Designing CIDS—The U. S. Army Chemical Information and Data System*

RUTH V. POWERS and HELEN N. HILL
Office of Engineering Research, University of Pennsylvania, Philadelphia, Pa. 19104

Received May 11, 1970

A total computerized system for the input, storage, and retrieval of chemical compounds and associated information is described. It features input via special chemical typewriters, computer analysis of the chemical record to provide a chemically verified connection table, elimination of duplicate compounds through a registry system, automatic assignment of fragment and generic search keys, and a real-time, time-shared retrieval system which permits search from remote console. The retrieval system utilizes an inverted key index and a three-level search strategy based on keys, molecular formula search, and atom-by-atom search, and provides output of structural formulas and text to a remote chemical line printer or on punched paper tape for printing on a chemical typewriter.

The U. S. Army CIDS is a total computerized system for the input, storage, and retrieval of chemical compounds and associated information. An experimental CIDS has been designed and implemented at the University of Pennsylvania. Compounds are entered via punched paper tape produced by typewriters with a chemical type font. The paper tape images are transcribed onto magnetic tape and analyzed mechanically to recognize the various types of information entered. At present these consist of the following: identification number, molecular formula(s), structure, nomenclature, and literature references and associated nonstructural keys.

The structure is converted to a connection table and verification of the structure and molecular formula is carried out to ascertain consistency and chemical veracity. The connection table is transformed to a compact linear code for storage. The structural representation and the associated text are also stored in a highly compacted form.

A registry system compares new entrants against compounds already in the file to prevent duplication and assigns a unique registry number to new compounds. This system may also be used to update or correct nomenclature and nonstructural information in previously registered compounds.

The search and retrieval strategy is based on a set of structural keys which has been designed to selectively retrieve compounds using a wide range of search criteria. These keys are automatically assigned to compounds on the basis of molecular formula and structure. Special data keys which have been typed in the input record are used to retrieve records containing references giving specific information described by the key (i.e., synthesis, infrared spectra, etc.). Registry number keys may be used to access a specific compound or a group of 1000 successive compounds in the file.

* This work was supported by the U. S. Army, Edgewood Arsenal, under contract No. DAAA15-69-C-0140. Presented before the Division of Chemical Documentation, Fifth Middle Atlantic Regional Meeting, ACS, University of Delaware, Newark, Del., April 2, 1970.

The keys are the basis for a three-level, inverted index to the file. Associated with each key in the index is the computer storage address of each compound to which the key was assigned.

The search system is primarily a special purpose, time-shared, multiterminal system. It utilizes two computers (Digital Equipment PDP-8 and IBM 7040) connected via a high-speed data channel, and communicates with teletypes, a cathode ray tube, and a remote high-speed, chemical line-printer. The PDP-8 serves as a terminal editor, multiplexor, and data concentrator, while the search system resides in the 7040.

The search consists of two phases. In the first, key-address lists corresponding to keys defined by a user's query are intersected or merged to determine the addresses of those compounds which satisfy the key requirements of the query. In the second phase, the selected compound records are accessed and may be tested further by molecular formula requirements or atom-by-atom search. At the completion of the first phase, the size of the accession list is relayed to the user, who has the option of asking that the search be terminated or continued. He also has the option of

requesting that the total stored records of all retrieved compounds be sent to the chemical line printer or punched on paper tape for printing on the chemical typewriter.

requesting only registry number output on either the teletype or chemical line printer.

requesting statistics only, in which case the number of answers after phase 2 processing (but not the answers themselves) are printed on the teletype.

CHEMTYPE SYSTEM: GENERAL DESIGN

The CHEMTYPE system of programs scans, interprets, and formats a typed chemical record which has been punched on paper tape and then transcribed onto magnetic tape.[1] It is designed to accept input from a variety of devices by having a replaceable front end which recreates the output of a particular device in the computer memory.

To date, two chemical typewriters with widely divergent characteristics have been used for input.

The Mergenthaler Chemical Typewriter has three case shifts, characters which type on the line, and one or two spaces above or below the line, and two correction modes. It punches coordinates for the location of every typed block (a block being a string of symbols typed between those control characters that change the position of the platen up, down, back a space, or back to the left margin).

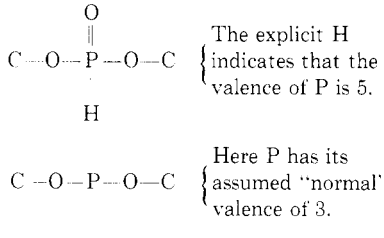
The Dura 1041 Chemical Typewriter has two cases, no coordinate facility, and no special correction mode.

CHEMTYPE programs allow the typist a maximum amount of freedom while maintaining extensive error checking and verification procedures to insure consistent and chemically valid input.

Compound records are typed as structured by a chemical editor according to CIDS structuring conventions. These conventions have been designed primarily to insure that any given substructural feature will be assigned its proper search key regardless of the type of compound in which the feature is present. Emphasis has been placed on creating a system of input and key assignment which is broadly conceived and can be extended to embrace a wide variety of compound types with little basic alteration. Careful monitoring of the structuring of input compounds has resulted in a file which is a more useful retrieval base.

Some chemical structuring conventions of primary importance to the system are:

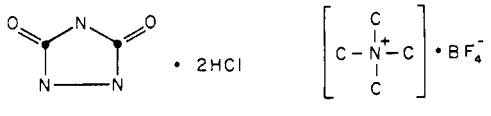
(1) Hydrogen atoms are not shown unless they are required either to specify a particular stereoisomer or to show that an atom is functioning with a higher valence than that assumed "normal" in CIDS (hydrogen never appears in the connection table). This assumed "normal" valence is the lowest common valence for each element; the only exception is carbon for which a valence of four is assumed. "Valence" here is the sum of the total number of charges plus the total number of bonds (including explicit and implicit bonds to H and counting double bonds as two, triple bonds as three, etc.).



The assumption of a "normal valence" makes it possible to scan the structural input and accumulate an "assumed hydrogen count" which can be compared with the molecular formula as a necessary part of effective chemical verification.

Deuterium and tritium atoms are displayed in the structure, but, as with hydrogen atoms, they are not listed in the CIDS node connection table. However, the attachment of deuterium or tritium to an atom is noted in the abnormality table associated with the connection table.

(2) Inorganic molecular or inorganic ionic components of an organic compound are represented by their conventional rational formulas and are not stored in the connection table, e.g.,



(3) Two molecular formulas may be present. The first, or Hill molecular formula, shows the total number of atoms of each element present in the anhydrous compound. The totals are presented with C and H first, followed by the other elements in alphabetical order according to symbol. Hydrates are portrayed in the usual way by dot-connecting the number of H_2O molecules to the anhydrous formula.

Since molecular formula keys are assigned on the basis of the anhydrous Hill formula, the exclusion of water from the summation allows clear and efficient use of those molecular formula keys which give the hydrogen and oxygen counts in the molecule. Thus, retrieval of a compound is not affected by its state of hydration. Hydrates may be retrieved separately, if desired, by specifying the parent compound and requiring the presence of water in the molecular formula.

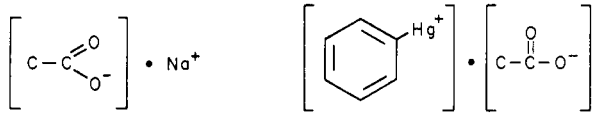
The second molecular formula is present if there is an addend other than or in addition to water. It gives the number of molecules and the formula of each molecular component, including water; formulas of organic components must be Hill ordered. The following example illustrates a case where both water and another addend are present:

$C_{36}H_{44}N_2O_{10}S \cdot 3H_2O$ (Hill formula)

$2C_{15}H_{22}NO_3 \cdot H_2SO_4 \cdot 3H_2O$ (added formula)

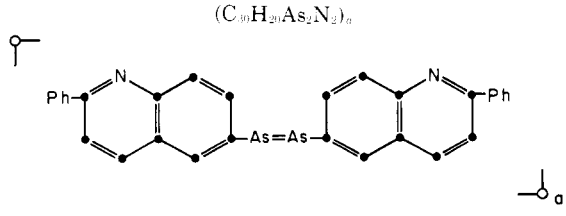
(4) The structuring of metal-containing organics varies according to the type of compound. All types in which the metal is attached directly to carbon only are structured in fully bonded fashion, e.g.,

Normal acid salts of hydrogen compounds which are named systematically as acids, and those organometallics in which the metal is unequivocally bonded directly to one or more carbon atoms and also to one or more inorganic or organic anions, are structured as split ionic formulas, e.g.,

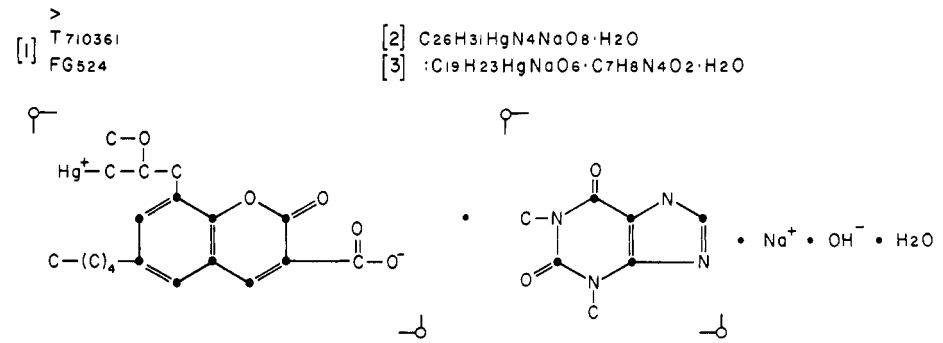


The rules for structuring and retrieving other types of metal-containing organics have been formulated and await implementation via the necessary computer programs. In general, the structuring is patterned after that conventionally employed with coordination compounds and consists of a linear portrayal in which the central atom is followed by summation formulas of the individual ligands, plus the bonded structure of either the total molecule or the individual organic ligands.

(5) Currently admissible polymers are limited to the addition type of homopolymer. The following example illustrates the form of the input:



The design of the treatment for all polymeric substances (other than any representative of natural classes) is complete and provides for search on the basis of either the various



8-(gamma-Hydroxymercuri-beta-methoxypropyl)-6-pentyl-3-coumarin carboxylic acid, sodium salt, compound with theophylline

[4] REF: SY/AD692666

**

Figure 1. Chemical record as typed for input

component parts of the substance, or the structural repeating unit, or both.

(6) The design of the structuring and retrieval techniques for polypeptides is also complete. It is patterned after the conventional biochemical technique whereby the commonly occurring peptide units and substituent groups are represented by letter abbreviations and any uncommon such units or groups are structured. The design permits search and retrieval in terms of any one, or any combination, of the usual parameters of polypeptide inquiry.

Thus CIDS has been designed so that its present capability to accommodate most of the classical types of organic compounds can be extended to include certain classes of currently inadmissible compounds, such as coordination complexes, polymers, and polypeptides, by the addition of auxiliary strategies for maximally effective search and retrieval.

INTERNAL SCANNING OF THE TYPED CHEMICAL RECORD

Figure 1 illustrates a typical candidate for admission to the CIDS file as it would be typed on a chemical typewriter. The CHEMTYPE scan allows any number of local control numbers (see [1] in the figure). The Hill molecular formula [2] is typed on the same line as one of the local control numbers. The addend molecular formula [3] (if present) is typed below the Hill formula.

The brackets in the structure are typed only as upper left and lower right bracket corners to conserve time and computer file space, but the complete brackets are created for output using the corner coordinates as guides. Carbon chains with three or more carbons may be compressed as shown and phenyl rings may be shown as Ph. A carbon atom may be represented by C or a "carbon dot." All inorganic ions are typed in a line outside the last lower right bracket corner. Nomenclature and other text may or may not be present.

Corrections and deletions may be made to any portion of the typed record or the entire record may be deleted by typing a special character at any time before the double asterisks (shown at the end of the record) are typed.

CHEMTYPE scans the record as it is recreated in a rectangular array in core, and formats for storage the local control numbers and molecular formulas. It forms a structural formula image which contains each typed character in the total molecule and its relative location in the core array. This is used to recreate the structure for output in the code of any output device specified by the retrieval system. The nomenclature and text, if present, are formatted to be used for output and for assignment of nonstructural data keys in the case of references [4].

A node connection table is formed from the structured portion of the record. All compressed chains and Ph radicals are expanded. If there is more than one structured organic fragment in a molecule, the atoms in each fragment will appear contiguously in the connection table.

An abnormality table is formed which indicates the valence of each atom and any charge or abnormal mass. In addition, the presence and number of deuteriums or tritiums attached to an atom are indicated.

Chemical verification procedures detect chemical errors as follows:

Elements in the molecular formula and connection table are checked for validity.

The valence of each atom in the structure is checked to see whether it is a valid valence for that element.

An assumed total hydrogen count is computed using the valence of each atom, the number of bonds to nonhydrogen atoms, and the total charges. Counts for all other elements in the connection table are totaled (taking into account any subscripts shown in the structure). All atoms present outside the connection table (with the exception of water) are added. These totals are compared with the totals in the Hill molecular formula.

If an addend molecular formula is present, the total of the atom counts in all addends (with the exception of water) are compared with the totals in the Hill molecular formula.

The total number of plus charges in the molecule must equal the total minus charges.

If water is present as an addend in the structure it must also be addended in the molecular formula.

If any inconsistencies are found at any stage during verification, the compound is rejected. When it has been ascertained that a compound is chemically valid, the node connection table and abnormality table are converted for storage to the Mechanical Chemical Code (MCC) representation which was developed at the University of

Pennsylvania by Lefkovitz under an NSF contract.[2] This linear representation requires an average of 2.4 characters per nonhydrogen atom (when ring and resonance indicators are included) and can be blown up into a connection table equivalent to the original for key assignment, registry, and search.

The CHEMTYPE system also assigns certain keys:

Polymer key when an indefinite molecular formula is found.

Abnormal mass key on detecting abnormal mass (including D or T) in the structure.

The resulting formated CHEMTYPE tape output is the basis for key assignment, registry, and search.

Although CHEMTYPE rejects compounds for 69 different errors, this has not alleviated the difficult problem of maintaining correct nomenclature and text input. These nonstructural fields must be checked by human editors, and even this does not insure accuracy, since the machines may type correctly but mispunch the paper tape. A great deal of effort has been made to maximize the amount of automatic error checking. However, even in the best designed system, the amount of human effort required to structure compounds, edit typed text, check rejects, file copy, etc. is gargantuan where a sizable number of compounds is concerned. To date about 70,000 compounds have been processed through CHEMTYPE resulting in two registered files of 33,400 and 14,300 compounds. These 70,000 include duplicates which were later removed by the registry process and compounds which were rejected for various reasons and had to be reentered at least once. Hardware problems (mispunching of paper tape, tearing of paper tape, errors in transcribing the paper tape images onto magnetic tape) have been found to be a fair portion of the total reject problem.

REGISTRY SYSTEM

The registry system determines if an input compound is unique or if it is the same as a previously registered compound. The procedure is primarily automatic, but calls for human participation at a critical point. Input compounds are sorted by molecular formula and compared against the master registry file which is also in molecular formula order. An atom-by-atom search is performed between an input structure and all registered structures which have the same molecular formula. If an input compound fails all atom-by-atom searches to which it is subjected, it is determined to be a new compound and is automatically assigned a registry number. If the input compound matches a registered compound, as determined by the atom-by-atom search, and the remaining input text also matches the additional stored text for the registered compound, the two records are identical and the input record is automatically discarded. If, however, connection table matches are encountered between the input compound and one or more registered compounds, but no complete record matches occur, an intellectual decision is then required to determine subsequent action. In this case, the input compound's record and those of all matches are printed for review. A chemist must decide if the input compound is to be registered or discarded, or if part of the input text is to be used to update the text of one of the registered compounds it matched. Two or more registered compounds will have equivalent connection tables in the CIDS system if the compounds are stereoisomers (since the connection table is only a two-dimensional representation). It is assumed, in this case, that there will be some difference in the nomenclature to identify the stereoisomer and to prevent the new compound from being discarded.

SEARCH KEYS

CIDS utilizes a set of chemical screens, or keys, which are automatically assigned on the basis of computer examination of the molecular formula and connection table of a compound. These screens identify functional groups, hydrocarbon radicals, and cyclic nuclei. Additional generic screens identify less common structural features and specify the elementary content of a compound based on the molecular formula. These screens are keys in the search system. They define the compositional and structural features which serve to discriminate among chemical compounds and are therefore of interest to the chemist and provide the primary basis on which queries addressed to the system are formulated.

Some of the keys are of the type which indicate whether a certain structural feature is present in a compound, such as a specific cyclic nucleus, or a functional group. The set of these keys is completely predetermined. The key assignment programs merely decide whether a given key is applicable to a compound or not. Other keys, however, are analytically assigned. For example, one program determines the skeleton molecular formulas of any cyclic nuclei that occur in a compound. A key is automatically generated to record each formula, whatever it may be. The number of keys of this type is obviously open-ended and depends only on the types of compounds that occur in the file.

One of the problems to which a great deal of effort was devoted was the selection of the rings to be identified by elementary ring population (GCN3) keys. The identification of only those rings specified by the Ring Index[5] rules was deemed insufficient because these rings conform to a prescribed orientation of a cyclic nucleus and thus frequently do not encompass other equally valid orientations. On the other hand, keys could not be assigned to all possible rings (or closed paths) of a structure, as this would result in too many "false drops" when retrieving on the basis of these rings. The following criteria for ring selection provide an adequate solution to the problem:

all rings that are members of *any* SSSR (smallest set of smallest rings) for a nucleus (all Ring Index rings are necessarily included in this group).

in addition, *all* rings of 8 or less atoms (even if they form a boundary for two smaller rings).

Computer programs were written to examine a connection table, determine the desired set of rings, assign the corresponding cyclic keys, mark the ring atoms and bonds, and identify resonant rings and mark their bonds as resonant, for use in further processing. It should be noted that all the generic cyclic nuclei keys defined below are analytically assigned.

A brief description is given below of the CIDS chemical keys. A more complete specification is given in another paper.[7]

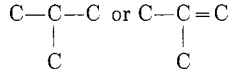
Molecular formula (MF) keys. These keys enable the user to require the presence of a specified element in the total (Hill) molecular formula. Qualitative keys are assigned for all elements except C, H, N, and O. In addition, quantitative keys are assigned for the 11 most frequently occurring elements. Keys also specify the absence of N or the absence of O.

Extracyclic (EC) keys. These keys describe the following characteristics of the acyclic portion of a structure:

EC1: number of extracyclic double bonds between carbon atoms

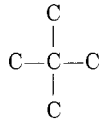
EC2: number of extracyclic triple bonds between carbon atoms

EC3: number of extracyclic



configurations regardless of other attachments

EC4: number of extracyclic



configurations regardless of other attachments

Specific cyclic nuclei (SCN) keys. These keys identify 137 cyclic nuclei which are expected to occur with relatively high frequency in queries.

The following set of generic cyclic nuclei keys which permit the specification of cyclic features without specifying the total ring system:

NCN: number of cyclic nuclei in a structure

DACN: number of direct nonhydrogen attachments to all cyclic nuclei in a structure

GCN1: number of rings in each cyclic nucleus

GCN2: numerical ring population of each cyclic nucleus

GCN3: elementary ring population of the rings of a compound (the rings so identified are described above)

GCN4: skeleton molecular formula of each cyclic nucleus

GCN5: number of double bonds in each cyclic nucleus

GCN6: relative position of heteroatoms in each one-ring nucleus containing two or more heteroatoms

Specific functional group (FG) keys. 271 functional group fragments which are expected to occur frequently in queries are identified. Except in a few special cases, no atom in these fragments can be part of a ring.

Nonspecific functional group keys. These keys specify the presence of an acyclic functional group which was not one of the 271 selected for a specific key and which contains one or more of the 11 commonly occurring heteroelements, viz., the nonmetals in Periodic Groups III through VII. The two types of these keys are:

Nonspecific diatomic functional group (ND) keys. These 66 keys are assigned in instances where any combination of two of the above 11 heteroelements are bonded together. The bond may be single or multiple and the two heteroelements may have other attachments.

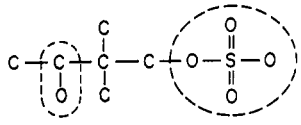
Nonspecific monatomic functional group (NM) keys. These 11 keys are assigned to all occurrences of these heteroelements which are not part of an FG key or an ND key.

Specific hydrocarbon radical (HR) keys. 61 acyclic hydrocarbon fragments are identified.

Nonspecific hydrocarbon radical (HRG) keys. 15 generic radical keys identify saturated hydrocarbon radicals containing a specified number (5 or more) of carbon atoms in any configuration.

Miscellaneous keys. These keys identify compounds containing a metal, a metal cation, an inorganic anion, or an isotope. They also identify classes of compounds, such as indefinite polymers.

Two examples given below illustrate the assignment of most types of structural keys. The keys assigned are listed below the structure.



6 molecular formula keys:

MF C 6	MF O 5
MF H 14	MF S 1
MF N 0	MF S (qualitative)

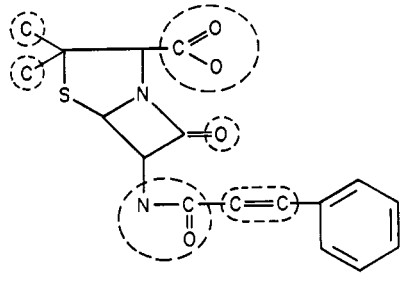
4 extracyclic carbon keys:

EC1 = 0	EC3 = 0
EC2 = 0	EC4 = 1

1 generic cyclic nuclei key:

NCN = 0

2 specific functional group keys for fragments circled in the structure above.



6 molecular formula keys:

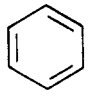
MF C 17	MF O 4
MF H 18	MF S 1
MF N 2	MF S (qualitative)

4 extracyclic carbon keys:

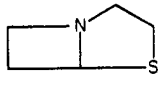
EC1 = 1	EC3 = 0
EC2 = 0	EC4 = 0

2 specific cyclic nuclei keys:

SCN48:



SCN61:



14 generic cyclic nuclei keys:

NCN = 2	GCN3 = C_3N_1
DACN = 6	GCN3 = $C_5N_1S_1$
GCN1 = 2	GCN3 = C_6
GCN1 = 1	GCN4 = $C_5N_1S_1$
GCN2 = 4,5	GCN4 = C_6
GCN2 = 6	GCN5 = 0
GCN3 = $C_3N_1S_1$	GCN5 = 3

4 functional group fragments circled in the structure above.
2 hydrocarbon radical keys (2 occurrences of the same radical) circled in the structure above.

It is obvious from the examples given that some of the keys assigned to a compound are redundant. For instance, if an SCN key exists to completely describe a cyclic nucleus, one might question the need for the GCN keys being assigned. This redundancy, however, is a deliberate part of the system design. It makes application of the system possible to queries posed in a variety of ways. The querist, of course, should use the most specific keys available which accurately define his question. For example, if an SCN key describes a cyclic nucleus in which he's interested, he would not use the GCN keys. Similarly, if an FG key specifies a desired functional group containing Br, he uses that key and does not utilize the qualitative MF key stipulating the presence of Br.

In addition to the chemical keys, CIDS also utilizes special data keys which are typed in the compound input record (Fig. 1, [4]). These keys currently identify 60 types of nonstructural information, including such categories as applications, derivatives, melting point, and suppliers. A series of mnemonic alphanumeric codes representing these keys are typed in the input record followed by a source reference which gives specific information of the type(s) specified by the key(s). A compound record may contain several of these key-reference entries. In the example given, the code SY indicates that synthesis information is provided in the reference cited. The nonstructural data keys function exactly like the structural keys during search. A query may contain any combination of structural and nonstructural keys or may contain only one type or the other.

In the current assemblage of chemical and nonstructural search keys 458,779 key assignments were made to a test file of 14,307 compounds for an average of 32 keys per compound, approximately 26 of which were of the chemical variety. That this provides a high degree of structural discrimination is evident from an early study involving a family of 78 isomers of molecular formula $C_{10}H_{16}O_2$. Using even a smaller assemblage of search keys than that now in the system, the study showed that a unique set of keys would be assigned to each of 63 of the isomers, and the remaining 15 isomers would be assigned seven different sets of keys, two or three isomers per set.

THREE LEVEL SEARCH STRATEGY

The CIDS system utilizes an inverted file with an index whose entries are the keys that have been assigned to the file. With each key is listed the random-access storage address of each compound to which the key was assigned. Through the use of the keys in a query, the number of compound records which must be accessed for retrieval is reduced.

The retrieval system employs a three level search strategy of which the first level is required and the other two levels are optional:

A key specification, the required first level, allows the system to obtain a list of all those compounds that are potential responses to a query.

A molecular formula search may be performed on the compounds that pass the key requirements.

A substructure or atom-by-atom search can be used to further reduce the number of retrievals in cases where the first two levels are not adequate. For example, this provides the ability to distinguish among positional isomers.

The keys provide the basic means for selecting structural features of interest when addressing a query to the system. Since CIDS is a real time system, it is particularly important that the keys in a query effectively reduce (through the use of the key index) the number of compound records which must be examined as possible responses to the query. In order to provide this capability, the set of CIDS keys has been made as comprehensive as possible through the use of both specific and generic keys and through the utilization of analytic keys whenever feasible. Keys assigned analytically reflect the actual chemical makeup of the file at any given time, rather that any preconceived ideas of the features likely to be present in the file.

A query may define a group of keys and then specify their combination in a parenthesized logical expression. For example, three keys may be specified, key1, key2, and key3, with the requirement that the desired compounds must satisfy the expression

key1(2key2 or not key3)

This means that the desired compounds must have one or more instances of key1 and two or more instances of key2, or one or more instances of key1 and no instances of key3. The conversion of the parenthesized expression to disjunctive form is performed automatically by the programs in the system.

In processing the key logical expression, the key index is utilized to access the list of storage addresses for each key defined in the query. These lists are combined as specified in the logical expression to produce an accession list which contains the addresses of all compounds which satisfy the key requirements of a query. If two keys are combined by an "AND" operator, the two lists are intersected. The resulting list contains all addresses which were present on both lists, as these were the compounds to which both keys were assigned. If two keys are combined by the "OR" operator, the two lists are merged. If a key is negated by the "NOT" operator, the addresses for the negated key are removed from the list corresponding to the non-negated key(s).

It must be noted here that each query must specify at least one key and that every minterm (expression between OR's) in the internally expanded logical expression must contain at least one non-negated key. The reason for this is that there must be some "positive" list of addresses from which to delete the negated addresses. There is no universal list which is utilized by the retrieval system for this purpose.

The second level of search, which is optional, allows the compound molecular formula of those compounds which responded to the key specification to be matched with requirements set forth in the query. The molecular formula statement permits imposing the following kinds of restrictions on the total (Hill) molecular formula, on one or more parts of a dot connected (addend) molecular formula, or on both:

The exact count or an upper and/or lower bound on the number of atoms of a given element, or simply the qualitative presence of an element, can be specified. For example, two to four N atoms could be required, or just the presence of N.

The element types which appear in the molecular formula may be limited to those which are specifically enumerated in the formula statement.

An algebraic relationship between the counts of any two elements may be specified.

The third level of search, which is also optional, allows the querist to search for the presence of substructures that are not completely specifiable in terms of keys. In specifying a substructure for atom-by-atom search, the following options are available:

A node, or atom, may be required to be (a) any given element type, (b) any element except C or H, (c) any element except H, or (d) any halogen. Or, a given node may be (e) required to be one of a listed group of elements or it may be (f) forbidden to be one of a group of elements, e.g.,

(e) E1 = N,P,As

(f) E1 = NOT N,P,As

In example (e) above, the specified node E1 must be either N, P, or As. In example (f), it may not be any of these.

A node may be required to be in a ring, not in a ring, or "don't care."

Abnormal mass, charge, or valence, or the number of deuterium or tritium attachments may be specified for any node. The specification may state a particular value for the abnormality, or in all cases except valence, an abnormality may be specified without stating its value.

A bond between two cited nodes may be a specific bond type, or a "don't care" bond.

Special "hanging" bonds, attached to only one cited node, specify the way the substructure is to be attached to the rest of the structure. These bonds may require attachment to C by a single bond, attachment to C or H by a single bond, or attachment to any node(s) by any type bond(s) ("don't care").

Each bond may be specified as in a ring, not in a ring, or "don't care." Each ring bond may be specified as resonant or nonresonant.

Several substructures may be combined in a logical expression with boolean logic, as in the case of keys.

If a query consists of keys alone, all compounds on the accession list are true responses to the query. The compound records must still be accessed to provide output data for each retrieval. If the query consists of keys and molecular formula statement only, all compounds on the accession list which satisfy the molecular formula requirements are true responses.

Figure 2 illustrates the use of two of the three levels of search. A query has requested all compounds which contain a benzene ring (key SCN48) and a nitro group attached to a ring (key FG154R), and whose molecular formula contains 10 to 12 C atoms, 11 to 15 H atoms, two N atoms, one S atom, two O atoms, and the number of H atoms must equal two times the number of C atoms minus eight. The intersection of the two key lists from the key index shows that compounds at locations 257 and 921 satisfy the key requirements of the query. These compound records are read into core for the second level of search—the molecular formula test. Compound 257 passes these requirements and is thus output as a response. The other compound shown fails the test. If an atom-by-atom search had been requested it would have been performed on compound 257 only.

A query can usually be answered more efficiently if it can be stated in terms of keys. Keys which are efficient in limiting the scope of a query reduce the length of the accession list of compounds which must be read into memory. Since disk accession time is relatively slow, the key requirements should be as specific as possible. The CIDS keys have been designed to provide efficient tools for specifying the majority of queries that will be posed to the system. When it is necessary to specify a structural fragment which is not one of the keys, it is worthwhile to specify key or molecular formula requirements that can act as a screen to reduce the number of atom-by-atom searches that must be made, since this is a relatively slow process. An earlier paper[6] develops more fully various techniques for efficient formulation of queries.

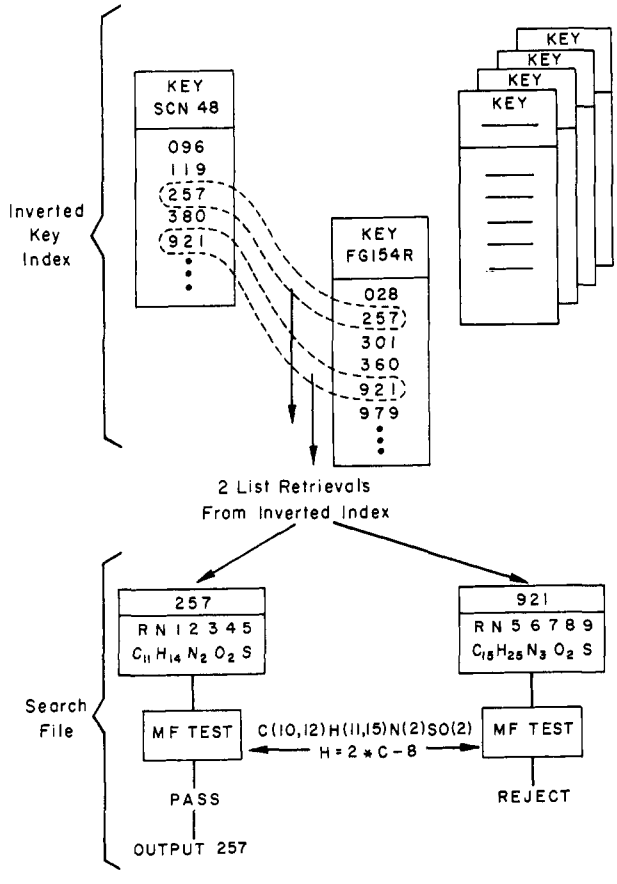


Figure 2. The inverted list search process

EQUIPMENT CONFIGURATION OF THE SEARCH SYSTEM

The primary work of the on-line, time-shared search system[8] is performed in an IBM 7040 equipped with two 7904 data channels. A 56 million character, IBM 1301 disk system, attached by data channel to the 7040, is used to store the retrieval file, two levels of the key index, programs in absolute (relocated) form, output stacks, and other pooled storage. Six tape drives are connected to the 7040 on another channel. On the same channel as the tapes, a Digital Equipment PDP-8 computer is attached for memory to memory transfers using a DEC-built interface. A direct transfer of 10-bit "sense line" messages, for which both computers are interrupt sensitive, is also provided. These "sense line" transfers

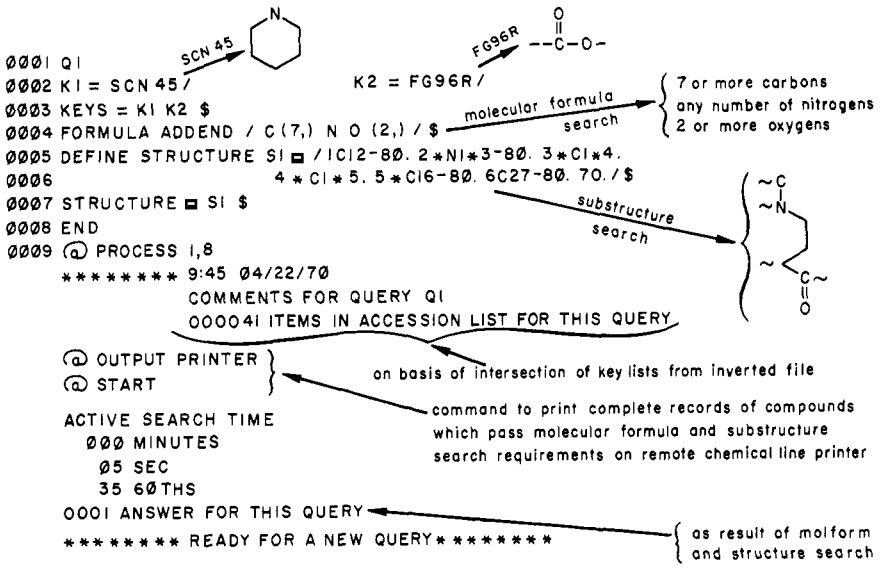


Figure 3. Query example as entered on teletype

provide a means of sending control information between 7040 and PDP-8 programs.

The PDP-8 is equipped with a small 32K disk. Other special equipment includes a series of interfaces for data lines. Low-speed data line interfaces are capable of servicing keyboard/printer devices operating at typewriter speeds. A high-speed interface (DEC 637), operating at 2000 bits per second, services the remote Data Products Chemical Line Printer and a DEC 338 Cathode Ray Tube Terminal. In this way, the PDP-8 acts as a multiplexor channel on the 7040's channel B, concentrating input data from the remote terminals and distributing output data to the proper terminal.

SEARCH SYSTEM CONSOLE OPERATION

The CIDS retrieval system allows the user to enter queries from a remote console using the CIDS retrieval language.[4] This language has been designed for flexibility in use and requires that:

A query must begin with a name of not more than 5 characters.

At least one non-negated key must be used to access the file.

A logical expression must be present to indicate how the key lists are to be combined. This may be a nested parenthesized expression which will be expanded and converted to disjunctive form internally.

An END statement must signal the end of a query.

All key definitions, structure definitions, molecular formula statements, and logical expressions may appear in any number and in any order in a query, provided that keys or structural fragments used in a logical expression are defined before the logical expression which references them occurs. More than one formula statement and/or logical expression may occur, with each occurrence replacing the previous one.

The editing facility permits the user to delete, insert, alter, and move lines in the query and print part or all of the input text on command. Any number of queries may be entered and any one of them processed. The user may stop a query that is being processed and destroy all record of it, or stop a query temporarily and then restart it. He may specify that he wants only registry numbers as output, or only the total number of actual answers, or the complete compound records. If his output is a complete record he may route it to the high speed remote Chemical Line Printer or have it punched on paper tape at his console to be printed later on a chemical typewriter. If his output is registry numbers or statistics he has the option of printing on the high speed printer or on his console.

The processing of a query is performed in two parts. First the query is checked for syntactic errors and all requirements are translated to an internal format. Key lists are intersected or merged on the basis of the key logical expression. The total number of accessions (the number of compounds that pass the key requirements) is returned to the user at his console. He now has the option of continuing further search, or, on the basis of the size of the accession list, he may decide to cancel further action on this query and try something else. At this point he also decides the type of output he wants and the device to which it will be sent.

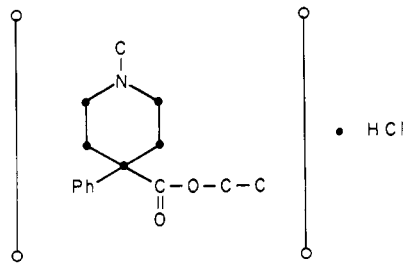
Figure 3 shows an example of a query which specifies the use of all three levels of search. The first level (on the basis of key lists) finds 41 compounds which are potential answers. The user then asked that searching continue and that output be sent to the remote chemical line printer. There was only one answer to his query on the basis of subsequent molecular formula and atom-by-atom search (Figure 4). The structure of the response is reproduced on output exactly as it was typed on input, including the use of Ph to indicate a phenyl group. It was found that when this same query was run without the atom-by-atom search requirement, there were 17 responses as a result of the molecular formula search alone.

QUERY NUMBER Q1

RN A0104602
TNT707397

$C_{15}H_{22}ClNO_2$

$C_{15}H_{21}NO_2 \cdot HCl$



Meperidine, hydrochloride
N-Methyl-4-phenyl-4-carbethoxypiperidine hydrochloride
REF: AN, AP, BS, LD, ME, MP, SO, SY / MI 680655

Figure 4. Query response on chemical line printer

SEARCH SYSTEM DESIGN

The design of the CIDS retrieval system is unique in that it involved the construction of a special-purpose, time-sharing system to perform the functions of scheduling retrievals, handling terminal input/output, and allocating the system's facilities among terminals. The most significant design objectives were

> A multiconsole operation which operates so that a user at a console is not aware of a significant slow down in response time when other consoles use the system.
>
> Minimization of motion of the access mechanism of the file storage device. In earlier efforts, the scheduling of retrieval accessions based primarily on the positioning of the file access mechansim had been shown to be beneficial. This feature, along with the fact that many searches are in progress simultaneously, significantly lowers the cost of individual retrievals.

The PDP-8 program is charged with the responsibility of catering to idiosyncrasies of terminal devices (e.g., speed, code differences, or logical record length). The scheduling algorithm in the PDP-8 provides an allocation of the PDP-8's facilities on a first-come-first-served basis for the consoles, except that priority is given to operations which transfer information through the 7040 data channel or onto the PDP-8 disk. The PDP-8 provides character-by-character input and output service for consoles, performs all editing functions and data conversion, and controls all communications between the search system in the 7040 and the remote consoles.

The search monitor in the 7040 continuously examines all job control blocks and, at any one time, may be consulting many accession lists and matching the requirements of many preprocessed queries. The activity of a search for any one console may be affected by the output accumulation for the console or by a weighting scheme intended to balance file accessions among consoles.

The system has a facility to adjust the allocation of accessions among consoles. Consoles can be made inactive for periods of time during the search process when certain conditions occur, such as:

> too much output has accumulated for the console.
>
> the console has received an inordinately large share of accessions in the given time interval.
>
> the console user has requested a suspension of execution to give him thinking time.

Execution can be restarted by program or by user request if conditions warrant. An efficient allocation algorithm can provide a number of users with almost uninterrupted search system response.

ACKNOWLEDGMENT

The authors wish to acknowledge the chemical guidance of the University of Pennsylvania Project CIDS Director, Clarence T. Van Meter, and the technical contributions of David Lefkovitz and the principal analysts and programmers—Paul Weinberg, Bonnie Sherr, Richard Haber, and Mary Milne.

LITERATURE CITED

(1) Lefkovitz, David, R. V. Powers, and H. Hill, "Computer Programming for an Experimental Chemical Information and Data System," CIDS No. 5 Status Report, University of Pennsylvania, Philadelphia, Pa., June 1968, AD-838725.

(2) Lefkovitz, David, "A Chemical Notation and Code for Computer Manipulation," J. CHEM. DOC. 7, 186 (1967).

(3) Lefkovitz, David, and R. V. Powers, "A List Structured Chemical Information Retrieval System," *Proceedings of the Third National Colloquium on Information Retrieval*, pp. 109–129, May 1966.

(4) Milne, Mary and P. R. Weinberg, "Query Formulation and Encoding," CIDS No. 7 Status Report, University of Pennsylvania, Philadelphia, Pa., November 1969, AD-869857.

(5) Patterson, A. M., L. T. Capell, and D. F. Walker, "The Ring Index," 2nd ed., ACS, Washington, D.C., November 1969, and Supplements.

(6) Powers, R. V., "Querying a Real Time Chemical Information Retrieval System," Master's Thesis, Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa., May 1969.

(7) Van Meter, C. T., E. N. Goldschmidt, and Mary Milne, "Handbook of CIDS Chemical Search Components," CIDS No. 6 Status Report, University of Pennsylvania, Philadelphia, Pa., December 1968, AD-851126.

(8) Weinberg, P. R., "A Time Sharing Chemical Information Retrieval System," PhD dissertation, University of Pennsylvania, Philadelphia, Pa., May 1969.