CURRENT RESEARCH AT CHEMICAL ABSTRACTS*

By G. MALCOLM DYSON

Chemical Abstracts Service, Ohio State University, Columbus, Ohio

The principal research projects which have been inaugurated at Chemical Abstracts during the past year are: (1) the study of annual index production; (2) the study of cumulative index production; (3) the organization of chemical knowledge with special regard to storage and retrieval; (4) the study of new journals and services not covered by (3); (5) a study of chemical semantics; (6) the production of a register of chemical compounds. In addition, there is a joint project centered at Chemical Abstracts and sponsored by the Synthetic Chemical Manufacturers Association (SOCMA) and the American Chemical Society, for the production of a "Lexicon of Non-systematic and Trade Names for Organic Compounds." I propose to discuss each of these, in turn, paying most attention to those which involve fundamental research considerations.

1. There is no need to stress the necessity for a study of index production. Chemical Abstracts indexes are probably of higher indexing quality and greater indexing depth than any other scientific indexes. The thoroughness with which they are prepared and the careful revision which they are given necessitate expenditure of considerable time. The economical use of time in connection with indexes already has been the subject of a communication (Chemical & Engineering News, 38, 70 (1960)) elsewhere; what we are most concerned with in our annual indexes is the use of modern methods of cold type and line-camera to the actual production of the index volumes. We have proved experimentally by the use of Varityper and Fotolister (both of which machines have been installed at Columbus), that a perfectly acceptable index page can be produced by this method, and that the price of printing such work by the litho-offset process is at least as economical as the traditional method. Indeed, apart from the potential saving in cost, certain other advantages of this method have proved apparent: namely, greater readability of subscript numerals (always somewhat painful in our traditional indexes) and a cleaner-looking and more readable page, stemming from the use of two columns instead of one. This does not, as might be expected, lead to the use of more pages; in three-column format there are so many short lines that transference to two-column format often gives fewer lines; that is, a three-column page of one hundred lines per column often yields only one hundred and eighty to one hundred and ninety lines instead of the theoretical two hundred; in certain cases the reverse is true, but the surprising fact emerges that the choice between two- and three-column format makes little difference to the number of pages required for an index. Some observers have formed the opinion that the two-column format is more easily read, especially where the subject matter consists of organic names.

The organizational difficulties in using this method are those associated with the conversion of a batch process to continuous production. A method is being worked out by which as fast as indexing is done the entries are set in cold type, accumulated in proper indexing order and new entries melded with the old, superfluous headings thrown out so that the index is being continually set and proofed. This involves much collaborative investigation with our colleagues in the indexing and editing departments, and when this problem is completely solved, as it will be fairly soon, we shall have saved many months in the time elapsing between the end of a given year and the appearance of the indexes, without sacrificing anything in quality that is of value to the user.

2. The extension of these principles to the production of cumulative indexes should be richly rewarding; the annual index cards which, in the future, will carry the volume number, can be interfiled; unnecessary headings can be cut, and identical subject headings coalesced by exactly the same techniques as will be used for the annual indexes so that, apart from editorial revision for errors and consistency, no additional setting will be required. This should not only eliminate most of the cost of re-setting but should considerably cut the delay in issue.

3. Information Storage and Retrieval

Extensive preliminary work with various media and methods has established that machines are not the difficulty in this field; suitable devices exist for handling and selecting data from any store, provided that the latter is suitably organized. No new and wonderful machines are needed at present. The problems -- and they are serious ones -- lie in the organization of the entry of data into the store, the "good housekeeping" of the data once inside the system, and the method of recall, especially in relation to the design of enquiries posed to the system.

3.1. Modus of Entry.--It is desirable that all information be entered into the system by a simple device. Philosophically, this "device" will be in two parts: a code or series of codes

*Presented before the Division of Chemical Literature, New York ACS Meeting, September 13, 1960.

which symbolize the data, and a permanent record of the data so coded. The former will be discussed later; for the latter we have selected a normal IBM card, punched with the standard IBM026 Keypunch. Slight alterations are made to the punch so that signs such as #$&*, not used in coding, are replaced by more useful symbols. This does not alter the punching pattern, but merely the symbols associated with that pattern.

3.2. Modus of Decoding.—To translate the encoded information back into a printed form, the card is run through the sensing keypunch which is connected to a Document-writing feature — an electric typewriter. This has upper and lower case letters, normal subscript and superscript numerals and a variety of brackets and punctuation marks, etc., actuated by a double-shift solenoid-operated key-bank. Three of the symbols on the punch are used as monitors to set the shifts or to underline. Thus, for example, if we punch *D into the card * actuates the first shift but prints nothing until the next column is read so that lower-case "d" is printed. Thus if "2" is punched alone it is so printed; if however, it is preceded by "*," a subscript "2" is printed. In this way, with three monitoring signs, the whole of the above symbols can be used.

3.3. The Primary Record.—The first card prepared relates to the chemical nature of the substance. Thus, a card is being prepared for each chemical compound of definite composition. The record for such a card might read:

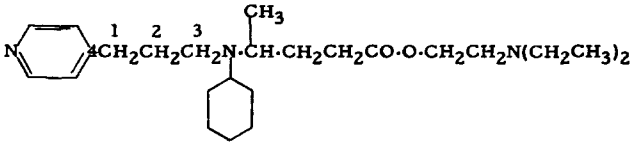
$$322_5 43{:}B6ZN{:}4/C_3{:}3N(A6)/2C_5{:}5X/C_2{:}2N(C_2)_2$$

207022

This record is interpreted as follows:

3.31.—The series of numbers up to the first colon represents the molecular formula $C_{25}H_{43}N_3O_2$.

3.32.—The structure code represents the structure:



3.33.—The final figure—207022— is the CA Register Number. This number is unique for the compound, and although it conveys no information concerning the nature of the structure, it forms the "address" by which the compound is found throughout our internal system. Thus, in concept files (see below), where the properties and attributes of compounds are filed, the register number of the compound is used rather than its chemical code.

In this way, it is proposed to build up a complete Registry of all compounds (starting with the Organic division) which will be set in cold type and printed by line-camera from time to time until a complete record is obtained.

It may be added that the example above is merely illustrative of our present pilot-plant experiments, other codes or number series may be used on the main project. So far, we have mounted all the compounds (both organic and inorganic) mentioned in the Formula Indexes of Chemical Abstracts according to their molecular formula on edge-notched cards; all these cards have been punched with the nature and number of elements present, so that it is possible to subdivide this pack according to any element or combination of elements. As a pilot run we have selected the organic fluorine compounds and have built a subsidiary pack of some 14,000 of these, which will serve for a pilot run through the system. From this pack, IBM cards of the nature described above are being prepared. These, when complete, will constitute one stage in the development of the pilot run.

3.4. The Direct Access Files.--For internal use, and even when dealing with external inquiries where only a single compound is in question, it is important to have a file that can be consulted manually. The file described above is not suitable for this purpose, so that its information is replicated in a second file of IBM cards containing apertures. The aperture of any given card is filled with a single frame of 16 mm. film showing (a) the Registry number, (b) the structural formula drawn out and numbered, (c) the name used officially in CA Indexes, and (d) a lead number to the bibliographic file which is also punched in the card. A user identifies the required compound by Registry number (if already known to him) or via the structure code/to registry number handbook, and then proceeds to the correct number in the Direct Access files, where all compounds are filed in order of their register numbers. The material in the aperture can be viewed on a simple reader and copies made by M.M.M. reader-printer.

Thus, an indexer or other enquirer can make a rapid approach to data on a few selected enquiries anywhere in the files. Since the files are kept in order of the Register Numbers, additions from current work going through the shop can be made without disorganizing the files, merely by adding at the end. Experiments are being made, with quite successful results so far, to use a semiautomatic device for photographing the structural formulas. These are photographed at high speed and at great reduction from adjustable cards with which the formulas are built up. The photocopies are mounted automatically in the apertures of the IBM cards.

3.5. Concept Code Files.--The fundamental unit of the concept code card pack is a single

abstract, the reference number of which is placed on the card together with the register numbers of the compounds mentioned in the abstract and one concept related to the compounds mentioned. Thus, supposing that a single abstract (No. 3435367) mentions the compounds 20027, 20028, 20029, 43789, and 68574 all in connection with their use in trypanosomiasis in man, all substances having a moderate positive curative action using the intravenous route. The IBM card for these data will have the entries:

20027	2.745.66	3435367	JACS59:2345
20028			
20029			
43789			
68574			

The first group of numbers represents the compounds, the second is an arbitrary codification of the pharmacological data and the third entry is the abstract. The final entry is the reference to the original communication. The characteristic number "7" of the 2.745.66 indicates that the physiological/pharmacological code is being used; "45" indicates human trypanosomiasis and the ".66" the route by which the drug was administered in the experiments. The initial "2" is a record, on an arbitrary scale, of degree of activity. Compounds in which no activity had been detected would have been entered on another card and the entry for the physiological data would have read 0.745.66 and "0" indicated null activity. Similar card packs are being built for each phase of the properties of chemical structure -- physical properties, patents, industrial applications, etc. One problem that has had to be solved is a means of citing any of the 9,000 journals in which chemical data may be found, without using too much of the card. This has been done by constructing a four-letter code for all the journals. A considerable degree of mnemonic quality has been retained in these abbreviations -- thus, JACS for the Journal of the American Chemical Society, JCSL for the Journal of the Chemical Society of London, JORG for the Journal of Organic Chemistry recall to mind the journal in question, often without recourse to the handbook. The full set of these abbreviations will, when checked, be published.

3.6. Methods of Retrieval.--The retrieval system is based on the IBM 1401 computer, which we hope to have delivered shortly. The card entries from the main structure file, duly sorted into sections representing the elements and important groups will be converted to magnetic tapes, as also will the cards representing the concepts and references.

Let us consider a problem such as the preparation of a bibliography of references dealing with those halogenated pyridine derivatives which have shown positive tuberculostatic activity in experimental animals. The halogen/nitrogen section of the main file is searched for such entries as contain entries for pyridine and halogen; the answer is a series of register numbers. This search is made, of course, on the magnetic tape and not on the generating cards. The numbers obtained in this search are then set to search for the same numbers in the physiological/pharmacological file, but the search is planned to record only those instances where the register number is identical and where the physiological code figure is of the form a.bcd.mno where a is not zero and cs are 45 (the code for tuberculosis). The answer to this search is a tape showing all compounds of pyridine with halogens having positive pharmacological action in tuberculosis. The output can be converted to a printed list showing all the data of the concept file cards concerned. From this the abstracts or original references can be taken off. This in itself may be the answer required, or the enquiree may prefer aperture cards of the abstracts themselves, or even copies of the abstracts in chronological order printed xerographically. All these forms are now the subject of experimental evaluation.

It will have been noted that the answer to the question set out in the previous paragraph is on a somewhat broader basis than the question itself; the question included the phrase "in experimental animals" and the answer covered all observations on the tuberculostatic activity of the compounds in question. This provision of rather more material than the question appears to demand -- sometimes called the provision of "display" -- is necessary to make certain that nothing of interest to the enquirer has been overlooked. The subject is one that cannot be discussed fully here; it might well be the subject of a separate talk.

4. New Journals and Services

Plans already have been made to introduce a new journal, Chemical Titles, as from January of 1961. This journal lists the latest titles received at Chemical Abstracts according to the Keyword-in-Context system and provides a quick approach to current chemical literature. This journal has been sufficiently introduced to members of the Society not to need further discussion here. From the experimental point of view, one of the most interesting phases of the development of Chemical Titles has been the programming of the computer and building up of a memory stop-list to prevent useless index entries being included in the publication. It is believed that this is the first computer-produced journal of any magnitude to be issued.

Naturally, our interest in the research division has moved to other types of new journals and services. There appears to be a definite

need for a classified list of all organic compounds described and examined in a given period on as current a basis as possible. We have been experimenting for some considerable time with the product of such a list. The material for a given issue would correspond to the titles of the corresponding issue of Chemical Titles; this would mean that the condensed references for the new Restricted express lists (REL) would be used as the "trace-back" for relating compound and literature. Entries would be made under the molecular formulas, subdivision under each molecular formula heading would be by structure notation and, in addition to these, the register numbers and references to original titles will be used. A distinguishing mark will be placed on compounds which are new, in the sense that they have not appeared previously in our records. We regard the use of this mark as a distinct advance; heretofore there has been nothing in any index to indicate the newness of a compound; should this mark appear on a compound it will immediately convey to the observer that there is no necessity to search the earlier literature for additional information.

4.1. Lexicography.--The pilot run -- comprising all data on the organic fluorine compounds -- will bring together a mass of material which has not previously been selected from the chemical record. In a case such as this, it is very probable that the collected data will have a wide interest; it is certain that the obtaining of such data from the record would be a long and tedious task, and that having been done once should not be done again. With this in mind it is proposed to use the material of this section to produce a "Lexicon of Organic Fluorine Compounds," which will show for each compound: (1) register number; (2) molecular formula, (3) structure notation, (4) bibliography, (5) physical properties, and such other data as may appear to be generally useful. Studies in the publication of such volumes will indicate where and how the principle may be applied effectively to other groups of compounds. So far we have made no decisions as to whether to choose the next pilot group on the basis of a single element -- the bromine compounds, the organo-tin compounds, the organo-phosphorus compounds -- or on the basis of a significant group -- as the nitro-compounds, the hydrocarbons, or some similar aggregate.

5. Chemical Semantics

One is plunged into semantic problems immediately research in the field of chemical documentation is started. One of our earliest studies, on Chemical Titles, posed the question as to which words in a title are significant and which are not worth inclusion in an index. This is not the simple problem that some imagine; whether to include "analysis," "detection," and "determination" as indexing points is not to be decided by some simple rule, but is to be considered in relation to the users' needs. In the same way we have decided not to use "synthesis" but to retain "syntheses"; to reject "acid" but to retain "acids" on the grounds, not of any rule, but of the consideration of each case in relation to the users' problems. This may perhaps result in better titles; we have already had one instance of an irate chemist whose paper on "An Introduction to the Study of Organic Bases" was completely unindexed because all of the words in the title were judged by the computer to be non-significant.

We have started a study of the language used by chemists -- that is, the language they use in formal printed records -- original communications and abstracts. A method has been devised, using only the simple standard IBM equipment, by which words and phrases used in text can be segregated and listed in context; this is leading us to a study of synonyms and what is wrongly called the "thesauric content" of language used in this field. So far we have restricted our work to Section 11 of Chemical Abstracts -- the Section dealing with physiological matters. This choice was made partly to avoid the masses of structural names of the organic division and partly because our immediate needs for concept coding lay in this area. We regard this as long-term research; we are not solving an immediate problem, but surveying the general sub-discipline. It will be some time, I expect, before we can demonstrate the effect of this work in our day-to-day routines. Nevertheless, since the whole of our work is fundamentally based on language, this study cannot help but be rewarding.

6. The CA Register of Compounds

Reference already has been made to the preparation of this register. For our own internal purposes it is necessary to have such a register, and we have decided to give every structure -- organic and inorganic -- a seven-digit number, as an "address." These numbers are strictly given at random; they are "idiot" numbers in the sense that they have no significance but their uniqueness. This is necessary in order easily to manage files and to allow for continual updating without destroying previous order. It is hoped that others may find such numbers useful in communicating structures by written or telegraphic means. With this in view we shall publish, in due time, a list of such numbers, together with the structures to which they refer. This dictionary will, of course, be "two-way" in the sense that one can look up both numbers and structures. This presents various problems which we are endeavoring

to solve; it is, however, too early as yet to report on the list.

7. Lexicon of Non-systematic and Trade Names for Organic Compounds

The American Chemical Society in collaboration with the Synthetic Organic Chemical Manufacturers Association has agreed to produce a lexicon with the above title. There has been need for a long time now of some compilation of non-systematic names used in organic chemistry with their official equivalents. The work of this compilation is being undertaken by the Research Division of Chemical Abstracts. There will be two separate editions — one a full scientific lexicon containing all the names found, and a second containing those names used in industry. It seems that the former will contain some 35,000 entries and the latter about 7,000 entries. Each edition will be in two parts; the first will contain names, arranged alphabetically and Register Numbers. Every name whether synonymous or not will appear in this list. The second part of the work will contain the Register Number (in order of which the entries in this section will be arranged); the molecular formula, the structure notation, the drawn and numbered structure of the compounds, together with the official CA name, the other scientific names which have been given to the compounds and any other "trivial" or trade names. Thus, a user ascertains from the first list the Register Number of the compound he is interested in; looks up that number in the second part and finds displayed all the nomenclatural data on the compound that is available. So far some 18,000 entries have been collected.

This record shows what has been done in the Division during the last twelve months; I hope from time to time to add further progress reports either here or in Chemical & Engineering News.