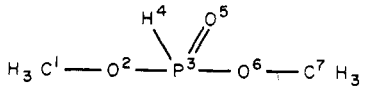
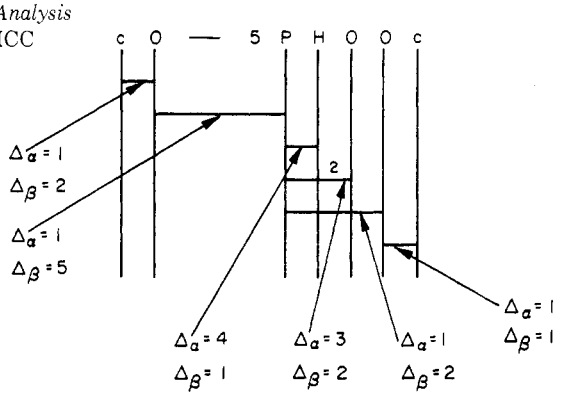
Example (A4)



Bond Analysis

MCC



ACKNOWLEDGMENT

The author gratefully acknowledges the suggestions of and consultations with the following persons. D. Loev, M. Plotkin and J. Munz, and C. T. Van Meter of the University of Pennsylvania, A. Genarro of the Philadelphia College of Pharmacy and Science, Sylvan Eisman of the Frankford Arsenal, William Wiswesser of Fort Detrick, P. Olejar, S. Rhodes, and T. Quigly of the National Science Foundation, J. Mitchell of the Edgewood Arsenal, and Fred Tate and his staff of the Chemical Abstracts Service.

LITERATURE CITED

(1) Hiz, J., J. CHEM. DOC. 4, 173 (1964).
(2) Sussenguth, E. H., *Ibid.*, 5, 36 (1965).

Use of a Nonunique Notation in a Large-Scale Chemical Information System*†

DAVID LEFKOVITZ
The Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa. 19104

Received May 4, 1967

This paper examines the functional requirements of an automated chemical information system as it might be implemented on third generation ADP hardware. Of primary concern in this examination is the formal representation of the structural formula within the automated files and the requirements placed upon this representation by the five system functions of Input, Registry, Storage for Search, Search, and Display. The various representations are divided first into the two broad categories of connection tables and notations. These are then broken into categories of unique and nonunique representations. Also examined are ease of automatic generation and manipulation of the various representations. The paper also presents a discussion of a desired systems approach to registry and search for real time, interactive operation and a final recommendation for structural formula representation in this type of system.

The primary objective of this paper is to discuss the combination of the concept of a nonunique notation or code for representing chemical structural formulas and the functional characteristics of third generation ADP equipment as they relate to the design of a large scale automated chemical information system. The most significant feature of such a system, which sets it apart from most other information storage and retrieval systems, is the requirement to search a large number of chemical structures, wherein the decision process to accept or reject a given structure for a given query can be complicated enough to require several seconds of processor time. Because a large scale system would ultimately have to handle about 3 million compounds, retrieval on a serial search basis would be prohibitive; therefore, screens, or retrieval keys, are assigned to each compound, analogous to the retrieval keys of a document in a document retrieval system, so that a more rapid decision can be made on a large percentage of the file, simply on the basis of a match between the keys in the query and corresponding keys in a given file compound.

The new generation of computer hardware provides direct access storage for several hundred million characters of information. That is, access may be had to any single record among a several-hundred-million character store

*Based in part on the paper "The Impact of Third Generation ADP Equipment on Alternative Chemical Structure Information System" presented before the Division of Chemical Literature, 153rd Meeting, ACS, Miami Beach, Fla., April 1967.

†This study was supported by the National Science Foundation under Contract No. NSF-C467.

within approximately 100 milliseconds; therefore, if the total file could be stored in such a medium, and if there existed an extremely efficient set of retrieval keys that would partition the file, for a given query, into a relatively small subfile, then a system could be designed that could search the file on a random access basis in real time more economically than considerably smaller files are presently being searched on a batched basis.

A second feature of third generation hardware, which differentiates it from the present line of computer hardware is that a relatively large number of remote consoles can communicate simultaneously with the central processor that is performing the above described search, so that many users could query the real time search system simultaneously. Typically, the number of such consoles that could be connected to a system would be on the order of several hundred. To provide real time service to all of the consoles, the compound files in the direct access memory would have to be used on a time shared basis under the control of an executive program in the central processor in a manner similar to that described in Reference (*1*). A few additional details on file organization, beyond those given in (*1*), are provided in The Hybrid System section of this paper.

There are three basic reasons for considering the implementation of a real time retrieval system in third generation hardware:

> Assuming the existence of an efficient set of retrieval keys, it will be far more economical to search the file randomly than serially.
>
> The system will have greater utility to research chemists if answers to individual questions, or more particularly, to a series of questions, can be retrieved immediately.
>
> The economics of retrieval may very well depend upon an interchange of information between the chemist who has posited the query and the automatic system which is attempting to search in answer of the query. The necessity for such an interchange increases as the structure of the key system and the file organization becomes more sophisticated.

A third hardware feature that will add to the effectiveness of a large scale chemical information system is the utilization of a wider range of input and output devices. The cathode ray tube (CRT), for example, can be used as an output medium, especially in a real time system, for the display of structural formulas as well as for the display of textual information. Furthermore, the CRT can be programmed to display perspective in order to better illustrate three-dimensional molecular features. The light pen can be used with the CRT as an input device for entering structural formulas in the query, or for modifying structural formula responses to present the system with newly created queries. The CRT can also be coupled with photographic or xerographic equipment in order to produce hard copy or for photocomposition. High speed line printers and chemical typewriters may also be used; the former for large volume, relatively high quality output, and the latter for low cost on-line, low volume, bi-directional communication between the querist and the automated system.

In general a chemical information system, whether it is designed to operate in a batched or an on-line real time mode, has all of the functional requirements of a document information system, and it is not the purpose of this paper to discuss these requirements and the design principles which follow therefrom, because they can be found elsewhere throughout the more recent IS&R literature. However, as stated above, the crucial problem for the large-scale chemical information system is storage and retrieval based upon the chemical structural formula. The problem is that of storing a representation for each of approximately three million chemical compounds that may eventually come into such a system, plus the storage of the associated retrieval key system, in such a form that the system is economically feasible from a storage space and file maintenance standpoint as well as from a standpoint of efficient search.

At present there appear to be two mutually exclusive schools of thought in the development of automatic chemical information systems. One is developed completely around the use of a connection table (CT) as the representation of chemical structures; the other is developed around the line notation (LN) as the representation. Table I presents five functional requirements of an automated chemical information system, and all of these

Table I. Functional Requirements of an Automated Chemical Information System

Functional Requirement of System	Description
Input	Structural formulas are input from a drawn diagram and encoded into CT or LN. Input medium may be: (1) chemical typewriter (2) CRT/light pen (3) pen writer (4) optical scanner (5) keypunch machine
Storage	CT or LN is stored for registry and search in magnetic mass storage medium. For registry, where file is ordered, magnetic tape may be used; for search, medium should be direct access medium such as disk or strip (data cell, race) storage.
Registry	Compounds are automatically registered into system by search through sequenced file of canonical structural representations.
Search	System performs full and substructure search on CT or LN in direct access storage.
Display	System displays structural formulas from search file either via computation upon CT or LN, or VIA structural formula image (SFI) retained a compressed from input device.

Table II. Comparison of CT and LN with Respect to Functional System Requirements

Representation Code—	R1	R2	R3	R4	R5	
Functional requirement of system	CT	Manually generated ULN	Manually generated NULN	Automatically generated ULN from CT	Automatically generated NULN from CT	Recommended for Hybrid System
input	Via mechanical input device	Manual, using rule book	R2	R1	R1	R1
Storage in search file	Least concise. 6–12 chars./ Non H atom	Most concise. 1 char./Non H atom	R2	R2	R2	R5
Registry	Can be made canonical	Canonical by virtue of rules, if uniformly applied and checked	Inadequate	Canonical by virtue of rules and guaranteed uniformity; however, program will be difficult to write	Inadequate	R1
Search (substructure)	Atom-by-atom search	Symbol connectivity search	R2	R2	R2	R5
Display	Key coordinate assignment on SFI	R1	R1	R1	R1	R5 (Based upon storage)

must be satisfied by the respective structural formula representation. The fact that no system in existence today combines the use of CT and LN in order to allocate to that representation those system functions to which it is better suited, does not indicate that such a "hybrid" system approach is inappropriate, or technically infeasible or unsound; the reason they have not been used in combination can probably be traced to the fact that the operational demands of batch processing systems, which are the type that have thus far been developed, are not so great as to require optimum efficiency with respect to each of the five functional requirements listed in Table I. Secondary reasons for the nonapplication of the hybrid approach could possibly be traced to economics and speed of implementation, and to the training and background of those who have designed and implemented the respective systems.

In the following section a comparison is made between the CT and LN with respect to these five system functions. A conclusion is drawn from this comparison which leads to the proposed hybrid system configuration, which is then extended to the use of a code which possesses all of the representational properties of a connection table and the conciseness property of the notation. This code is called the Mechanical Chemical Code (MCC) and is fully specified in Reference (*2*). Its development follows five principles which are a direct consequence of the five functional system requirements enumerated in Table I. The code itself represents a synthesis of ideas contained in existing line notations, and its form is based upon a code suggested by Hiz (*3*) in 1964. In essence the code could replace both the CT and the LN in the hybrid system thereby returning to a singular representation that can most efficiently serve all of the system's functional requirements.

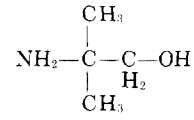
THE HYBRID SYSTEM

Comparison of CT and LN. Table II presents a comparison between the system's utilization of connection tables *vs.* various forms of the line notation. Four cases are enumerated for the use of a line notation, each being assigned a representation code of R2 through R5 in Table II. The CT is assigned the code R1.

The first LN case (R2) is the currently used unique line notation (ULN), such as the Wiswesser Line Notation (*4*), which is manually encoded by persons trained in the application of the rules. A cipher is produced by the ULN which can be used for registry as well as generic search.

The second case (R3) is the manually encoded nonunique line notation (NULN) which is easier to encipher though not as versatile as ULN since it cannot be used for registration. No current use is made of this type of notation. An example of a nonunique WLN is presented in Figure 1. The NULN may be used as readily as ULN for generic search as it unambiguously describes the same structure (if the notation language is in fact unambiguous).

The third case (R4) is the automatically generated ULN from a CT, wherein a computer program translates from a CT to ULN. The purpose of this approach would be



WLN (without methyl contractions) ZX1&1&1Q

NUWLN Q1X1&1&Z

Figure I. Example of a nonunique linear notation (NUWLN)

to standardize the rules through the medium of the computer to assure uniformity in the generation of the notations. There are no systems in use today that utilize this approach.

The fourth case (R5) is the automatically generated NULN, wherein a computer program automatically generates nonunique notations. Such a program would be considerably less complex than the ULN program of R4, but no current system utilizes this approach either.

Table II lists the five system functional requirements in the first column, and a comment indicating the mode of application or the applicability of each representation to each function is given in the respective block of the table. Where a representation code (R1 through R5) is given in a block, the application is the same as for the indicated code. For example, the *Input* procedure for R3 is the same as for R2—namely, *manual, using a rule book*.

The recommendations of the last column in Table II are based upon the respective qualities of each representation.

Starting with *Registry* the CT (R1) is recommended because its uniqueness (canonicity) is a relatively simple consequence of a computerized algorithm. In one instance is the canonical CT numbering algorithm of CAS (*5*), and in the other is the isomer sort registration (*6*) (wherein isomers are differentiated via atom-by-atom search) of the Army Chemical Information and Data System (CIDS). In the line notation, only ULN approaches are applicable to Registry, but these are dependent upon the correct application and interpretation of the rules, by people in the case of R2 and by machine in the case of R4; this program, however, might be very difficult to write. The human interaction in the CT registry is at a clerical level (*Input*), whereas in the LN registry it is at a higher intellectual level, and programs to detect errors in lower (intellectual) level human functions have, in the past, been more successful than for higher level functions. To detect chemical inaccuracy in a LN (by comparing element counts with the molecular formula) is relatively easy, but to detect errors in the application of the uniqueness rules (as illustrated in Figure 1) might be tantamount to writing the R4 program itself.

In summary, *Registry* by CT is relatively automatic, certain, and is based upon the lowest intellectual level of human participation, and hence is recommended for a large scale system.

The *Input* procedure follows from the requirements of the Registry and therefore is likewise CT.

The specific advantages of the LN are its conciseness, and, as recently shown by Hyde (*7*), its capability for a symbol connectivity search, which, if the notation is properly designed, can be as effective as the CT atom-by-atom search. As will be shown, subsequently, conciseness is a prime factor in the requirements for a real time system because of the expense and limited capacity of direct access storage, and as the LN can also be used for detailed substructure search, Table II recommends that a LN be used for *File Storage* and *Search*. Finally, since it is already in the search file storage, it should also be used for *Display*. [It is not essential for display if the typewriter structural formula image (SFI) is also to be stored.] However, because (1) the LNs would most reliably and economically be produced by a conversion program from CT, (2) it is easier to convert from CT to NULN, and (3) the requirements of storage and search can be satisfied equally as well by NULN, Table II recommends that R5, the automatically generated NULN, be utilized.

File Construction and Search. Figure 2 presents a block diagram which summarizes the recommendation column of Table II.

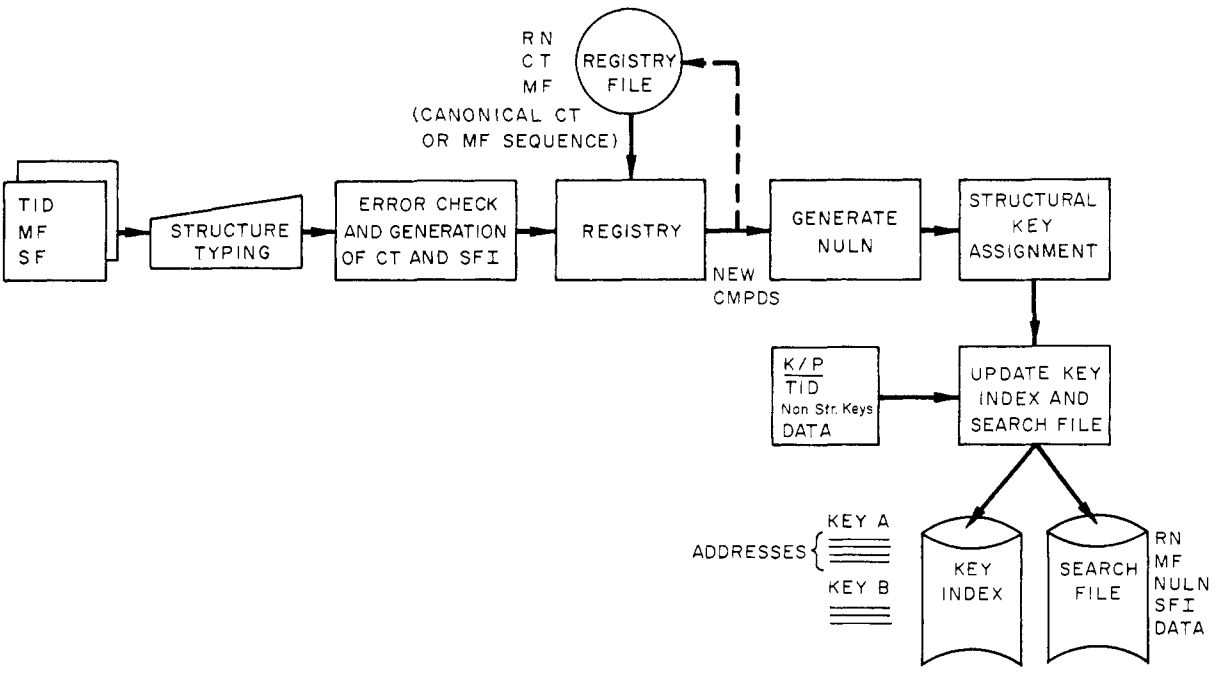


Figure 2. File construction in hybrid system

The input to the system is hard copy containing a temporary identification number of the compound, the molecular formula, and the structural formula. These data are mechanically encoded by means of a chemical typewriter, an optical scanner, or possibly a cathode ray tube and light pen combination. The output of any of these devices is convertible to computer magnetic tape, whereby it is first processed by an error checking program to detect errors, or, as far as can be determined, chemical errors in the structural formula or the molecular formula. A connection table is then generated, and if the display of the structures in the retrieval system is dependent upon a direct image of the encoded structural formula, this image with complete or with key coordinates of the SF is retained (herein called an SFI). The CT or the CT/MF combination serves as the basis for unique registry. In the CAS system a canonical CT is generated by means of an algorithm which canonically numbers the atoms in the connection table. The registry file is then sequenced alphanumerically by the resulting canonical CT so that a potential registrant is admitted as a new compound or not, depending upon whether its canonical CT matches or does not match an existing one in the registry file. The Army CIDS, on the other hand, sorts the registry file by molecular formula and will search for a potential registrant within an isomeric group. Either of these techniques is certain to produce an accurate registration if connection tables are accurately generated from the system input. After registry the new compounds would pass through a program that generates a nonunique line notation from the connection table. Retrieval keys may then be assigned, based either upon the NULN or upon the CT, whichever functions more effectively within the key assignment program. The structural keys would then be combined with nonstructural keys which are entered by means of a keypunch process and identified by corresponding TID numbers. TIDs would be cross-referenced to the appropriate registry numbers and the structural and nonstructural keys under a given registry number would then be used to update an inverted key index, which is stored in a direct access disk memory.

Each key in the system may be identified by an integer code number, and a table is provided which translates from the English or natural language form of the key to the code number. In the inverted index, under each key (for example, key A, as shown in the diagram) are listed in sequence all of the addresses of compounds in the search file which contain the given key. A certain amount of reserve space is left after the sequence of addresses for updating purposes. If this reserve space should be consumed the update program can link to another reserve block in the key index. Periodically a file maintenance program reconstructs the key index and collects all of the linked blocks under a single block, once again associated with its appropriate key. Since the addresses are maintained in sequence it is relatively easy to perform the Boolean function of 'AND' or the intersection of two or more such key listings by reading blocks of addresses into the core memory of the processor, performing the appropriate intersection in a single pass of two lists, and continuing to read more blocks, if required, until all of the required keys in the conjunction have been intersected. Likewise the Boolean negation is performed by ensuring that no address on a negated key listing appears in the final intersected list. The Boolean 'OR' function is a simple merge of two or more lists. Thus queries containing Boolean expressions of structural formulas which produce structural keys, structural keys themselves as they may appear directly in a query, and nonstructural keys may be searched directly in the random access key index before entering the search file. The number of addresses which result from this search will indicate an upper bound on the actual retrieval in the search file.

The search file itself is also stored in direct access disk or strip (Data Cell or RACE) memory. This file can become quite large, particularly if a large amount of compound-associated data is to be stored. In Figure 2 the components of a record in this file are indicated as a registry number (RN), molecular formula (MF), nonunique line notation (NULN), structural formula image (SFI), and data which may be printed but not searched, such as literature references. If the data file were to become very large and if it were not always the case that these data were a retrieval requirement, then it would be possible to split the search file and store the bulk data in a separate, less expensive storage medium. In addition, the requirement for an explicit SFI may in time disappear, when it becomes possible either to print the structural formula directly from the line notation or at least to store key coordinates with the line notation and to recompute the original formula based upon these coordinates.

Table III presents an estimate of the storage requirement for the search file, excluding the print data such as bibliographic references, and assuming that key coordinates are used instead of a compressed structural formula image. Three million compounds it is estimated would require approximately 330,000,000 characters, which is 1½ IBM 2314 direct access (disk) storages. Also, half of the storage requirement is required by the key coordinates, and if in the future it were to prove feasible to compute a structural formula directly from the stored line notation, then this file requirement would be halved.

The bulk print data could be stored in the replaceable Data Cell cartridges or the replaceable disk pack cartridges.

Figure 3 illustrates the search of these files in the hybrid system. The queries are presumed to come either from remote consoles which could be as simple a device as a teletypewriter or a chemical typewriter, or could be the more sophisticated buffered CRT with light pen combination. Alternatively a query could be submitted on-site by the card reader.

Table III. Estimate of Storage Requirement for Search File (Excluding Print Data)

Record Component	No. of Characters (Av.)
RN	9
MF	12
NULN	25*
(Key Coordinates)	60
(Record Control Data)	5
	111 ≈ 110
*An average sized compound is assumed to be 25 non H atoms	× 3 million compounds = 330 million characters

The query is described in this diagram as consisting of two parts, A and B. Part A is the initial query specification and part B provides further information to guide the retrieval and output after the system has responded with initial retrieval statistics, based upon the key index search. A given query pattern could repeat part A a number of times, each time with a new modification, based upon the retrieval statistics, before proceeding to a part B.

Part A of a query consists of the following: It would be headed by an ID number so as to distinguish it from other queries with which it was being batched or time shared, and to route the responses back to the appropriate query source. Part A of the query might consist of as many as 4 subparts.

The first could be a structural formula or a set of structural formulas with an indicated Boolean relationship among them (indicated by B in the figure). The structural formula could be entered by a chemical typewriter or input via the light pen/CRT combination, or, if the structural input devices were not available, it could be typed as a connection table or typed as a NULN.

The second subpart could be a molecular formula or a molecular formula range specification, such as "six to ten carbon atoms, any number of hydrogens, no more than two oxygens and anything else."

The third subpart could be structural keys in a Boolean expression, assuming the querist had access to a thesaurus of these keys and desired to use them in addition to or in lieu of a structural formula.

The fourth subpart would consist of nonstructural keys in a Boolean expression, also selected from a controlled system thesaurus. A given query could consist of any one or a combination of these four subparts.

Part A of the query is first processed by the automatic key assignment program if there is a SF in the query. This program would assign from the same structural key vocabulary as was used in the structural key assignments to the compounds of the file. These keys are then combined with any other keys of the query and decoded in the inverted key index. At this point the number of addresses in the decoded listing are returned to the querist as retrieval statistics. If the query is entered from a remote console then the retrieval statistics are returned in real time. If the query has been entered on-site by the card reader then the retrieval statistics would be printed and instructions from the querist with regard to possible actions based upon the retrieval statistics would subsequently be followed in part B of the query. In either case the response to the retrieval statistics may be a modification of query part A, which in essence is another query, or the entry of query part B.

Part B consists of the same ID number followed by three retrieval specifications. The first specifies an output device. This choice would usually depend upon the magnitude of the retrieval statistic. In a browsing mode, either where the retrieval statistic is small or where the querist simply wants to view some of the retrievals immediately, the console typewriter or CRT may be designated as the output device. More voluminous retrievals, as indicated by a larger retrieval statistic, would be designated for output to a line printer that might be located at the computer site or at a remote location. In the former case the output would be to a magnetic tape which would subsequently be printed, and in the latter case the output would be over telecommunication lines through a high speed data set to the remote line printer. In addition it should be possible to interrupt any output from the remote console and either switch the output device or request that all retrievals, including those already printed, be output on a high speed output device.

The second retrieval specification relates to the level of inquiry in the search file. Three such levels are here suggested. The first is to retrieve, without further examination, all addresses decoded in the index. This would be

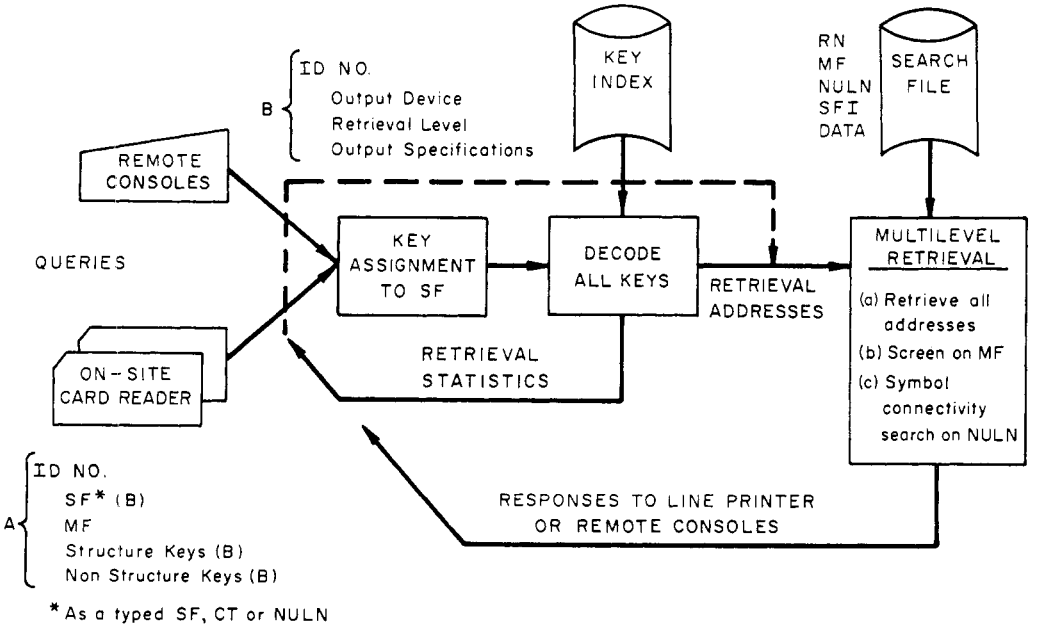


Figure 3. File search in hybrid system

used if the retrieval statistic were relatively small or if the querist regarded his part A key specification as very exact. This level would, of course, be the most economic from the standpoint of processor time but might be the least economic from an over-all systems point of view in terms of the amount of data to be printed or transmitted. The second retrieval level would be to screen on the molecular formula specification, if one existed in part A of the query, and the third level would be to perform a symbol connectivity search on the nonunique line notation.

The third output specification in part B of the query designates the specific data elements to be printed, such as the molecular formula, perhaps the line notation, the structural formula, and the print data. In all cases the registry number would be printed. A description of the executive system that would be required to control such a search system in a real time, time shared environment is described in Reference (*1*).

THE MECHANICAL CHEMICAL CODE: AN EXTENSION OF THE HYBRID SYSTEM CONCEPT

It was recommended in Table II that (1) the system input be via the connection table because it is a mechanical process requiring the least intellectual effort on the part of the human beings who are a part of the process and is therefore most susceptive to automatic error checking by computer; (2) storage for search be via line notation because it is the most concise and therefore makes the use of direct access storage economically feasible; (3) the connection table be used for registry because it is most susceptive of automatic (computerized) canonical encoding (or can be used in conjunction with the molecular formula for registry); (4) substructure search be via LN because the search file storage is in LN and an atom-by-atom search in the form of symbol connectivity search is feasible and either equally or more efficient than atom by atom search on a connection table; (5) line notation be used to display the structural formula possibly in conjunction with key coordinates, again because the search file stores the LN.

The nonunique line notation was recommended for the reason that it would be easier to write a program to translate from connection table to a nonunique rather than a unique line notation, and because a unique line notation is required only for registry and not for search. However, one might question the process of converting from a connection table representation to a line notation, if it is the case that there does not exist a one-to-one relationship between the two—i.e., if all of the information contained in the connection table representation (which is judged to be the most complete representation of the structural formula) cannot be represented in the language of the line notation. Furthermore, if it is not the case that there is a one-to-one relationship, then there may be ambiguities inherent in the language of the line notation which will produce ambiguous notations, and which would partially negate the search effectiveness of the file. A further benefit that could be derived from a line notation that stood in a one-to-one relationship with a CT would be that just as the LN contains all of the CT information for the purpose of search, so it would contain all of the same information for the purpose of registry and could therefore replace the CT in the registry file. This latter advantage would not be as significant as in the search file because registration will probably continue to be a serial process, and hence, minimizing the size of the registry files is not as critical to the operation of the system as it is in random access search. However, some small economies could be gained by a more concise registry file.

At this point in the discussion a change in terminology is to be made. Since it is the case that the LN, as it is utilized in the envisioned hybrid system, is a storage and processing mechanism used exclusively by the computing system and is never used externally either for input or for output, it would seem appropriate to drop the nomenclature of *notation*, since this may connote manual encipherment, and to substitute the term *code*. Furthermore, as it is to be recommended that the conversion in either direction from CT to the *code*, or vice versa, be completely automatic, without requiring any chemical interpretation, this *code fix* should be modified by the adjective *mechanical*. And finally, as the *code* represents *chemical* structures it should be called a Mechanical Chemical Code.

It functions completely internally to the machine; it is designed to optimize machine functions, and the rules of formation of this *code*, either for the purpose of encoding or decoding from or to connection tables or structural formulas need never be the concern of a chemist, but should be completely and readily specifiable in terms of a computer algorithm.

In summary, the desired properties of this code should be as follows:

> It should correspond one-to-one with the CT.
>
> The conversion in either direction, from or to a CT, is completely automatic, without requiring any chemical interpretation.
>
> It should be more concise than a CT, in the sense that it requires fewer characters (or bits) in digital storage.
>
> It should be searchable directly by either an atom- or a symbol-connectivity search.
>
> It would be desirable, though not essential, if a structural formula could be displayed directly from it.

File Construction in the MCC System. Given a Mechanical Chemical Code with the above described properties, the file construction block diagram of Figure 2 (the Hybrid System) could be redrawn as shown in Figure 4. The MCC has replaced the CT in the registry file and the CMF [Coded Molecular Formula, described in Reference (*2*)] replaces the MF. Two possible registry systems have been diagrammed. In one, the file is in CMF sequence and the registry process is analagous to isomer sort registration. In the other, the file is maintained in MCC sequence and the registry is the same as now exists at CAS. In the upper registry system of the diagram, the typed structures are generated originally in MCC and thereby registered. In the alternate registry system the canonical CT of CAS would be their basis of registration, and a conversion from CT to MCC is performed for registry and storage. After the registry process, structural representation in either case is via MCC and the generation

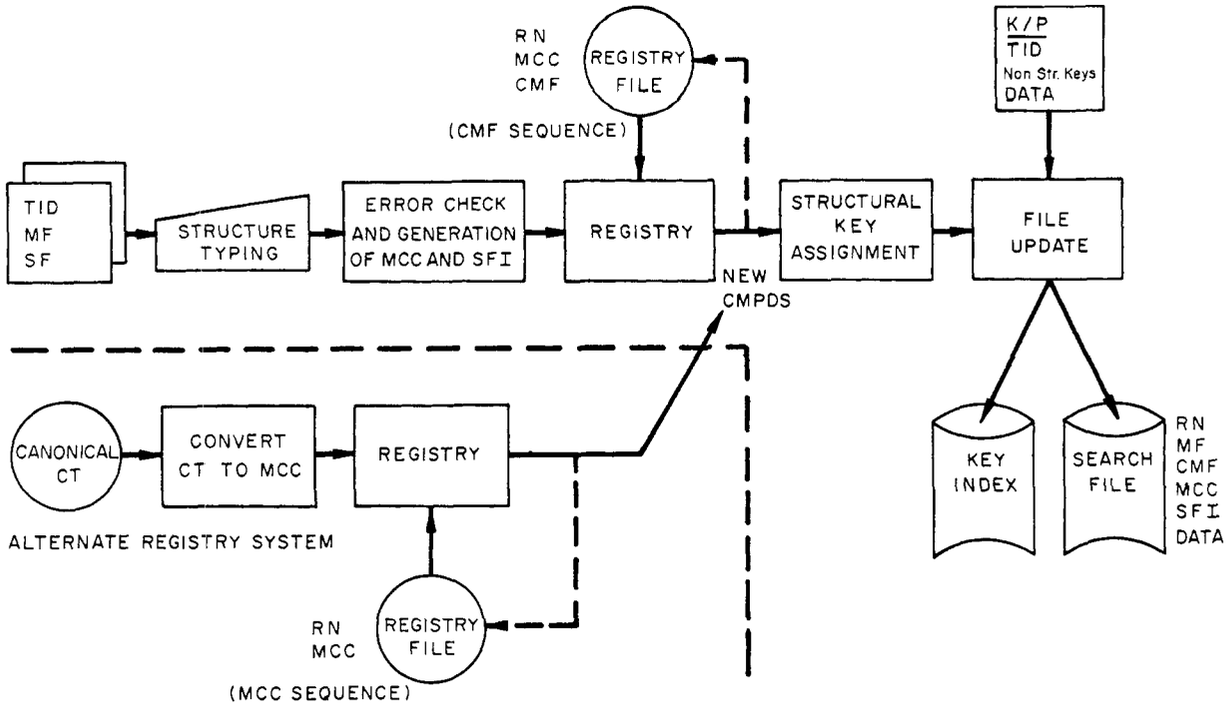


Figure 4. File construction in MCC system

of NULN, as shown in Figure 2, is not required. From this point on the two systems are identical except that MCC has replaced NULN and CMF is stored in addition to MF. Therefore, in one form of this system the only representation of the chemical structure throughout the entire system is MCC. In the alternate system the canonical CT is still used as the basis for registration, but the registry file itself contains the canonical MCC representation.

Summary. Figure 5 presents a schematic diagram of the logical process that has led to the synthesis of the MCC specified in (2). Starting at the top are the five system functions presented at the beginning of this paper in Table I. These in turn led through a description of the Hybrid System to a set of *code* requirements. These requirements were in turn imposed upon the form of the MCC itself. The properties of other chemical structure languages were examined and two were found to have certain formal aspects which were amenable to the *code* requirements. Most significantly, the form of the *code* follows that published by Hiz (*3*) and Eisman (*8*) in 1964. Certain conventions of the Wiswesser and Hayward Line Notations have also been adopted. The fourth major ingredient in the synthesis is the connection table and abnormality table as it has been implemented in the system at Chemical Abstracts Service and by the Army CIDS.

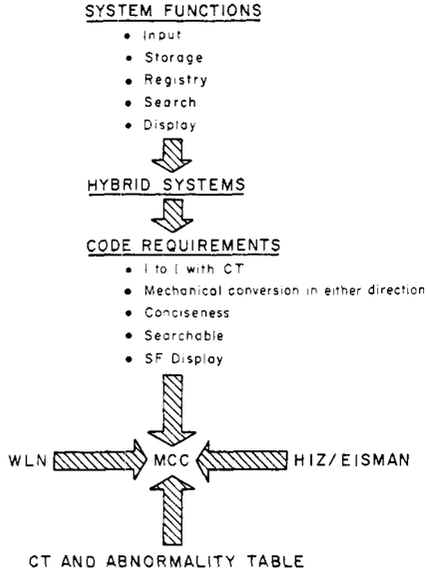


Figure 5. Synthesis of MCC

ACKNOWLEDGMENT

The author gratefully acknowledges the suggestions of and consultations with the following persons. C. T. Van Meter and the University of Pennsylvania CIDS staff, Paul D. Olejar, Sarah M. Rhodes, and Thomas W. Quigley of the National Science Foundation, and James P. Mitchell and his staff of the Edgewood Arsenal.

LITERATURE CITED

(1) Lefkovitz, D., and R. V. Powers, "A List Structured Chemical Information Retrieval System," *Proc. 3rd Ann. Colloq. Inform. Retrieval*, May 1966.
(2) Lefkovitz, D., "A Chemical Notation and Code for Computer Manipulation," J. CHEM. DOC. 7, 186 (1967).
(3) Hiz, H., "A Linearization of Chemical Graphs," *Ibid.*, 4, 173 (1964).

(4) Smith, E. G., *et al.*, "W. J. Wiswesser's Line-Formula Chemical Notation," unpublished data, 1966.
(5) Morgan, H. L., "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service," J. CHEM. DOC. **5**, 107 (1965).
(6) Lefkovitz, D., and C. T. Van Meter, "An Experimental Real Time Chemical Information System," J. CHEM. DOC. **6**, 173 (1966).
(7) Hyde, E., and L. Thompson, "Organic Search and Display Using A Connectivity Matrix Derived from Wiswesser Notation," Division of Chemical Literature, 153rd Meeting, ACS, Miami Beach, Fla., April 1967.
(8) Eisman, S., A Polish Type Notation for Chemical Structures," J. CHEM. DOC. **4**, 186 (1964).

Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds*

E. HYDE†, F. W. MATTHEWS, LUCILLE H. THOMSON and W. J. WISWESSER††
Canadian Industries Limited, Central Research Laboratory, McMasterville, Quebec

Received July 26, 1967

A computer program is described which generates a connectivity matrix using as input an unmodified Wiswesser notation. This program records the topology of a molecule as a statement of the atoms and their connectivity. One symbol is used to represent each atom and this symbol is descriptive of the atom and its bonds. The network of a complex molecule is recorded as a series of interruptions in an assumed linear path. The application of this matrix to information handling of chemical structures is described in a subsequent paper.

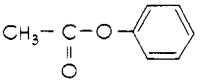
An investigation has been initiated by Imperial Chemical Industries Ltd. to establish a mechanized system for the retrieval and analysis of chemical information. An atom-by-atom connectivity system based on mathematically derived matrixes was considered, but the investigation showed this method to be too cumbersome for the proposed system. Furthermore, in many cases this method destroyed the record of the molecular arrangement of organic compounds.

Having also considered the work reported on the generation of a matrix from the I.U.P.A.C. notation, (*1*), we decided to investigate the usefulness of the Wiswesser notation for this purpose. These investigations have shown that the notation effectively describes the chemistry and the topology required for mechanized retrieval and analysis of chemical information. A computer method has been devised for producing a matrix directly from the notation. This matrix when compacted for tape storage, constitutes a record averaging 60 characters, and is in a form suitable for search and correlation purposes.

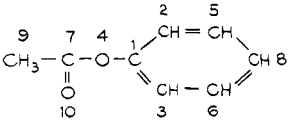
ATOM-BY-ATOM CONNECTIVITY (*2, 3*)

There are two problems associated with any atom-by-atom approach. Firstly, the vast majority of single atoms in any molecule have no descriptive value for search purposes, and secondly an atom-by-atom matrix is a bulky record which is made up of descriptions of atoms and bonds. If the next step is a mathematically generated matrix in order to ensure a canonical ordering of the atoms, then the chemically significant ordering of the atoms is destroyed in the resulting element listing.

The following example will give a clear picture of the disruption of the record of a simple molecule.



The canonical ordering of the atoms derived on a mathematical basis for ultimate magnetic tape storage is as follows:



Thus the record states:

Atom No.	Element	Bond	Connection
1	C	-	-
2	C	L	1
3	C	L	1
4	O	1	1
5	C	L	2
6	C	L	3
7	C	1	4
8	C	L	5
9	C	1	7
10	O	2	7

Ring Closure 8–6
L = Alternating bond.

*Presented before the Division of Chemical Literature, 153rd Meeting, ACS, Miami Beach, Fla., April 1967.
†Present address, Imperial Chemical Industries Limited, Pharmaceuticals Division, P.O. Box 25, Alderley Park, Macclesfield, Cheshire, England.
††Present address, U. S. Army, Fort Detrick, Frederick, Md.