Classification and Mechanization*

By J. LEIBOWITZ
Office of Research and Development, U. S. Department of Commerce, Patent Office, Washington, D. C
Received February 20, 1963

The purpose of this paper is to elucidate the subject of mechanized information retrieval for the newcomer to the field. This will be done by discussing the problem of generic retrieval of chemical compounds and the application of a machine system to this problem. The situation to be presented involves the case where a user is concerned with knowing what other subjects may be related to a specific inquiry as well as being able to find the subjects thus related.

The choice of this particular example is governed by the fact that it serves to point out the respective roles of classification and the machine in mechanized systems. Moreover, it emphasizes the intellectual aspect of information retrieval and is not only indicative of the difficulties in the field, but is also predictive, at least by implication, of potentially fruitful research areas.

This discussion is based upon experience in the Office of Research and Development in the U. S. Patent Office.

General Discussion.—Classification is the intellectual process which is concerned with the generic relation. A classification system reflects this intellectual process in its expression of generic concepts and in its organization of subject matter in accordance with these concepts.

It is important to take note of the intellectual nature of classification systems in the study of information retrieval. It will help, for instance, in differentiating among retrieval systems. Each system has its purpose and the classification is a reflection of that purpose. If the user is concerned with specific or unique subjects, he has "little interest in the particular relations expressed, and less in the formal pattern"[1] of such relations among these subjects. These are useful to him only insofar as they enable him to find his specific subject. Any other alternative form or basis of classification may be used so long as it achieves the desired result.

The user who, on the other hand, is concerned with the generic associations of things, is also concerned with the basis for such associations. A classification based on size or weight of chemical compounds will not satisfy a requirement for generic relationships based on structure. The user is as much interested in the generic patterns displayed in the classification as he is in retrieval on the basis of such patterns.

Generic concepts can be formulated, apart from knowledge of any particular embodiment of the generic idea and apart from any ability to retrieve such embodiment. One can envision, for example, a particular chemical structure configuration without knowing whether or not anything exists which exhibits this configuration or where to find it, if it does exist. The limitations on the generic expression in a classification system are not due to any limitations on the intellectual capabilities of classification so much as they are due to the practical problems of organizing and retrieving information which embodies the generic expression.

It is important to keep these features in mind, as one studies information retrieval systems. The role of the machine can be better understood as a device to overcome the practical limitations of classification and thus extend its intellectial capabilities.

Genus-Class-Species.—It will be useful to adopt the concepts of Broadfield[2] with respect to "genus," "class," and "species." According to these concepts, the genus is a meaning which is realized in the species; the class is a collection of the species. "The genus is that part of the essence of x, which is predicable also of y and z, differing from x specifically, as distinct from the collection xyz. The group is an assemblage, the genus a unifying principle."

Applied to chemical compounds, the term "aromatic amine" is a genus or, synonymously, a class concept. Also the term, "aromatic primary amine," expresses a class concept, subordinate to and included by "aromatic amine." The class of "aromatic amines" contains, collectively, all specific members such as "aniline," "diphenylamine," "procaine," "sulfanilamide," and others. "Diphenylamine" is excluded from the class of aromatic primary amines while the others remain.

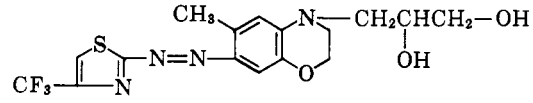
A typical classification schedule for chemical compounds contains class concepts of this nature, in hierarchical subdivision. When one considers the fact that the class concept is a mental concept which defines a characteristic of a chemical compound which may or may not exist, and also that each class concept may be subdivided an indeterminate number of times, it is easy to see that in principle there are no definite limits to the scope of classification schedules.

Generic Retrieval.—Generic retrieval involves retrieval where the search question expresses a class concept and the answers pertain to the class as represented by any and all of its members. For instance, if an inquiry concerns "amino" compounds, the answers pertain to the class of "amines" which includes such members as "procaine," "sulfanilamide," "dimethylamine," and others. This inquiry is too general to be realistic but the class can be delimited by specifying further characteristics, as, "an azo linkage between the aromatic nucleus and an anthraquinone group" and the like.

It should be mentioned that, in the system to be discussed, the answers are obtained in the form of an identification of the documents containing the pertinent information.

* Presented before the Division of Chemical Literature, 142nd National Meeting of the American Chemical Society, Atlantic City, N. J., September, 1962.

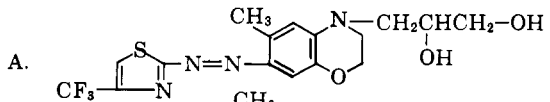
Purpose of Generic Retrieval.—We are familiar with the case where a requirement for generic retrieval occurs when one is faced with a generic class to begin with. For example, in the Patent Office, an examiner may be presented with a claim concerning a generic class of compounds, and he is required to find information pertaining to any and all members to meet that claim. But a requirement for generic retrieval can arise even though the subject initially presented is specific. To render this statement more concrete, consider the following formula as an example of a specific subject.

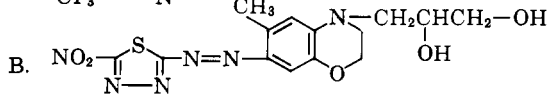
A. 

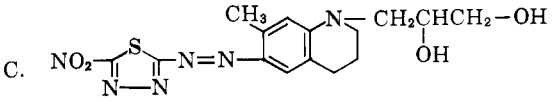
Assume that this is presented for a "prior art" search such as is made in connection with the patentability problem in the Patent Office. It would then be necessary to find a disclosure not only of the compound *per se* but, in case of failure to find such disclosure, to find related or equivalent compounds. The question of equivalence is bound to the patentability problem and to a complex of technical and legal considerations. This topic as such is outside the scope of this paper. But the point is that the examiner seeks, in view of these particular considerations, for a class of compounds which will include the specific compound A and also those other compounds which he regards as related to compound A. Which class he selects depends on, and varies, with the particular problem involved.

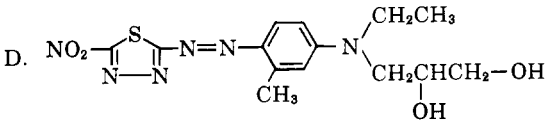
It is probably safe to say that this situation, wherein a search question is generalized from a specific instance, is not peculiar to patent searching. Thus a research chemist, given a specific compound, may wish to find other compounds which bear some relation to it from the point of view of some property or reaction, or any other selected aspect. Which class is selected will vary, as in the examiner's case, with the problem involved. For a particular reaction, an azo group may be equivalent to an acylamino. From the point of view of dyeing properties, these may be unrelated. But from the step where a class concept is selected and generic retrieval is required, the problem of retrieval becomes a common problem, regardless of the original basis for selection and the particular choice made.

Problems.—The formation of a class concept is an act of classification. In the case given, this act is performed by the user of the system. In order to provide retrieval based on the concept formed by the user, it is necessary for the classifier or designer of the system to conceive the same concept and, additionally, to organize the subject matter into the class corresponding to said concept. To consider the problem entailed, a few more formulas are presented (including A).

A. 

B. 

C. 

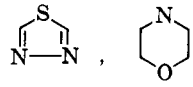
D. 

Structure Characteristics.—We can use chemical substructures to represent the defining characteristics of class concepts concerning chemical compounds. For instance, an "azo" compound is one which contains the structure —N=N—. Compound A, accordingly, belongs to the class of azo compounds, along with B, C, and D. Compound B is also a "nitro" (NO_2) compound, along with C and D, also a "thiadiazole"

together with C and D, and so on.

Combination Structural Characteristics.—Class concepts can be expressed as combinations of these characteristics, *e.g.*, "nitro azo" (NO_2, —N=N—) defines a class which includes B, C, D, but not A; "azo thiadiazole" (—N=N—),

includes B, C, D, but not A; "azo thiadiazole morpholine" (—N=N—),



includes B but not A, C, and D.

Higher Levels of Generality.—Terms defining a higher level of generality may be used, for example, "6-membered ring," "nitrogen ring," "oxygen ring," "5-membered ring," "sulfur ring," and the like. The term "6-membered ring" (compound) includes A, B, C, and D; "nitrogen ring" (compound) includes A, B, C, D, but "6-membered nitrogen ring" (compound) includes A, B, C, but not D.

These terms can be combined, of course, in the same way as in the case of the substructures. In addition, terms at one level can be combined with terms at another level, *e.g.*, "a 6-membered nitrogen ring compound having a nitro group, etc."

It is not difficult to see the problem of generic or classificatory retrieval. There are, of course, innumerable generic classes which could be thus formulated. How can a classification achieve a congruence, in type and number of generic concepts, with the users' requirements, and also provide access to the class members denoted by these genera? Compounds have joint attributes and belong to a multiplicity of classes. Additionally, the classes established in the classification include a multiplicity of common members. A compound in the "azo thiadiazole morpholine" class would also be placed among the "azo morpholines" class, the "azo thiadiazoles," the "nitrogen heterocyclic azo" compounds, and so on. Considering that there may be disclosures of a number of different compounds in each document and that it is the document which is ordinarily the item which is classified in manual

retrieval systems, the task of providing adequate generic retrieval by manual methods is, as a practical matter, well-nigh impossible.

Mechanized System.—The role of mechanization with respect to this problem can be illustrated with a typical machine search system within the author's experience.

A classification system can be constructed from "above" or from "below,"[3] that is, from the general to the specific or from the specific to the general. In the case of the particular system under consideration, it was constructed, open-endedly from "below," that is, from the specific compound to more general class descriptions (Fig. 1, read upward from A, B, C).

This open-ended method achieves a correspondence between the terms in the classification and the subjects classified. There are no superfluous terms in the schedule in that in each class there is at least one disclosed member. Also, each species disclosed has a home in at least one class represented in the schedule.

By taking advantage of the fact that chemical compounds can be depicted by structural diagrams, correspondence is also achieved between user and classifier insofar as definition and scope of the class concept is concerned. The substructure as a definition of the class concept is immediately recognizable in the members of the class. We know the genus "by seeing how it is revealed" in the species or the "modes of its realization".[4] This, needless to say, is not so readily achieved with respect to nonstructural concepts, such as those which deal with such factors as the behavior, reactions, and properties of chemical compounds.

Retrieval.—The following diagram is presented to illustrate the mechanism of retrieval with respect to the system constructed.

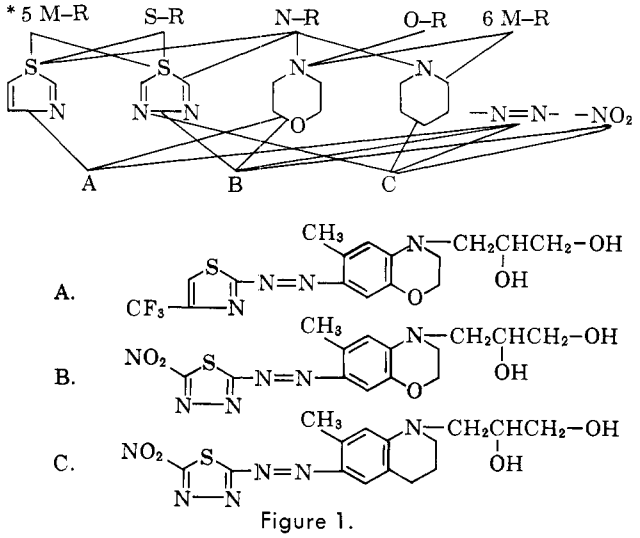


Figure 1.

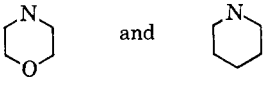
*"M" represents "membered," "R" represents "ring," so that "5 M-R" is read as "five-membered ring," "S-R" as "sulfur ring," and so on.

A few terms only are shown, derived from compounds A, B, and C, in the manner previously described. The lines indicate their derivation, if they are read from below, upward.

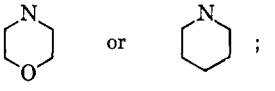
It will be seen that this structure constitutes, in effect, a number of small individual classifications, each individually derived from a specific subject. If the construction of the system is from "below", upward, its use is from "above", downward. For instance, if one were to seek all "oxygen ring" (O-R) compounds, entry into the system is at this point, and we can see that this term includes

compounds, which, as indicated, by the diagram are A and B.

The effectiveness of machine processing in connection with this kind of searching becomes more readily apparent by considering combinations of class terms. Thus, a search expressing a conjunction of (6 M-R. N-R) requires that the compounds found possess a ring structure having 6 members and also nitrogen in the same ring. The machine program directs the operation to the two addresses which correspond to "6 membered" and "nitrogen ring," finds all rings which satisfy the conjunction of these, *i.e.*,



then finds all compounds which possess



in this case A or B or C. If the class is (6 M-R. N-R. O-R) it would locate

and find A and B, only.

Details of this operation are given in a previous report.[5]

Many complex questions involving logical products and logical sums, combined at any level can be created, *e.g.*, "a compound characterized by (a and b) or (c) and a compound characterized by (a and b) and (c or d)," and the like (the alphabetical letters represent the same or different class concepts).

What is done by the machine can, of course, be done by a human searcher. One could scan sets of lists of terms and perform the same logical operations done with the machine. The difficulty, however, resides in the practicality of such a procedure. The lists are long and the logical operations varied. Such procedure can be time-consuming, inaccurate, and inconsistent.

Conclusion.—It will be seen that the machine is useful in connection with combinations of classes. These do not have to be prescribed in the classification but can be created as required and when required. Conjunction and disjunction of classes involves logical operations, and the machine performs these logical operations very effectively. By delegating these operations to machines, the system designer or classifier is given freedom from the burden of prescribing combination classes and can concentrate on more fundamental problems of classification.

It is interesting to speculate about the use of these systems. In a sense, the user participates in the making of the classification system. He can now create and use

many more combination classes than were available in the prescribed classification. Does he create the same class of equivalents with respect to a specific inquiry as he would have selected from the prescribed manual classification? The manual classification does provide guidelines on the basis of the past. It reveals the "pattern of knowledge"[6] to the user. How does the use of a machine system add to and cut across this pattern of knowledge?

There is a great deal of research to be done before these questions can be answered. The making of a mechanized system is only a very small part of much more fundamental studies of classificatory information retrieval systems and their performance in a man-machine environment.

REFERENCES

(1) A. Broadfield, "The Philosophy of Classification," 1946, p. 1, referring to G. O. Kelly, "Classification of Books," 1937, p. 72.
(2) A. Broadfield, *ibid.*, pp. 25–27.
(3) B. C., Vickery, "Classification and Indexing in Science," 1958, p. 11.
(4) A. Broadfield, ref. 1, p. 7.
(5) J. Leibowitz, J. Frome, and F. D. Hamilton, "Chemical Language Coding for Machine Searching," R. & D Technical Notes #3, Office of Research and Development, Patent Office, U. S. Department of Commerce, March 28, 1962.
(6) B. C. Vickery, ref. 3, Introduction by D. J. Foskett, XVII.

An Introduction to Deep (Coordinate) Indexes*

By JOHN C. COSTELLO, Jr.
Information Research Division, Battelle Memorial Institute, Columbus, Ohio
Received February 20, 1963

I. What is an Index?—An index is a guide to the information in a document or in a collection of documents. The word has its origin in Latin and means "to point out, to guide, to direct, to locate." A searcher in need of information consults the index available to him and, through proper use of it, is guided or directed to the source or has the source pointed out to him or located for him. When an collection exists without an index, retrieval of information must be accomplished by searching the documents themselves. Generally the index to the information in a file or in a collection of documents is external to and separate from the file or collection itself. However, the index to the information in a file of documents *can* be incorporated with the information in a number of ways. For example, each document may have entered on its face sheet or cover the index entries for the information it contains, in which case the searcher evaluates each document not by examining the body or text, but rather by examining the index entries. The searcher examines *serially* or *sequentially* the potential sources of information to identify those which are pertinent by the process of comparing description of each with description he is seeking. Separate entry of these index data on cards produces a card catalog, a familiar form of external index or locator device.

There are a number of systems in which either the complete documents themselves or their abstracts have been stored on film or computer tapes. The index entries have been stored with each document to produce a machine-searchable serial file of the collection. The index has been incorporated with and is integral with the file.

The more familiar relationship of index to file is that exemplified by books, in which the searcher is directed to specific page locations in the text, and by card catalogs, in which the searcher is directed to specific shelf locations for sources. In card catalogs, each card represents a document. Each document in the collection is accessible by a number of cards. Typical catalogs include cards for each author and for the several broad classifications to which a cataloger has decided the document properly belongs. The description on each card usually takes the form of a subject heading, a multiconcept statement of information content.

Other familiar forms of conventional indexes classify documents according to broad subcategories of disciplines, such as chemistry or biology. Each reference included in the index is entered usually under only one subcategory, such as Agricultural Chemistry, Dye Chemistry, or Food Chemistry. The bibliographic data may be presented by themselves or an abstract may be included. Access to information in this type of index is more severely limited than in card catalogs since references or access points are much fewer in number.

The filing schemes maintained by most professional persons for their own collections actually are forms of classification indexes. The file folder labels are index descriptions and indexing for these schemes consists of filing each document in that folder on which the index description is considered most appropriate for content.

II. What is a Coordinate Index?—Unlike indexes based on limited information description such as classification

* Presented before the Division of Chemical Literature, 142nd National Meeting, of the American Chemical Society, Atlantic City, N. J., September 10, 1962.