At this point we had been learning how to retrieve information from this bank of literature for 13 months. We have used it as the principal means of satisfying FDA requirements for bibliographies on our drugs. To date we have submitted about 300 questions to the computer. For the FDA application the results seem to be entirely satisfactory. On research questions, aside from such conspicuous limitations as the short time span of publications involved and restricted journal base, we have found others. For instance, the code does not permit one to distinguish between α- and β-adrenergic blockers.

In evaluating results with our non-FDA retrieval efforts, there appear to be about 15% failures. A further 15% gave only fair results, in that part of the material retrieved was highly pertinent, but some other valid material known to be present was not recovered.

Among the failures were such topics as

(a) factors affecting blood flow in adipose tissue;
(b) studies concerned with the electrical charges on or electrophoresis of the formed elements of the blood;
(c) tranquilizers used in preoperative therapy of asthmatic patients.

We suspect it was lack of data rather than faulty search strategy which accounted for no response in this last case. However, a "no response" result in many of those searches was quite an appropriate reply, and often comforting.

A primary advantage of the Ringdoc-ARCS system is its timeliness in making published information available to users, and especially its machine retrievability in different contexts. The timetable which we and Derwent strive to maintain is:

Abstracts issued four weeks after original journal publication. Ringdoc cards shipped at six weeks via air freight, and introduced into ARCS tape by seven weeks, so that the material selected becomes retrievable at the eighth week following publication.

In actual practice, there have been delays of one sort or another, which have prevented achievement of that schedule entirely (shipping strikes, loss of material, faulty processing). But it is run by human beings for other no less fallible human beings with not all conditions subject to exact control.

Of course, our own present lack of full grasp of the system's capacities, and the machine's intransigence toward inaccurate, incomplete, or otherwise inadequate instructions, has pulled us up short at embarrassingly frequent intervals, too. We do feel that we are learning, that we can live with this system, refine it, and continue to perfect our use of it.

LITERATURE CITED

(1) "Manual of the International Statistical Classification of Diseases, Injuries or Causes of Death," 1955 revision; World Health Organization, Geneva, 1959.
(2) Epizootiology Section, National Cancer Institute, Bethesda, Md., 1966; Available from Superintendent of Documents, Government Printing Office, Washington, D. C., $3.50.

A Chemically Oriented Information Storage and Retrieval System. I. Storage and Verification of Structural Information

CARLOS M. BOWMAN, FRANC A. LANDEE, and MARY H. RESLOCK
Computation Research Laboratory, The Dow Chemical Company, Midland, Michigan

Received December 29, 1966

A computer-based system has been designed to handle chemically oriented files. The Wiswesser line-formula chemical notation is a practical method for representing structural formulas for input. The file organization and methods of verifying the accuracy of the notation as well as cost are described.

The problem of organizing and indexing chemically oriented data and information is a difficult one. The literature itself is quite voluminous on the subject. In 1964 an excellent survey of the various methods was published (*1*, *2*). Most of the systems described reported their result in the retrieval of the names, structures, or identification numbers of compounds which satisfy certain structural relationships. A computer-based system has been devised which will allow searches to be made not only on structural considerations but also on properties and other pertinent information about chemical compounds (*3*). The system will be described in a series of papers, of which this is the first.

This paper will discuss the establishment of the section of the file which contains the structural considerations—*i.e.*, name, structural configuration, molecular formula, etc. It will also discuss a computer program which has been written to check the accuracy of structural and molecular formula information. Finally, some typical costs of input will be presented.

FILE ORGANIZATION

An information system usually has two types of files: a document file which contains the information and an index file which permits the location of items in the document file. In this chemically oriented system the information is collected around each chemical compound, so that there will be one section for each compound in the file. In each one of these sections will be found all the information and data which have been stored for that compound. Since the document file is recorded on magnetic tape, economic considerations require that the information be stored in a compact form.

A typical record for a given compound will consist of:

1. Accession number. A sequential number is assigned automatically by the computer to each new compound entered into the system.
2. Computer record data. This section contains a series of pointers and counters which allows the computer program to determine the presence and size of the various parts of the record.
3. Chemical structure. A total description of the structural configuration of the compound is stored in the form of the Wiswesser line-formula chemical notation.
4. Molecular formula.
5. Name. The systematic name as used by *Chemical Abstracts* is used to name the compound.
6. Physical properties. In some cases the actual data are stored. In other cases references are made to separate files containing more detailed information, such as infrared spectra, vapor pressure, and temperature relationships.
7. Biological properties. This section contains information on the extent of testing, gross results, and references to detailed data.
8. Bibliographical references. References to other documents such as company reports, bulletins, and literature references are also included in the file.

The document file has been designed in such a way that additional information may be added at any time. The computer pointer system has been devised to permit the indefinite expansion of any record without having to reprogram or redesign the file. The index file is a standard coordinate index and its composition and preparation will be described in a future paper.

CHEMICAL STRUCTURE IDENTIFICATION

The structural identification of the compound in the document file should completely describe the relationships of the atoms in the molecule. In this way the structural information need only be recorded once and searches and other manipulations can be carried out by appropriate computer programs. The Wiswesser line formula notation (*4*) was chosen to represent the structure. A notation system should have several features to be useful. It should be complete, unique, unambiguous, concise, readable, economical, and capable of being manipulated by existing computer equipment. The Wiswesser notation was the best over-all choice that could be made from those notations and topological representations available.

The notation uses only 40 characters in its vocabulary—the 26 letters of the alphabet, the 10 digits, the space, ampersand, hyphen, and slash. Consequently, most typewriters, computer input-output devices, and computers can handle the notation easily. On the other hand, the restricted symbol list requires the notation to be a very redundant language, each symbol having several meanings dependent on how it is used. The notation has undergone extensive revision since its original publication (*5*) and is now much more complete. Several redundancies have been introduced to facilitate computer processing.

This extensive use of the notation revealed some inadequacies and ambiguities in the rules. It also pointed out areas which were not considered by the rules, such as stereoisomerism, metallocenes, and polymers. As the Wiswesser notation became more widely used, the deficiencies became apparent to others. This resulted in the organization of a group of users who meet frequently to discuss problems, make suggestions, and in general improve the notation system. The revised manual is a product of this interchange of ideas and experiences, and it is expected that, through this association of users, additions and changes to the notation rules will continue to be made, but only after considerable experimentation and testing. Thus, this notation system embodies rules which reflect the current needs of actual operating systems. The notation need not become obsolete as new structural relationships are brought to light.

NOTATION CHECKING

Several approaches have been used to check the accuracy of the notation. Some have a second encoder look at the structure and the already written notation; others have a second person encode the structure on a separate sheet and then have a clerk make a visual comparison. Comparison has also been done by having punched cards made from one input sheet and verification of the cards from a second sheet prepared by another encoder. The coding time is by far the most expensive step in the input process, so it was decided to write a computer program which would automatically check the notations and detect errors.

In addition to the notation, the molecular formula also provides a point from which classifications, searches, and other such manipulations can be made. The formula is used to check the validity of the notation, as will be described later.

CODING THE STRUCTURAL FORMULA

The notation is produced manually by an individual familiar with structural formulas and diagrams, usually a college graduate with a general organic chemistry background. The encoder is presented with a structural diagram and a molecular formula, and he writes the notation and the molecular formula on a form suitable for keypunching. The encoding time varies, of course, with the complexity of the molecule, but generally it can be done at the writing speed of the encoder. Initially the molecular formula was checked against the structure, but experience with the checking program made this step unnecessary.

The revised manual for the Wiswesser line formula notation (*4*) is written in such a way that a formal training period is not necessary. After an individual has read the manual and worked the examples and exercises he should be ready to code compounds. Initially his work is supervised to assure that the rules are being interpreted correctly. Errors detected by the checking program are referred back to the original encoder, thus providing him with an indication of the quality of his performance. Complex structures require more time, and at times it is best to have several persons encode the same structure and consult with each other on the resulting notations.

The checking process is carried out in such a way that the original card file of notations, molecular formulas, and reference numbers is read through the computer and emerges from the process unchanged in order or content. The computer program, however, looks at each card and checks the data in many ways for correctness and consistency. Whenever something is found that is known to be wrong, or which cannot be explained, the card in question is copied on a new punched card and an error message is printed on the line printer. The program processes between 200 and 300 cards a minute.

The original deck is saved and the error cards and messages are returned, preferably to the person who did the original encoding, for correction. The corrected cards are recycled through the computer until no errors are detected. These corrected cards are then collated with the original deck. The collator removes the erroneous cards from the original deck and replaces them with the corrected cards to give a total file of corrected cards.

The checking program is written in extended ALGOL-60 for the Burroughs B-5500 computer and is divided into three sections. The first section deals with the format of the cards and the molecular formula, the second section calculates the molecular formula from the notation and compares it with the input notation, and the third section is a syntactical check of the notation.

The first part of the first section of the program consists of a series of bookkeeping checks to determine if the card has been punched correctly and that, if there is more than one card for a given entry, the cards are in the proper sequence. The remainder of the section checks the molecular formula. The first check made is the order of the elements in the formula. The prescribed order is C, H, N, O, P, S, F, Cl, Br, I, followed by any other elements in alphabetical order. Any deviation from this order will result in an error notification.

As the elements are being checked, a table is prepared in the computer memory which lists the number of atoms for each element as indicated by the formula. This table will be used later in the second section of the program. Many molecular formulas are written as the formula of the original compound coupled with water of hydration, or as hydrochlorides, or other such combinations. The encoder is not required to merge the formulas of the two compounds into one formula. Provisions have been made to handle the most common addends such as HCl, HBr, H_2SO_4, H_2O, HBF_4. Thus, the molecular formula for ethylamine hydrochloride may be written as $C_2H_7N \cdot HCl$. The program does combine all similar elements in the element table regardless of how the formula has been written.

The second section of the program analyzes the notation and constructs a similar table which contains the number of atoms of each element as indicated by the notation. This table is compared to the original elemental table created for the molecular formula. The contents of the tables should match if the notation and formula are correct. If a match is found at this point in the program, the third section of the program is then used to make a syntactical check of the notation. A formula mismatch will cause an error message to be printed, a card to be punched, and the analysis of that notation to be terminated. The program then reads the next card and repeats the process.

As would be expected, the analysis of the notation to obtain a molecular formula will detect syntactical errors. These are noted and appropriate messages are printed. As this section of the program was being developed, it became apparent that the so-called methyl contractions in the notation were a source of many difficulties. To avoid this problem a small routine was inserted at the beginning which scans the notation for methyl contractions and expands the notation at this point.

It should be noted that whereas most of the structure handling systems reported to date tend to ignore hydrogen and hydrogen counts, many notation errors are detected in this program only by a discrepancy in the hydrogen count. Thus, the redundancy introduced by the hydrogen count has become a very useful additional checking tool.

In the course of writing the analysis of ring systems, an algorithm was developed to calculate the number of atoms in a given ring system. The equation:

$$N = S - 2(r - 1) - b - m + x - l$$

where

- N = the number of ring atoms in a ring system.
- S = the sum of the ring numerals in the notation.
- r = the number of ring numerals.
- b = the number of shared bridge locants.
- m = the number of multicyclic point locants cited.
- x = the number of x symbols and other cited ring segments bonded to four other ring atoms.
- l = the number of –& spiro ring linking signs.

was incorporated into the program and is used to determine the number of atoms in any ring system. It was in the development of the ring analysis portion of the program that several inconsistencies in the rules of the notation were found. They were brought to the attention of the users and corrective action was instituted. Typical examples of such rule changes include: the citing of a multicyclic point locant twice if the atom occurs in four cited rings, and the differentiation between shared and unshared bridges. The latter change became very useful in the decoding of complicated notations.

The third section of the checking program was written to detect errors which would cause the notation not to be unique or canonical. The notation rules contain certain seniority rules which specify the order in which groups or rings should be cited. Deviations from these rules will not obscure the structural configuration of the molecule being coded, but will give rise to more than one notation for the same structural diagram. Since one of the uses of the notation is to ascertain, by computer, the previous

existence of the compound in the file, nonunique notations must be avoided. At present this section of the program checks the citing order of the substituents on a benzene ring and the starting point of the notation. Further work on this type of manipulation and analysis of the notation is continuing and will be reported later.

ERRORS DETECTED

The types of errors which are detected by this checking program can be summarized as follows:

1. Sequence errors. These are simple errors usually caused by improper sorting of the input cards and are confined to notations which require more than one punched card to record the information.
2. Explicit formula errors. These errors are deviations from the established sequence for elements in the molecular formula. The use of symbols not recognized as element symbols is also detected and noted.
3. Notation and implicit formula errors. There are many notation errors which can occur and most of them are detected when the elemental tables are compared as discussed earlier. Also, errors in the number of atoms in a formula are detected at this time.
4. Canonical errors. These are errors in the application of the seniority rules of the notation.
5. Program deficiencies. These are notations which cannot be handled by the program and will be further discussed later.

The frequency of occurrence of these various types of errors varies, depending on the type of compounds being coded. Over 100,000 compounds have been coded and checked by the checking program. The number of sequence errors detected is negligible, since the number of notations requiring more than one card has been a very small proportion of the total. Molecular formula format errors are also very few. Inasmuch as the canonical check has not been fully implemented, no data will be reported on this occurrence frequency.

About 7% of the compounds checked are rejected. Approximately one half of these rejections are due to errors either in the notation or the molecular formula. These are errors which are detected at the time the input formula is compared to the formula calculated from the notation. About half of these are due to incorrect notations and half to mistakes in the molecular formula. Most such errors can be corrected easily and recycled as described earlier.

The remaining compounds rejected represent those notations which the program is not able to process. These are usually fairly complex structures which require rather complicated analysis. In practice these errors are segregated from run to run and cataloged according to the particular feature that is causing the rejection. This tabulation of trouble spots is reviewed periodically. When a significant number of compounds has been rejected because of a given program deficiency, the program is revised and enlarged to take care of that difficulty. In this way the number of rejections due to program deficiencies has been reduced from over 10% to about 2%. It is felt that this practical approach will focus the effort on the more significant problem areas.

Recently, it was decided to eliminate the verification of punched cards by keypunch operators, as keypunch errors are detected by the computer checking program. This increased the number of rejections about 3%, but it cut the operator time almost 50%. Only those data which cannot be checked by the computer, such as compound number, are verified.

FILE UPDATING

Once the notations and molecular formulas have been checked and corrected they are ready to be read into the master compound file. All corrections are recycled through the checking program to make sure that other errors have not been introduced at the time the corrections are made. Those notations which are rejected because of program deficiencies are checked manually before placing them in the permanent file. The correct notations, formulas, and reference numbers are read onto magnetic tape and sorted in alphabetic sequence according to the notation. This alphabetically ordered file is then compared against a master file which is also ordered by notation and contains only the notation formula. Compounds which are not found in the master file are given a unique registration number by the computer and are added to the master file. Duplicate compounds are not entered in the file, but are printed out on the line printer along with the information already recorded on magnetic tape for that compound.

At the same time as the registry file is being updated, the new compounds are added to a master data file which contains not only structural information but other data and references to information pertaining to that compound. If at updating time it is known that a compound already exists in the file it is flagged and the program picks up the registry number and then later adds the information to the master data file under the proper compound.

COSTS

An analysis of the costs of the various operations reported in this paper has shown this method of inputting chemical structure information into a computer system to be economical and practical. Table I summarizes the time and cost of coding and checking the notations and molecular formulas. It should be pointed out that the encoder starts with a legible structural diagram and a molecular formula printed on a card.

Table I. Time and Cost of Coding and Checking Notations and Formulas for 1000 Compounds

	Time	Cost
Coding (technical person)	20 hr.	$100.00
Keypunch (operator)	7 hr.	28.00
Checking (computer)	4 min.	22.00
Corrections		
Correcting and recoding	2 hr.	20.00
Keypunching	1 hr.	4.00
Checking (computer)	14 sec.	2.00
Total		$176.00

The registration and file updating time and cost will depend on the size of the file which is to be updated. The larger the original file, the longer it will take to update it. In Table II are shown some costs and times accumulated when 15,000 compounds were added to a file containing 45,000 compounds. Reducing the total figures to the same basis as the preparation figures gives $11.20 as the cost of registering 1000 compounds.

Table II. Time and Cost for Registration
(Computer Operation)

	Time	Cost
Reading onto tape and sorting notations alphabetically	11.75 min.	$72.00
Registration and file updating (15,000 into 45,000)	41 min.	96.00
Total	52.75 min.	$168.00

The average cost then of inputting a compound into the computer system described is approximately $0.19 per compound.

LITERATURE CITED

(1) "Survey of Chemical Notation Systems," Publication 1150, National Academy of Sciences–National Research Council, Washington, D. C., 1964.
(2) "Survey of European Non-Conventional Chemical Notation Systems," Publication 1278, National Academy of Sciences–National Research Council, Washington, D. C., 1965.
(3) Landee, F. A., "Computer Programs for Handling Chemical Structures Expressed in the Wiswesser Notation," Presented before the Division of Chemical Literature, 147th National Meeting of the American Chemical Society, Philadelphia, Pa., April 8, 1964.
(4) Smith, E. G., "Line-Formula Chemical Notation," in press, McGraw-Hill Book Co., New York, N. Y.
(5) Wiswesser, W. J., "A Line-Formula Chemical Notation," Thomas Y. Crowell Co., New York, N. Y., 1954.

Early Experience with a Technical Correspondence Center*

B. F. CLARK, D. J. FLOTO, and R. E. MAIZELL
Olin Mathieson Chemical Corporation, New Haven, Connecticut 06504

Received August 25, 1966

Experiences associated with the establishment and first year's operation of a technical correspondence center are discussed. Methods of retrieval of shelf-filed documents *via* a punch card file, cross indexed in depth, are described. Present procedures and planned techniques for maintaining and improving distribution of information are also presented.

Management's desire to provide an improvement over individual filing systems in our Chemicals Group research center resulted in a decision to establish a central technical correspondence file. After conferences with key personnel plus field trips for outside study, we devised a system intended to serve our present needs and to allow for future development. The Center became operational in September 1965, only two months after management's decision to go ahead.

A set of ground rules was written which were introduced by a list of anticipated benefits, information on selection and submission of material, and other details. These rules were distributed to all technical personnel, accompanied by a cover letter from the Director of Research indicating his approval and requesting cooperation. As indicated in the ground rules, the purposes of this center, were to assure that all technical correspondence pertaining to company business be brought together, adequately safeguarded, and made accessible for quick use as required. The specific goals included:

1. Speedier and more positive access to necessary information.
2. Elimination of need for maintaining large personal files.
3. More productive time utilization.
4. Cross-fertilization of ideas.
5. Ensuring safekeeping of company information regardless of personnel changes.

As an added incentive, management decided that no new filing cabinets would be purchased for use by individuals.

The ground rules specify that correspondence and related documents (preferably originals) received from whatever source are sent to the Technical Correspondence Center (TCC) within five working days of receipt. For memoranda and letters directed to individuals or organiza-

* Presented before the Division of Chemical Literature, 152nd National Meeting of the American Chemical Society, New York, N. Y., Sept. 13, 1966