r. INSTITUTE FOR SCIENTIFIC INFORMATION

Interview on July 26, 1962, at the Institute for Scientific Information, Philadelphia, Pennsylvania

People interviewed: Dr. Eugene Garfield, Director; and Dr. Irving H. Sher, Director of Research

Interviewer: I. Moyer Hunsberger

Summary

In addition to publishing Current Contents and Index Chemicus, the Institute for Scientific Information is performing sponsored research on "citation indexes," one of the potential uses of which is in conducting generic searches for certain classes of organic compounds. A research proposal submitted to NSF in 1960 by the Institute requested, among other things, financial support for indexing Index Chemicus in both the IUPAC and Wiswesser notations. Certain more specific budgetary information was requested by NSF, but the Institute never resubmitted the proposal. In order to publicize the ideas concerning chemical notations and chemical codes contained therein, parts of this proposal are reproduced in an Addendum to this report.

The RotaForm Index, a new computer-produced rotated molecular formula index to Index Chemicus, permits certain types of generic searches to be performed. It is planned to supplant this index with RotaDex—a fully rotated molecular formula index coupled with a simple four- or five-character generic code which describes 20 or 25 structural attributes with combinations of 32 alphanumeric symbols. This structural code is designed to distinguish between most isomers in the molecular formula index and to permit relatively simple generic searches. The first column of this code describes any homocyclic rings present and indicates the presence of spiro configurations, while the second column gives analogous information for heterocyclic rings and bridges. Column 3 describes five classes of functional groups containing oxygen (or sulfur), while Column 4 indicates the presence of non-resonating unsaturation as well as four classes of functional groups containing nitrogen (or phosphorus). Of the 26 test questions in the interview questionnaire, the RotaDex code was able to distinguish 10 completely and four partially.

As a follow-up of his doctoral dissertation, Dr. Garfield would like to complete a grammar containing algorithms for converting chemical names to Dyson notation, to Wiswesser notation, to generic codes such as RotaDex, to CBCC-type codes, or to structural diagrams.

Historical Background

As publisher of Current Contents and Index Chemicus, the Institute for Scientific Information has a vital interest in the retrieval of scientific, and particularly chemical, information. Under a contract with the Pharmaceutical Manufacturers Association, the

Institute has coded all literature (from 1958 to 1960) on new steroids for the U. S. Patent Office. As a consultant to Smith, Kline, and French (SKF), Dr. Garfield worked with Dr. George Hager on the development of SKF's modified version of the CBCC chemical code.

Citation Indexes

Certain basic papers in any field of science are cited in many subsequent papers, and each of the latter cites many other papers in this same field. If a "citation index" is formed from all of these literature references, one has a very powerful "zig-zag" coverage of the literature in the given field (128a). If the field in question is a certain class of related organic compounds, examination of the citation index affords one method of performing a generic search. Incidentally, such a search also would yield data on any pharmacological activity exhibited by the compounds in question. The completeness of such a search obviously is a function of the comprehensiveness of the index and of the quality of the citations processed. The citation index can be arranged in a variety of ways so as to specify systematically the authors and journals cited and the citing authors and journals. The Institute for Scientific Information is studying citation indexes under contracts with the National Science Foundation and the National Institutes of Health. Computer print-outs of several hundred thousand reference citations have been prepared. Over 1.3 million citations from 1961 journals have been processed, including the leading chemical journals.

Suggested Comparison of Dyson and Wiswesser Notations

Shortly before the first issue of Index Chemicus was published, the Institute for Scientific Information submitted a proposal (dated February 15, 1960) to the National Science Foundation (NSF) for, among other things, an experimental comparative study of the IUPAC and Wiswesser notations via notation indexes to Index Chemicus. It was hoped that the experience of encoding large numbers of compounds into each notation plus the reactions of the index users would provide valuable evidence as to the relative merits of the two notations. In addition, a generic index based on the CBCC chemical code was proposed; the use of structural formulas of reduced size (Miniprint) also was suggested as a means for distinguishing isomers in a molecular formula index.

The above proposal also requested funds from NSF to make possible a reduction in the proposed subscription price of the then non-existent Index Chemicus to non-profit institutions and to individuals. NSF would not supply funds for this purpose for the reasons given in a letter (dated February 29, 1960) from Dr. Dwight E. Gray (of NSF) to Dr. Garfield; specific changes in the proposal also were suggested in this letter. Dr. Garfield never resubmitted this proposal, as Index Chemicus was launched by obtaining initial support from industry. He told the interviewer he would be glad to resubmit it if there still is serious interest in the facets which have not been tested. He also stated that he would give permission to prepare IUPAC and Wiswesser notation indexes of Index Chemicus; in fact, he said he would distribute such experimental indexes free of charge to subscribers to Index Chemicus, even though this might obligate ISI to continue such a service. However, he added that he personally believes that suscribers would prefer a simple alphabetic and/or CBCC-type generic index to any kind of notation index.

Since Dr. Garfield kindly made available the above proposal and his file of pertinent correspondence, the interviewer was able to arrive at the following conclusions: (1) NSF had no choice but to refuse to fund the proposal in question, (2) NSF clearly explained the basis for rejection and the changes needed to make the proposal suitable for consideration, and (3) Dr. Garfield's idea to compare the IUPAC and Wiswesser notations on a use basis was indeed meritorious; his decision not to supply a more detailed budget breakdown can only be regretted, since the chemical community has been the chief loser. His reasons for withdrawing the proposal were given in his letter of March 16, 1960, to Dr. Gray. Dr. Garfield felt that he could not do the research that would be necessary to supply NSF with the requested budget breakdown.

Because the above proposal contains so many interesting ideas in the field of chemical indexing, the title page and pages 5-8 of this proposal are reproduced in an Addendum to this interview report.

RotaForm Index

An experimental rotated molecular formula index, called RotaForm Index,* covering the first 42 issues of Index Chemicus recently has been distributed free of charge to all subscribers. This Index, produced by special programming of the IBM 1401 and 7090 computers, makes it easy to locate all organic compounds containing a given element which have appeared in Index Chemicus. In addition, the user can perform certain limited types of generic searches, though usually a fairly large number of unwanted compounds may be obtained. For example, one can readily locate all organic compounds containing such elements as cobalt or beryllium. If such a group is small enough, one can, by inspection pick out finer subdivisions of the group in question. This Index appears to be particularly useful in the field of organometallics, where the number of compounds is not especially large.

On the other hand, it would not be profitable to use the RotaForm Index to search for all compounds containing, for example, two benzene rings and 2 nitro groups anywhere in the molecule. A search like this would leave the searcher with a large number of entries to scan visually, and, even then, he could not tell the structure from the listing. However, if this search were made much more specific by adding the requirement that phosphorus also must be present, a much smaller list for visual scanning would result.

Although it may seem rather strange, Dr. Garfield stated that some users of Index Chemicus perform generic searches by scanning through this publication issue by issue. The large number of structural formulas makes such scanning less tedious than it would otherwise be. Some users report that they can scan an entire issue in 15 minutes. In the opinion of the interviewer, resort to such scanning merely indicates the need for a good subject or generic index.

*Garfield, E., J. Chem. Documentation, 3:97-103, 1963. This article was in press at the time of the interview.

RotaDex

This system has been developed by Dr. Sher and presently is in an experimental stage. A very simple, four-column generic code is coupled with a rotated molecular formula index to distinguish between isomers. Such a purpose does not require a detailed structural code, if only because there are, on the average, only 3 different structural isomers represented by a given molecular formula. This generic code also would make possible a generic index, if that were desired. Even without a generic index, this code would enable one to obtain more satisfactory results from a generic search than now are possible with the RotaForm Index. In fact, this four-column code would be printed opposite to the molecular formula entry.

The print-out of this code would be columnar, i.e., the same kind of structural information would always be in the same column; this would facilitate visual scanning. The code uses 32 of the 36 alphanumeric characters (the digits 1, 2, 5, 0 are not used) to describe 20 structural attributes of a given compound. Mnemonic code terms facilitate learning of the code. Any one of the 32 characters, when used in the structural code, signifies the presence or absence of any combination of 5 chemical features.

The main features of the code now will be described.

In Column 1 the five chemical features are: (1) homocyclic ring (any size and any degree of saturation), (2) fusion of two homo-rings, (3) fusion of three homo-rings, (4) fusion of four or more homo-rings, (5) spiro configuration. Any of the letters from A to P indicate that at least one ring is present. For example, the letter A indicates that the only homocyclic rings present in a compound are single; if a compound simultaneously contains a single ring, two ring fusions, three ring fusions, four ring fusions, and a spiro configuration, the letter K is used.

Column 2 gives similar information to that in Column 1, but for heterocyclic rings. Bridged carbocyclic or heterocyclic configurations are indicated instead of spiro configurations.

Column 3 is reserved for oxygen and sulfur groups, but no distinction is made between a given oxygen compound and its thio analog. The following mnemonic symbols are used for the indicated structures:

ac = free acid, salt, ester, amide or mono- or dithio analog.

one = aldehyde or ketone carbonyl or thio analog.

ol = hydroxyl or thiol.

ox = −O− or −S−

diox = $Y\overset{=O}{\underset{=O}{}}$ or thio analog, where Y might be S, N, etc.

Column 4 is reserved for groups containing nitrogen or phosphorus and for unsaturated linkages. The following mnemonic symbols are used:

am = N in primary or secondary amines.

ene = any non-resonating double or triple bond between any elements.

az = uncharged N with no H attached.

plus = ammonium N, sulfonium S, phosphonium P, etc.

diaz = N to N bond.

Demonstration on Compound List and Test Questions

Of the 26 questions listed in the interview questionnaire the RotaDex code was able to distinguish 10 completely and four partially, while it was not able to distingiush the following 12: (a) unbranched and branched chains; (c) alcohols and phenols; (f) primary, secondary, and tertiary alcohols; (g) cis-trans isomers; (h) d, l, dl, and meso forms; (i) alpha and beta forms of steroids; (k) ring position isomers; (l) resonating unsaturation in rings; (n) quinones; (p) caged rings; (q) three-dimensional structures; and (v) partially known and partially unknown structures. Search questions (aa) and part 1 of (cc) could be answered; (bb) and parts 2, 3, 4, and 7 of (cc) could be partially answered, while parts 5 and 6 of (cc) could not be answered.

Suggested Mechanical Conversion of Chemical Names to Chemical Notations

As a follow-up of his doctoral dissertation (128), Garfield would like to do further research on what he calls "Chemtran;" for example, he would like to try to devise algorithms to convert chemical names to IUPAC notation, to Wiswesser notation, to the CBCC or other fragmentation code, and to structural diagrams. Thus, Chemtran may be considered as a grammar of organic nomenclature. The picturesque term, "Chemonym," has been suggested by Dr. Garfield for any of the above products from the algorithm; in this terminology the IUPAC and Wiswesser notations for each compound would be chemonyms. He believes that, at least with respect to the older literature, one virtually is forced to start with names. Among other problems, it would be necessary to devise a system for recognizing different names used for the same compound.

An algorithm for converting a chemical name to a notation would permit a computer to perform the operation, thus making the conversion cheap and removing all subjective influences from encoding. Dr. Garfield believes that different chemists at present will, in many cases, write different notations for the same compound.

Dr. Garfield believes that if a computer which could "read" structural formulas were available the need for a notation would be removed. Such computers, he thinks, are not far in the future, since great progress is being made in the field of character and pattern recognition.

Parenthetically Dr. Garfield called attention to his Institute's "Copywriter"—presently developed to an experimental stage through a grant from the Council on Library Resources. This device probably will be made into a character recognition device. At present it is a facsimile device which makes possible the selective and instantaneous copying of printed matter (words, phrases, sentences, or small structural formulas) on electrosensitive paper tape. Such an instrument could have important applications in the removal from the literature of the information needed for citation indexes and in connection with Miniprint structural formulas.

Returning to the subject of Chemtran, Dr. Garfield said that it should be possible to convert one notation to another by use of a computer, particularly because at present a structural formula can be stored, and presumably examined in a computer.

General Views Expressed

Dr. Garfield thinks that every chemical company will continue with its own fragmentation code until a very detailed, universally applicable code becomes available. He thinks that a sufficiently detailed fragmentation code would be usable on a file containing more than 2 million compounds. If too many false drops are obtained from a generic search for a certain substructure, the code need only be made more specific.

Dr. Garfield asserted that serial numbers suffice for identification of the compounds located by a generic search. If these serial numbers and the corresponding structural formulas can be viewed on microfilm, a notation is not necessary in order to learn the identity of the compounds. Furthermore, he thinks it is cheaper to use the serial number than the notation.

Dr. Garfield expressed the opinion that Dyson's computer searches for substructures will be very expensive from a programming standpoint; he thinks prompt programming also will be difficult to maintain. Dr. Garfield thinks the Gordon-Kendall-Davison notation might be better suited for computer searches than either the IUPAC or Wiswesser notation. Both Drs. Garfield and Sher believe that a retrieval system such as Dyson envisions for CA (in which a search for any conceivable substructure can be performed) will be exceedingly expensive. The possibility of infinite flexibility also was questioned by both. They believe that it would be much more practical and certainly cheaper to have a system which can perform a wide variety of searches, but not necessarily any conceivable search.

Dr. Garfield emphasized that coordinate (Uniterm) indexes can be adapted to any code, but he believes the mechanics involved would be extremely tedious with a file of over two million compounds.

Although CA covers some 9,000 journals, Dr. Garfield pointed out that about 25 journals contain 90 per cent of all the new organic compounds. He thinks that one cannot justify the tremendous cost of examining 9,000 journals merely in the name of completeness, particularly since he believes that virtually all reliable new information is published in one of the 25 journals. He thinks it is valid to wonder if the work appearing in obscure journals is worth all the effort needed to locate it. Citation indexing will surely pick up this remaining residue.

ADDENDUM

TO

Interview on July 26, 1962, at the Institute for Scientific Information, Philadelphia, Pennsylvania

PART A: COVER PAGE OF NSF PROPOSAL

A PROPOSAL CONCERNING THE PROMPT DISSEMINATION & RETRIEVAL OF INFORMATION ON NEW CHEMICAL COMPOUNDS INCLUDING A COMPARATIVE STUDY OF MOLECULAR FORMULA, "MINIPRINT", ALPHA-GENERIC AND LINE NOTATION (CIPHER) INDEXES.

Submitted to the

National Science Foundation

by the

Institute for Scientific Information
Philadelphia 22, Pa.

February 15, 1960

SUMMARY

A new monthly publication, the INDEX CHEMICUS, is proposed. The INDEX CHEMICUS will include a register of all newly synthesized chemicals reported in the scientific journals. For each article complete bibliographical information will be given as well as pertinent structural diagrams, chemical names, and unique page and serial number identification for individual compounds. The format of the proposed pocket-size publication is illustrated. Examples are shown for the register as well as the monthly molecular formula, author, institutional and journal indexes. These indexes will be cumulated quarterly and yearly.

In addition, it is proposed that the Foundation support the publication, for one year for experimental purposes, of several quarterly and yearly indexes including two line-notation (cipher) indexes (Dyson and Wiswesser); a "Miniprint" molecular formula index which would include reproductions of an "LC" chemical compound card containing structural diagrams, chemical names, ciphers, etc.; and an alphabetically arranged generic index prepared by use of the Chemical-Biological Coordination Center (CBCC) chemical code.

It is suggested that these indexes and the INDEX CHEMICUS would be distributed gratis to a sample of 1,000 chemists, each of whom would receive three monthly issues as well as the quarterly molecular formula, line-notation and generic indexes. In addition, all subscribers to the INDEX CHEMICUS would receive the yearly cumulations. In this connection, support from the Foundation would enable academic subscribers to receive the INDEX CHEMICUS at a reduced cost of $100 per year.

PART B: PAGES 5-8 OF NSF PROPOSAL

G. NATURE OF FINANCIAL SUPPORT REQUESTED

Two basic factors will be affected by NSF support. These are: 1) Rate Structure; 2) Additional features and indexes.

1. Rate Structure of the INDEX CHEMICUS without support from NSF will be $500 per year for industrial subscribers and $250 for educational institutions. We believe this will seriously affect the availability to academic research workers. It is obviously beyond the reach of individuals, small research organizations, and many colleges and universities. With NSF support the price structure will be established at $100 for non-profit organizations including educational institutions, government organizations, public libraries, hospitals, etc. We do not wish to have an artificially low or high price. However, at this price the INDEX CHEMICUS will be within the reach of all who need it.

2. Additional Features and Indexes.

   a) STRUCTURAL DIAGRAMS. Without initial support from NSF the INDEX CHEMICUS will include only one structural diagram for each paper indexed. This would be either a parent compound or a typical compound as illustrated in the example in paragraph B. However, we fully recognize that this is short of the ideal, that it would be preferable to provide a structural diagram at least for all compounds in a paper which are different enough to create confusion in the minds of the searcher. It is questionable whether a structural diagram should be provided for a long series of compounds in which only one R group is modified.

   b) LINE-NOTATION (CIPHER) INDEXES
   Much has been said about the use of the Dyson and Wiswesser systems for indexing chemical compounds. However, an adequate operational test from the user's point of view has never been conducted. With support from the NSF both quarterly and yearly indexes by both systems would be prepared. This would, in effect, be equivalent to preparing a Library of Congress (LC) card for each compound. The individual user will make the decision to use either, both, or neither cipher system. We believe that one full year's experience indexing compounds by both systems will provide useful information on both systems particularly from the indexer's viewpoint. We have made arrangements with the authors of both systems to make sure that each method receives fair treatment. We also believe that a survey of users, at the end of the first year, may be indicated. If a formal survey seems appropriate a supplementary request would be prepared. To make certain that a sufficiently large enough sample of users is available to survey we propose to distribute the INDEX CHEMICUS gratis to 1000 chemists for a period of three months. Each chemist would receive three issues of the INDEX CHEMICUS and the corresponding quarterly indexes. In addition, the yearly cumulations would be distributed to all INDEX CHEMICUS subscribers.

c) ALPHABETIC INDEXES

We believe that an alphabetical word index to key functional groups and generic types in a line notation index would not only greatly facilitate the location of compounds sought by the inexperienced reader but would also help to familiarize him with the notation system without the necessity of frequent referral back to the book of rules. For this reason an alphabetic index would be added to help the reader find commonly occurring configurations without learning the entire cipher system. If cipher indexes are to be used for generic searches then such common headings as thiazines, sulfonamides, androstanes, nitrofurans, etc., would be helpful to the reader as he scans each monthly issue. These alphabetic indexes would be comparable to the LC Subject Heading List and the Dewey Relativ Index.

d) MOLECULAR FORMULA INDEXES CONTAINING STRUCTURAL DIAGRAMS

It is quite generally known that in searching CA for a particular compound one may start with a name or a structure. Rather than attempt to determine the correct CA nomenclature for this compound it is simpler to compute the molecular formula. Having done this it is then possible to check the molecular formula indexes.

However, when there are many isomers for the same formula it is then necessary to decipher the various names provided. This could be eliminated by providing in the molecular formula indexes a structural diagram for each compound.

Obviously this would increase the cost and size of the conventional molecular formula index.

We believe that there may be a reasonable compromise to this problem through "Miniprint". This term was coined to distinguish low reduction ratio photography from high reduction ratio such as Microprint cards. In our experiments with "Miniprint" (see attached sample pages) we have found that one can save 90 per cent of space and printing costs. In addition, structural diagrams are quite discernible to the naked eye. Further, with the use of the simplest inexpensive magnifying glass printed words become quite legible. We believe experiments on an operational level with a molecular formula index containing structural diagrams is now justified. A "Miniprint" index of complete structural diagrams could be printed on 250 pages and cover over 50,000 compounds.

There are two practical reasons for adding a "Miniprint" molecular formula index if the Dyson and Wiswesser cipher indexes are included.

1. In order to efficiently and accurately construct a line notation cipher for an individual compound it is first necessary to draw the structural diagram. This step, however, is one of the most expensive facets of chemical indexing. If the structural diagram must be drawn then the raw material for a formula index containing structures has been created. It is then a relatively simple operation to prepare the "Miniprint" index by sorting the 3" x 5" slips containing the structural diagrams, chemical names, ciphers and journal

references. Indeed, this 3" x 5" slip is the direct counterpart of the LC card mentioned above. Its use would enable the reader to determine the specific cipher for a particular compound, further facilitating the use of the cipher indexes. It would also be of great value to chemists who use their own special codes—just as many special librarians modify the Library of Congress classificiation for their own collections.

2. It would not be an entirely fair test of the value of cipher indexes if the chemist did not have alternative methods of searching at his disposal. Since structural diagrams are the "daily bread" of the organic chemist, provision of these in a molecular formula index may be sufficient for a large number of searches. Since the "Miniprint" index would include the Dyson and Wiswesser ciphers it may prove to stimulate their use. It would provide the searcher a simple starting point to have the cipher for one known compound immediately available.

e) ALPHA GENERIC INDEXES

While the LC card approach is extremely desirable, the molecular formula index can never be a complete solution to the problem of the generic search. The molecular formula is popular amongst chemists because it is either known or easily determined. It is as yet uncertain whether chemists can "compute" a Dyson or a Wiswesser cipher as easily. In addition, most of the existing literature of chemistry includes molecular formula data. It is still the practice of journals to include molecular formulas for new compounds. It may be many years, if ever, that either cipher system is adopted by individual journals.

For this reason the primary use of the cipher indexes will be in generic searches. This would be especially the case if molecular formula indexes are provided which contain structural diagrams. Such an index eliminates the primary criticism of the molecular formula index, i.e., that more than one compound is described by the same molecular formula.

There are a number of approaches to the problem of generic indexing. These include machine and manual methods. For the INDEX CHEMICUS machine methods are not feasible as the number of users without equipment would be high. As an alternative approach to the use of cipher indexes for generic indexing we propose the use of the well known Chemical-Biological-Coordination Center (CBCC) chemical code. This code was designed for generic searching and has never been adequately tested on a large scale in a printed index. We wish to emphasize that CBCC coding would be employed as a device for generic indexing only. Aside from the inclusion of the CBCC code on our LC card the reader would never realize that it had been used to compile the generic index. The generic index would be arranged alphabetically and include such commonly used generic groupings such as thiazines, sulfonamides, etc., as is the case of the alphabetic index to the cipher systems. However, inclusion of the CBCC cipher on the 3" x 5" LC slip would be an advantage to many chemists familiar with this published code. It should be stressed that the addition of this generic index does not add

much additional expense to the project for the same reasons discussed above. Once the structural diagram is drawn the time required to "encode" is relatively small. The inclusion of this generic index, however, will give the users an alternative to cipher indexes. Only the inclusion of a molecular formula index and a generic index represents a complete alternative to a cipher index.