Acid". *Nature (London)* **1953**, *171*, 737-738.
(22) Garfield, E. "*Citation Classics*—Four Years of the Human Side of Science". In "Essays of an Information Scientist"; ISI Press: Philadelphia, 1983; Vol. 5, pp 123-134.
(23) Pauling, L. Commentary on "The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry". *Curr. Contents/Phys., Chem. Earth Sci.* **1985**, *25* (4), 16.
(24) Ortega y Gasset, J. "The Revolt of the Masses"; Norton: New York, 1957; pp 110-111.

Peculiarities of Chemical Information from a Theoretical Viewpoint

ROBERT FUGMANN

Hoechst AG, 6230 Frankfurt/M 80, Federal Republic of Germany

Received January 25, 1985

Chemistry is exceptional among the sciences in that information in this field is, owing to the structural formulas, extraordinarily clearly and lucidly defined, particularly durable, and intensively used. Exceptionally high demands can and must therefore be made on the accuracy of the information supply in this field, and its indexing and retrieval techniques are correspondingly advanced. This is the outcome of the unusually great efforts that have been undertaken in chemical information science. Lawful connections between parameters that exert an influence on the accuracy of information supply *in general* are revealed most clearly in chemistry. Thus, chemical information science can also contribute to the development of more advanced techniques for a more accurate supply of information in other fields.

INTRODUCTION

In the field of chemistry today one can observe an intensively used worldwide information network. Large national information systems have also arisen and are beginning to cooperate at the international level. They exhibit a great capacity to bring together relevant information in response to a majority of the practitioners' search requests. Wherever their performance is not yet adequate, we know fairly well how to effect the needed improvement, and we can also quite accurately estimate the cost of the improvement. Local activities have developed where this need is particularly great or specific, and in some cases they have even taken the form of joint endeavors by chemical companies, although they are competitors on the market.

All this raises the question of what features are peculiar to chemical information to enable it to achieve a position of such prominence in science and technology. The answer to this question can lead to a better understanding and even better utilization of chemical information, and also to improvement in teaching in this field. It can also lead to cross-fertilization of the information supply process in other fields. In particular, theoretical principles that have been established in the area of chemical information could also bear fruit in other areas. In this paper we shall investigate from a theoretical standpoint some phenomena that are peculiar to chemical information and study the ways in which they have either constituted an extraordinary challenge to or considerably facilitated the development of adequate information services in chemistry. We shall also investigate some current and hitherto unresolved issues in information science and attempt clarification from a theoretical viewpoint.

These deliberations will substantially be based on the "Five Axiom Theory", essential features of which have already been published in several journals, including this one.[1-3] A detailed discussion of this theory is beyond the scope of this paper, and we must refer to the referenced papers when questions concerning the theory itself, its axioms, and the definition of the terms arise. The meaning of some core terms used in this theory can be looked up in the table on page 119 of reference 2. These terms are repeatedly marked with an asterisk in the text.

CLARITY AND INTERNATIONAL UNIFORMITY OF THE STRUCTURAL FORMULA LANGUAGE

Chemistry has at its disposal a language of such uniformity at the international level and of such clarity as is scarcely found in any other discipline. It consists of the structural formula. Even texts in an unfamiliar foreign language are relatively comprehensible for the chemist, if structural formulas are amply included in the text. Through their pictorial character they convey a structure concept in its greatest conceivable comprehensibility, clarity, and definitiveness. The structural formula also lucidly displays every kind of structural relationship between more or less closely related substances. This alerts the chemist to analogy inferences and at the same time overcomes the barriers to international communication that exist in other fields of knowledge.

Vagueness in the meaning of terms has often constituted a serious obstacle to communication among humans, and expert terminology does not constitute an exception to this rule. In the literature we encounter at least a dozen different definitions of the term "corrosion". When an inquirer, with one of these definitions in mind, searches the literature with this term as a search parameter, he will be directed to many publications that he will rate as "irrelevant" because they do not deal with the subject of special interest to him. If, in another example, "polyolefin" means both a polymer prepared from olefins and also a hydrocarbon with several olefinic double bonds, the difference in meaning is so great that an indexer must not permit the term to enter a search file without clarification and, sometimes, revision, i.e., translation into another term. Otherwise, the accuracy* of searches in the information system would be intolerably low. The various problems that are involved in any kind of meaning perception and translation are, fortunately, absent from the core of chemical information, the structural formula.

THE LONG-LASTING VALUE OF CHEMICAL INFORMATION

The development of chemical information* has attracted particularly great efforts and occasioned unusually large expenditures on the part of the scientific community. The long-lasting value of chemical information, one of its most

0095-2338/85/1625-0174$01.50/0 © 1985 American Chemical Society

Robert Fugmann obtained a Doctor's degree from the University of Hamburg in 1951, having worked in the fields of steroids and triterpenes. In 1952, he joined the Hoechst AG where he worked on organic syntheses in the pharmaceutical research laboratories and pilot plants. He became head of the company's central department of scientific documentation in 1960. He developed the documentation systems GREMAS and TOSAR and the Five Axiom Theory of indexing and information supply. He initiated the Division of Chemical Information in the German Association of Chemists (GDCh) and is presently chairman of their board. He received the Herman Skolnik Award of the American Chemical Society's Division of Chemical Information and the Gmelin-Beilstein Medal of the GDCh. He is a member of the advisory board of the IDC and of the board of the German Classification Society. He has lectured widely at home and abroad and has 55 publications in the scientific information field to his credit.

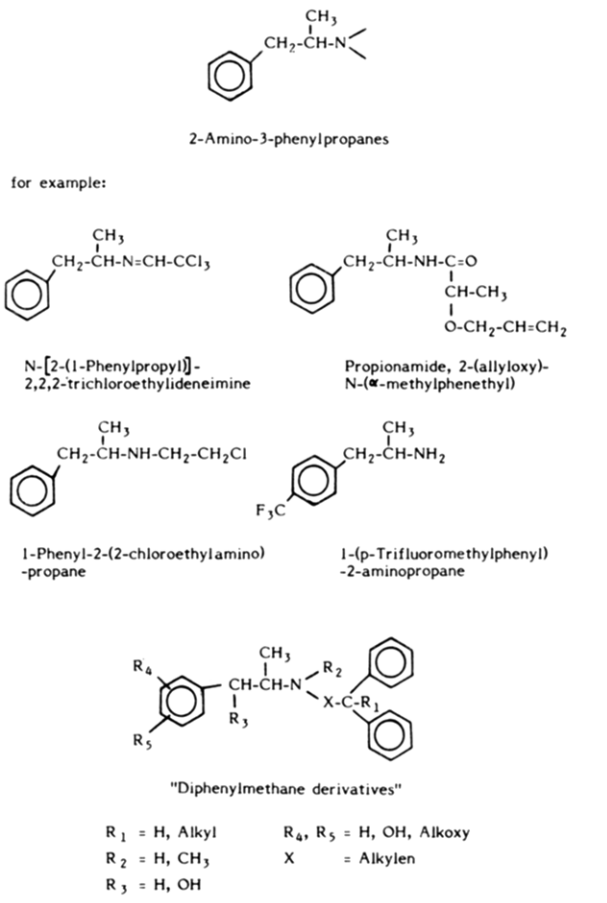

Figure 1. Heterogeneity of the names for compounds of a common compound class.

prominent features, has been the motivating force behind this effort. In chemistry it is worthwhile to recover the knowledge gained through experimentation even in times long past. As a rule, the description of experiments has been so accurate, at least in the last 100 years, that they can be repeated at any time and can be used as a foundation on which to base further research. The result is a vast saving of experimental work that would otherwise have to be undertaken. And, when new theoretical ideas have been developed, for example, concerning relations between the structure and the reactivity of chemical compounds, much experimental material is required to confirm, improve, or refute a tentative theory, and, thus, to achieve progress. No one could develop all this knowledge and carry out the required experiments through his own efforts. Here, the possibility of utilizing experience gained by others, even long ago, is of great value.

Consequently, search files of chemical information are particularly large. The largest abstracting organization in the world is Chemical Abstracts Service, which over the years published over 10 million abstracts. Not only are the chemical information files extraordinarily large, they are also normally searched from beginning to end, which is not the case in many other fields. Searching a file for only the recent past or only for the latest additions to the file significantly simplifies a search. In chemistry, however, only the most sophisticated search techniques can satisfy the demand.

DEMANDS MADE OF RETRIEVAL ACCURACY

Chemical information files are probably more frequently and intensively searched than any others. This is due to the intensive research work being conducted in chemistry. Only electronics and electrotechnology can compete with chemistry with respect to the proportion of sales revenues that is currently reinvested in research work.

It addes to the intensity of literature usage in chemistry that information* that is published in highly specialized fields may be useful to specialists in quite different fields. A chemist who works, for example, in the alkaloid field may encounter in the dye literature a cyclization reaction that may be of great value to him. Furthermore, chemists change the area in which they work fairly frequently, at least in industry. In order to familiarize himself quickly with a new area of work, the chemist requires rapid access to and accurate retrieval of the relevant literature. Ten thousand literature searches a year is not unusual in a large chemical company, provided the technical and personnel resources are available. Correspondingly high requirements are placed on the precision ratios of the searches. Otherwise, the searcher (or the intermediary to whom the searches are delegated) will be overtaxed and have to forgo searches and/or restrict them to smaller and smaller parts of the search file.

Particularly high demands are made on the order*-creating power of a retrieval system if high precision must be coupled with high recall ratios, as for instance in the field of chemical patents. But basic research in chemistry also demands the most complete information* possible if there are only few relevant documents in the file. Although it is sometimes stated that only *a majority* of the relevant references need be located, this attitude would quickly change if the inquirer had an opportunity to see the responses that had escaped his attention. Often an inquirer is convinced he has retrieved "complete" information from his file and is satisfied with the result. But such a statement should only be made if he has had an opportunity to compare his results with those from other files

or with those of a different query formulation from the same file. The mere "completeness" of the file searched can by no means be taken as a guarantee for correspondingly "complete" search responses.

It is unique to chemical documentation that loss- and noise-avoiding retrieval (cf. page 122 of reference 2) is not only a requirement but has already been achieved, at least in searches for structural formulas. This is made possible by the exceptional clarity of the structure concept and by particularly highly developed (even if relatively costly) indexing and retrieval techniques.

ANALYTIC–SYNTHETIC PRINCIPLE IN CHEMICAL INDEXING LANGUAGES

In the Five Axiom Theory precision and recall (and hence retrieval accuracy*) are related to representational fidelity and predictability (cf. pages 122 and 123 of reference 2). Uncontrolled natural language cannot provide a degree of predictability that satisfies even moderate requirements, at least as far as *general* concepts* and statements in storage and/or retrieval are concerned. This became apparent and was taken into consideration early in chemical documentation. The necessity of representing concepts in *indexing language* has therefore always been recognized.

For example, the uncontrolled natural langugage input of the names or nomenclature of chemical compounds would result in a file that could hardly be successfully searched for the members of compound classes.

All compounds in Figure 1 are members of the class of 2-amino-3-phenylpropanes, but no string of characters is common to them which could be used to retrieve them from a file of this type. It is also impossible to compile the (possible) names of all conceivable compounds of this class and to use them as search alternatives. In other words, the strings of characters used to represent this general concept in such a file are unpredictable, but it is exactly these representations that must be known in order to phrase a query for a file like this.

That concepts other than molecular structures should also be translated into an indexing language mode of expression for the sake of improved predictability is obvious from the following example. The preparation of a substance may be expressed in an original text as follows:

"The standard procedure was also applied to substance 11a which yielded the compound mentioned in the title" (namely, lysine).

"Several amino acids were isolated from culture broths of *bacillus subtilis*...The process proved superior to the conventional one...Prices for lysine are expected to drop as a consequence of this new process...."

"Decarboxylation and removal of the protective group led to lysine."

In other words, the concept of preparation can be presented in a text to be indexed not only through *lexical* expression such as "preparation" or "synthesis" but also through a great variety of *nonlexical* modes of expressions. It is difficult to imagine an algorithm that could correctly manage all these inherently unpredictable text passages that stand for the concept of preparation. It is not the computer but the human that would be overtaxed with such a task, for he would have to write a program to cope with an almost unlimited number of unpredictable situations. Not surprisingly, the programs so far proposed for this purpose have always proved inadequate on closer examination.

As far as the indexing of molecular structures is concerned, it is very informative to investigate why an approach that works with descriptors such as "phenothiazines", "steroids", "malic acid esters", "aliphatic 2-hydroxy carboxylic acid esters", etc. is inferior to the topological indexing approach, as will be apparent from the following consideration.

Our theory requires that an indexer should always use those descriptors from his vocabulary that *most appropriately* represent the concepts to be indexed (cf. page 123 of reference 2). Only through this kind of "mandatory" indexing* can that degree of representational predictability and fidelity be attained which an indexing language promises and which is expected by an inquirer. Merely "controlled" indexing* does not meet this requirement, because it gives freedom to an indexer to choose *any* more or less appropriate vocabulary descriptor. The consequences of this kind of freedom are similar to those that would prevail in a department store where personnel feel free to display incoming articles at *any* more or less appropriate location and to disregard the directory displayed to customers.

The search for the most appropriate descriptors in a vocabulary may be seriously impeded or even rendered impossible in everyday practice if the vocabulary and the relational network that is woven into it are large and difficult to survey (cf. page 10 of reference 1). Then the indexers will have to content themselves with *some* more or less appropriate descriptor that happens to be encountered or happens to come to the indexer's mind. Keeping a vocabulary within the boundaries of its manageability is therefore a requirement for good indexing and for achieving the degree of search accuracy that is promised by the indexing language and expected by the searcher.

What contributes most to the size of a vocabulary and of the relational network woven into it (and thus to its conceptual opacity) are those (composite) concepts that comprise several conceptual constituents that are themselves already represented in the vocabulary. These concepts are also responsible for the continuous growth of a vocabulary, an occurrence that will be observed if *every* concept of interest has access and claims a descriptor of its own. The alternative of *analyzing* these concepts into their conceptual constituents (semantic factoring) *and* at the same time representing the connectedness that prevails among these constituents through a successive conceptual *synthesis* constitutes an effective countermeasure against vocabulary hypertrophy. This "analytic–synthetic" approach has been emphatically recommended as a generally applicable procedure for all fields of science by Ranganathan and his Indian school.[4]

The topological approach to the documentation of molecular structures constitutes a classic example of the analytic–synthetic principle. In chemistry, this approach is literally predesigned by nature. The conceptual constituents of a molecule can only be its atoms, which are represented in the "*vocabulary*" of the periodic table of the elements. The *grammar* of the topological indexing language is represented by the connection table, which displays the syntactic relations prevailing among the individual atoms.

One obstacle to the widespread acceptance and application of this principle in other fields is that the conceptual constituents are less obvious than in those chemistry. But the device of employing semantic categories (cf. page 123, chapter on the axiom of representational fidelity, of reference 3) can also serve in these other fields as a guide to indicate the direction and the extent that conceptual analysis should take. In principle, all descriptors that stand for concepts comprising constituents from more than one single category should be excluded from the vocabulary. It is exactly this principle that is followed in the topological indexing language (though largely subconsciously), if we consider, as we do here, the chemical elements as belonging to an individual category of their own. Categories other than that of the substances that we could recommend for the field of applied chemistry are those for living entities, apparatus, processes, and operations, as well as the properties of these objects, processes, and operations.

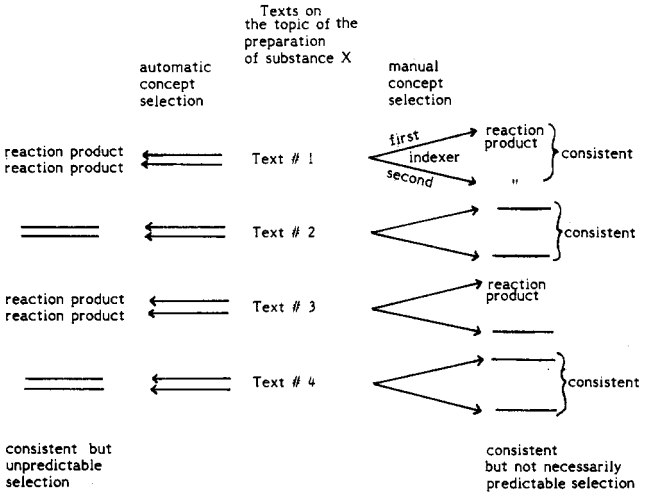

Figure 2. Consistency vs. predictability in concept selection during indexing.

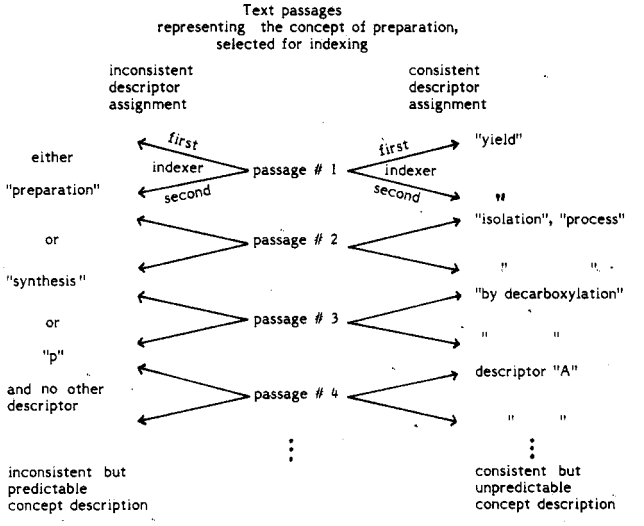

Figure 3. Consistency vs. predictability in concept description during indexing.

Thus, chemistry has been a pioneer in demonstrating the capabilities of the analytic–synthetic principle by developing it to a high level of maturity for use on a large scale. But even in chemistry the capabilities of this principle have not yet been fully exhausted. For the documentation of chemical reactions in applied chemistry, the analytic–synthetic principle is still awaiting employment.

CONSISTENCY VS. PREDICTABILITY IN INDEXING

In the Five Axiom Theory, indexing* is defined as the process of (a) discerning the essence of a document and (b) representing this essence in an indexing language mode of expression, i.e., with a sufficient degree of predictability and fidelity.

Indexing serves the purpose of searching, and, when an inquirer phrases a search for his topic of interest, he must know if he can expect the concepts of his search to be selected for the index by the indexers. Only then can he unhesitatingly phrase these concepts as search parameters; as otherwise, he could lose relevant* responses in his search. For example, if the searcher cannot rely on the indexers (or on the indexing algorithm) to have assigned the descriptor "product in a chemical reaction" (e.g., "P") to each substance having this function, then the searcher would be well advised to omit this search parameter from the query.

One kind of indexing consistency that has often been considered as a measure of indexing quality is consistency among indexers in selecting the same concepts from the same document.[5-7] However, one particular text, e.g., number 1 in Figure 2, may be treated perfectly consistently by several indexers, for example, when the expression "reaction product" expressly occurs in the text (cf. Figure 2, right-hand column). In another text (number 2) this context may not be expressly stated, and the indexers (or the indexing algorithm) may, again, perfectly consistently fail to assign this descriptor. An algorithmic selection of concepts may, in spite of its inherently perfect *selection consistency* (cf. Figure 2, left-hand column), also fail to assign this descriptor in certain cases.

Following this kind of consistent indexing a search with (among others) the parameter "reaction product" will exclude several relevant* documents, e.g., number 2. Obviously, this kind of indexing is not conducive to accurate retrieval. If, on the other hand, the indexers disagree even in the selection of concepts from the same document (e.g., number 3 in Figure 2), they will certainly do so in the case of different documents. Consistency in the selection of concepts for the index is therefore a necessary but not a sufficient condition for accurate retrieval. Instead, it is *selection predictability* that should be aimed at. This can be achieved by making it obligatory for the indexers to select the concepts from certain predetermined semantic categories, as has been mentioned.

Full-text storage would not solve the problem of appropriate selection, because the concept "reaction product" may be implied in a text in an entirely unpredictable mode of expression and would often escape detection in retrieval. Examples of such cases were given earlier. Our present and foreseeable techniques of text analyses are far from being able to promise an algorithmic solution that satisfies more than merely moderate demands.

The second step in indexing, namely, the *description* of the previously selected concepts, could easily be rendered most consistent by merely adopting the modes of expression encountered in the original text (cf. Figure 3, right-hand column). Then, for example, the descriptors* "isolation" and "process" would be consistently assigned to the second of the above-mentioned original texts, as would perhaps "decarboxylation" to the third, differently phrased text, and so on. After this, the searcher can only guess which particular descriptors would have to be used in the query in order to address each of the many, differently phrased and differently and unpredictably (though most consistently) indexed texts in the file, all dealing with the same topic, namely, the preparation of a certain compound.

Hence, a kind of *description consistency*, which is at least achieved in part by the adoption of expressions from the original text, is not only not conducive but even detrimental to accurate retrieval because it impairs representational predictability. This holds true at least for the indexing of general concepts* and statements. Inconsistent indexing of the kind in which an indexer has a choice among a limited number of predetermined (controlled) descriptors would be preferable, because it would render the representation of concepts more predictable (cf. Figure 3, left-hand column). Thus, *description consistency* is found to be neither a necessary nor a sufficient precondition for accurate retrieval. It is even harmful if achieved by the mere extraction of words from a text.

Consequently, overall indexing consistency, applied to both indexing steps, cannot be a useful measure of indexing quality, if "consistency" means that the repetition of a process will always yield the same result. Instead, *representational predictability* should be aimed at in both indexing steps (in combination with a sufficient degree of representational fidelity) and should be used as a criterion for good indexing instead of indexing consistency.

Quantifying representational predictability for the purpose of assessing and comparing indexing quality is a task still to be accomplished.

"INVERSE RELATION" BETWEEN PRECISION AND RECALL

It has been repeatedly asserted that some kind of "inverse relation" exists between the precision and recall ratios such that enhancement of recall necessarily results in deterioration in precision and vice versa. Scepticism has been expressed concerning such a postulate, in particular from the statistical viewpoint with regard to the veracity of the recall-precision graphs.[8-10] This issue is far from being settled.

In the Five Axiom Theory precision and recall are related to representational fidelity and predictability in indexing, respectively. An inverse relation between precision and recall should inversely link fidelity with predictability. Again, it is chemistry that convincingly teaches us that no lawful relation such as this can exist, although such a relation seems to occur in practice, at least under certain circumstances.

Molecular formulas, for example, may be looked upon as an indexing language*for structural formulas. As a language, it suffers from an underdeveloped syntax* (which would have to represent the connectivity prevailing between the individual atoms of a molecule) and, hence, exhibits fairly low representational fidelity. Low precision ratios are the consequence in searches for molecular structures on the basis of molecular formulas. But without any decrease in representational predictability, the fidelity of this kind of indexing can be drastically improved by the introduction of more grammar into the indexing language, i.e., through transition to topological representations. The consequence is drastically improved precision ratios at no decrease in recall. Hence, precision can be increased without at the same time impairing recall.

Of course, the syntactic device used for this purpose must also meet the requirement of predictability. Otherwise, the use of syntactical search parameters would necessarily cause a loss of relevant responses, and the undesirable, empirically observed inverse effect on recall may well occur.

An obvious example is a link, the use of which is dependent on the contingencies and unpredictabilities of natural language text structure. Substance and substance property, for example, may be widely separated in an original text, and their spatial proximity or separation is far from being an indicator of their logical relationship. Requiring the two concepts to co-occur in a common sentence or even within a minimum distance in a sentence will necessarily lead to loss of relevant* texts when more precision is sought by the use of this syntactic device.

On the other hand, in the light of our theory several examples come to mind in which representational predictability can be increased without at the same time impairing representational fidelity. One way is to base one's indexing work on a set of semantic categories and thus make the *selection* of concepts for the index more reliable and predictable. Then, a searcher will less frequently use search parameters for concepts that are only occasionally selected and entered in the search file, and which, when phrased as search parameters, will cause a loss of relevant documents. The consequence is an increase in recall without a concomitant decrease in precision.

If indexers are experts in the field of the literature to be indexed and/or if they can devote more time to their work, then they will more reliably trace and use the most appropriate descriptors in their vocabulary. They will understand better the meaning of each descriptor and make sure, before deciding in favor of a certain descriptor, that a more appropriate descriptor reliable not available in the vocabulary, one which an inquirer may detect and be apt to use unhesitatingly. The consequence is an increase both in representational fidelity and predictability. From this perspective, any indexing language that is not used in the mandatory mode deceptively promises to provide, through the specificity of is descriptors, a search accuracy that is not attainable in practice, and it offers descriptors that an inquirer would do better to avoid, due to their lack of assignment reliability.

Hence, when precision and recall have been found empirically to be inversely related, this is not due to a lawful relationship but merely to the use of defective indexing devices or, as is the case of nonmandatory indexing, to the inadequate application of an adequate device. Analogously, an "inverse" relationship between the weight of a load and the safety of its transport will only be found where the use of suitable means has been disregarded. "The faster one moves the more dangerous it will be" and "the brighter the sunshine the more one will suffer from sunburn" are also examples of a wrongly stated and purely "empirically" determined inverse relationship.

ROLES

Back in the early days of the mechanized supply of chemical information, a characteristic cause of deficient representational fidelity had already been recognized. It appeared when substances were indexed without any indication of the role the substance played in the context. Roles were then introduced to indicate whether, for example, a substance was manufactured or subject to analysis, etc. However, the desired increase in representational fidelity was frequently not achieved because the use of these roles was impeded by various constraints. For example, the assignment of roles was artificially limited to only one per substance. When several roles should be assigned to a single substance, one of them had to be arbitrarily selected and the remaining ones disregarded. In other words, the predictability of role assignment was significantly reduced. Moreover, representational fidelity was raised only slightly when the number of roles provided was relatively small. For the vast majority of the possible functions of a substance there were no roles available at all, so that they have often not fulfilled the expectations of their users.

In the IDC system[11] the predictability of role assignment is assured by a special indexing convention. Here, and in an important subarea, it is not permissible to index an object without also indicating the process in which this object participates and the manner in which this happens. For example, to any substance participating in a chemical reaction its role as a reaction product, starting material, or auxiliary must be assigned. Additionally, and to further improve representational fidelity, it is obligatory for the indexer to indicate the substructure that was formed in this reaction.

This kind of "relation indicators" can also be used profitably for nonchemical processes to express, with a high degree of representational fidelity *and* predictability, which object participated in a process and the nature of its participation. For each process in which an object participates, a corresponding relation indicator must be assigned to the object. Thus, in many fields outside chemistry these relation indicators could cause an increase in precision without at the same time impairing recall (cf. page 12 of reference 1).

SEARCHING WITH NEGATED CONCEPTS

When an inquirer requests a search with negations, he rarely wishes all texts in which the negated concept occurs in *some* context to be rejected. Rather, he wishes to restrict the negation to the specific context of his interest.

If, for example, a search is requested for esters of aliphatic diamino carboxylic acids of the lysine type, without further substituents on the amino acid chain, it is highly desirable to restrict this negation precisely to this chain and to admit any

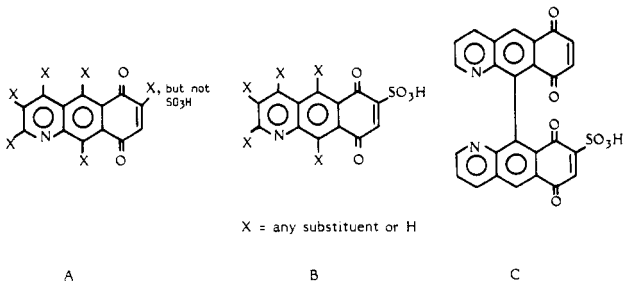

Figure 4. Undesirable course of a logical negation.

kind of substitution that may occur on other chains in the rest of the ester molecule, e.g., in the alkyl group of the ester, for such a substitution would not render the molecule irrelevant to the inquiry.

In spite of the exhaustively precise atom syntax of our present topological systems, they are, strange to say, weak in performing versatile logical negations in a query. If, for example, references are sought for substances of type A in Figure 4, then the mere exclusion of all responses to query B does not meet the requirements of the search. In logical operations of this kind substances of type C will also be excluded, although they are, strictly speaking, perfectly relevant to the topic of the search.

In another example, documents may be sought that report on properties of a substance other than its long-known bactericidal activity. Then again it is not permissible generally to exclude any appearance of the descriptor for bactericidal activity in the same text in which the substance in question is mentioned. A perfectly relevant passage could appear in a stored text along with some *other* substance for which its bactericidal activity is mentioned.

A search could also be requested for copolymers that consist of butadiene and styrene and contain no acrylonitrile as a comonomer. Again, it would not be permissible entirely to forbid the appearance of acrylonitrile in any part of the text. It could be mentioned in a peripheral or completely different context not affecting the relevance of the text. It is even possible that the *absence* of acrylonitrile is expressly stated. Again, its exclusion if it is expressed in the query should be restricted precisely to the polymer with the monomer components butadiene and styrene.

All these negations require an accurate presentation of the syntactic relations that prevail among the concepts in a document, at least as far as they might be subjected to negations in a query, as is possible, for example, in the TOSAR system of the IDC (cf. page 13 of reference 1).

It is true that indicating the number of monomers forming a polymer can, to a certain extent, serve the same purpose. But this device will not work satisfactorily in a search for "copolymers containing *among others* styrene and butadiene, but no acrylonitrile".

There is hardly any field other than chemistry in which so highly syntactic indexing languages are used and where, consequently, both positive and negative syntactic search parameters are possible. This is only another manifestation of representational fidelity, and again, chemistry seems unique among the sciences in that here, too, with regard to syntax, extraordinarily high search precision is possible.

TWO WAYS OF GENERALIZING A MOLECULAR STRUCTURE

The aforementioned concept of semantic categories helps us to distinguish between two ways of generalizing a molecular structure. This distinction is of considerable practical importance in chemical documentation.

From logic, we have learned that within an individual semantic category we can proceed to more general concepts by successively dropping conceptual constituents from a concept. If we take the example of lysine, we can obtain the following hierarchies:

lysine

aliphatic α-amino carboxylic acids	primary aliphatic 1,5-diamines
α-amino carboxylic acids	aliphatic 1,5-diamines

In this kind of generalization an exhaustively defined substructure is always retained. Here, the requirements for the applicability of the topological method are fulfilled. We will call this type of generalization the "substructure generalization".

A second type of generalization is involved when we proceed from

lysine
to
aliphatic diamino carboxylic acids
and to
amino carboxylic acids
etc.

Here, the substitution positions are interminate. Consequently, such a structure cannot be represented topologically, at least not without distortion of the general concept*, although the task is well within the capabilities of a well-designed fragment code. We will call this type of generalization of the "global generalization".

It would not constitute a solution for the topological method to establish the indexing convention that, as in the above-mentioned global generalization, always "lysine" or "glycine" has to be assumed. First, the number of conventions to be memorized would soon become so large that they could no longer be reliably applied by the indexers. Second, the consequence of such a (distorting) representation is that each search for the assumed compound (here for lysine or glycine) would unavoidably retrieve all generalizations for which the specific compound was agreed upon to stand. Similarly, if the ethyl group was assumed to stand for "alkyl" and "hydrocarbyl", then all these general structures will necessarily and undesirably be retrieved in a search for a compound specifically containing the ethyl group. This is equivalent to deficient representational fidelity and, in an advanced state, also to deficient representational predictability, which is due to the increasing disregard of the increasing number of conventions to be stated.

Here, we can clearly see the boundary separating the topological method and the fragment-code approach. This analysis also reveals that the strengths of a good fragment code persist even in the age of topological methods, in that they provide the undistorted and predictable representation of global structure generalizations.

CONCLUSIONS

Chemical information files are particularly large, they continue to expand at a fast rate, and they are also used very intensively. Consequently, there is a demand in chemistry for particularly accurate information retrieval. The requirements of patent law have contributed greatly to the urgency of this demand. Research and development work in the area of chemical information has been carried out with extraordinary intensity, not least through the great efforts that the chemical industry, with good reason, has devoted to this field. Thus problems that still lie ahead in other fields have already appeared and been solved in chemical information systems.

It is inherent in the nature of technical concepts in chemistry that they reveal with exceptional clarity the essential features of two basic types of lingual modes of expression, namely, the lexical and the nonlexical modes. Here, too, the relation between the kind of expression on the one hand and accuracy

of retrieval on the other is especially evident. This has led to various theoretical insights, which are also valid in other fields, where they could render much experimentation superfluous. In particular, it is the analytic–synthetic principle of classification to which chemical information owes its great success. This principle also promises to bring about major progress in other fields, although its applicability is not as apparent in these fields as in chemistry. Chemical information could profit still more from this fruitful principle if its strength were more clearly realized and if it were more consciously employed. In particular, the documentation of chemical reactions and of fields closely related to chemistry, such as pest control, plant protection, and pharmacy, could be made more effective and durable.

If, as theory requires, a distinction is drawn between the mode of expression and the concept to be conveyed, then the example of chemistry very clearly shows the necessity of expert and reliable manual indexing, if accurate searches are to be expected of an information system.[12,13] It shows us, too, how far distant we are from the replacement of good manual indexing by any type of automatic, algorithmic indexing. The crucial obstacle here is the inherent unpredictability of an author's and inquirer's natural language expressions for *general concepts* and statements. This was hitherto prevented their reliable algorithmic detection in a text and also their reliable algorithmic translation into an indexing-language mode of expression.

Theory also teaches us that concept detection and translation seem programmable in principle in the case of *individual concepts*, because they are nearly always encountered only in the (fairly predictable) lexical mode of expression in the original texts. Any claims of success for automatic indexing should be viewed with scepticism whenever no distinction has been made between individual and general concepts.

All these interrelationships are revealed especially clearly in chemical information science. Hence, it has been and can also in the future be a pioneer for other fields. However, some fields of pure and applied chemistry, too, could profit from a more consistent and conscious application of the theoretical principles that have been used successfully in chemistry for the last 2 decades.

*The meaning in which the terms with an asterisk are used in this paper can be looked up in the table on page 119 of reference 2.

REFERENCES AND NOTES

(1) Fugmann, R. "Toward a Theory of Information Supply and Indexing". *Int. Classif.* **1979**, *6*, 3–15.
(2) Fugmann, R. "Role of Theory in Chemical Information Systems". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 118–125.
(3) Fugmann, R. "The Five-Axiom Theory of Indexing and Information Supply". *J. Am. Soc. Inf. Sci.* **1985**, *36*, 116–129.
(4) Ranganathan, S. R.; Gopinath, M. A. "Prolegomena to Library Classification", 3rd ed.; ASIA: London, 1967.
(5) Rolling, L. "Indexing Consistency, Quality and Efficiency". *Inf. Process. Manage.* **1981**, *17*, 69–76.
(6) Cooper, W. S. "Is Interindexer Consistency a Hobgoblin?" *Am. Doc.* **1969**, *20*, 268–278.
(7) Zunde, P.; Dexter, M. E. "Indexing Consistency and Quality". *Am. Doc.* **1969**, *20*, 259–267.
(8) Ellis, D. "Theory and Explanation in Information Retrieval Work". *J. Inf. Sci.* **1984**, *8*, 25–38.
(9) Farradane, J. "The Evaluation of Information Retrieval Systems". *J. Doc.* **1974**, *30*, 195–209.
(10) Bollmann, P. "Probleme des Messens und Bewertens beim Information Retrieval". "Deutscher Dokumentartag 1981"; Saur: Munich, 1982; p 421.
(11) Fugmann, R. "The IDC System". In "Chemical Information Systems"; Ash, J. E.; Hyde, E., Eds.; Horwood: Chichester, U.K., 1975; pp 195–226.
(12) Moses, P. B.; Nelson, L. E. "Indexing and Abstracting Chemical Information: The View of Two Industrial Chemists". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 189–190.
(13) Rowlett, R. J., Jr. "An Interpretation of Chemical Abstracts Service Indexing Policies". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 152–154.

Principles for the Continuing Development of Organic Nomenclature

N. LOZAC'H

Institut des Sciences de la Matière et du Rayonnement, Université de Caen, Caen Cedex, France

Received December 10, 1984

Complexity of organic nomenclature increases not only because more and more complicated structures are to be named but also because, for lack of a generally accepted policy, several linguistic procedures are used for describing the same structural feature. A prerequisite for any progress in systematic nomenclature is a precise definition of nomenclature operations through which structural features are represented by linguistic procedures. Progress is to be sought mainly in the elimination of alternative procedures, new methods being introduced only when no existing method seems satisfactory. Nodal nomenclature, which inserts, as far as possible, existing methods in a general logical frame, shows that it is really possible to simplify systematic nomenclature even if for practical—mainly financial—reasons such a change may be long and difficult to introduce.

Nomenclature of organic chemistry, because of the considerable number of organic compounds, constitutes a particularly interesting subject of study. The very magnitude of the number of organic compounds has created a situation whose complexity has probably no equal in the field of scientific language. The quantity of available information being very large in organic chemistry, a significant part of most research projects should generally consist of a careful investigation of previous results.

Although other methods for representing chemical structures, such as line notations or connection tables, have been developed, organic nomenclature remains an indispensable tool for current communication and for access to alphabetical indexes. It is therefore timely to evaluate existing nomenclature methods and, as far as possible, to improve them so that they are able to cope with the future needs of the chemical community.

As happens for any language, future development of chemical nomenclature is largely unpredictable because nomenclature will have to deal with various problems, some of which are not yet known. Nevertheless, what can certainly be improved is the methodology to be used for solving no-

0095-2338/85/1625-0180$01.50/0 © 1985 American Chemical Society