NEW KINDS OF INDEXES

By CHARLES L. BERNIER

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio

Chemists have become well accustomed to five kinds of indexes. They may find it important to consider the possibility of new indexes in addition to the usual subject, molecular-formula, author, numerical-patent, and organic-ring indexes. A number of new kinds of chemical indexes have been proposed. Some are in limited use.

GROUP INDEXES

The chemical-group index,[1] of these new kinds, is one of the more promising. This kind of index enables the searcher to find references to compounds containing a common chemical group or combination of groups. By its use the selection of references to classes of compounds is possible. Generic searches for chemical structures are facilitated.

A number of these chemical-group indexes ("group indexes" for short) in punched-card or computer-tape form are now in use. Those at the Dow Chemical Company,[2] The National Institutes of Health,[3] and The Patent Office[4] come to mind. A group index was used by the Chemical-Biological Coordination Center in Washington.[5]

The searcher of these group indexes selects from a list of chemical groups, such as "amino," "iodo," "ethyl," and "phenyl," those found in the class of compounds in which he is interested. The symbols for the chemical groups chosen are correlated to select all indexed structures containing these groups. If the searcher is interested in parts of standard groups, e.g., the carbonyl oxygen of the carboxyl group, or in groups not on the list, he chooses all listed groups containing those parts or including the unlisted groups. For example, if he wished to find references to chemical structures which contain a 3-carbon ring fragment attached by its central carbon atom to an isopropyl group, he would select "isopropyl group" from the list as well as all carbon-ring groups and correlate these in pairs. The documents or references selected would be examined for relevant material and the remainder rejected.

Mechanized group indexes which correlate structure with properties also have been built.

More elaborate devices for selecting references to chemical structures have been designed. In Mooer's Zatopleg,[6] for example, symbols for all atoms and bonds of molecules are listed in such a way that they can be stored on a computer tape for selection of every combination and permutation of atoms, bonds, and groups in all structures indexed. Mooers' system is probably more flexible with regard to selection of parts of structures than any other yet devised. Experimental work on this system is being carried out by The Patent Office. The structure designations produced by this method are not unique and require a computer or the equivalent for proper use. They are not designed for visual searching.

Another effective group index for searching by computers has been designed by Norton and Opler.[7] This system also includes coding of position isomerism. Thus, 1,2- can be distinguished from 1,3-dichlorobenzene.

Notations, such as the International Notation, can be searched for groups and correlations of groups by computers and even simpler machines. Notations will be discussed a little later in relation to cipher indexes. In this connection, Opler has suggested[8] that it should be possible to program a computer to translate one group index or notation into another group index or notation. If this suggestion proves to be true, as seems likely, then choice of different successful group indexes and notations becomes of less importance because computers can take the searcher rapidly from one notation to the other without laborious manual conversion.

It seems possible now to produce a type of group index in published book form.[1] Possible advantages include: wide, economical distribution to chemists, no waiting for answers from a centralized question-answering service, use by those who do not own or have access to a computer, and no waiting for the computer to be available, programmed, and operated. The results would be available without machine operation. For simple searches involving correlation of but two listed chemical groups the book-form group index would undoubtedly be faster as well as more convenient and less expensive than a computer index. For searches involving correlation of three or more listed groups, the computer would be more accurate and might be faster, but it would nearly always be less convenient. For searches on unlisted groups, parts of groups, and position isomerism, a computer system would perhaps usually be faster and more accurate although the economics would still probably be in favor of the book-form index at this date. The proportional frequency of these different types of search remains to be measured as does the effect of the availability of various kinds of group indexes upon the frequency.

The user of a book-form group index who wished to know, for example, about all steroids containing fluorine could readily locate references to such compounds indexed. This book-form index would enable the searcher to locate references to compounds containing, for example, plutonium, periodate ions, etc., as complete, and usually homogeneous collections. The searcher who used such group indexes would select names or symbols for one or more atoms

or chemical groups found in the compound or class of compounds in which he was interested. Searching in the index by use of the names or symbols for atoms or groups he would discover references to any entered specific compound in which he was interested by its molecular formula associated with its groups. Or, if the index lacked the exact compound, he very likely would be led to discover references to analogous and closely related compounds. It is this last feature of easily selecting analogous compounds in the absence of indexed information on the precise compound desired that makes group indexes in book form especially attractive. Searches of discovery rather than recall will often be directed towards compounds yet unprepared. This is so because the number of known compounds (about a million) is very small when compared with the number for those for which the structures can be drawn. The searcher should always expect to find analogous rather than identical structures for questions involving discovery rather than recall. Knowledge of these analogous structures will usually be useful; occasionally it will be more useful than the information originally sought. For questions of recall, in which information previously read is sought, the searcher should, in contrast, expect always to find the precise reference sought. Group indexes of this nature would make knowledge of systematic organic nomenclature either unnecessary or of minor importance in locating information about compounds. This should be good news for those who do not have the time to learn systematic nomenclature and to keep fully informed thereon.

A group index in book form might have the names of the elements as major subdivisions. Thus, under the "nitrogen" division would be found references to all indexed compounds which contained nitrogen. As page headings under each of these major element divisions there would be all of the groups containing, for example, nitrogen. Thus, there would be group names such as "amino," "azo," "hydrazino," "hydrazo," "nitro," and "nitroso." Under these major page headings for groups there would be subheadings for all of the other groups found in the specific compounds indexed. The index entries would be found under these subheadings and would be easily accessible through all of the groups in the molecules indexed. Finally, the molecular formulas and notations (or names) of the specific structures indexed would be included in order to pin-point the specific compounds for those who required this highly specific type of search. Such indexes could be published to refer to a definite unit of literature, e.g., book, volume of a journal, issue of the journal, etc.

Present-day molecular-formula indexes enable selection of references to individual compounds and also selection of references to one or only a few homogeneous classes of compounds depending on the arrangement of the symbols within the molecular formulas. For example, the Hill system of placing carbon first, hydrogen second, and the rest of the elements alphabetically thereafter enables the searcher easily to find all compounds containing the same number of carbon atoms. The Hill system index also enables rapid location of information about all isomers of a given compound. It was not designed to help the searcher with other kinds of generic questions, i.e., with those involving selection of references to all compounds containing the same atoms or chemical groups. These same statements also apply to formula indexes which use different orders of symbols in the molecular formulas. The Richter system, which employs more classification and is more complicated than is the Hill system, gives some aid in the selection of references to certain classes of compounds.

Other orders of symbols in molecular formulas have been suggested: for example, all elements alphabetically with the exception of carbon (placed next to last) and hydrogen (placed last). Another order (used by G. M. Dyson[9]), is S, N, O, C, H, and all other elements alphabetically preceding these. Each of these arrangements gives one homogeneous type of classification for the molecular formulas in the index. If a different type of classification is sought, these indexes may prove to be of little use. The group index, however, brings together hundreds of complete, homogeneous classes if the index is in book form, and makes literally all classes sought complete and homogeneous, if the index is in the form of punched cards or computer tape. By "complete" is meant that the class contains all references to compounds with given groups. By "homogeneous" is meant that the class contains references to only those compounds with these groups; there is no extraneous material.

One of the more interesting features of the group index in bound form is that it would not be, according to present calculations, an indefinitely large volume. It probably would be, at most, only five or six times the size of the present molecular-formula indexes, since the average molecule studied at present has fewer than six different groups in it.

NOTATION INDEXES

Another type of index possible is the notation or cipher index.[10] Chemists have, at times, experienced difficulty in locating organic compounds by names in subject indexes, particularly the more complex compounds. When the name of the compound is not well known, many chemists approach organic nomenclature and location of information about

compounds through the molecular-formula index. They calculate the molecular formula and then search among the names of the isomers in the formula index in order to select one which they know to be (or which seems to be) the compound in which they are interested. From the formula index they then turn to the subject index. There they may find entries with modifying phrases which increase selectivity of references. They may also find references to useful analogous compounds. This rather roundabout procedure is made necessary by the time required to learn systematic nomenclature. Searching would be more direct if they could turn to a cipher index to locate the compound of interest.

Deriving ciphers from organic compounds has been proved much easier to learn than naming them. An index based upon ciphering should, correspondingly, be much easier to construct and to use than one based upon names. The International Notation derived from the Dyson notation can be learned for effective use, according to some tests run at Chemical Abstracts, in less than one twelfth the time needed for learning systematic organic nomenclature. This advantage of better than twelve to one is somewhat equivalent to buying a $6,000 automobile for less than $500. The cost of producing such a notation index in conjunction with a subject index covering the same material should be somewhat less than the cost of producing a molecular formula index in conjunction with a subject index including systematic nomenclature because of the saving of space by the shorter ciphers and because the cost of naming compounds is greater than the cost of ciphering. I am confident that the differential of at least twelve to one in learning time in favor of notations for organic compounds over nomenclature will induce chemists to use notations in their notebooks -- perhaps in publications -- and to accept them in indexes. Ciphers, as well as names of compounds, can be spoken, for example, into a dictating machine.

Just as going from the name to the structure is easier than the reverse process, so it is much easier to go from the cipher to the structure. This is true because knowledge of numbering and precedence is unnecessary in drawing the structures. Thus, a cipher index could be used by those who knew less about ciphering than would be required for producing a cipher index.

As I see it, the most important function of the cipher or notation is to provide a way of listing organic structures that is more rapidly learned than is the use of a list based upon systematic nomenclature.

CORRELATIVE TROPE INDEXES

Another type of index which has come into existence in the last decade is the correlative trope index. Extensive development of this type of index has been carried out largely by Mr. Calvin Mooers.[11]

Correlative trope indexes[12] are so named because unrelated terms are correlated in the selection of documents. The terms of the vocabulary used in indexing documents are necessarily of a broad, generic nature in order to reduce their number. The terms used for trope indexing are usually generically related to the specific terms that would have been used in standard subject indexing, which is normally to the maximum specificity. These generic terms are closely similar to rhetorical tropes of the usual specific terms, if not actually identical with such tropes. For example, "olefin" might be the trope term used instead of "1,3-butadiene," "thermodynamic properties" might be used in place of "entropy," and "aircraft in place of "heliocopters." The purpose of reducing the size of the indexing vocabulary is to help the searcher find all terms for comprehensive and proper selection. In the use of correlative-trope indexes, two or more words selected from the small generic vocabulary are correlated to select documents or references to documents.

Correlative indexing is, of course, not limited to the use of trope or generic terms in the vocabulary. I have been told that punched cards were used by the Mayo Clinic early in this century for correlating the data of medical case histories.[13] The punched-card indexes that have come into prominence in the last decade are largely correlative. A few, notably those of Mooers,[11] are correlative trope indexes. Difficulties that some index searchers have experienced in answering generic questions probably have been the source of most of the present-day interest in mechanized documentation. Correlation of two or more terms independent in meaning and alphabetization was seen to enable increased selectivity of documents. At one time, it was believed that correlation of any terms, specific or not, would give true generic indexing. Actually, it is now seen that the ability of a correlative index to handle generic questions is a function of the vocabulary and not of the correlation. For example, correlation of the words "nickel" and "electroplating" does not give all information about "coating with metals." It requires correlation of the more general words "coating" and "metals" to produce this result.

The selectivity of correlative indexes is increased by correlation of more terms simultaneously. In fact, correlation of too many terms, or correlation of the wrong terms, increases the selectivity so greatly that blank sorts and loss of relevant information may become serious problems.

Because of the enormous number of permutations and combinations of terms possible from even a modest vocabulary, some of those

documentalists interested in mechanization have argued (probably incorrectly) that correlative indexes in book form could not be produced. It has been thought that the book would be too large. Some studies I have made seem to indicate that, by simple techniques, it is now possible to produce correlative indexes in book form. Independent studies by others have indicated similar possibilities.[14] Research is going ahead in this area.

Correlative indexes in book form have the tremendous advantage of automatically providing the searcher quickly with closely related and analogous information if the precise information sought is unavailable. The index suggests related information during the search just as a mail-order catalog suggests new products by pictures and descriptions found near to the description of the product sought.

Correlative trope indexes, in either book or mechanized form, facilitate generic searches because of the generic nature of the terms and because of the selectivity provided by their correlation. An example of a generic question is, "What reactions of Group-Three elements with chalcogenides have been studied?"

FORMULA-STRUCTURE INDEX

The formula-structure index which we are now studying at Chemical Abstracts is like the standard formula index except that, substituted for the systematic name, is a structure code indicating some important structural features of the compounds indexed.

One simple structure code, The Dyson index[15] (not to be confused with the Dyson cipher), indicates the number of rings of different sizes, the number of methyl and ethyl groups, the number of double and triple bonds, the presence of functional groups, etc. Another structure code, developed by Luhn,[16] indicates the number of "nodes" or atoms attached to 2, 3, and 4 other atoms.

The purpose of these codes is to help distinguish among isomers which a molecular formula, by itself, is unable to do. While some of these codes are not so precise as are systematic names, they usually are adequate for differentiating among the various known isomers, and are very much simpler to learn to use than is systematic nomenclature. In fact, the learning time for the Dyson-index code (not the Dyson cipher) has been shown to be substantially zero; the searcher simply starts off by using it.

An example of a formula-structure-index entry for reference to the compound ethyl-2-nitrobenzene follows: $C_8H_{10}NO_2$ 00011 00010/ 6.8. This code following the molecular formula means that there is one six-membered ring, one aromatic ring, one ethyl group, a nitro group, and more than one ring substituent. The symbols for the chemical elements of molecular formulas in this formula-structure index need not be arranged in the usual order. It may be found more helpful, for example, to have an order in which all elements are placed alphabetically with carbon next to last and hydrogen last.[17] This particular order has advantages, the best of which is that the number of hydrogen atoms in the structure has the least control over the position of the entries in the index. Thus, miscounting or errors in determination of the number of hydrogen atoms then has least effect.

The purpose of studying formula-structure indexes at this time is to enable the construction of interim indexes giving access to references for compounds before the annual formula indexes are available. Personnel to supply systematic names for such interim indexes will be unavailable for some time yet. The training time required for selecting and calculating the molecular formulas and for deriving the structure codes are much less than the time needed to learn systematic organic nomenclature.

A formula-notation index also is possible. It can be used exactly as a formula-structure index simply by ignoring the position isomerism and order of operations.

CITATION INDEX

Another type of index, new to chemists, has been suggested by Eugene Garfield. This is the citation index.[18] It enables the searcher to go from an earlier paper to all later papers which have cited it. In effect, it enables one to discover the descendents of an idea. At present, it is possible to start with a later paper and by means of the references cited to discover earlier related papers. The citation index would enable the searcher to reverse this process. In the legal field, citation indexes have proved to be vitally necessary in locating later legal decisions related to an earlier one. For the field of chemistry the ability to go from earlier papers to later ones citing them may turn out to be much more useful than it now seems. I believe that it would be of more use than simply a device for flattering the earlier author. The citation index would be derived from references cited in chemical papers and would lead the searcher from the earlier papers to all of the later papers related to it by citation and also, usually, by related subject interest.

CONCORDANCE

The concordance, or word index, while not exactly new to chemists, and certainly not new to Biblical scholars, would be of help if the searcher used the exact words of the author to lead him back to the original document. The

concordance might be called a word index since exact words, and their contexts, rather than subjects are indexed. A concordance for an annual volume of Chemical Abstracts would be much larger than the Biblical Concordance and probably would be considerably less useful because word indexing of this nature leads to scattering, omissions, a flood of trivial entries, and bulk.

REACTION INDEX

Reaction indexes have been compiled. Theilheimer[19] is an example. It should be possible to produce reaction indexes for the periodical or abstract literature also.

The kind of reaction is so indexed that if the specific reagents and products sought cannot be located, then analogous reagents and products which undergo the same reaction can be discovered. Analogous reactants and products sometimes may be very difficult to locate in standard subject indexes.

TAXONOMIC OR BIOLOGICAL-NAME INDEX

Chemists study organisms. They are interested in biological reactions, control of pests, chemotherapy, pharmacology, etc. In all of these studies, named organisms usually are of interest.

A taxonomic or biological index would give references to all documents (e.g., abstracts) in which the same organism was studied. Biological Abstracts has such an index. Perhaps chemists would find one useful also.

Such an index could be arranged taxonomically to bring related organisms together in order to make it easy for the searcher to discover studies about related organisms in the event that the specific organism originally of interest had not been studied.

MACHINE INDEXING AND SEARCHING

Another type of searching system has been suggested by H. P. Luhn.[20] In this system, important sentences are extracted from documents by a computer. The searcher does not use an index, but instead, writes an essay about the question. This essay is then fed into a computer which makes a comparison with the extracts of documents already produced and selects those related on the basis of probable likeness to the essay question. If this system proves to be successful, it will constitute a new type of document selector which is neither an index nor a classification.

The extraction of important sentences from the original documents is carried out by the computer which counts and records the number of times each word occurs in the document, eliminates the simple and technically uninformative words, and detects and selects sentences containing most of the more important words located within a certain distance apart.

CLASSIFICATIONS

Classifications and indexes overlap. Subject indexes can be viewed to be alphabetical classifications although they are definitely not hierarchical classifications. Classification is used occasionally in subject indexes as a tool.

Considerable important work now is being done on classifications. Ranganathan's colon classification[21] and the Universal Decimal Classification[22] are two examples. The works of Vickery[23] and others have yielded notations and arrangements of words that are, in effect, standard sentences identifying documents classified. Standard languages have been created. If these standardized sentences, associated with document references, are arranged into an order, say alphabetical, and perhaps permuted, they constitute an index to the material.

The technical thesaurus,[24] which is not an index, is, in effect, a comprehensive classification of terms which may make indexing and the use of indexes easier and more precise. The use of a technical thesaurus, were one built, would lead the searcher from those terms which he knew to all of those which he needed to know in making a complete search in an index or classification. The current state of knowledge in a field could be gained in outline form by the use of such a thesaurus.

SUMMARY

The future importance of these new indexes is difficult to evaluate. I like to think that nothing in the way of effective new chemical documentation is too good for chemists. Certainly chemists need generic searching aids. Group indexes and correlative trope indexes should serve as such aids. Whether these new indexes will appear in the form of books, cards (unpunched or punched), or computer tape is largely, if not entirely, a matter of efficiency and economics, which requires more study. Economics, efficiency, and speed of service needed will largely control whether searching means are to be placed on the desk of the chemist or maintained in a central information-searching agency.

Many chemists have experienced difficulty in learning and using a systematic organic nomenclature. There is probably no one "correct"

name for any organic compound. Systematic organic nomenclature is growing. Time is required to keep up with this growth. Group indexes, notation indexes, and coded formula indexes would help those who have difficulty with systematic nomenclature. So would the use of a notation [10] which is much easier to learn.

The system of machine abstracting under development and studied by Luhn [20] would, if successful, probably make documentation centers practical for answering certain types of questions difficult to handle with conventional indexes. This system would make subject indexing in such a documentation center unnecessary.

Citation indexes [18] sometimes would lead searchers into strange and unpredictable relations. It is very difficult to picture just how useful such indexes would be in the field of chemistry. They probably would be expensive but not especially difficult to produce.

The concordance might, as I have mentioned above, be the least useful to chemists of all the indexes that I have discussed.

Reaction and taxonomic indexes should prove useful, in my opinion, but probably could be incorporated effectively into current subject indexes rather than existing as separate entities.

CONCLUSION

In conclusion, I find the subject of indexes fascinating. I am convinced that chemists need new kinds of indexes. There is promise and encouragement in the number of capable scientists who have been attracted to study index design. Certainly the future for new document selectors, indexes or not, looks very bright.

REFERENCES

1. Bernier, Charles L., "Correlative Indexes I. Correlative Chemical Group Indexes," American Documentation, 8, 306-13 (1957).
2. Nutting, Howard S., Private communication
3. Gamble, Dean F., "A Coordinate Index of Organic Compounds," paper presented at ACS Meeting, March 31, 1955.
4. Frome, Julius, and Leibowitz, Jacob, "A Punched Card System for Searching Steroid Compounds," Patent Office Research and Development Report No. 7.
5. "A Method of Coding Chemicals for Correlation and Classification," Chemical-Biological Coordination Center, National Research Council, Washington, 1950.
6. Mooers, Calvin N., "Ciphering Structural Formulas—The Zatopleg System," Zator Technical Bulletin No. 59, Zator Company, Boston, Mass., 1951.
7. Norton, T. R., and Opler, A., "A Manual for Coding Organic Compounds for Use With a Mechanized Searching System," The Dow Chemical Company, Midland, Michigan, 1956.
8. Opler, A., suggested in a conversation.
9. Dyson, G. Malcolm, "Studies in Chemical Documentation," Chemistry and Industry, 676 (1952).
10. Dyson, G. Malcolm, "Some Applications of the Dysonian Notation of Organic Compounds," paper presented at ACS Meeting, April, 1947.
11. Mooers, Calvin N., "A New Semantic Principle in Information Retrieval Systems," a paper presented before the American Documentation Institute, November 5, 1954.
12. Bernier, Charles L., "Correlative Indexes II. Correlative Trope Indexes," American Documentation, 8, 47-50 (1957).
13. Garfield, Eugene, private communication.
14. O'Connor, John J., private communication.
15. Dyson, G. Malcolm, "Studies in Chemical Documentation," Chemistry and Industry, 676-84 (1952).
16. Luhn, H. P., "A Serial Notation for Describing the Topology of Multidimensional Branched Structures (Nodal Index for Branched Structures)," Research Laboratory, IBM, Poughkeepsie, New York.
17. Skolnik, Herman, and Hopkins, Jane K., "A Simplified Stoichiometric Formula Index," paper presented at ACS Meeting in Miami, 1957.
18. Garfield, Eugene, "Citation Indexes for Science," Science, 122, 108-11 (1955).
19. Theilheimer, W., "Synthetic Methods of Organic Chemistry," Series 2, Vol. X, Interscience Publishers, Inc., New York, N. Y., 1956.
20. Luhn, H. P., "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, 2, 159-65 (1958).
21. Ranganathan, S. R., "Colon Classification and its Approach to Documentation," in Bibliographic Organization, edited by J. H. Shera and M. E. Egan, (1951); "Library Classification as a Discipline," Classification Research Group Bulletin, No. 2, (1957).
22. Bradford, S. C., "The Universal Decimal Classification, 62-86," in Documentation, 2nd Ed., Crosby Lockward and Son Ltd., London, 1953.
23. Vickery, B. C., "Notational Symbols in Classification," The Journal of Documentation, 8, 14-32 (1952).
24. Bernier, Charles L., and Heumann, Karl F., "Correlative Indexes. III. Semantic Relations Among Semantemes—The Technical Thesaurus," American Documentation, 8, 211-20 (1957).