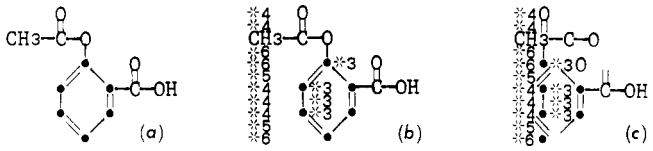

Figure 16. Computer program condensation of structure. Aspirin as typed (*a*), coded (*b*), and coded and condensed (*c*)

the character on the line except where there are no more significant characters until the right hand margin. With all of the left side blanks coded, we can, in effect, push the structure to the left, squeezing out the left side blanks, and then drop the right side blanks. By performing this process, we can reduce the number of positions required to store this structure from the 154 we started with to 61 without record marks and 72 with—a condensation of more than 50%. Figure 16,*c* shows aspirin coded and condensed. Actual storage of the structure is in one continuous record of the lines concatenated, with demarcation characters between the lines.

ACKNOWLEDGMENT

The author gratefully acknowledges the important contributions to this work made by Maxwell Gordon of SK&F and the late J. M. Mullen of Shell Development Co. Thanks are also due Paul Craig, Helen Ebert, and Marianna White of SK&F for their technical and editorial assistance and encouragement.

LITERATURE CITED

(1) Craig, P. N., and H. M. Ebert, "Eleven Years of Structure Retrieval Using the SK&F Fragment Codes," J. CHEM. DOC. **9**, 141–6 (1969).
(2) "Chemical Typewriter Prepares Ring Structures, Complex Formulas," *Chemical and Engineering News* **30**, 2622 (1952).
(3) Feldman, A. P., D. B. Holland, and D. P. Jacobus, "The Automatic Encoding of Chemical Structures," J. CHEM. DOC. **3**, 187–9 (1963); "Survey of Chemical Notation Systems," NAS-NRC Publ. **1150**, p. 424, Washington, D.C., 1964.
(4) Mullen, J. M., "Atom-by-Atom Typewriter Input for Computerized Storage and Retrieval of Chemical Structures," J. CHEM. DOC. **7**, 88–9 (1967).
(5) Gordon, M., "The Potential Impact of Chemical Typewriters on Documentation," *Pharm. Ind.* **28**, 893–7 (1966).
(6) Rice, C. N., K. D. Ofer, R. B. Bourne, and S. W. Logan, "A Pilot Study for the Input to a Chemical-Structure Retrieval System," *Abstracts of Papers*, B14, 151st Meeting, ACS, Pittsburgh, Pa., March 1966.
(7) Jacobus, D. P., K. H. Zabriskie, and M. Gordon, "Compatability in Chemical Information Systems," J. CHEM. DOC. **9**, 118–25 (1969).
(8) Waldo, W. H., and M. DeBaker, "Printing Chemical Structures Electronically: Encoded Compounds Searched Generically with IBM-702," *Proc. Int. Conf. Scientific Inform.*, Washington, D. C., November 16–21, 1958, NAS-NRC, Washington, D. C., 1959; J. CHEM. DOC. **2**, 1–2 (1962).
(9) Hyde, E., and L. Thomson, "Structure Display," J. CHEM. DOC. **8**, 138–46 (1968).

# The Multiterm Index: A New Concept in Information Storage and Retrieval*†

HERMAN SKOLNIK
Hercules Incorporated, Hercules Research Center, Wilmington, Del. 19899

Received September 2, 1969

An index not only *can* be a creative communication medium, it *needs* to be in a research and development environment. A creative index is achievable if the relationship and association of things and actions, one to another, can be communicated as a continuous function vis-a-vis the real world of science and technology.

A chemist does not think of a chemical, for example, ethyl alcohol, in isolation. Ethyl alcohol is not merely a word or a term without dimensions to a chemist. It is a concept that he associates with or relates to a product, a reactant, a solvent in a reaction, a use, a property, etc. It is within the semantics of his conceptual needs that he would like to use an index to retrieve those documents he needs. He wants more than documents, however, from the index. He wants the index to direct him to only those documents which are pertinent to his problem. He wants the index to help him to generate thoughts and to suggest new combinations. He wants the index to help him in terms of his language, logic, and semantics and through a generic or specific approach, whichever occurs to him first. He wants the ability to browse among the terms to discover the term that is on the tip of his tongue or recessed in his memory. These are the criteria an index must satisfy if it is to be a creative medium of communication.

Indexing via a strictly dictionary logic is the most prevalent noncreative system. For example, in this type of index, LIGHTNING and LIGHTNING BUG must be placed in a strictly alphabetical order within the L's, and there is no other alternative. The uniterm index, probably the most popular of the dictionary types, in its simplest form would post a document concerned with lightning bug under the separate terms LIGHTNING and BUG (or INSECT, if so directed by a thesaurus). More sophisticated uniterm indexes would employ roles and links to differentiate and to relate some terms. Semantic control in most uniterm indexes, if exercised at all, is through the use of roles and a thesaurus.

An index which has a purpose and which relates with

*Presented before the Division of Chemical Literature, 158th Meeting, ACS, New York, September 1969.

†Contribution No. 1487 from the Research Department of Hercules Incorporated.

HERMAN SKOLNIK

**A new indexing system, the "multiterm" system, is described. It was conceived and developed to achieve through assigned terms and combination of terms a communication of relatively high informational content and discriminatory power. In its simplest form, a multiterm is a combination of correlatable terms, such as**

**C/R/P/A/**

**in which C may be a chemical prepared from reactant R by process P using catalyst A. The oblique stroke or virgule following each of the four terms is the program code for the computer to permute the multiterm; in the above example, three additional multiterms would be generated:**

**R/P/A/C/**
**P/A/C/R/**
**A/C/R/P/**

**The importance of logic and semantics in the assignment of each term, the importance of employing a defined directional order for the terms in each multiterm, and the necessity for imposing fixed punctuation in a computer operation are stressed.**

the information needs of users is deeply concerned with the logic and semantics of the terms used. This consideration might result in the use of LIGHTNING and LIGHTNING BUG, but most likely, depending upon the subject areas of interest to the users, other alternatives would be chosen, such as

WEATHER:LIGHTNING
or ENERGY:LIGHTNING

INSECT:LIGHTNING BUG
or INSECT:FIREFLY
or LUMINESCENCE:FIREFLY

Logic and semantics are thus the characteristics that distinguish between a subject and a word in an index.

Another consideration in the design of an index is the need for each term to be an optimum communication. This is important to the user who needs a direct match between his interest and a document, and who wants to read only those documents he needs to for his given purpose. A single word in an index carries the least amount of information and the maximum amount of noise. For example, the term ALLOY means too little or too much to a searcher. The least the index system should do is to differentiate ALLOY documents in terms of preparation, reaction, property, use, etc. In anticipation of computerization, and the need for input and processing economy for computer operations, let us indicate differentiating roles of terms by the following mnemonics:

| | |
|---|---|
| A = Analysis | R = Reaction |
| E = Effect | T = Treatment |
| P = Preparation | U = Use |
| Q = Quality or Property | |

Then,

ALLOY -Q

tells the searcher that the document is concerned with the property or properties of an alloy or alloys. Considerably more information is imparted if ALLOY is treated as a class and members of the class are delineated as:

ALLOY:Al -Q

This treatment implies, however, that alloys are more important than aluminum to the users of the index. Assuming that such is the case, the class:member relationship allows the searcher to browse among all the specific alloy terms or to find directly all documents concerned with the properties of aluminum alloys. The simple expediency of crossing ALUMINUM ALLOY to ALLOY:Al, as long as such crosses are consistent and logical with the purpose of the index, serves the needs of the user.

The searcher, however, needs a still more discriminating communication than ALLOY:Al -Q to be able to separate desired documents from the total on properties of aluminum alloys. His interest, for example, may be in the specific property, stress corrosion, and, more specifically in test methods for studying the stress corrosion of aluminum alloys or for any alloy. If a document is concerned with test methods for studying the stress corrosion characteristics of aluminum alloys, the following combination of terms constitutes a communication with a large amount of information:

ALLOY:Al -Q / STRESS CORROSION / TEST METHOD /

We call this system of indexing multiterm indexing. The oblique stroke or virgule (/) following each term is our code for the computer to permute the multiterm so that the following two entries, in addition to the one above, will be alphabetized automatically and correctly in the printout:

STRESS CORROSION / TEST METHOD / ALLOY:Al -Q /

and

TEST METHOD / ALLOY:Al -Q / STRESS CORROSION /

Consequently, this document is retrievable from each of the three terms, and each of the three multiterms communicates the same amount of information. Indeed, the information content of a multiterm is essentially that of a compressed abstract. Because the terms are assigned or selected by their meaning and relationship to the needs of the users, the multiterm can convey considerably more information than authors generally put into titles of documents. An author writes a title to convey the concepts he considers to be important to his objectives; an indexer needs to select terms that convey the concepts that are important to the users of the index. The difference can be considerable.

In a very direct manner, a multiterm is an association of terms in a coordinate relationship. It communicates not only a complex relationship, but it communicates an amount of information considerably greater than that contained in the sum of each of the single terms alone in contrast to the coordination of uniterms.

It is seldom that a chemist wants to retrieve all documents on a single subject, such as ETHYLENE GLYCOL. He probably has something more specific in mind, such as all documents on the preparation, reaction, property, or use, and, most often, within this specificity, his interest may be even more specific, such as documents on the preparation of ETHYLENE GLYCOL from ETHYLENE by one of several processes, such as EPOXIDATION. The multiterm of interest would be:

ETHYLENE GLYCOL -P / ETHYLENE -R / EPOXIDATION / CATALYST:X -U /

or any of its permuted multiterms:

ETHYLENE -R / EPOXIDATION / CATALYST:X -U / ETHYLENE GLYCOL -P /
EPOXIDATION / CATALYST:X -U / ETHYLENE GLYCOL -P / ETHYLENE -R /
CATALYST:X -U / ETHYLENE GLYCOL -P / ETHYLENE -R / EPOXIDATION /

In a uniterm system, even in one with roles and links, the searcher tends to stop with one term if the number of documents posted is not very large; the probability of going beyond the combination of more than two terms is exceedingly low. Furthermore, the probability of the indexer's posting the document under OXIDATION is generally determined by the emphasis of the author of the document, and the probability of the searcher's using OXIDATION or EPOXIDATION is determined solely by how he regards the reaction at the time of search. It is highly unlikely that both the indexer and searcher would have equal appreciation of the catalyst term. Experienced users of uniterm systems tend to have a Pavlovian complex in combining too many terms for they have learned how easy it is for coordination of terms to result in *no documents* of interest.

In the multiterm system, the searcher in the above example has four equal entry points that can result in a retrieval. Furthermore, each of the four multiterms gives him more information than he might have before entering the index. For example, the multiterm index lists for him the various catalysts that the documents report as being used in the epoxidation of ethylene for the eventual preparation of ethylene glycol; it reveals to him those documents which the indexer (or author) may have considered as OXIDATION rather than as EPOXIDATION, either correctly or erroneously. Furthermore, it reveals to him, on browsing, those documents that report the preparation of ethylene glycol directly from ETHYLENE OXIDE.

Discriminatory power is the major advantage of the multiterm system and browsability is a close second advantage. If terms are selected within a prescribed grammar, logic, and semantics, we need worry no longer about recall or precision.

To attain optimum discrimination and browsability in the multiterm system, it is important to establish a logic of directional order. The most reasonable directional order is from generic to specific. If there is no generic:specific relationship, as is the case in a chemical reaction, the directional order needs to be defined carefully. Thus, in the ETHYLENE GLYCOL example above, and in all chemical reactions, the directional order is: chemical being prepared → reactant(s) → process → reaction conditions (catalyst, solvent, etc.) → equipment → use of chemical prepared → property of chemical prepared.

Permutation of a multiterm with an imposed directional order yields multiterms with directional order. For example, the following multiterm:

PRODUCT -PQ / REACTANT -R / PROCESS / CATALYST / USE / PROPERTY /

not only allows controlled browsing among the PRODUCT multiterms so that the searcher can discover all documents which are concerned with the preparation of the chemical of interest from any specific reactant, but allows him to discern the preparation from other reactants and by other processes, and to ascertain the validity of other terms with which he might extend his search meaningfully and profitably. Thus, the searcher has available for browsing purposes five additional multiterms:

REACTANT -R / PROCESS / CATALYST / USE / PROPERTY / PRODUCT -PQ /
PROCESS / CATALYST / USE / PROPERTY / PRODUCT -PQ / REACTANT -R /
CATALYST / USE / PROPERTY / PRODUCT -PQ / REACTANT -R / PROCESS /
USE / PROPERTY / PRODUCT -PQ / REACTANT -R / PROCESS / CATALYST /
PROPERTY / PRODUCT -PQ / REACTANT -R / PROCESS / CATALYST / USE /

Because of the directional order of the initial multiterm, permuted multiterms are produced in the computer printouts in the same directional order as the multiterms for other documents with fewer or more terms. Thus, if the searcher's primary interest were the preparation of glycols in general, he would find under ETHYLENE GLYCOL -P, other processes, such as HYDRATION and HYPOCHLORINATION with ETHYLENE as the starting material. If EPOXIDATION were the searcher's primary interest, the epoxidation multiterms would lead him to the various starting olefins and the respective glycol products and, most important, the different oxidizers that had been used, such as *t*-butylhydroperoxide, hydrogen peroxide, peracetic acid, and perbenzoic acid. Consequently, the multiterm system provides the searcher with a means for probing into the literature with terms and concepts he might not think of otherwise, and, in this sense, the multiterm system enhances the searcher's creativity.

Just as the arrangement of terms in the multiterm must have a logic, so must the terms by themselves. For example, in the indexing of chemicals used as catalysts, a decision must be made on which is the more important concept: the chemical or the fact that it is a catalyst for oxidation, hydrogenation, polymerization, etc. From a chemist's viewpoint, it is more meaningful to have a generic:specific relationship for catalysts in association with a process. In the multiterm system, we indicate the generic:specific relationship with a colon,

CATALYST:BORON OXIDE
CATALYST:CUPRIC CHLORIDE
CATALYST:COBALT ACETATE
CATALYST:NICKEL OXIDE
CATALYST:PALLADIUM ACETATE

This generic:specific relationship in proximity to a process term, such as OXIDATION, allows the searcher, by merely scanning the multiterm printout, to extend his research by the alternatives suggested, and, to this degree, the multiterm system again enhances the searcher's creativity.

Imposing a logic on each category of terms also has the advantage of placing like things together. But "like things" are really determined by the needs of a community of scientists. If polymers are an important area of research to this community, it is a disservice to scatter polymer terms from ACRYLATE POLYMER to XYLENE POLYMER in an alphabetical arrangement. Such an arrangement does bring together acrylic acid and its esters with the polymers prepared from them. With respect to meaningful and associative relationships, however, it is far better to bring polymers together in a class. The prefix "poly" is an unsatisfactory mechanism in an alphabetical index as the "poly" term would be interrupted between the "poly(1...)" and "poly(met...)" terms with terms such as Polymer, Polymerization, Polymerization Catalyst, and Polymerization Inhibitor. This interruption can be many pages in a computer printout. The logic of generic:specific, on the other hand, satisfactorily places all polymers together as an associated whole which makes browsing easy and which permits the detection of relationships. Examples of generic:specific polymer terms are:

POLYMER:ACRYLATE
POLYMER:AMIDE
POLYMER:BENZIMIDAZOLE
POLYMER:BUTADIENE
POLYMER:ETHER
POLYMER:ETHYLENE
POLYMER:PROPYLENE
POLYMER:VINYL CHLORIDE

Another important point of decision is how to handle terms which are associated with term modifiers. For example, the stability of chemicals is an essential area of interests to many chemists, and, in particular, to polymer chemists. If the needs of the users require the index to stress the property STABILITY, then the kind of stability is probably best indicated as follows:

STABILITY, BURNING
STABILITY, HEAT
STABILITY, LIGHT
STABILITY, STORAGE

Thus, the comma is used to indicate the relationship between a term and its modification; a colon is used to indicate the generic:specific relationship; and a hyphen is used to indicate that the term is related to its preparation, reaction, use, or property. Another punctuation rule we have adopted, particularly because our multiterm system is computer processed, is the use of a space between a term and a hyphen, e.g., ETHYLENE -R. The space is necessary for control of the computer alphabetization: without the space ETHYLENE, BROMO and other ethylene derivatives could precede and be interspersed in the ETHYLENE-X terms. But this is essentially a matter of living with the characteristics of a computer.

The primary purpose of this paper has been to introduce and to describe broadly multiterm indexing. Subsequent papers will describe the system in more detail with examples of operating indexes which were designed and established by the concepts discussed in this paper.

(Reviewers of this paper have urged me to consider a comparative study between the multiterm system and the uniterm, facet, permuterm, and permuted title systems; also a comparative study with "articulated" subject indexing was suggested. These studies would be interesting and will be considered as time permits. One reviewer suggested an elaboration of the advantages of multiterms over KWIC and KWOC, as each involves permuting via computer. His comments do this quite adequately: [multiterm system has] "the great advantage of internal order. Thus, in keying in on ETHYLENE OXIDE, for example, you don't have to scan lines *left* and *right* for the starting material; you know *just* where it will be.")

In summary, multiterm indexing is a system for communicating the informational content of documents by coordination of terms in defined directional orders. The philosophy of assigning terms is that of convergence, so that "like things" are together and are logically related to a searcher's interests and information needs. The logic, however, is not a matter of a priori decisions, but rather of an evolving commitment to the needs of those who will be using the index.

## ACKNOWLEDGMENT

I gratefully acknowledge the contributions of the following: Ross H. Petty, Benn E. Clouser, W. R. Payson, Audrey G. Watson, Ruth E. Curtiss, and Barry J. Kocher, Hercules Research Center, Hercules Incorporated.