSAR graph is nothing other than a special kind of "structural formula", not for an individual compound, but for the contents of a document.

These graphs are so helpful in comprehending a text, especially of complex patent claims, that an indexer often diagrams the contents of a document in the form of such a graph, even when there is no intention of putting the graph itself into the search file.

The indexing of a document by means of a TOSAR graph bears a strong resemblance to a kind of indexing recommended by Skolnik.[6] This is obvious from the fact that his "multiterms" could be derived by a purely mechanical linearization process from our graphical representation.

NONSTRUCTURAL CONCEPTS

The third major part of POLIDCASYR is the vocabulary of nonstructural terms. The indexing of nonstructural concepts in POLIDCASYR is based on a vocabulary consisting of two different parts designed to complement each other (cf. ref 4). One is a preestablished, hierarchically structured vocabulary of about 2000 coded terms. As a supplement to this vocabulary, there is an open-ended vocabulary with particularly specific terms. They are freely taken from natural language and subject to continual revision. These terms are used in addition to the preestablished ones. But this use is restricted to cases in which the specificity of the preestablished terms does not suffice to represent a specific concept or subject. An example is the encoded vocabulary term "isomerization" and some of its more specific natural language terms, e.g., "racemization", "tautomerization", "rotation isomerization", etc.

Reliable searching of the patent literature requires a particularly reliable kind of indexing. Specifically, it must be guaranteed that the terms which the vocabulary provides and even suggests for the search were, in fact, assigned reliably to the relevant documents in the file. We call this kind of indexing "*mandatory* indexing" to distinguish it from merely controlled or even free indexing. It is typical of mandatory indexing that the indexer continually search for the most appropriate terms in the vocabulary. This requires that the vocabulary be organized in the utmost conceivable degree of order and clarity. Otherwise, the indexer would be overtaxed if expected always to select the most appropriate vocabulary terms for a subject or concept to be indexed (cf. ref 7).

On the other hand, an open-ended vocabulary is inherently a nonmandatory one. But it can efficiently compensate for the unavoidable lack of specificity in a mandatory vocabulary should a loss of relevant information be of minor concern to an inquirer. POLIDCASYR presents the possibility of using both kinds of vocabulary and thus meets the divergent requirements in a community of users.

In the following we discuss two of several measures taken in the POLIDCASYR system with the aim of establishing order in a preestablished vocabulary so as to enable the indexer to use it in a mandatory mode, even under the pressure of time.

```
1. Substances
2. Apparatus
3. Living entities
4. Processes, reactions
5. Properties
     process-related
     non-process-related
6. Fields of application
```

Figure 7. Semantic categories in the "mandatory" vocabulary of POLIDCASYR.

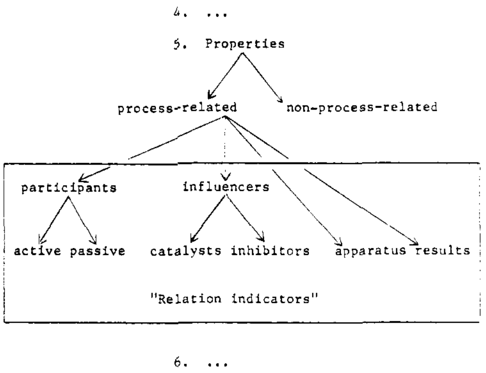


Figure 8. Subdivision of the properties in the mandatory vocabulary.

SEMANTIC CATEGORIES

One of these order-creating measures is to arrange the vocabulary according to a small number of particularly important semantic categories. In the POLIDCASYR system a set of semantic categories is in operation as depicted in Figure 7. These semantic categories serve not only as an order-creating principle for the vocabulary, but they are also useful as a guide in deciding which kind of term can be admitted without reservation to the vocabulary and which different kind of term should be rejected. If a term represents a combination of categorical concepts, then the term should be admitted to the vocabulary only in exceptional cases. This rule prevents excessive growth in a vocabulary and in its network of concept relations (cf. ref. 7).

THE RELATION INDICATOR

Another order-creating principle is the so-called deep casus or relation indicator. This principle was discovered almost simultaneously and independently by Fillmore,[8] Diemer and Henrichs (cf. ref 7), and our working group at Hoechst.[7] This principle becomes obvious if we subdivide the property category into process-related and non-process-related subcategories. The process-related ones can then be regarded as various manifestations of the corresponding process. For each process represented in the vocabulary we can, after some consideration, recognize a family of closely related concepts. This family comprises the active and passive participants (those that influence the process positively and negatively), the apparatus, and, most important, the results of the process under consideration (see Figure 8). For example, the family of the concept "oxidation" consists of the oxidant, the object that is being oxidized, the catalysts and inhibitors, the property of oxidation resistance, and finally the product of the oxidation (Figure 9).

Codes have been introduced for these relation indicators to ensure that any conceptual member of the family can indeed be represented, independent of whether natural language has already coined a concise term for such a concept. These relation indicators have proved invaluable for the construction and reliable handling of a vocabulary. For example, some consideration would reveal that they comprise simultaneously

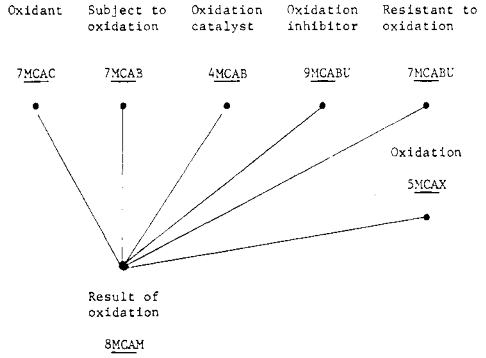


Figure 9. Notations for a family of relation indicators.

the functions of roles and links, and that they do this in a particularly efficient manner.

CONCLUSION

The POLIDCASYR system is another manifestation of the general philosophy of IDC, namely, to relieve the searcher from expending a large and continually increasing effort in scanning through myriads of irrelevant responses to the machine searches. Rather, the expenditure is shifted to a meticulous document analysis. Thus, the intellectual effort can be kept nearly constant in the course of time and also relatively low even if the system is used heavily.

Through the development of the POLIDCASYR system another goal was achieved, namely, to cover the entire field of organic chemistry with a *single* indexing language that is sufficiently general and effective. Admittedly, the full payoff of this approach will be realized only in the relatively distant future, and considerable far-sightedness is required to undertake such an investment in information science. We have, however, determined that such an approach is fully justified from an economic viewpoint, when it is concentrated on the literature of particular interest to a group of institutions, and when the input workload is shared by these institutions.

REFERENCES AND NOTES

(1) Suhr, C. "Neue Wege in der Dokumentation der Makromolekularchemie", *Chem.-Ztg.* **1972**, *96*, 342–347.

(2) Chemical Abstracts Service, Structure Input Manual, Chapter L (Polymers), Feb 27, 1975.

(3) DERWENT Publications Ltd.: Plasdoc Punch Code Manual 1975, Instructions for Coding and Searching.

(4) Fugmann, R. "The IDC- Systems". In "Chemical Information Systems", Ash, J. E.; Hyde, E., Eds., Ellis Horwood Lt., Chichester, England, 1975.

(5) Fugmann, R. "Representation of Concept Relations Using the TOSAR – System of the IDC", *J. Am. Soc. Inf. Sci.* **1974**, *25* 287–307.

(6) Skolnik, H. "A Multilingual Index Via the Multiterm System", *J. Chem. Doc.* **1972**, *12*, 128–132.

(7) Fugmann, R. "The Glamour and the Misery of the Thesaurus Approach", *Int. Classification* **1974**, *1*, 76–86.

(8) Fillmore, Ch. "The Case for Case". In Bach & Harms, "Universals in Linguistic Theory", Holt, Rinehart & Winston, New York, 1968; pp 1–88.

(9) Skolnik, H; Hays, J. T. "A New Notation System for Indexing Polymers", *J. Chem. Doc.* **1970**, *10*, 243–247.

(10) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Registry System", *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111–121.

(11) Loening, K. L.; Metanomski, W.; Powell, W. H. "Indexing of Polymers in Chemical Abstracts", *J. Chem. Doc.* **1969**, *9*, 248–251.

(12) Hoffman, W. S. "An Integrated Chemical Structure Storage and Search System Operating at Du Pont", *J. Chem. Doc.* **1968**, *8*, 3–13.

(13) Duffey, M. M.; Klanberg, I. M.; Mahr, S. C., Meier, L. L.; Romstad, J. L. "Computer Indexing of Polymer Patents", *J. Chem. Doc.* **1968**, *8*, 85–88.

Some Problems Encountered in Interdisciplinary Searches of the Polymer Literature[†]

L. GUY DONARUMA

New Mexico Institute of Mining and Technology, Socorro, New Mexico 87801

Received January 23, 1979

In a number of areas of polymer science (particularly those of a nontraditional nature), the handling of polymer information needs reassessment, redefinition, or definition. Some problem situations are the interfaces between (1) polymers and medicine and (2) polymer science and engineering as well as subfields within polymer science. Actual examples of problems encountered when attempting to retrieve information are cited, and the relationship of these problem situations to potential or pending solutions is discussed. Specific topics include polymeric drugs and other biomedical uses of polymers, polymer blends, polymer characterization, and inorganic polymers among others.

In order to make use of accumulated knowledge on such a vast scale as that collected in modern chemistry, ever more sophisticated indexing and abstracting systems have been developed and continue to develop. Apace with the development of these has been the construction of nomenclature systems. The accomplishments in indexing and abstracting have more or less taken the route of utilizing the nomenclatures of the various subdisciplines within chemistry. Divergencies exist in the nomenclature, and one of the most striking lies in comparison of the organic and the inorganic nomenclatures. The former is substitutive and the latter additive. That is to say that in devising a name for an organic compound, one does so by adding to a parent name (more often than not that of a hydrocarbon, but not completely so) the name of an atom or group of atoms taking the place of a hydrogen atom present in the parent, e.g., 2-chloropropane, chlorobenzene, etc. In an additive system one devises the name in the same fashion by which many German verbs and nouns are constructed, that is, by stringing together the names (or a close facsimile thereof) of the component parts of the compound, e.g., lead chloride, hexaamminecobalt(3+) chloride, etc. This type of nomenclature works well particularly when applied within the pertinent subdiscipline. However, when one moves into an interdisciplinary subdiscipline, such as polymer science, difficulties are encountered in devising nomenclature systems since we must take into account the fact that there are, for example, both inorganic and organic polymers and even semiinorganic or semiorganic polymers. Polymer charac-

[†] Presented before the Division of Chemical Information Symposium, "Retrieval of Polymer Information", 176th National Meeting of the American Chemical Society, Miami Beach, Fla., Sept 13, 1978.

0095-2338/79/1619-0068$01.00/0