The Role of a Central Chemical-Biological Records Office in a Pharmaceutical Company*

By HARRIET A. GEER and C. CECILY HOWARD
Parke Davis and Co., Detroit, Michigan
Received June 5, 1961

It is common practice for pharmaceutical companies to maintain a central office where all test data are collected. The services performed vary from company to company, but in general the following are rendered: (1) a check for the presence of a single compound or type of compound in a chemical file; (2) location of the biological test results for a single compound or type of compound; (3) location of a sample of a desired compound; (4) distribution of information from the central file to the research staff.

To perform these duties, a central office must maintain files varying in complexity. A molecular formula file is essential to determine accurately whether a compound is in the file. A name or specific notation file can be used but the complexity of the rules leads to some uncertainty in always deriving the same name or notation for a compound. According to conventional methods, carbon and hydrogen precede other elements in the molecular formula. Recently several workers (*1, 2, 3*) have used a so-called inverted formula in which carbon and hydrogen follow the other elements present. Skolnik omits hydrogen completely. The inverted method is superior to arrangements based on carbon in bringing like compounds together.

The main reason for locating a compound in a pharmaceutical company's file is to gain access to its biological activity. For this purpose, a master file in which the chemical compound is the unit generally is maintained. Here the chemical, physical and biological data for each compound are assembled. The melting point, boiling point, color, state, and other identifying properties should be included as well as those which are useful to the tester, such as solubility and stability. Precautionary measures in handling should be recorded, too. The source of the compound should also be noted, including the notebook number when the compound is synthesized in the company laboratory.

The simplest way of arranging the master file is by number. Filing by number requires no preliminary training, is fast, and is less subject to error than filing by name or formula. A chronological number, assigned to each compound upon accession, is most frequently used. The information on each compound may be arranged in different ways in this file. Summarizing the data on a single card or collecting it in a folder are two methods. A summary card allows rapid examination of all results for a compound and is very useful. The main disadvantage is that this card must be pulled and refiled each time new results are added. Also, more detailed biological test data must be filed separately when a summary card is used.

In addition to a summary card file and a molecular formula file, Parke-Davis has a *Chemical Abstracts* name file. We also have a source file listing all compounds received from outside the company. Cards for these files are made by a duplicating process which eliminates typing of each card individually.

Summaries of results of individual compounds can be collected by machine methods. The Dow Chemical Company codes and punches their biological data on IBM cards. From a combination of these cards and a set punched with *Chemical Abstracts* names, lists of results can be printed with a tabulator. The Cancer Chemotherapy National Service Center stores their test results in the IBM RAMAC. They can prepare a list of test results by typing out the Center number of a compound. Because of the limited capacity and high cost of the RAMAC, the Cancer group uses it only for compounds currently under test.

Often, it is desirable to locate all compounds with similar structures. A *Chemical Abstracts* name file is very useful when making a quick check for certain types of compounds. Fletcher found that the inverted molecular formula file was helpful in this respect, too. Since these files are inadequate for making thorough searches of large numbers of compounds, chemical structure coding is necessary.

There are many structure codes in use at the present time. They are listed here, with some of their users; the list is not complete and is limited to the United States.

1. Individual Chemical Codes Developed by each User: Abbott Laboratories, Cancer Chemotherapy National Service Center, Ciba, du Pont (Chemical Dept.), ASTM Infra-Red Project, Lederle, Merrell, Monsanto, Parke-Davis (IBM code), U. S. Patent Office (Steroids), U. S. Patent Office-National Bureau of Standards.
2. Norton-Opler Code: Dow and Midwest Research Institute.
3. CBCC or Modified CBCC Code: Ethyl Corporation, Merck, Sharp and Dohme, Sloan-Kettering, Smith, Kline and French.
4. Wiselogle Code or Modified Wiselogle Code: du Pont (Grasselli Div.), Metal and Thermit, Parke-Davis (McBee File), Squibb, Warner-Lambert.
5. Wiswesser Notation: Searle, Army Chemical Center
6. Dyson Notation: Under study at Chemical Abstracts

The above listing includes codes which completely describe the compound as well as those which show only certain characteristic structural features. The Wiswesser, Norton-Opler, U. S. Patent Office-National Bureau of

*This paper was presented at a symposium on "Information Retrieval and Analysis" which was held jointly by the ACS Divisions of Medicinal Chemistry and Chemical Literature (New York, N. Y., September, 1960).

Standards, Dyson, and Monsanto codes describe the structure completely. The Searle Company is the only pharmaceutical company which uses a code from which the exact structure can be reconstructed. This cannot be used easily for mechanical searches on accounting-type equipment without additional coding.

Originally, Parke-Davis used the Wiselogle classification system with edge-notched cards. When this file became too large, a new code was developed before converting to IBM. This code contains many of the principles of the Wiselogle system as well as characteristics of the CBCC code. Frequently occurring functional groups are shown specifically while less common ones are represented by a general designation. The point of attachment of functional groups is not specified. In fused rings, the individual heterocyclic rings are coded separately as well as the total ring skeleton. In addition, certain spatial relationships between elements, functional groups, rings, and unsaturation are indicated.

The choice of a code is so dependent upon the choice of equipment that they should be considered together. In general, those codes which describe the structure nonspecifically are best for searching with accounting-type equipment, whereas those which describe a structure completely require more complex machines. Smith (*4*) uses additional columns on the IBM card to indicate by direct punching the presence of certain symbols or combination of symbols in the Wiswesser notation. In this way a search can be made on a simple sorter for the structural features so coded by a single pass of the cards. The Dow Chemical Company uses the Norton-Opler code with a simple IBM sorter by indicating the structural groups present but not their points of attachment (*5*). Waldo's (*6*) method of depicting structures was developed for computer use and can not be adapted easily to accounting-type equipment. This is also true of the specific code (*7*) developed jointly by the U. S. Patent Office and the National Bureau of Standards for the SEAC. This code is based on Mooers' topological approach to structural representation.

Equipment for structure searching ranges from edge-notched cards to all-purpose computers. Four important ways in which types of equipment vary are: amount of coding space available, speed of searching, cost, and delivery of information. Coding space, speed, and cost all increase from a minimum for edge-notched devices to a maximum for computers.

For delivery of information from a structure search, edge-notched cards are very satisfactory. The cards obtained show a structure, file number and any other information placed on the original card. The delivery of information from a search with machine-sorted punched cards is less satisfactory since structures cannot be printed from the cards obtained. One can pull copies of the structures from another file but this is a sizable task if a few hundred must be pulled. Another solution is to put the structure directly on the punched card. Dow uses Multilith and Parke-Davis uses Ozalid coating to reproduce a structure on the card. Merrell draws it on by hand. Unfortunately, this structure cannot be transferred automatically to a new card from the punched information. This is a definite disadvantage since machine-sorted cards do not last indefinitely. Microfilm inserts can be used for showing structures, but their cost is high and the cards containing them also wear out. The use of a line-formula notation such as that of Wiswesser or Dyson prints out an alphabetic-numeric representation of the structure. Although users claim that one soon develops the ability to visualize the structure from the notation, it is not as satisfactory as a conventional structure for most people. A perfect solution to the problem would be a clear, indestructible, plastic card from which the structure could be reproduced and which would not be as difficult to sort as the plastic cards used now.

Delivery of information from computers has been improved by the methods of Waldo and Opler for representing chemical structures. Waldo depicts the structure by a series of numbers and letters so that the computer prints out a recognizable structure. Rings are converted to rectangular figures which can be printed line by line on the high speed printer. The computer can be programmed to search these symbols for the desired structural characteristics. Opler (*8*) has devised a dot-display method using a cathode ray tube for computer output of chemical structures. Each symbol in a structure is depicted by dots in an 8 × 8 dot pattern. The display area for a structure is made up of 32 × 32 of these squares. A structure can be stored in the computer directly from a drawing of symbols on the pattern of squares by indicating the horizontal and vertical location of each functional group or element.

The number of compounds to be searched is an important factor in determining the type of equipment used. Other factors include the number of searches made and the availability of a piece of equipment. Thus, although the cost of computer time is usually prohibitive, a company with excess computer time might find the computer an ideal structure searching tool. However, it is best to use simpler devices unless one is willing eventually to assume computer costs.

In addition to locating chemical and biological information in the central files, there are other ways in which a records office can serve the research staff. For example, it should be possible to locate the samples after a structure search has been made. There is little gained if a structure search yields compounds which can not be located. At Parke-Davis, the samples are stored in the Central Records Office and the amount of each compound is recorded in a 3 inch × 5 inch card file. The process of sending samples with sheets for recording test data to the testing programs can also be handled by a records office.

Another service of a central records office can be the summarizing at regular intervals of results for compounds supplied by outside contacts. Since all results are sent to the office, it is the logical place for this information to be gathered together.

A central records office generally distributes results obtained in the testing programs. At Parke-Davis, a monthly bulletin is distributed which records the structures of compounds showing activity in a testing program. If the biological activity is coded on IBM punched cards as is done at Dow, it is possible to make summaries from these. Waldo devised a novel method for compiling reports of test results directly from a computer (*6*). The information is coded in the test laboratory and added to the computer tape. The computer arranges the results according to test programs and activity. By using the structures of the compounds already stored in the com-

puter, the high speed printer rapidly prints the report.

SUMMARY

The rapid growth of pharmaceutical research in the last twenty years has created a need for the collection and organization of the results of this research in a central location. In response to this need, the central records office has become an accepted part of the research division. The increase in the amount of information handled by this office has resulted in the development of machine methods for retrieving data. In addition to the collection and organization of research data, the central records office plays a vital role in the exchange of information between the sections of the research division.

REFERENCES

(1) Dyson, G.M., *Chem. and Ind., (London)* 676 (1952).
(2) Fletcher, J.H., Dubbs, D.S., *Chem. Eng. News*, **34**, 5888 (1956).
(3) Skolnik, H., Hopkins, J.K., *J. Chem. Educ.*, **35**, 150 (1958).
(4) Smith, E.G., *Science*, **131**, 142 (1960).
(5) Nutting, H.S., and Klesney, S.P., paper presented before the American Chemical Society Division of Chemical Literature, Pittsburgh, Pennsylvania, Meeting Jan. 20, 1958.
(6) Waldo, W.H., Gordon, R.S., and Porter, J.D., *Am. Document.*, **9**, 28 (1958).
(7) Ray, L.C., and Kirsch, R.A., *Science*, **126**, 814 (1957).
(8) Opler, A., and Baird, N., *Am. Document.*, **10**, 59 (1959).

BOOK REVIEW

Chemical Dictionary—Russian-English. Eugene A. Carpovich. Technical Dictionaries Co., Box 144, New York 31, N. Y., 1961. 352 pp. $14.00.

This new Russian-English dictionary covers a broad range of chemistry from Materials through Theoretical Chemistry and into Engineering. Very recent source material is required to accommodate so many fields. This is especially desirable for a Russian dictionary since some branches of chemistry have only recently become active there. Apparently some attention was given this point in the preparation of this book.

Many chemical formulas have been included. This is somewhat unusual for a Russian dictionary but it will save the translator some time, especially if he is translating outside his own particular field.

Partly because many cognates are omitted, the word list is more highly specialized than most other Russian Chemical dictionaries available now.

In the reviewer's opinion this dictionary does not meet the standards of general usefulness set by Callaham, but then Carpovich is not quite as expensive.

This new dictionary should be satisfactory for working chemists and translators alike. The printing is larger than usual, and the cover is rugged enough to stand a certain amount of abuse.

Research Center
Hercules Powder Company
Wilmington, Del.

Robert H. Saunders

BOOK REVIEW

Critical Solution Temperatures. Alfred W. Francis. Advances in Chemistry Series, No. 31, American Chemical Society, Washington, D. C., 1961. 246 + vi pp. $5.00.

The critical solution temperature is an important property. It answers the question: "Is A soluble in or miscible with B?" Critical solution temperatures answer this question for any temperature and for any composition of two materials, for above the critical solution temperature curve (composition *vs.* temperature), the components are miscible in all proportions unless one crystallizes out. These data are used for choosing solvents for extractions, for screening solvents, and for studying molecular structures.

This book is a compilation of these data which are scattered through the literature for over 6000 binary mixtures. Methods are described for determining this property. How to choose solvents for extraction and how to estimate this property for untested systems are described. The data are listed alphabetically under each solvent. Aniline and furfural points are listed in a special table as well as alphabetically under aniline and furfural. A separate table is also included for lower critical solution temperatures. The compilation is based on 495 references.

Dr. Alfred W. Francis has made a real contribution in compiling these data. No chemical library can afford to be without this book.