Introduction to the Symposium on the Employment of Grammar in Indexing Languages[†]

ROBERT FUGMANN

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

Received April 23, 1982

I am very happy to receive the Herman Skolnik Award. This is a strong encouragement for me and my co-workers to continue our line of work and to provide our experience to anybody who can make use of it.

When I discontinued my synthetic work in the laboratory 25 years ago and set out to work in the field of chemical information in my company I did not realize that I was going to continue synthesis, although on a purely conceptual plane. At that time I looked 4000 miles to the East and watched in India Ranganathan's work developing in the field of universal library classification. Looking another 4000 miles to the West I realized in the U.S the advent of computer technology and the emergence of the topological representation of molecular structures, for which Calvin Mooers had laid the corner stone, a line of work largely promoted through Chemical Abstracts Service and also by several other individual contributors in the early stages of the development.

Both these successful approaches are seemingly very dissimilar in appearance, but it was not difficult to realize the hidden analogy prevailing between them. We adopted their essential principles and techniques both in our theoretical reasoning and in our operational procedures. We owe much to the Indian and the American line of working, and it makes me happy that through this award both lines of working are recognized as well. Let us as an introduction into the topic of our symposium recall a few essential features of Ranganathan's "analytico-synthetic" approach to universal library classification.[1] First, the subjects presented by an author or inquirer are analyzed into their meaningful conceptual constituents, which are represented in a vocabulary of class numbers. This analysis is guided by principles and postulates, in particular by a small set of fundamental categories, into which any subject of interest is broken down. Second, the connectivity of these constituents is represented by means of a special grammar, in particular its syntax. This syntax is expressed through the facet formula which is characteristic of this approach.

The topological approach also works with meaningful conceptual constituents, namely, the atoms of the molecular structure. The periodic table of the elements constitutes, so to speak, the vocabulary of this language. It is deliberately dispensed with the storage of the more or less systematic names for compunds as often used by the authors in their publications. In spite of the smallness of the vocabulary used, an infinitely large variety of molecules can be represented very precisely. This is due to the large contribution made by the grammatical syntax, which is inherent in a topological representation. It expresses very precisely the connections existing between the atoms in the form of their chemical bonds. Seen from this viewpoint the topological language is one with an extraordinarily highly developed syntax.

What both approaches have in common is their clear-cut borderline between the task of the vocabulary on the one hand and that of the syntax on the other. The success of both of these approaches certainly rests on the extensive and well-planned use of grammatical-syntactical devices.

Many contemporary indexing languages display a deplorably low search accuracy. Closer inspection would reveal that this is largely due to an underdeveloped or even entirely missing grammar. In such a case the vocabulary is burdened and often overtaxed with a task for which grammar is more competent. As a consequence, the vocabulary and in particular the network of relations in the vocabulary become too large and unmanageably ramified. This constitutes a steadily increasing obstacle to their reliable employment. Great progress could be made if these indexing languages were equipped with more grammar.

Several workers have already recommended this line of work, among them Skolnik with his "multiterms",[2] Farradane with his "relators",[3] Fillmore with his "deep casus",[4] Bhattacharyya with his "POPSI" system,[5] Austin with his "PRECIS"[6] system, and Craven with the system "NEPHIS",[7] to mention only a few. Roles and links,[8] the "TOSAR" system,[9] and "relation indicators"[10] also aim at representing concept relations and, thus, at improving search accuracy.

The object of this symposium was to promulgate the idea of employing more grammar in indexing languages. What the seemingly so diverse papers in this symposium have in common is the topic of syntactical devices. Gopinath, one of Ranganathan's earliest collaborators, presented a paper on some principles of the analytico-synthetic approach to library classification. Vladutz made a name for himself 20 years ago with his approach to reaction documentation. This approach was essentially, although not obviously, an analytico-synthetic one. In subsequent years Vladutz has continued to work intensively on the subject of grammar in indexing languages. Kolb reported on favorable experiences gained in the IDC with the nontopological, fragmentational representation of structures and reactions. He, too, placed emphasis on syntactical devices and on the combination of a fragmentation code with the topological approach. In my own contribution to this symposium I tried to show under what circumstances a highly developed indexing language is indispensable in the long run, if high demands are made on the accuracy of the retrieval and on the survival power of the entire information system.

Especially in our era of large and rapidly expanding information systems we should be fully aware of the capabilities and limitations of indexing languages and in fact of those which are not restricted to their vocabulary. If we correlate the extraordinarily good performance of topological methods in chemical documentation on the one hand with their richness in grammar on the other, then we hold in our hands the key to the understanding of many deficiencies in our contemporary information systems and also the key to their substantial improvement.

[†] Presented before the Division of Chemical Information, 183rd National Meeting of the American Chemical Society, Las Vegas, Mar 30, 1982.

0095-2338/82/1622-0117$01.25/0

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.