A Chemically Oriented Information Storage and Retrieval System. III. Searching a Wiswesser Line Notation File*

CARLOS M. BOWMAN, FRANC A. LANDEE, NANCY W. LEE, MARY H. RESLOCK, and BETSY P. SMITH
Computation Research Laboratory, The Dow Chemical Co., Midland, Mich. 48640

Received June 9, 1969

A chemical structure search system based on a stored data file of Wiswesser line formula notations uses the search capabilities of a flexible, general information search program. Searching is executed on inverted index files of assigned fragment codes, open-end connected symbol fragment codes, and element-symbol count codes. The transition from a partial to a complete computerized system is described.

Construction of a computer-based system for handling files of chemically oriented information is the main subject of a continuing series of papers, each of which describes details of some one facet of the system. Earlier papers have covered the general plan of organization and the development of a structure information file based on Wiswesser line notations.[9] This paper will describe the present status of utilization of this file—i.e., current tools and techniques for structure and substructure searching.

FILE DEVELOPMENT

A brief review of tape-stored document file development is necessary to indicate the sources of index search tools. The file record for each structure includes a document number, a complete structure description in Wiswesser line notation form, and a molecular formula. Computer check programs help to assure the accuracy of the descriptive records. Two of these programs were described in earlier papers. One checks the notation accuracy[1] by comparing the submitted molecular formula with a molecular count calculated by computer from an analysis of the notation. Discrepancies flag errors in the notation or document, or indicate inadequacies in the check program. The other program[2] analyzes complex polycyclic structures, chooses, by Wiswesser rules, the canonical path, and writes the correct notation.

Additional accuracy checks detect errors in input card sequence numbers and in document number duplication. The data are sorted (on tape) in alphabetic order by notation and checked for duplicate description entries. The verified data are entered on magnetic tape. This tape-stored document master file is the source of various search tools.

SEARCH TOOLS

Retrieval of the information stored in the master file is dependent on the indexing methods devised for selectively accessing it. The data records are computer-manipulated and computer-analyzed to create two types of search indexes, desk tools and computer tools.

Desk search tools are simply printed listings of computer-manipulated data records, and include:

1. An alphabetic index (notation)
2. A molecular formula index
3. A numeric index (document number)
4. A permuted index (notation symbol)

These listings are valuable as complements and/or supplements to the computer search tools. For certain kinds of questions, they provide immediate and sufficient answers. A particularly useful auxiliary listing is a permuted index of current document additions—i.e., those which have not as yet been fully processed and added to the tape-stored master file. This index (created from cards) serves as an interim search tool between file updatings.

Computer search tools are coded indexes created by computer analysis of the records on the document file tape. These tools are part of a developing series of coded indexes, the development being determined partially by an over-all master plan and partially by continuing evaluation of the codes currently in use. A standardization of code number/document number format has been maintained throughout this series. The advantages of this are:

1. Economy in creation of the search tools and in utilization of the general search program, and
2. Flexibility in using combinations of codes and in integration with other facets of computer-stored and searchable information.

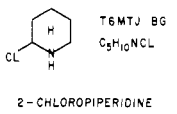
The code numbers, with their associated document numbers, are stored on tape. For each code, sort programs create an inverted index of code terms.

Each of the index codes is designed to meet a specific kind of search requirement. However, the broad range of detail within each permits overlapping and redundancy. This redundancy, which helps to ensure comprehensive retrieval, is apparent in the following brief descriptions of the four index codes now in use:

*Presented before the Division of Chemical Literature, 157th Meeting, ACS, Minneapolis, Minn., April 16, 1969.

00jjjj - Molecular formula
1jjjj - Notation suffixes
2jjjj - Acyclic and sidechains
3jjjj - Ring combinations
4jjjj - Chelate rings
5jjjj - Carbocycles
6jjjj - Heterocycles
7jjjj - In-ring analysis
8jjjj - In-ring substitution

Figure 1. Assigned fragment codes



2-CHLOROPIPERIDINE

00217	04101	61062
01105	05201	80055
02110		26042

Figure 2. Assigned fragment codes for 2-chloropiperidine

Assigned Fragment Code. This is a 6-digit classification index code[3] in which the first two digits are decimal, the remaining four octal (Figure 1). The first digit, which identifies the index, is zero (0) and is usually suppressed. Each of the second decimal digits is assigned to a broad descriptive feature of common occurrence in chemical structures, e.g., the molecular formula, types of rings, or acyclic groups. Each of the octal digits is assigned to descriptive details within each class.

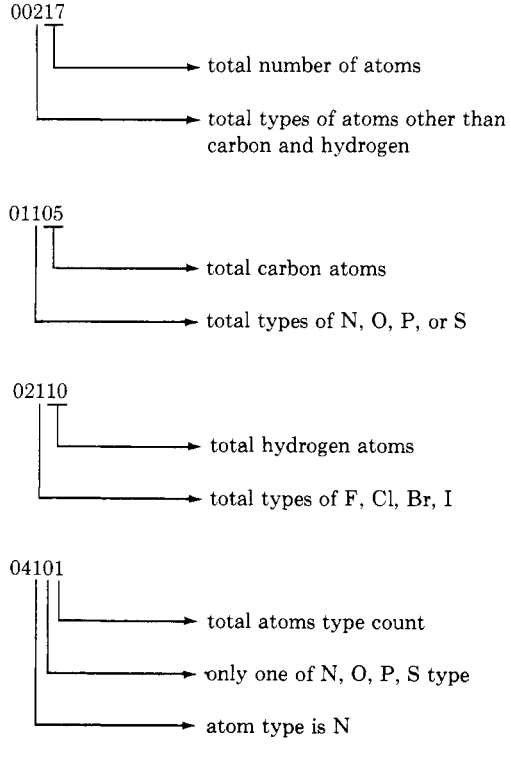
A computer program has been written to analyze each record in the structure data file and generate the correct code for each of its program-defined descriptive features. The number of codes per record obviously varies with the complexity of the structure described. Almost 41,000 code terms have been assigned. It would be impossible in this paper to describe all the structure features included in this index, but the kinds of detail covered may be illustrated by examining some of the descriptive codes for 2-chloropiperidine (Figure 2).

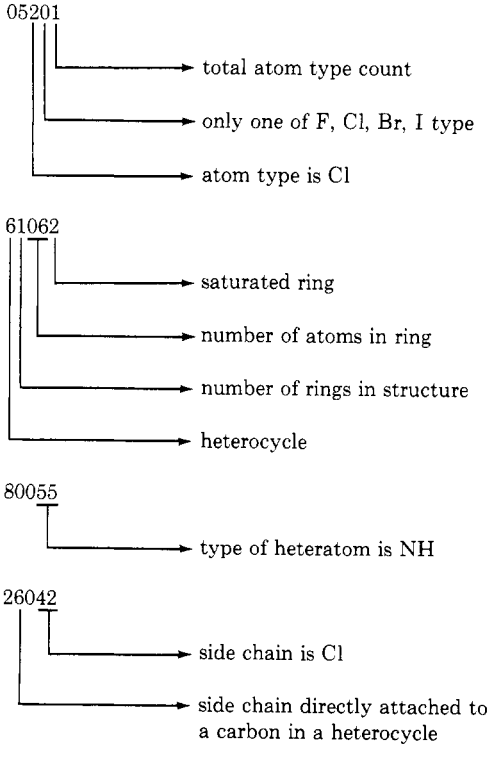
Descriptive index terms for 2-chloropiperidine fall into four of the nine classes of structure information. In this case, only the molecular formula (zero) class requires more than one index code term. The assigned codes have the following meanings:

One additional useful feature of this code should be mentioned. The use of octal value digits permits an inherent broadening of search question terms based on this index whenever the expansion symbols, *, 8, or 9, are substituted for any of the assigned octal digits. These symbols have the following meanings:

* indicates values 0 to 7
8 indicates values 0 to 3
9 indicates values 4 to 7

Connected Symbol Code. This is an open-ended, computer-generated code created from the data record by computer recognition of connected notation symbol fragments within program-defined parameters. Code development is similar to that reported by Hyde *et al.* in 1967.[5] Detailed analysis of in-ring connected symbols is in progress. The remainder of the index code is operational. Except for the code identifying number, the initial 4, the sequentially-generated 6-digit codes have no significance in themselves. The guide to their use is a permuted index of fragment symbols in which each index entry is listed with its code and a frequency count (Figure 3).





Code		Total
403133	TMSWA	2
404155	AOSWA	32
404865	WSAYSSWA-NA-	1
402045	RSWF	76
402415	TSWL	9

*

Figure 3. Permuted index of connected symbol fragments (Example)

Element-Symbol Count Code. This index code is based on the actual count for each element in the molecular formula and for each symbol occurrence in the notation. Evaluation of searches based on the assigned fragment and connected symbol fragment indexes prompted the introduction of this code. It provides a simple, coarse screening, useful either as an initial query step or as a later discriminating query on broad search results.

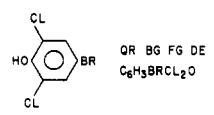
Each file record is computer-analyzed, and for each element except carbon in the molecular formula and each symbol in the notation, two 6-digit codes are generated.

The first two digits of the codes are 50 for the elements, 51 for the notation symbols. The next two digits identify the element (atomic number) or the symbol (assigned number, based in part on alphabet position). The last two digits indicate the actual count for the element or symbol in the record. Two zeros in the last positions simply indicate a count of one or more. For compounds containing no carbon in the molecular formula, a code 500600 is generated.

A few of the codes generated for 2,6-dichloro, 4-bromophenol illustrate these coded index terms (Figure 4).

Saved Search Results. An option of our general search program permits search results to be stored on tape. These stored document numbers, in effect, constitute a selective searchable subject file. The file has a name and a computer-assigned code number by which it may accessed in later searches. The first digit of this 6-digit code is 9; the remaining digits provide for a sequential search number. This option is used to save only those search results which are broad enough to be useful in future searches.

From each of the inverted index tapes, a frequency count of code terms has been made. These printed listings are arranged in term number order. The connected symbol fragments are listed with their corresponding index codes; the other codes are self-defining to the searcher. These frequency counts serve as a key to the use of the codes.



2,6-DICHLORO, 4-BROMOPHENOL

500606 - indicates 6 carbons in the molecular formula
501700 - indicates chlorine in the molecular formula
501702 - indicates 2 chlorines in the molecular formula
511700 - indicates hydroxy group (Q) in the notation
511701 - indicates 1 hydroxy group in the notation
510702 - indicates 2 chlorines (G) in the notation

Figure 4. Element-symbol count codes

SEARCHING

General Search Program. The long-range plans of our information groups include efficient automation of individual information services and integration of individually developed services.

The general search program[4] is designed to achieve these goals. Each information group maintains a degree of standardization in indexing format. The program itself incorporates flexibility through options to handle diverse demands of each group. Search question format is standardized for all groups.

Actual search questions are coded index terms linked with standard intersection, union and/or exclusion logic (AND, OR, NOT). The search program, using Polish notation strategy for searching the inverted indexes, compares matching terms and selects hits. Standard retrieved output is an ordered listing of document numbers. Output options (in addition to the searchable saved search tapes described above) include:

a. A tape-stored listing of retrieved document numbers.
b. A printed notation-molecular formula listing for each document number.
c. A tally of search results—the printed document number listing is suppressed.

The current output options available to the user are:

a. A listing of document numbers.
b. A listing of chemical names, authors, and document numbers.
c. A listing of comprehensive or selective biological screening data.

Files for the last two user options are currently accessible through the document number.

Search Strategy. Maximum search efficiency is measured by time and cost figures, as well as by the comprehension and relevance of retrieved information. These criteria dictate the choice of search tools. For some questions, the choice is obvious. For instance, it would be unnecessary and extravagant to run a computer search for all file compounds containing a rarely-occurring element such as thorium—a permuted index quickly and accurately answers such a question.

On the other hand, a permuted index would be an impractical tool to choose for locating all the phenols in a large file, if computer search tools are available. Between two such extremes are the searches which require a careful comparative evaluation of search methods before choosing a tool or combination of tools.

The following examples illustrate a few cases of tool selection:

1. A question asking for all file compounds containing one sulfur could be answered by one of several methods:

The molecular formula or permuted (notation symbol) listings are possible tools. Either would result in a long, difficult, and probably inaccurate search.

A better choice is the assigned fragment index code which requires four terms, all of them containing an expand symbol (*) to produce rapidly a complete, 100% relevant answer. The coded terms are:

```
( 042*1
OR 044*1
OR 045*1
OR 047*1 )     (Expanded total: 32 terms)
```

The best choice is the element-symbol count code which requires only one term, 501601, to attain the same satisfactory answer. Computer process time and cost are greatly reduced.

2. A question asking for all 1,3,4-oxadiazoles and/or -thiadiazoles in the file, a much more specific request, could be answered satisfactorily with a permuted index. If the file is a large one, however, the search is long and prone to human error. Rapid, accurate results can be obtained through a computer run of two searches using assigned fragment code terms. Both searches use index terms asking for a 5-membered unsaturated heterocycle containing a diazo (—NN—) group in the ring structure. Each adds a further condition, search 1 requesting oxygen, search 2 requesting sulfur (Figure 5).

A more efficient method of achieving the same results would search first for the features common to both structures and use the results of this search for subsequent screening. A run of three searches would be necessary, but repetitious work would be eliminated (Figure 6).

3. Combinations of codes are sometimes the most efficient method. For example, a save search tape, 900031, was created in a search for all phenols in the files. A subsequent search for all *tert*-butyl substituted phenols could be resolved by two code terms:

```
900031
AND  22033
```

The assigned fragment code 22033 means a tertiary butyl group (. . . 33) directly attached to a benzene ring (22. . .).

EVALUATION

Established search facilities based on systematic chemical nomenclature and a modified Opler-Norton[8] code indexing were maintained during the period of notation file and retrieval tool development. For several months, comparative searches were run and results evaluated. Document retrieval is essentially equivalent. Existing programs for users' options (listings of chemical names or biological data) have not been altered; therefore, time-cost figures for these remain unchanged. The significant improvement is in the reduction of actual searching time and in the bulk of material to be handled to achieve comparable results. The computer-based program has been in operation for about six months, handling an average of 2 to 3 computer searches per week.

Search 1	Search 2
(61051	(61051
OR 61055)	OR 61055)
AND 80105	AND 80105
AND 80064	AND 80046

Figure 5. Search conditions using fragment code terms

Search 1	Search 2	Search 3
(61051	1#	1#
OR 61055)	AND 80064	AND 80046
AND 80105		

Figure 6. Elimination of repetition

Realistic time-cost figures are difficult to ascertain for several reasons. During this time, major changes were made to augment computer hardware, with consequent effect on software. A high percentage of irrelevant search answers revealed the need for a supplementary coarse screening device—i.e., the element-symbol code. The effectiveness of this index code in reducing search time and increasing relevancy of search answers is yet to be evaluated. Ring position relationships cannot be determined by existing search methods; therefore time-consuming visual screening of search results (notations) is necessary.

Actual computer process time for searching the index tapes created for approximately 100,000 structures ranges from <2 minutes to ± 12 minutes, varying with the complexity and number of searches per run, and with the number of search options requested. Recharge rates are 10–12¢ per second for process time and 2¢ per second for input/output time. Thus, a search run using 300 seconds process time and 2000 seconds I/O time would have a computer charge of $76 for document number retrieval. The charge per search is prorated by the number of searches in the run. It is reasonable to expect $\leqq$24-hour turn-around time for this option.

Separate tape files of chemical names and authors ($\sim$100,000 records) and biological screening data (5,000,000 entries on 32 tapes) are accessible through document numbers. Costs for these auxiliary output options vary from $\sim$$85 per search run for chemical names to $\geqq$$700 for the biological data. Batch runs prorate, by the number of searches, the cost of retrieving this auxiliary data. These batch runs include searches for other information groups.

All programs have been written in Extended ALGOL-60 for the Burroughs B-5500 computer. Programs now in progress or included in master plan projections are designed to further increase relevancy and scope of search results (including ring position relationships and structure display), reduce search time, and provide expanded accessibility to auxiliary information.

SUMMARY

The utilization of information in a computer-based chemical structure file has been described. Current structure and substructure search tools include printed indexes (desk tools) and tape-stored indexes (computer tools). The latter use the capabilities of a general information search program. This program is designed to meet the search needs of individual information groups, as well as to implement integration of related files of information.

LITERATURE CITED

(1) Bowman, C. M., F. A. Landee, and M. H. Reslock, "A Chemically Oriented Information Storage and Retrieval System. I. Storage and Verification of Structural Information," J. CHEM. DOC. 7, 43–47 (1967).

(2) Bowman, C. M., F. A. Landee, N. W. Lee, and M. H. Reslock, "A Chemically Oriented Information Storage and Retrieval System. II. Computer Generation of the Wiswesser Notations of Complex Polycyclic Structures," *Ibid.*, 8, 133–137 (1968).

(3) Bowman, C. M., F. A. Landee, M. H. Reslock, and B. P. Smith, "Automatic Generation of Structural Fragment Codes From the Wiswesser Line Notation for Rapid Structure Searches," Proceedings of the Wiswesser Line Notation Meeting of the Army Chemical Information and Data Systems, James P. Mitchell, Ed., pp. 49–56, **EASP 400-8**, Edgewood Arsenal, Md., 1968.
(4) Farris, R. N. "Computers Cut the Cost of Literature Searches," *Chem. Eng. Progr.* **62** (5), 89–91 (1963).
(5) Hyde, E., F. W. Matthews, L. H. Thomson, and W. J. Wiswesser, "Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds," J. CHEM. DOC. **7**, 200–204 (1967).
(6) Landee, Franc A., "Computer Methods of Handling Files of Chemically Oriented Information," unpublished paper presented in Moscow, USSR, Oct. 1965.
(7) Landee, Franc A., "Computer Programs for Handling Chemical Structures Expressed in Wiswesser Notation," Presented before the Division of Chemical Literature, 147th Meeting ACS, April 8, 1964.
(8) Opler, A., and T. R. Norton, "A Manual for Programming Computers for Use with a Mechanized System for Searching Organic Compounds," The Dow Chemical Co., Western Division, Pittsburgh, Calif., 1956.
(9) Smith, E. G., The Wiswesser Line-Formula Chemical Notation, McGraw-Hill, New York, 1968.
(10) Smith, E. G., "Machine Sorting for Chemical Structures," *Science* **131**, 142–146 (1960).

Index Chemicus Registry System: Pragmatic Approach to Substructure Chemical Retrieval*

EUGENE GARFIELD, GABRIELLE S. REVESZ, CHARLES E. GRANITO, HAYES A. DORR, MARIA M. CALDERON, and ANDREA WARNER
Institute for Scientific Information (ISI), Philadelphia, Pa. 19106

Received July 21, 1969

The *Index Chemicus Registry System (ICRS)*, launched in 1968 with the support of a dozen industrial and government organizations, is now a current operational monthly service. Subscribers receive magnetic tapes and printouts, in which the weekly issues of *Index Chemicus (IC)* have been encoded in Wiswesser Line Notations (WLN). Over 13,000 compounds per month are provided in machine language. The canonical WLN is also provided in alphabetized printouts. Encoding of over 400,000 new chemical compounds from *IC* has already been completed, including all those reported in 1967, 1968, and 1969. Since the tapes also include title and other bibliographic information, this paper describes the use of supporting software provided for SDI search systems employing "word" and other searching terms, in addition to the WLN fragments. Use of the monthly and annual printouts are illustrated for those searches which do not require computer manipulation.

The *ICRS* is designed to provide chemists with current and retrospective chemical information reported in the *IC*.

As *IC* has been described elsewhere,[1] it is sufficient to state that *IC* provides detailed abstracts of journal articles which report new chemical compounds or new chemical reactions.

The *ICRS* has, as yet, not been described in the literature, and a brief description of its main characteristics is necessary to enable one to understand how to search for substructures, both currently and retrospectively.

ICRS consists essentially of four data files: WLN magnetic tapes, *IC* bibliographic tapes, WLN printouts, and *IC* weekly issues.

The WLN magnetic tapes contain unique WLN structural descriptions of all new compounds reported in the *Index Chemicus* and are arranged in abstract number sequence. The WLN tapes also contain molecular formulas and *IC* registry numbers, which identify a specific line in the numbered *IC* abstract where a structural diagram and other information is given.

The *IC* bibliographic tape provides, in machine language, most of the information provided in the printed *IC*: bibliographic citations; codes for new reactions and analytical instrumentation; subject-index terms which are assigned by chemists and include terms related to the properties, uses, and biological activity of the compounds.

The WLN printout version is alphabetized according to the WLN, to provide easy scanning for similar type compounds. The corresponding article from the *IC* can be identified through the registry numbers associated with the notation.

Many searches can be done by simply referring to the monthly or annual *ICRS* printouts. The search for substituted adamantanes is shown in Figure 1. The WLN notation for adamantanes is L66 B6, etc. The *ICRS* printout identifies abstract number 101318, which contains several adamantanes, each of which is separately encoded. The *IC* abstract is shown in the lower portion of Figure 1.

The printouts are also used to formulate machine-search questions.

*Presented in part at the ACS MARM Meeting, Washington, D. C., February 14, 1969, and before the Division of Chemical Literature, 157th Meeting, ACS, Minneapolis, Minn., April 16, 1969.