REFERENCES AND NOTES

(1) Ranganathan, S. R.; Gopinath, M. A. "Prolegomena to Library Classification", 3rd ed.; Asia Publishing House: London, 1967.
(2) Skolnik, H. "The Multiterm Index: A New Concept in Information Storage and Retrieval". *J. Chem. Doc.* **1971**, *10*, 81.
(3) Farradane, J.; Gulutzan, P. "A Test of Relational Indexing Integrity by Conversion to a Permuted Alphabetical Index". *Intern. Classific.* **1977**, *4*, 20.
(4) Fillmore, C. J. "The Case for Case". In "Universals in Linguistic Theory". Bach, R., Harms, R. T., Eds.; Holt, Rinehart, and Winston: New York, 1968; pp 1-88.
(5) Bhattacharyya, G. "POPSI, Its Fundamentals and Procedure Based on a General Theory of Subject Indexing Languages". *Lib. Sci., Slant Doc.* **1979** *16*, 1.
(6) Austin, D. "PRECIS", *Lib. Sci., Slant Doc.* **1975**, *12*, 89.
(7) Craven, T. C. "NEPHIS: A Nested Phrase Indexing System". *J. Am. Soc. Inf. Sci.* **1977**, *28*, 107.
(8) Spang-Hanssen "Roles and Links Compared with Grammatical Relations in Natural Languages". *Dausk Teknisk Literatursels Kab Skriftserie* **1976**, *No. 40*, ISBN 87-7426-013-8.
(9) Fugmann, R.; Nickelsen, H.; Nickelsen, I.; Winter, J. H. "Representation of Concept Relations Using the TOSAR System of the IDC". *J. Am. Soc. Inf. Sci.* **1974**, *25*, 287.
(10) Fugmann, R. "Toward a Theory of Information Supply and Indexing". *Intern. Classific.* **1979**, *6*, 3.

Role of Theory in Chemical Information Systems

ROBERT FUGMANN†

Hoechst AG, 6230 Frankfurt am Main, Federal Republic of Germany

Received April 20, 1982

Continued lack of a well-established, accepted theory has seriously impeded development of effective and efficient information systems. Chemical information has always been exceptional with respect to clarity, amount, usefulness, permanence, and, hence, maturity of organization. Experience in this field suggests a tentative theory that rests on the following five axioms. Definability: the compilation of information relevant to a topic can be delegated only to the extent to which the topic can be defined. Order: any compilation of information relevant to a topic is an order-creating process. Sufficient degree of order: demands made on the degree of order increase as the size of the collection and/or the frequency of the searches increase. Predictability: success of any directed search for relevant information hinges on how readily predictable are the modes of expression for concepts and statements in the search file. Fidelity: success of any directed search for relevant information hinges on the fidelity with which concepts and statements are expressed in the search file. The observance of these axioms could generally assist in the design and improvement of information systems. These axioms could have settled many controversies among scientists concerning the question of the necessity or dispensability of natural and indexing language.

INTRODUCTION

The design and use of information systems have developed into fields of great scientific and economic importance, but the various processes, which are involved in supplying relevant information to an inquirer as a response to his search request, are still not yet fully understood. As a result, many information systems have failed or have at least not been as effective as they were expected to be. New information systems have again and again been designed without any certainty that they may in the long run display an essentially higher survival power than those for which they are intended as a substitute. Great benefit could be gained from a theory that could explain and even predict the behavior of an information system over time and under the constraints of its continual growth with respect to file size, use, and conceptual volume. The lack of such a theory has often been stated and deplored (cf., e.g., ref 1).

The accuracy and economics of the supply of relevant information and, hence, the survival power of the entire information system depend heavily on the way in which information was previously indexed or classified, on the indexing language employed (if any), and on the mechanical tools used for the handling of the indexing language and of the search file. In our company and also in the IDC (International Documentation in Chemistry, Federal Republic of Germany) firms a comprehensive information system for the chemical literature has been use for many years. In the design and further development of this system we were guided by a number of principles which manifested themselves in the formulation of a small set of axioms in a new theory of information supply and indexing[2,3] first published in 1972. The statements expressed in these axioms have occasionally been set forth by other workers as well, but these statements have been published in widely scattered and unconnected literature. Of the various theoretical approaches to a theory of information systems, that of Rush and Landry[4] resembles ours most closely. We also lean on the "analytico-synthetic" approach of Ranganathan and his Indian school (cf., e.g., ref 5; see also "Introduction to the Symposium on the Employment of Grammar in Indexing Languages," preceding paper in this issue).

It is in the nature of axioms that they are self-evident, require no proof, and are even unprovable. Nevertheless, it is useful to compile them and to phrase them explicitely, for the conclusions that can be inferred from them will stand on especially firm ground. Our five-axiom theory has hitherto served us well in explaining, controlling, and predicting the response of a large and operational chemical information system to the everchanging demands that are made on it. We shall discuss this theory and some of its implications in the following. The meaning in which several terms are used in this theory is shown in Table I.

FIVE-AXIOM THEORY OF INFORMATION SUPPLY AND INDEXING

Table II provides an overview of the entire five-axiom theory. Let us assume a large collection of documents, the

† 1982 Herman Skolnik Award address, American Chemical Society, Division of Chemical Information, Las Vegas, Mar 30, 1982.

0095-2338/82/1622-0118$01.25/0

Table I. Explanation of Some Terms Used in the Five-Axiom Theory

1. *Accuracy:* The accuracy of an information supply (20, 21) is determined by the extent to which the *noise* of irrelevant responses and the *loss* of relevant (33) information (19) has been avoided.- A highly accurate information supply is characterized by high ratios of precision and recall.

2. *Associative relation:* Any relation which is not hierarchical (13) and not purely logical (11, 25).

3. *Author-lingual expression:* Natural-language expression (26) of an author or information seeker.- *Uncontrolled* use (8, 16) of these expressions is typical of author-lingual expressions.

4. *Concept:* The entirety of the true and essential statements that can be made about a referent (32). Each statement constitutes a conceptual feature of the concept.

5. *Concept, general:* Concept (4) with respect to which at least *one more specific concept* (7) exists which is meaningful from the perspective of the experts in a subject field.- It includes only few conceptual features.

6. *Concept, individual:* Concept (4) with respect to which *no more specific concept* (7) exists which is meaningful from the perspective of the experts in a subject field (17).

7. *Concept, specific:* Concept (4) which comprises a fairly large number of conceptual features.

8. *Controlled vocabulary:* Vocabulary which contains those descriptors which are *permitted* for indexing (14, 16).

9. *Descriptor:* Lexical unit, (24) i.e. string of characters, agreed upon to denote a concept in an information (19) system.

10. *Descriptor, composite:* Descriptor (9) whose meaning includes the *meaning of at least two descriptors* in the same vocabulary and also the non-hierarchical relation (13) between them (28, 29, 35).

11. *Descriptor, elemental:* Descriptor (9) whose meaning does not include *the meaning of another,* associatively (2) (i.e. non-hierarchically (13)) related descriptor.

12. *Generic relation:* Two concepts (4) are in generic relation if they share *at least one (essential!) conceptual feature* and if one of the concepts has at least one additional conceptual feature.

13. *Hierarchical relations:* The type of concept relation that exists in a system consisting solely of *generically* (12) *related concepts.*- It excludes the associative (2) and the purely logical relation (10, 11, 25).

14. *Indexing:* The process of a) *discerning the essence* of a document and
b) *representing this essence* in an indexing-language mode of expression (8, 15), i.e. with a sufficient degree of predictability and fidelity.

15. *Indexing language:* Language which represents concepts (4) and statements in a body of documents with a sufficient degree of predictability and fidelity (14).

16. *Indexing, controlled:* Indexing with the use of a controlled vocabulary (8), i.e. in which any more or less appropriate vocabulary descriptor is *permitted* to represent a concept.

17. *Indexing, free:* Indexing in which common *natural-language expressions* (26) are used *without restriction* to represent the concepts (4) and statements of a document.- This representation is sufficiently predictable only in the case of individual concepts (6).

18. *Indexing, mandatory:* Indexing in which it is *obligatory* for the indexer to use those descriptors (9) in the vocabulary which *most appropriately* represent the topic under consideration.

19. *Information:* Message that proves to be *of interest to a recipient.*- It is immaterial whether the type of message is defined beforehand and whether the message causes the recipient to make decisions.

20. *Information supply, directed:* Type of information (19) supply in which the selection is accomplished *without the personal intervention* of the information seeker.- It is based on a more as less exact, predetermined definition of the topic of interest. It is delegated at least to the extent to which previous indexing (by somebody else) is involved.

21. *Information supply, undirected:* Type of information (19) supply in which the information seeker himself selects the documents of interest from a body of candidate documents which has not (or at most very superficially) been preselected.- The undirected information supply does not require a predetermined definition of the information seeker's interest profile.

22. *Inquiry:* An information seeker's specification of what he is interested in and what he expects to be supplied to him by the information system.

23. *Lexical expression:* Use of a lexical unit (24) to denote a concept.

24. *Lexical unit:* Linear string of characters specifically agreed upon to denote a concept.- A unique location can always be assigned to a lexical unit where it can be entered or looked up in an alpha-numeric arrangement (9, 23, 34).

25. *Logical relation:* Syntactical relation expressed by the Boolean operators AND, inclusive OR, exclusive OR, NOT (cf. 2).

26. *Natural-language expression:* Expression which does (or could) occur in the original text of documents or in the inquiry (22) of an information (19) seeker.- Natural-language terms may be freely chosen or subject to controlled use (3, 8, 16, 17).

27. *Order:* The meaningful proximity of the parts of a whole at a foreseeable place.- Anyone who requests order to be established in such a *"whole"* (e.g. in a collection of documents or descriptors) will have to specify in advance what kind of arrangement he will consider *"meaningful".*- This reveals the inherently subjective nature of the concept of order.
The *foreseeable place* may be a distinct location in a shelf of books or the responses in a retrieval operation. *"Proximity"* means that closely related items are brought together and that non-related ones are largely kept outside the foreseeable place.

28. *Postcoordination:* Avoidance of composite descriptors (10).

29. *Precoordination:* Use of composite descriptors (10).

30. *Pertinence:* Any message that proves to be of interest, irrespective of whether it satisfies the parameters of an inquiry, ranks as pertinent, and, hence, as information.- Even irrelevant (33) responses may turn out to be pertinent. An inquirer is unable to define his entire field of (potential) interest. He can therefore not expect an information system to supply him with all those messages in the system that he would find interesting if he encountered them and no other message but those.

31. *Query:* Entirety of the parameters which are effective in the (mechanized) selection of responses to an inquiry (22).

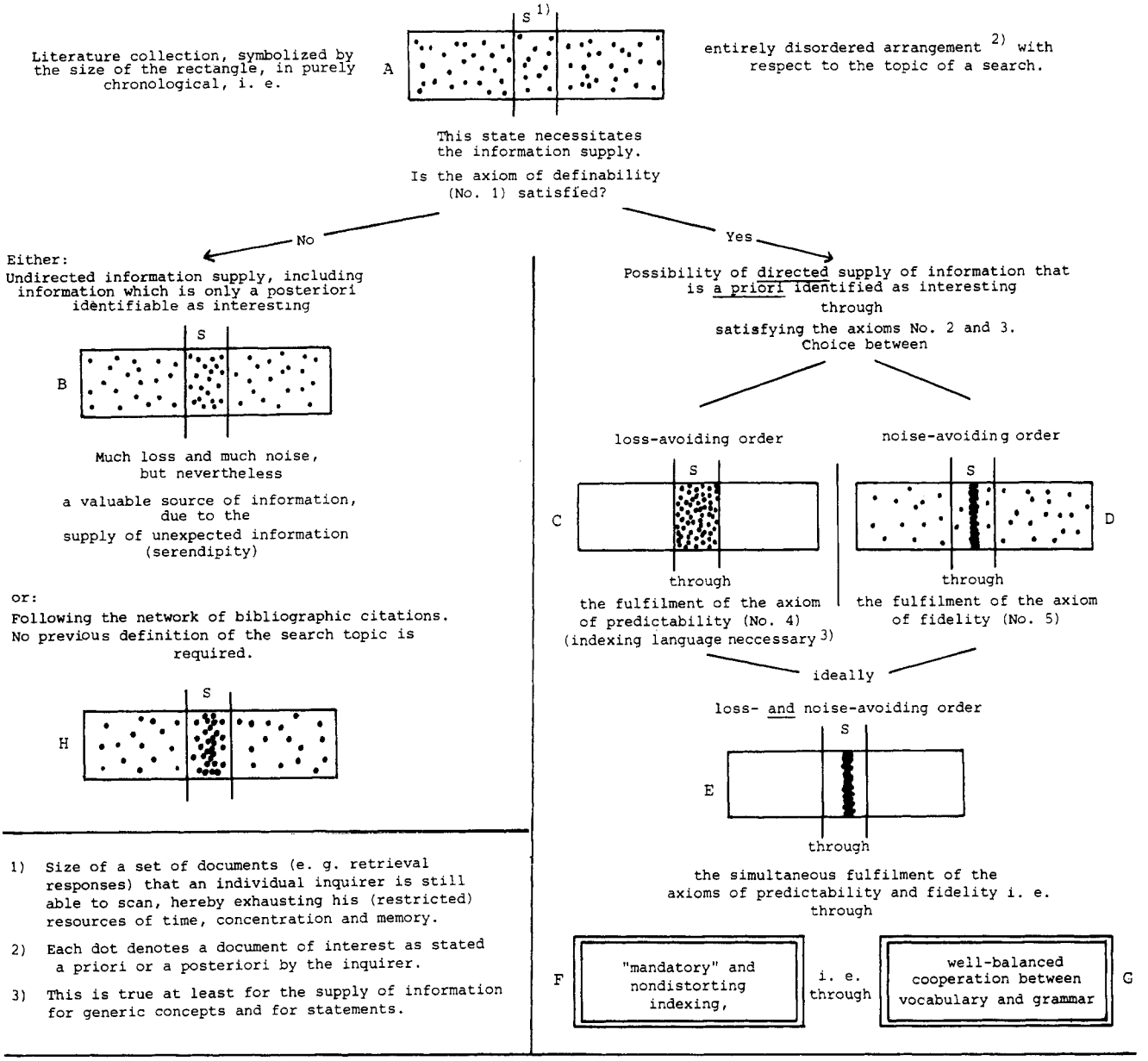
32. *Referent:* Anything about which statements ("predications") can by made (cf. 4).

33. *Relevance:* Any message is considered relevant if it comprises all the concepts (4) and concept relations (as far as they can both be defined in the inquiry) in the desired or in a higher degree of specificity.- The occurrence of the required concepts and relations must be clearly recognizable to an expert in the field, irrespective of the mode of expression chosen by an author and irrespective of whether they are expressly stated or only implied. A relevant message may or may not rank as an interesting (pertinent (30)) one for the inquirer. It may already have been known to him or he may consider it useless for several other reasons.

34. *Syntactical expression:* Use of several lexical units (24) and syntactical (35) devices to represent a concept (4).- Definitions, paraphrasing sentences and groups of sentences are examples of syntactical eypressions.

35. *Syntax:* Any way of representing non-hierarchical concept relations other than incorporating them into a composite descriptor (10, 34).

Table II. Overview of the Theory of Information Supply Based on Five Axioms



1) Size of a set of documents (e. g. retrieval responses) that an individual inquirer is still able to scan, hereby exhausting his (restricted) resources of time, concentration and memory.

2) Each dot denotes a document of interest as stated a priori or a posteriori by the inquirer.

3) This is true at least for the supply of information for generic concepts and for statements.

Axioms

1. Axiom of definability

The compilation of information relevant to a topic can be delegated only to the extent to which the inquirer can define the topic in terms of concepts and concept relations.

2. Axiom of order

Any compilation of information relevant to a topic is an order-creating process.

3. Axiom of the sufficient degree of order

The demands made on the degree of order increase as the size of a collection and/or the frequency of the searches increase.

4. Axiom of predictability

The accuracy of any directed search for relevant information depends on the predictability of the modes of expression for concepts and statements in the search file.

5. Axiom of fidelity

The accuracy of any directed search for relevant information depends on the fidelity with which concept and statements are expressed in the search file.

size of which is symbolized by the rectangles in Table II and in which each dot stands for a document of interest to an information seeker. Frame A represents a literature collection that is in a purely chronological arrangement, i.e., one that is in its original state of almost complete disorder with respect to the topics of the documents. Any kind of disorder can be overcome through the expenditure of time, concentration, patience, and memory. However, these natural resources are limited. We can therefore delimit a subset S of our imagined literature collection as being of such a size that a searcher is able and willing to scan through it from one end to the other, thereby depleting the resources of time and concentration that

he can devote to this search. We shall repeatedly have to refer to this maximum manageable document quantity S.

AXIOM OF DEFINABILITY

An information system user expects to be spared the work of scanning unmanageably large document collections. He wants instead to be directed to a subset of documents of at most size S. This means that *someone else* must have done the work of bringing together the documents in which the inquirer is interested. But to ask someone else to perform a task in which one is interested always requires some kind of *definition* of the task to be done. An inquirer may, for example, be interested in literature about the "treatment of leprosy with diphenyl sulfones". Then this definition of the search topic may well serve as the basis for arranging a collection in an orderly fashion such that the literature on the topic concerned is more or less accurately brought together without the personal intervention of the inquirer. This "directed information supply" as we shall call it is symbolized by frame C.

However, the searcher will be overtaxed if required to *define* everything that might be of interest. Although working for the moment in the field of chemotherapy, one might well be interested in an article on new energy-saving routes to some key chemicals or on the use of liquid xenon as a solvent in spectroscopy when they are encountered in the current literature or elsewhere. Tomorrow it may be an article of quite a different topic, one that is impossible to define beforehand.

If an interest profile completely lacks definability, only the information seeker can select documents of momentary interest. He must therefore put up with a rather disordered literature collection through which he will have to browse. This situation is symbolized by frame B. Creative humans have always met a considerable part of their information needs by way of this "undirected information supply".

Another way of being directed to documents of interest without the necessity of a previous definition of the subject of interest is that of pursuing a network of citations. Dependent on the homogeneity of the subjects in a document under consideration and also dependent on the completeness of an author's citations, relevant documents will appear more or less concentrated in the set compiled in this manner (frame H).

The axiom of definability summarizes these considerations and states the following:

(1) The compilation of information relevant to a topic can be delegated only to the extent to which an inquirer can define the topic in terms of concepts and concept relations.

Close examination reveals that a large part of the information required by a chemist consists of a definable and undefinable part. Examples are as follows: "How superior is our product X to competing products already on the market?" "What is the best synthetic route to product X?"

What an inquirer can demand of an information system is merely literature on substance X. A comparison with competing products or of different synthetic routes must be left to the inquirer. The choice of comparison criteria and the judgement of their weights relative to one another are too subjective.

The axiom of definability gives a clear answer to the question of whether the future will bring an end to the demand for printed information, so that we might look forward to a "computerized, paperless information age". It is hard to imagine how the undefinable part of the information needs of the creative scientist could be met without the distribution of printed matter.

The axiom of definability can warn information scientists against attempts to satisfy demands that are inherently unsatisfiable due to their subjectivity and undefinability. Such attempts are doomed to failure from the outset and only serve to make the information scientist the target for criticism. In practice the information scientist may often be tempted to ask a disappointed inquirer "Do you expect me to read your mind?" How much more tactful and convincing to refer instead to the axiom of definability!

AXIOM OF ORDER

As previously stated, an ordered collection of documents can be attained if the inquirer can, at least partially, define his topic of interest. It is immaterial to our consideration whether this ordered arrangement is in permanent existence, as on the shelves of a library, or whether this arrangement is created on demand by a retrieval process. Here the collection of documents is partitioned into two parts. In one of them, namely, in the retrieval responses, material relevant to the inquiry is isolated or at least concentrated, as depicted in frames C–E. At least the inquirer is relieved of trying to scan the entire collection, a task which would overtax his resources of time and concentration in most cases. He can instead concentrate his attention on a limited, isolated section of the collection. This idea is expressed by our second axiom:

(2) Any compilation of information relevant to the topic of an inquiry is an order-creating process.

It is true that this statement has often been expressed in the past, perhaps most emphatically by Rush.[4] But one finds that this idea has often been disregarded if "order" is defined as the "meaningful proximity of the parts of a whole at a predetermiable place". This definition is derived from one that is ascribed to the philosopher Driesch.

AXIOM OF SUFFICIENT DEGREE OF ORDER

In Table II the lowest tolerable degree of order is depicted in frame C. By no means should the degree of order, which prevails in a collection or can be established in it on demand, be lower than indicated and the relevant documents be more scattered, as otherwise a number of relevant references would fall outside the maximum manageable document quantity S and thus escape the inquirer's attention.

The opposite extreme is the high degree of order depicted in frame E. All relevant material is localized in a very narrow range and is entirely free of irrelevant responses and omissions. Search time, attention, and memory are no longer occupied. This is the optimum case of purely directed information supply. Examples are the search for an individual chemical compound or for a clearly defined class of compounds.

An information system that is limited to the mediocre degree of order C not only is rather unwieldy but will also rapidly become more so and will soon cease to function satisfactorily. During the continued growth of the file the text quantity to be scanned will soon exceed the maximum size S that a human can manage. From this point onward more and more relevant information will be lost.

The optimum degree of order, however, as symbolized by frame E, endows an information system with great survival power. The collection can continue to grow and so can the frequency of the searches, and only at some distant point in time (if at all) will the critical document quantity S be reached. This consideration leads us to our third axiom:

(3) The demands made on the degree of order increase as the size of the collection or the frequency of the searches increases.

Many information systems have already failed because this law was not taken seriously in the early stages of the system's development. In the early stages almost any information system will perform well, because low demands are made on the degree of order attainable in it. This was often mistakenly

seen as an indication of the soundness of the system's entire conception.

TWO DIFFRENT KINDS OF ORDER

In a collection of documents two kinds of order may be desirable, depending on the purpose the ordered arrangement is to serve. In the kind of order we have already discussed, the avoidance of a loss of relevant information is intended, and noise, i.e., lack of precision, is tolerated. This situation is indicated by the relatively large distances between the individual dots in frame C, for these distances dilute the relevant information compiled. Instances in which loss-avoiding order is paramount are patent-infringement searches and searches for the side effects of pharmaceuticals. The second kind of order is the noise-avoiding one as depicted in frame D. Here precision is aimed at, even at the expense of some omissions. This is the kind of order frequently preferred in research work. The research chemist, for example, does not normally need to know all the many variations of a particular reaction type when he wants to use this reaction for the preparation of a certain substance.

In practice neither of these basic forms of order is normally encountered alone. In most cases they are superimposed on one another. Nevertheless, it is instructive to investigate them separately, because then their inherent features become most obvious. Loss-avoiding order in its pure form is attained through fulfilment of the axiom of predictability and noise-avoiding order through fulfilment of the axiom of fidelity. These are the last two axioms remaining to be discussed.

AXIOM OF PREDICTABILITY

It is inherent in uncontrolled, natural language that it often provides a large variety of modes of expression for a concept or a statement. Let us imagine, for example, a few possibilities of common language expressions for the subject of pesticides for plant protection: "...fungicides for the control of mildew in vineyards..."; "...and through the application of this emulsion piricularia orycae was thus completely suppressed in rice plantations..."; "...through the use of chemical traps against the oriental fruit fly...". If the original texts are entered into the file without any control or revision, then the concept of plant protection will be represented in very many different modes of expression in the search file.

Let us now consider the unfortunate position of an inquirer who is interested in retrieving literature on chemical plant protection. In order to phrase a query, he (or his machine program) will have to know in advance the expressions through which the concept of his interest is represented in the search file, for just these expressions will have to be phrased as search alternatives. If one disregarded only one single mode of these expressions, then this would inevitably lead to loss of relevant information contained in the file. This raises the important question of how is one to know in advance just which lingual representations of the search topic have entered the file and should therefore be included in the query as search alternatives. Certainly they cannot be taken from the relevant documents contained in the file, for these are the target of the search and cannot be assumed to be known beforehand. They are the ones still to be discovered by the proposed search. Nor can the inquirer rely on his memory and imagination; i.e., these expressions cannot be expected to come to mind sufficiently fast and completely. Instead, one must be able reliably to reconstruct or *predict* by which modes of expressions a topic under consideration is represented in the file. Only in these circumstances can all the modes of expression for a concept or a statement in the file be covered in the phrasing of a query. Only in these circumstances can an exhaustively complete query be phrased and therefore loss of relevant information be avoided. We can therefore formulate our fourth postulate:

(4) The accuracy of any directed search for relevant information depends on the predictability of the modes of expression for concepts and statements in the search file.

Our considerations reveal that it is not (at least not primarily and solely) the ambiguity and vagueness of uncontrolled natural language text that render it inadequate for satisfactory retrieval, for in the above examples the precision of the author's lingual modes of expression leaves nothing to be desired. It is their lack of predictability which constitutes a serious impediment to an accurate retrieval in an author lingual file.

As far as the predictability of the natural-language expressions of the authors is concerned, a fairly clear dividing line can be drawn between what philosophers call "individual concepts" on the one hand and "general concepts" and "statements" on the other.

Individual concepts refer to persons, institutions, individual chemical compounds, individual publications, or individual patent specifications. Distinctive lexical units are always used to denote them, for example, a proper name or a unique number. Only a single name or a few names at most will be in common use for an individual concept. These names are easy to recall or look up in dictionaries. They can, consequently, be easily and completely compiled and used as search alternatives in a query. In other words, the natural language expressions are highly predictable for individual concepts. If in an information system only individual concepts and no others are stored and sought, there is therefore little need to represent these concepts by expressions other than those used by the authors of texts. In other words, author lingual expressions are quite adequate for the retrieval with individual concepts as search parameters. This is in sharp contrast with general concepts such as processes of any kind and the properties and uses of substances. It is typical of these general concepts that they are frequently not expressed by lexical units of their own such as names and numbers but instead in a nonlexical manner, i.e., by definitions, statements, or even groups of sentences. Here the plurality of possible natural-language expressions is almost infinitely large.

Chemical reactions are another example. They are rarely expressed by a reaction name. Usually they are represented by their typical reactants and products connected by plus signs and the reaction arrow as syntactical tools. For example, very many different reactants and products could conceivably have been used by authors to represent the Friedel–Crafts or Diels–Alder reaction. Everyone of these combinations could appear in an uncontrolled natural-language file and would therefore have to be phrased as an alternative search parameter in a query. Since this is an impossible task, only very incomplete queries could be phrased for author lingual files of chemical reactions. Loss of relevant information is therefore inevitable in such a search. Again, this is due to the unpredictability of the reactant and product names used in the file to represent the reaction being sought.

A similar situation prevails in the case of general concepts which refer to classes of interrelated chemical compounds. Again, the number of names for the individual compounds of such a class is infinitely large, and it is unpredictable which of them have in fact entered the file. They can therefore not be phrased as alternative search parameters in the desired completeness, and loss of relevant information is inevitable.

Only in passing we mention here that this difficulty cannot be overcome by any kind of "truncation". Close examination would reveal that normally the names or nomenclatures of interrelated compounds do not share a common string of characters that could profitably be used as some kind of general search parameter for a class of compounds. Considered from this point of view, fragmentation codes, topological

devices, classifications, and thesauri are tools that precisely aim at making it possible to predict which expressions have been used to represent a certain concept in the search file.

We have now deduced an important conclusion from our axioms: If accurate searches for general concepts and statements are demanded of an information system, then author lingual expressions are inadequate in a search file because of the unpredictability of their modes of expression. For the same reason full-text storage and retrieval is also inadequate for accurate searches, for this method is only a variation of the technique of entering natural lingual expressions uncontrolled into the file. If searches for general concepts and statements are to be performed with passable accuracy, the use of an indexing language or classification is therefore indispensable because of the vast improvement in predictability it affords.

The axiom of predictability also sheds light on certain assertions occasionally encountered in the literature, stating that natural language is or will soon be absolutely preferable to indexing language in information systems or stating that the reverse will be true. The author of such an assertion should always indicate whether it is intended to apply to systems for individual concepts or general concepts or perhaps even to those for statements. He should also specify what degree of order is regarded as sufficient, now and in the future, and what kind of order is being demanded of the information system. Depending on the individual circumstances, each of these seemingly contradictory assertions may be perfectly correct. Frequently, a combination of both kinds of language would optimally meet the demands made on an information system. The axiom of predictability could have settled many fierce controversies that have flared up among information scientists in the past concerning the question of the necessity or dispensability of natural or indexing language.

AXIOM OF FIDELITY

Let us now turn to the last of our axioms, that of representational fidelity. We need not discuss this axiom in greater detail because it is perhaps the most obvious one in our set of five. This holds true in particular if one visualizes it from the viewpoint of the inquirer. The less precise one can phrase the topic of an inquiry, using the vocabulary and syntax of an indexing language, the less precise will inevitably be the responses to such a query, no matter which retrieval tools are employed. No search mechanism can be expected to work more precisely than it was instructed to do through the phrasing of the query. We can therefore phrase our fifth axiom as follows:

(5) The accuracy of any directed search depends on the fidelity with which concepts and statements are expressed in the search file.

Representational fidelity manifests itself not only in the size and specificity of the vocabulary of an indexing language but also in the expressiveness of its grammar, particularly in the syntax of this grammar. In some fields of chemistry such as polymer chemistry and in chemical engineering a great deal of information is conveyed through the associative and logical relations prevailing among the individual substances and processes. It is therefore much in the interest of representational fidelity, and hence search precision, if these relations can be phrased as search parameters, along with the vocabulary descriptors. Examples of synthetic devices in operational systems are roles and links, a predetermined sequence of facets, and a kind of syntax which is expressed in standardized noun phrases. They can, however, serve their purpose only if they also satisfy the axiom of predictability, which has often not been the case for some linking devices and which have therefore not yielded the expected results.

It is most informative to compare topological representations with fragmentation codes with respect to their representational fidelity. Topological representations are unbeatable if a molecular structure or substructure can be defined atom by atom. A correspondingly high degree of precision is exhibited by the responses to such an inquiry. Frequently, however, a molecular fragment is expressed only in a generalized form such as "alkyl", "cycloaliphatic rings", or "olefinic aliphatic dicarboxylic acids". Forcing generalized concepts like these into topological representations always entails more or less serious distortion. What is expressed is either too specific or too general. The responses to such a distorted inquiry will therefore be correspondingly inaccurate.

It is inherent in fragmentation codes, on the other hand, that they are capable of expressing generalized concepts such as those mentioned in the foregoing in an adequate level of generality, without any distortion. In the case of generalized structural concepts they may therefore well be superior to topological representations with respect to representational fidelity. They can therefore very effectively complement a topological system. Good results have been obtained in our company and in IDC by combining both kinds of structural representation. Under ideal circumstances both the axiom of predictability and the axiom of fidelity are satisfied to perfection at the same time. As a result, both the loss of relevant information and the noise of irrelevant information are totally avoided, and the idelaly high degree of order is achieved as depicted in frame E. An example is the topological search for a structure or substructure which is defined atom by atom.

It has often been discussed whether there is some kind of inverse correlation between the ratios of precision and recall if queries of variable generality are posed to an information system under consideration or to different information systems. It is obvious in the light of our axioms that an increase in precision can adversely affect recall only if an enhancement in representational fidelity at the same time impairs predictability or if the reverse is true. It is evident in chemical documentation that there is by no means a necessary correlation of the postulated kind. We need only compare a kind of query which contents itself with merely enumerating the atoms of a structure, e.g., the empirical formula, with another one which utilizes the topological syntax between the atoms of the structure in addition. By no means does the greater representational fidelity of topological representation entail a decrease in predictability, and by no means does the much more precise topological search display a lower recall ratio than the empirical formula search. Closer investigation would reveal that the cases of this inverse correlation are largely produced through a defective, namely, syntax-deficient, indexing language. Here, an increase in fidelity is possible only through an expansion of the indexing language vocabulary. It is only through this makeshift method that the reliability (and, hence, the predictability) of the vocabulary usage during indexing is impaired, which necessarily decreases the recall ratios.

We recall that the ultimate effectiveness of any tool depends not only on the tool itself but also on its knowledgeable and skillful use. Indexing languages as tools for achieving order in document collections do not constitute an exception to this rule. It is even inherent in certain types of indexing languages that they cannot possibly be employed reliably and consistently in daily practice by the indexer.

MANDATORY INDEXING

Let us assume that an indexing language contains hierarchically related descriptors such as "chemical reaction", "oxidation", and "autoxidation". In a markedly liberal kind of indexing the indexers may be permitted to use any of these more or less general descriptors for a specific topic, provided the descriptor is taken from a controlled vocabulary.

An example of a topic to be indexed may be "chemical reaction of linseed oil when exposed to air". Then, an indexer may well be tempted to use the descriptor "chemical reaction", because this term occurs in the text, although "autoxidation" would be a more appropriate descriptor in the vocabulary. It is in fact the descriptor which an inquirer would certainly choose to retrieve this document. In other words, if such a liberal kind of indexing is permitted, one and the same topic in different documents will be indexed with varying specificity, depending on which descriptor an indexer encountered first or memorized first and also depending on the mode of expression that an author had chosen.

Under these circumstances an inquirer can later only *guess* which descriptors of a more or less general kind might have been used by the indexers. The inquirer is therefore compelled to include some of the less appropriate descriptors as alternative search parameters in the query in order to counteract loss of relevant information which otherwise would occur. These less appropriate descriptors will, however, inevitably produce correspondingly irrelevant responses.

Hence, the representational fidelity which an indexing language promises and displays may well be a deceptive one. The specificity of the descriptors cannot be exploited if liberal indexing is exercised and if loss of relevant information is to be avoided. This holds true because the inquirer cannot rely on the specific descriptors to have been used by the indexers. He can therefore use these descriptors only in alternativity with more general ones. Thus, the representational fidelity of the topic of an inquiry is reduced below what the vocabulary suggests, and the responses to a search with such an undesirably generalized query will be correspondingly imprecise. This uncertainty can be countered only by a kind of indexing in which it is *mandatory* for the indexers to find and to use those vocabulary descriptors that most appropriately represent the essentials of a document. It is these descriptors on which the inquirer will later rely, for they promise to retrieve the documents of interest most precisely. Obviously, only this kind of "mandatory indexing" can provide that degree of predictability and fidelity which is promised by the indexer and his indexing language and which is expected by the system user (see frame F).

Seen in this light, any accurate search in a document file is preceded by another kind of search, one incessantly and latently performed by indexers, namely, by the search for the most appropriate descriptors that an indexing language provides for a topic under consideration. In the last part of our discourse we shall investigate the pecularities of this special search more closely.

NECESSITY OF INDEXING LANGUAGE GRAMMAR IN LARGE INFORMATION SYSTEMS

We have already stated in axiom 3 that for any kind of search to be successful, the human's natural resources of time and concentration must not be overtaxed; this same rule also holds true for searches in a vocabulary. As far as this search is concerned, it always proceeds along the network of relations which is woven into the vocabulary and which should lead the searcher and indexer unerringly and rapidly from any less appropriate descriptor to increasingly appropriate ones. For example, the indexer should be safely guided from "chemical reaction" to "oxidation" and, finally, to "autoxidation" if this descriptor is appropriate for the topic under consideration. The success of these vocabulary searches will greatly depend on the smallness and conceptual transparency of the vocabulary's network. A vocabulary of, for instance, 10 000 general descriptors will contain a correspondingly large and highly ramified network of concept relations. Too much time and effort of concentration and memory would have to be expended for finding one's way from a less appropriate descriptor to the most appropriate one, which the inquirer expects to have been assigned to the pertaining documents.

The idea behind vocabularies of this size, complexity, and relational ramification is to provide high specificity and expressiveness, which is in the interest of representational fidelity. But every language owes its expressiveness to the interplay of vocabulary and grammar and to the well-balanced cooperation of both lingual tools. It is typical of a syntax-deficient indexing language that it needs, for example, the composite descriptor "propylene oxidation" for achieving high representational fidelity, although the elementary concepts propylene and oxidation are represented in the vocabulary also. Indexing language grammar could, however, precisely represent the syntactical relation between oxidation and propylene in a document or inquiry and thus make the composite descriptor "propylene oxidation" entirely superfluous in the vocabulary.

Close inspection of excessively large vocabularies (or of those that are growing rapidly) reveals that a vast majority of their descriptors are of the composite type like our example "propylene oxidation". These composite descriptors contribute most to the size and complexity of the relational network in the vocabulary and consequently also represent the greatest obstacle in the search for the most appropriate descriptors. In other words, indexing languages should follow the model of any other expressive language and should not try to solve their task solely through their vocabulary and in disregard of grammatical syntax. The success of the topological approach constitutes a splendid corroboration of this general principle (see frame G).

An argument in favor of preserving author lingual expressions in the search file is that any translation into another language requires expert knowledge and has always met with difficulties, especially where the meaning of the text to be translated is vague. Translation into an indexing language does not constitute an exception to this rule. In cases of ambiguity it is therefore advisable to preserve the original, author lingual mode of expression in the search file and to keep this text accessible for the corresponding natural-language search parameter as well, at least in addition to an attempted interpretation and translation into the indexing language. Fortunately, chemistry is exceptional among the sciences in that a vast majority of its terms are fairly well defined.

Another justification of author lingual expressions in the search file is that indexing always involves the selection of what appears essential from a certain viewpoint and during a certain period of time. This subjective selection may well prove inadequate from a different viewpoint. If, then, the full text of the documents is additionally accessible for a mechanized search, the recall ratios for certain inquiries may well be improved beyond those which are attainable solely through an indexing-language file. This holds true at least if the wording chosen by the authors is sufficiently predictable.

CONCLUSION

Chemists have always been extraordinarily receptive to conceptual analysis and resynthesis, for their daily experimental work is literally of an analytic and synthetic nature. They have fully accepted this approach for the documentation of their molecular structures. For the field of chemical reactions and for nonstructural concepts, this promising approach is still awaiting application. Admittedly, this approach is, at least in the input stage, more expensive than the mere entering of the wording of uncontrolled natural-language text. But again, chemical literature is exceptional because of its volume and its usefulness over the ages. This has, in turn, justified extraordinarily large efforts to keep up its particularly good accessibility and to use highly advanced technology for this

purpose. In hardly any other field of science are the capabilities and limitations of natural language on the one hand and indexing language on the other so clearly revealed as in chemistry. The key to the understanding of the pecularities of both kinds of language is representational predictability and, hence, the differentiation between individual and general concepts. The answer to the everlasting question of whether to prefer natural or indexing language is found to be not one of "either/or" but rather one of "both/and". Chemistry, in several respects, has thus been a pioneer in information science. Through a consistent continuation of the analytico-synthetic approach the accessibility of the chemical literature can further be improved. This would also exert a beneficial influence on the documentation of the literature of other fields.

REFERENCES AND NOTES

(1) Borko, H. "Toward a Theory of Indexing". *Inf. Process. Manage.* **1977**, *13*, 355.
(2) Fugmann, R. "The Theoretical Foundation of the IDC System". *Aslib Proc.* **1972**, *24*, 123.
(3) Fugmann, R. "Toward a Theory of Information Supply and Indexing". *Int. Classif.* **1979**, *6*, 3.
(4) Landry, B. C.; Rush, J. E.: "Toward a Theory of Indexing II". *J. Am. Soc. Inf. Sci.* **1970**, *21*, 358.
(5) Ranganathan, S. R.; Gopinath, M. A. "Prolegomena to Library Classification", 3rd ed.; Asia Publishing House: London, 1967.

Searching the Literature To Learn How the Term "Ligand" Became a Part of the Chemical Language

WILLIAM H. BROCK*
Victorian Studies Centre, The University, Leicester LE1 7RH, England

K. A. JENSEN
Department of General and Organic Chemistry, University of Copenhagen, The H. C. Ørsted Institute, DK-2100 Copenhagen, Denmark

CHRISTIAN KLIXBÜLL JØRGENSEN
Département de Chimie Minérale Analytique et Appliquée de l'Université, Section Chimie–Sciences II, 1211 Genève 4, Switzerland

GEORGE B. KAUFFMAN
Department of Chemistry, California State University, Fresno, California 93740

Received September 22, 1981

An examination is made of the origins, dissemination, and eventual general adoption of the term "ligand", which is used by chemists to refer to atoms or groups attached to a central atom in coordination compounds or organometallic compounds. The term, first proposed by the German chemist Alfred Stock in 1916, was not readily accepted into other languages, particularly English. The first influential use of ligand in English did not occur until 1941 in Jannik Bjerrum's doctoral dissertation "Metal Ammine Formation in Aqueous Solution". Suggestions are made as to criteria and factors facilitating the adoption and acceptance of neologisms into scientific language and literature.

INTRODUCTION

In the course of editing an encyclopedia article by Kauffman on "Coordination Chemistry",[1] in which the word ligand[2] appeared, Brock became interested in the origin of the term. Letters by Kauffman and Jensen directed to authorities who were active in the field of coordination chemistry during the period from the 1930s to the 1950s elicited varying replies based on memory as well as documentation. Jørgensen located a reference to the origin in Fritz Ephraim's "Anorganische Chemie".[3] Ephraim attributed the word to Alfred Stock (1876–1946), the inventor of the well-known oxidation number system of inorganic nomenclature. Although Stock proposed the term ligand in 1916,[4] it did not come into extensive usage among English-speaking chemists until the 1940s and 1950s. Jensen has used his long-term expertise in matters of nomenclature to follow the progress of the term from Stock's original proposal to its almost universal use today in English and other languages. This article, then, is a brief account of our joint efforts to trace the origin and dissemination of a common chemical term and to offer some speculations on the factors governing the acceptance of such a term into scientific nomenclature.[5]

ORIGIN

During his pioneering experimental work on the boron hydrides during his World War I sojourn at the Kaiser-Wilhelm-Institut für Chemie in Berlin, Alfred Stock directed his attention to the analogous hydrides of silicon, thinking that their highly reactive and volatile character might have military applications. At a K-W-I meeting on Nov 27, 1916, he discussed the similarities between carbon and silicon chemistry first noticed as early as 1857 by Wöhler and Buff. For Stock, these analogies were made more evident and interesting by the deeper knowledge of atomic structure that was then becoming available to chemists:

> The surprisingly rapid development of our knowledge of the nature of atoms promises that in the not too distant future, it will be possible to develop an atomic–

0095-2338/82/1622-0125$01.25/0