computer series. 16

edited by
JOHN W. MOORE
Eastern Michigan University, Ypsilanti, MI 48197

# Introduction to Chemical Information Storage and Retrieval

**Peter F. Rusch**
Lockheed Information Systems, 3460 Hillview Ave., Palo Alto, CA 94304

Other articles in this series have undertaken an explanation of what digital computers are, how they operate, communicate and manipulate various forms of *data*. This article will introduce the topic of chemical information handling using computers. As chemists, we have a long-established heritage of chemical information because previous generations of chemists have recognized the need for recording and conveying chemical information. We rely upon this store of information, use it, interpret it, benefit from it and contribute to it. Ultimately, the cycle repeats and each cycle ends with an even larger body of chemical information to enter into the next cycle. Thus, this article shall be concerned with chemical information as embodied by the chemical literature and its content.

The content of the chemical literature as data lends itself to a variety of computer-based manipulations. Collection, formatting, distribution, storage, search, and retrieval of the chemical literature are all aided by computer processing. Sections of this article will cover the following topics in computer based chemical information processing: codes, the representations of the information to be processed; operations, the computer-based manipulations of the coded information; file structures, the ways in which information may be stored for processing; databases, the sources of information for processing by search for retrieval; telecommunications, the means used to interface between a computer and the outside world.

To establish an initial frame-of-reference we must consider the properties of a computer to understand its capabilities. Previous articles in this series have been for that exact purpose, and they are recommended to the reader (*1*).

Digital computers are perceived to be prodigious arithmetic machines or calculators. In fundamental terms this is quite correct. How, then, does a digital computer apply fundamentally arithmetic tasks to textual chemical information? The information is transformed into digital quantities upon which a computer can operate.

## Fundamentals

Much of the modern, technological world uses a relatively small set of characters to transfer information. Twenty-six (or fewer) Roman letters, ten Arabic numerals, and a handful of punctuation symbols are all that is required.

Computers and peripheral devices (*1*) work with binary representations grouping binary digits (or bits) into larger, more meaningful quantities. There are many code systems for representing the character set composed of Roman letters, Arabic numerals, and punctuation. These code systems are usually referred to by their acronyms such as ASCII (pronounced as-key) or EBCDIC (pronounced eb-see-dik) or any one of many others. Code systems are created by assigning a numeric value to each character to be represented. The numeric values assigned for many characters in ASCII (American Standard Code for Information Interchange) and EBCDIC (Extended Binary Coded Decimal Interchange Code) are in Table 1.

It can be inferred from Table 1 that the assignment of numeric codes for the Roman letters A through Z is contiguous in the ASCII character set and not in EBCDIC. Other differences in definition are also noted: the numerals 0 through 9 have values which precede the letters in ASCII and follow the letters in EBCDIC; lower-case letters have values which are greater than upper-case in ASCII and less than upper-case in EBCDIC.

With codes such as ASCII or EBCDIC it is possible to encode alphanumeric text such as this article itself into a machine (computer) readable form.

Code systems such as ASCII or EBCDIC are termed "instantaneous" (*2*). That is, any given, fixed length segment (for example, 8 bits) uniquely represents a single member of the character set encoded. A sequence of bits is "quantized" into discrete packets of bits each commonly called a "byte." A byte may have different definitions depending on the hardware for which it is defined but in most cases consists of eight bits. Each byte (sometimes called a "character" for obvious reasons) may be interpretted independent of other bytes in the sequence.[1]

[1] Coding schemes which are not instantaneous play an important role in information processing in general. Their description is outside the scope of this article and the interested reader is referred to the quite readable monograph in reference 2.

The Computer Series attempts to delineate the current state of the art of computer usage in chemical education while accommodating the needs and background of readers who have little or no computer expertise. In addition to full-length articles, short descriptions of specific applications of computers in classrooms or laboratories are published as "Bits and Pieces," so that readers who have appropriate computer systems can obtain and use the programs described. The editor's intention is that the Computer Series be understandable for beginners but at the same time interesting for experts.

**John W. Moore** received his AB from Franklin and Marshall College and his PhD from Northwestern University, concentrating in physical inorganic chemistry. Following NSF-sponsored postdoctoral work at the University of Copenhagen he has taught at Indiana University and Eastern Michigan University. He is co-author of *"Environmental Chemistry," "Chemistry,"* and the soon-to-appear third edition of *"Kinetics and Mechanism,"* as well as numerous journal articles. In 1977 he received a Distinguished Faculty Award from Eastern Michigan University, and in 1979 he was appointed by Governor William G. Milliken to the Michigan Environmental Review Board. Dr. Moore has produced numerous audio-visual teaching aids including films, computer-generated overhead transparencies, and other computer graphics. Currently his spare time is devoted to several PETs, an S-100 microcomputer system, and a shiny new Apple.

**Table 1. Partial List of Symbols and Coded Numeric Values in ASCII and EBCDIC**

| Symbol | ASCII | EBCDIC | Symbol | ASCII | EBCDIC |
|---|---|---|---|---|---|
| A | 65 | 193 | a | 97 | 129 |
| B | 66 | 194 | b | 98 | 130 |
| C | 67 | 195 | c | 99 | 131 |
| D | 68 | 196 | d | 100 | 132 |
| E | 69 | 197 | e | 101 | 133 |
| F | 70 | 198 | f | 102 | 134 |
| G | 71 | 199 | g | 103 | 135 |
| H | 72 | 200 | h | 104 | 136 |
| I | 73 | 201 | i | 105 | 137 |
| J | 74 | 209 | j | 106 | 145 |
| K | 75 | 210 | k | 107 | 146 |
| L | 76 | 211 | l | 108 | 147 |
| M | 77 | 212 | m | 109 | 148 |
| N | 78 | 213 | n | 110 | 149 |
| O | 79 | 214 | o | 111 | 150 |
| P | 80 | 215 | p | 112 | 151 |
| Q | 81 | 216 | q | 113 | 152 |
| R | 82 | 217 | r | 114 | 153 |
| S | 83 | 226 | s | 115 | 162 |
| T | 84 | 227 | t | 116 | 163 |
| U | 85 | 228 | u | 117 | 164 |
| V | 86 | 229 | v | 118 | 165 |
| W | 87 | 230 | w | 119 | 166 |
| X | 88 | 231 | x | 120 | 167 |
| Y | 89 | 232 | y | 121 | 168 |
| Z | 90 | 233 | z | 122 | 169 |
| 0 | 48 | 240 | (SPACE) | 32 | 64 |
| 1 | 49 | 241 | . | 46 | 75 |
| 2 | 50 | 242 | , | 44 | 107 |
| 3 | 51 | 243 | ( | 40 | 77 |
| 4 | 52 | 244 | ) | 41 | 93 |
| 5 | 53 | 245 | - | 45 | 96 |
| 6 | 54 | 246 | $ | 36 | 91 |
| 7 | 55 | 247 | ? | 63 | 111 |
| 8 | 56 | 248 | | | |
| 9 | 57 | 249 | | | |

**Table 2. List of Relational Operators**

| | |
|---|---|
| $<$ | Less Than |
| $=$ | Equal to |
| $>$ | Greater Than |
| $\leq$ | Less Than or Equal To |
| $\geq$ | Greater Than or Equal To |
| $\neg$ | Not Equal To |

Punctuation symbols vary from one code system to another. Since computers operate literally on any data presented, it is necessary to have a unique representation for the space or blank symbol normally inferred in alphanumeric (alphabetic and numeric) sequences. This is similar to the need for a zero digit in a number system. In both ASCII and EBCDIC the space has a value less than that for any alphanumeric character; in other codes it has a value greater than that for the alphanumeric characters. The actual value of the space character within a code system can create differences in the presentation of chemical information.

Non-Roman characters can, of course, be similarly encoded. In Japan, at least one coding system has been developed using a sixteen-bit representation to encode approximately 32,000 Kanji characters (*3*). The Japanese Patent Office transfers information using a coded representation of the Kana character set, which is a relatively small phonetic alphabet.

## Operations

Computers are fundamentally capable of arithmetic operations. Thus, one could program a computer, for example, to add the ASCII coded values of A and Q to produce the binary result 10010010. Such an operation is perfectly legitimate although the result may not have a clear meaning. Of greater interest are the logical and relational operations, permitting comparisons such as "greater than," "less than or equal to," etc. Most of these operations are directly available in programming languages and correspond more or less directly to hardware functions of the computer itself. To illustrate the variety of such operations some of them are shown in Table 2. By using relational operations a computer can be programmed to manipulate alphabetic, numeric, and punctuation characters (including space) represented internally in a computer as coded, numeric sequences.

One widely used application is sorting. The ASCII representations for "dog" and "cat" are DOG = (0100 0100 0100 1111 0100 0111); CAT = (0100 0011 0100 0001 0101 0100). A simple program can be executed to arrange the data in ascending order producing an "alphabetical" listing:

CAT
DOG

This result depends directly upon a computer's ability to determine through a program that one coded sequence is greater than another. Sorting results from a comparison of the coded character strings from left to right one character at a time.

Longer words or even phrases can be similarly represented and sorted but an additional complication develops; the data may be of varying length. For example, the following character strings are of different lengths.

ELECTRON
ELECTRONIC
ELECTRON SPIN RESONANCE

What rules should be applied to sort such a list? Clearly, length becomes a part of the criteria so that the rule "nothing before something" can be applied. Thus, ELECTRON followed by nothing is sorted before ELECTRON followed by SPIN RESONANCE leading to the order:

ELECTRON
ELECTRON SPIN RESONANCE
ELECTRONIC

The term ELECTRONIC follows the multi-word term because the coded value of space (or blank) is less than that for any alphabetic or numeric character. By assigning a different, larger coded value to the space character the order of the last two terms would be reversed.

A practical example of the result of sorting a great many varying-length character strings is found in the alphabetical organization of the General Subject Headings in Chemical Abstracts (CA) Volume and Collective Indexes. Currently, approximately 40,000 General Subject Headings appear in boldface type in a CA Volume General Subject Index. The index headings for a document are assigned by document analysts at Chemical Abstracts Service (CAS). During a six-month volume period the index headings are encoded in computer-readable index entries which may be formatted for printing and also released in computer-readable form.

CA General Subject Index Headings are much like normal English text; with few exceptions they can be represented by a small character set. By contrast, nomenclature for chemical substances generally requires a larger character set including Roman letters, Greek letters, Arabic numerals, Roman numerals and special symbols including superscript and subscript notation. Some examples are

$\beta$, $\epsilon$-Caroten-3′-one
Tricyclo[2.2.1.$0^{2,6}$] heptane
2*H*-Azepin-2-one

The larger the set of symbols to be represented by codes, the larger values of the numeric codes assigned. Using an eight-bit byte it is possible to assign codes to up to 256 different symbols. In order to handle chemical nomenclature Chemical Abstracts Service has established just such a character set for internal use. One feature of this character set is the use of an

additional eight-bit byte for each symbol to represent the font and position of each symbol (*5*). Thus, there is a code for upper or lower case, Roman or Italic, superscript or subscript, etc. For external use this detailed character set is mapped into the less-detailed ASCII character set.

No discussion of sorting techniques would be complete without mentioning "sort keys." For most sorting applications the character data to be sorted is its own key. Thus, the numeric coded values for DOG and CAT can be used directly to produce the desired alphabetical order: CAT followed by DOG. In the sorting of systematic chemical nomenclature a sort key is used as a surrogate for the character data.

The need for a sort key is amply illustrated by the chemical substance names:

1-butene
2-butene
1-buten-2-ol
2-buten-1-ol

These chemical substance names as listed are in a chemically related order; the hydrocarbons are grouped and are followed by the derivatives. Using these character data as their own sort keys would result in a less meaningful sort order

1-buten-2-ol
1-butene
2-buten-1-ol
2-butene

In this order all items beginning with the numeral 1 are grouped and all items beginning with numeral 2 are grouped. The embedded hyphen has a coded value less than any alphanumeric character. To produce the desirable order from a random collection of input character data it is necessary to generate and use a sort key.

For the simple example shown here the derivation of sort keys is quite straightforward. Sort keys are composed of all alphabetic characters followed by numerals and ignoring punctuation such as hyphen. This leads to the following sort keys in ascending order

butene 1 (from 1-butene)
butene 2 (from 2-butene)
butenol 12 (from 1-buten-2-ol)
butenol 21 (from 2-buten-1-ol)

During the sorting process the original character data and the sort keys are connected. When the sort is complete only the original character data is output in the order determined by the sort keys. Extension of this simple sort key generation leads to the observed order of chemical substance names in the CAS Volume Chemical Substance Index. Details of this process may be found in the CA Chemical Substance Index Introduction.

Once procedures have been established to successfully sort character data, the manipulation of merging is easily added to the repertoire. The procedure of comparing character data proceeds with two or more incoming sets of items sorted in the same way (ascending or descending) as the desired merge. In general, a sorting process is more demanding of computer resources than is a merging process. Thus, in practice, a large sorting process is usually broken into several smaller processess the results of which are merged to form the final sorted data.

As one might expect the process of sorting data of various types has occupied the interests and talents of computer scientists. Numerous computer-based algorithms have been devised to efficiently sort varying amounts and types of data. One source of this information is found in Baase (*4*).

## Information Storage and Retrieval

Sorting and merging depend on a computer's ability to compare digital information and to proceed depending on the result. Another application of these processes leads to the fundamentals of information search and retrieval. Given a body of digitized character data a computer can be programmed to search the data and retrieve (output) all or part of it. This notion of search and retrieval requires three parts: a source of information (databases); a method of storage (file structure); a method of search.

### *Databases*

A database can be operationally defined as a collection of information. Most frequently a database is composed of some surrogate information for actual items of interest. A telephone book is a database composed of names, addresses and telephone numbers, not the people who are identified. An abstracting and indexing database, such as *Chemical Abstracts,* contains author names, journal titles, index entries (among other things) and not the articles, reports and patents that constitute the chemical literature. Such databases are often referred to as secondary information sources. What is chosen to document an item in a database is a matter of policy determined by the builders of the database. Should the telephone book include postal code as part of the address? Should the molecular formula for a chemical substance used as a solvent be entered? (The answer to the last and to all other chemical information questions is: "It depends.")

From a user's point of view one characteristic of any chemical literature database must be firmly in mind. Databases are descriptive not interpretive. A chemical literature database is historical and records descriptions of documents for purposes of access. Interpretation of chemical information accessed from the database is the realm of the chemist and must be done in the knowledge of the broader context of chemistry as a science.

Any database implies some form of organization and medium for preparation, storage and distribution. These aspects are important in a practical sense because they can determine the uses of a database but they are independent of the database content. For example, Chemical Abstracts information is available in several media (printed or computer-readable) and several organizations (author index, molecular formula index, etc.). These only affect the use of the Chemical Abstracts database, not its definition.

In chemical information there are two broad categories of items that can be documented to create a database; the chemical literature and chemical substances. Documentation of the chemical literature results in bibliographic, subject and chemical substance information. Among bibliographic information are author name, journal name, volume, issue, pages, and publication date. A title is classically considered to be bibliographic but also acts as an important part of the subject information that also includes keywords and general subject terms. Chemical substances reported in the chemical literature can be identified by one or more names, molecular formula, structure or physical properties collected in a database.

Users of computer-readable products must function the actual database content and the end user. With printed versions of a database no such interface is necessary but the access is limited. One obvious limit is the number of printed copies, each of which can be accessed by only one user at any time. Secondly the access is limited to single, predetermined access points. A computer-readable file can be made available to many users simultaneously and each user may have multiple access points some of which are not provided for in printed copy.

### *File Structures*

Computer-based chemical information systems imply some file structure or organization of the information. Two are of interest to us: sequential and inverted. One of the most confusing aspects of computer-based activities for the neophyte is to keep in mind that although a file structure may mask or enhance the access of the database content, it is separate from it.

A brief introduction to sequential and inverted file structures can be achieved by considering a book, with content organized sequentially by paragraphs and pages. Using this structure one can access the information sequentially from introduction to conclusion. Analogy exists with computer-readable chemical information which can be accessed from first record to last with stops along the way depending on some search criteria. Such file structures exist and are the basis for some chemical information systems, because they are easily processed and stored relatively inexpensively.

The other file structure used by chemical information search systems is the inverted file. By analogy with a book, the table of contents and the index can be used to access the sequential information first by subject, then by page number, whereas, sequential access is by page number then subject. Thus, the index of a book is an example of an inverted file; the access is the inverse of that available sequentially. This notion of "inversion" (or creating indexes) can be carried to greater levels of sophistication that are easily and quickly created and accessed by computers. Inverted files are relatively more expensive because of their size and the hardware needed to store them. Size is an important aspect since the ultimate inverted file would provide initial access to each and every word in the sequential file. A thorough introduction to many useful file structures can be found in the paper by Dessy and Starling (*6*).

### *Chemical Search Systems*

Chemical information systems provide a means for accessing file content derived from a database and made searchable using some file structure. Computer-based search systems are usually of two types broadly referred to as batch and online. As with other aspects of this topic some operational definitions from the user point of view will be useful. Batch search systems are characterized by the need to define *a priori* the search activity. A search query is formulated and submitted. Results are examined and, in many cases, a reformulation of the query is desirable or necessary. The batch process proceeds from beginning to end without intervention of the searcher. Any adjustment to the search must be in the form of repetitive searches with modifications.

The online search process more closely parallels that which we do manually using printed information sources. By contrast to the batch search the online search encourages and enables the searcher's intervention during the search process. A search query is formulated and executed using a computer-connected terminal. As results are derived they are reviewed and the search continues from beginning to end under the control and review of the searcher.

Both batch and online search systems permit use of Boolean[2] logic expressions that permit communication of the searcher's intent in unambiguous ways and they parallel thought processes used in conducting a search. As is often the case with any computer-based processing, the potential exists for rapidly generating an enormous amount of unintended output. This is particularly true in chemical information searching because the search terms, the words used, have inherent ambiguities that lead to unintended results. The greatest weapon that any searcher has to combat this problem is a good knowledge of the subject matter and the databases used to derive the search files.

Let us examine a simple search to see the use and result of the Boolean operations: AND, OR, NOT. A search for "sodium" AND "chloride" in a database of chemical literature could retrieve references to "sodium chloride" or "sodium and barium chloride" or "sodium fluoride and potassium chloride." The Boolean AND merely requires co-occurence of the search terms; it does not imply adjacency. Specific search systems have search techniques that assist the searcher with this problem of adjacency. The Boolean OR is usually inclusive in a chemical search system, thus, "sodium" OR "chloride" means either "sodium" or "chloride" or both. The Boolean NOT is simply that, a negation. A search for "sodium" NOT "chloride" will include any item identified by "sodium" except those that also have the term "chloride." Finally, the more restrictive exclusive OR can be produced by the nested combination of operations (A OR B NOT (A AND B)).

[2] Named after the 18th century mathematician George Boole.

### *Telecommunications*

For most computer-based online chemical information systems the user is remote from the computer site. In this context remote may mean anything from the next room to another continent. Thus, the communication between the user and the computer is another important aspect of any chemical information system.

A common configuration for this communication consists of a terminal at the user site and some device (an acoustic coupler or modem) to connect the terminal to a "voice grade" telephone line. At the computer site there is the necessary equipment to take the communicated signals into the computer system hardware itself.

This telecommunication system handles the problem of transmitting essentially character data between the computer system and user terminal, by sending and receiving a sequence of binary digits that are interpreted according to some coding scheme such as ASCII (Table 1). The actual bits are represented by state changes of some measurable quantity such as current, voltage, frequency or phase. For voice grade telephone lines the method of choice is frequency in the range of audible tones. Using two tones it is possible to assign the standard of a low tone as a zero bit and high tone as a one bit. Since any two consecutive bits may have the same value (the ASCII code for letter A is 01000011) a sampling interval must be established for each bit transmitted and received.

## Chemical Information

### *Databases*

Chemical information exists in great variety. In chemical research and teaching, databases could be created from: a set of student records; a collection of examination question and answers; results of experimental work; a subset of the chemical literature relevant to teaching or research interests; lists of chemical substances used or synthesized. Appropriate databases may exist in these areas or they may be created by their users.

Creation of any database is not a simple task; there are many long-range decisions to be made. Data collected must be complete but the collection process must not be so difficult as to render the database intractable or untimely. Updating by addition, deletion, or replacement must be permitted and the information stored must be complete to enable future use in ways not anticipated at the outset. Local databases of the types mentioned above are characteristically small and tailored to the needs of a specialized user community. Many time-sharing systems and personal computer operating systems have file creation and manipulation processes that facilitate local database creation, updating, search and retrieval.

Large databases with more global coverage and use usually contain hundreds of thousands of items and many are in excess of a million. Several important chemical literature and chemical substance databases are listed with brief descriptions of coverage in Table 3. Surely one of the most important databases in chemistry is that produced by Chemical Abstracts Service (CAS). It is by no means the only database important to chemistry but it is, by far, the most global in its coverage of the chemical literature and chemical substances. The familiar printed form of this database is distributed in several forms to provide access by a variety of individual search terms or access points. The computer-readable form of this database

is the basis for many search and retrieval operations in industry, academia and the commercial sector.

Growth of online commercial search services (7) has increased the awareness and use of computer-based chemical information search and retrieval. Compatible terminals needed to access these services are frequently already available at many sites and commercial telecommunications networks are widely available. Cost for these services is in direct proportion to use. There is no need to learn a programming language or to develop search and retrieval programs. Each system has its own "query language" designed to help users formulate queries in a natural way. Using these services chemical information can be quickly and inexpensively accessed by chemists who know best their own information needs. Student use of these systems can be incorporated into current course work or research activity (8).

### Subject Searching

At the risk of oversimplification, chemical subject searching may be viewed as using descriptive words to access documents in a database of the chemical literature. For any subject area there are words that occur frequently in conversation and written works that describe research in that subject area. Words used in communication with colleagues form an uncontrolled vocabulary that may have variant, synonymous forms to describe the same phenomenon. Most chemical literature databases include sources of uncontrolled vocabulary such as titles or keywords. In contrast, controlled vocabulary can be intellectually added as part of the indexing process and consists of preferred terminology with preferred spellings used for particular purpose and usually with intellectual relationships among the preferred terms. Chemical Abstracts (CA) General Subject Index Headings are a controlled vocabulary. For example, documents about the phenomenon commonly referred to as the "solvated electron" and its variant forms "hydrated electron," "ammoniated electron," etc. are indexed using the single CA General Subject Index Heading, "Polaron in Solution." Chemical literature databases from Chemical Abstracts Service contain both controlled and uncontrolled vocabulary offering a wide variety of terminology for access.

Subjects can also be represented by the names of authors or represented by the names of authors or journals publishing in a particular subject area. Thus, the bibliographic data can be searched in most chemical literature databases to provide subject access. A few databases also offer search of cited references to access documents of similar subject interest.

### Substance Searching

Descriptions of chemical substances permit several methods of full structure and substructure search. Nomenclature can be searched as complete names representing full structures or for significant portions representing substructure. Techniques of substructure searching via nomenclature require databases of rigorously controlled, systematic nomenclature (9). Molecular formulae permit substructure searching in a rudimentary way and full structure searching if the problem of isomerism can be suitably handled using other search terms such as nomenclature. Although lacking the detail desired for many substructure searches a combination of nomenclature and molecular formula data can provide practical substructure searching for large numbers of chemical substances.

A majority of known chemical substances contain at least one ring, thus, the number, size and elemental composition of rings are used to document chemical substances and provide access for searching. Ring data are commonly derived for unique ring skeletons independent of substitution on the rings or multiple occurences of the particular ring system in a chemical substance. Thus, ring data describe substructure and for uncommon ring systems may be used independently to great advantage. For more common ring systems, such as phenyl, ring data are best used in combination with nomenclature and molecular formula information.

**Table 3. Partial List of Chemical Information Databases Available Commercially**

| Name | Producer | Coverage |
|---|---|---|
| CA Search | Chemical Abstracts Service<br>Columbus, OH | Chemical Abstracts |
| BIOSIS | Biosciences Information Service<br>Philadelphia, PA | Biological Abstracts |
| NTIS | National Technical Information Service<br>Springfield, VA | Reports of Government-Sponsored Research |
| SCISEARCH | Institute for Scientific Information<br>Philadelphia, PA | Science Citation Index |
| SPIN | American Institute of Physics<br>New York, NY | selected physics journals |
| CAB | Commonwealth Agricultural Bureaux<br>Farnham Royal, Slough, England | 26 abstracting and indexing journals of CAB |
| IPA | American Society of Hospital Pharmacists<br>Washington, DC | International Pharmaceutical Abstracts |
| EM | Excerpta Medica<br>Amsterdam, The Netherlands | 43 abstracting and 2 indexing journals of EM |

Ring data are one method for partially describing the topology of a chemical substance. Other terms that describe structural features can be generated for particular atom-bond combinations such as functional groups and chains. Complete topological descriptions can be generated using either linear notations or graphical representations that describe all the atoms and bonds in a chemical substance. Working with such descriptions it is possible to execute substructure searches for highly localized features that cannot be uniquely identified by nomenclature. Nomenclature searching, however, will continue to be significant because all chemical substances have one or more names but not all chemical substances have either a molecular formula or a known structure.

Chemical substances may be searched in the chemical literature using nomenclature, molecular formula, ring data, linear notations and a variety of other search terms. A developing trend is to use the unique CAS Registry Number to represent specific chemical substances in the chemical literature. The recognized separation between chemical substances and the chemical literature has resulted in separate databases being created to satisfy separate search needs with the CAS Registry Number serving as a link between them.

### Sample Search

The following search will illustrate some of the possibilities for both chemical substance and chemical literature searching. The purpose of this sample search is to obtain chemical literature references to studies of toxicity of the class of chemical substances shown in Figure 1. The search strategy will be di-

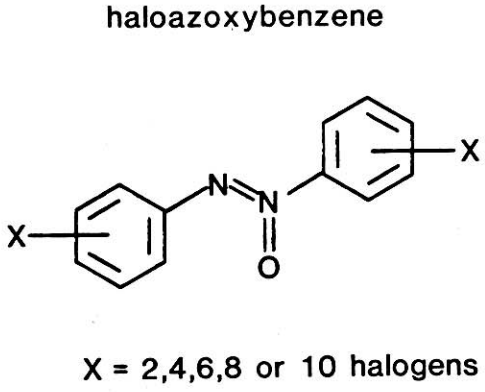


Figure 1. Class of chemical substances for which chemical literature references on toxicity are to be located.

CAS REGISTRY NUMBER: 495-48-7
FORMULA: C12H10N2O
ANALYSIS OF RINGS: C6
NUMBER OF RINGS: 1
SIZE OF RINGS: 6
CA NAME(s): HP=Azoxybenzene (8CI)
HP=Diazene (9CI),SB=diphenyl-,NM=1-oxide
SYNONYMS: Azobenzene,oxide;Azoxybenzide

Figure 2. Chemical substance database record for azoxybenzene.

rected first at chemical substance databases to derive CAS Registry Numbers for specific chemical substances of the class. Secondly, the search will be directed at chemical literature databases to coordinate chemical subjects and chemical substance search terms.

"Azoxybenzene" is commonly used to describe the unhalogenated ring assembly shown in Figure 1. From a complete name match search it is possible to derive a record (Figure 2) that describes the substance in an online chemical substance database (*7b*). Although "Azoxybenzene" was the systematic name for this chemical substance in the CA Eighth Collective Index (8CI) period its current systematic CA Index Name is "Diazene,diphenyl-,1-oxide" which completely describes the unhalogenated substructure. Expanding the search for other members of the class of substances requires use of less restrictive terminology and non-nomenclature search terms as follows. From nomenclature, the search term "Diazene" is used to describe the central chain.

Molecular formulae are searched for those chemical substances that contain twelve carbon atoms, and two nitrogens and one oxygen and either two, four, six or eight hydrogen atoms or ten fluorine, chlorine, bromine or iodine atoms. Finally, a specially derived search key is used to retrieve chemical substances containing only carbon, hydrogen and atoms from Groups VA, VIA and VIIA. The three groups of search terms are combined with a Boolean AND leading to nine chemical substances in the class from a database of more than 750,000 chemical substances. CA Index Names, CAS Registry Numbers and molecular formulae for these nine chemical substances are shown in Figure 3. This same search strategy applied to another database of more than 1,250,000 chemical substances not contained in the first provided eighteen additional chemical substances in the class.

With the set of CAS Registry Numbers representing specific chemical substances it is possible to search appropriate databases of the chemical literature for documents referencing one or more of the substances and the desired subject of "toxicity." Some of the variant forms of "toxicity" such as "toxic," "toxicology" and "toxicological" are easily retrieved using a search term with right truncation indicating specified leading characters followed by any other characters, for example, "toxic?". Examples of synonymous subject search terms related to toxicity are

Poisons
Toxins
Industrial Hygiene
Pollution

These related terms can be derived from authoritiative sources such as the CA Index Guide and used as appropriate to the goals of the search. Examples of document references retrieved using the combined search strategy are given in the references (*10, 11*).

## Conclusion

Chemical information in the form of descriptions of the chemical literature and chemical substances can be encoded as digital representations. These representations can be manipulated by computers to assist with collection, formatting and distribution. Computer-based search techniques offering advantages of speed and multiple access can be used by chemists to search databases that supplement traditional printed information products.

Diazene,bis(4-bromophenyl)-,1-oxide
1215-42-5 $C_{12}H_8Br_2N_2O$

Diazene,bis(2,5-dichlorophenyl)-,1-oxide
961-28-4 $C_{12}H_6Cl_4N_2O$

Diazene,bis(4-flourophenyl)-,1-oxide
51789-07-2 $C_{12}H_8F_2N_2O$

Diazene,bis(4-fluorophenyl)-,1-oxide,(E)-
32213-80-2 $C_{12}H_8F_2N_2O$

Diazene,bis(3,4-dichlorophenyl)-,1-oxide
21232-47-3 $C_{12}H_6Cl_4N_2O$

Diazene,bis(2-chlorophenyl)-,1-oxide
13556-84-8 $C_{12}H_8Cl_2N_2O$

Diazene,bis(pentafluorophenyl)-,1-oxide
1800-29-9 $C_{12}F_{10}N_2O$

Diazene,bis(4-chlorophenyl)-,1-oxide
614-26-6 $C_{12}H_8Cl_2N_2O$

Diazene,bis(3-chlorophenyl)-,1-oxide
139-24-2 $C_{12}H_8Cl_2N_2O$

Figure 3. Nine specific chemical substances in the class of substances shown in Figure 1.


## Acknowledgment

The author wishes to thank Ms. M. V. Reitano of the Health, Safety and Human Factors Laboratory, Eastman Kodak Company, Rochester, New York, for suggesting the sample search and for providing some of the search terms.


## Literature Cited


(1) Moore, John W., and Collins, Ronald W., "Computer Series, 1: A Tool-Not a Gimmick," J. CHEM. EDUC., *This Journal* **56**(3), 140–147 (1979).
(2) Aronson, Jules, "Data Compression-A Comparison of Methods," U.S. Dept. of Commerce, National Bureau of Standards, NBS Special Publications 500-12, June 1977 (Supt. of Docs. No.: C13-10:500-12).
(3) Nakayama, K., Ikeda, K., Sakaguchi, A., Oikawa, A., Ebihara Y., and Veda, S., "Online Information Retrieval at the University of Tsukuba," Proceedings of the Third International Online Information Meeting, London, **1979**.
(4) Baase, S., "Computer Algorithms: Introduction to Design and Analysis," Addison-Wesley Publishing Co., Reading, Mass., **1978**.
(5) Rule, D. F., "Character Sets," *J. Chem. Info. Comp. Sci.*, **15**[1], 31 (1975).
(6) Dessy, R. E., and Starling, M. K., "Information Retrieval and Laboratory Data Management," *Anal. Chem.*, **51**[9], 924A–948A (1979).
(7) (a) Bibliographic Retrieval Services, Inc., Corporation Park, Bldg. 702, Scotia, NY 12302; (b) Lockheed DIALOG Information Retrieval Service, 3460 Hillview Avenue, Palo Alto, CA 94304; (c) NIH-EPA Chemical Information System, Information Sciences Corporation, Suite 500, 918 16th St. NW, Washington, DC 20006; (d) SDC Search Service, System Development Corporation, 2500 Colorado Avenue, Santa Monica, CA 90406.
(8) (a) Krueger, G. L., and DesChene, D., "Introducing On-Line Information Retrieval Systems to the Undergraduate and Graduate Student in Chemistry," J. CHEM. EDUC. **57**[6], 457 (1980). (b) Drum, C. A., and Pope, N. F., "On-Line Databases in Chemistry Literature Education," J. CHEM. EDUC., **56**[9], 591–2 (1979).
(9) Fisanick, W., Mitchell, L. D., Scott, J. A., and VanderStouw, G. G., "Substructure Searching of Computer-Readable Chemical Abstracts Service Ninth Collective Index Chemical Nomenclature Files," *J. Chem. Info. Comp. Sci.*, **15**[2], 73–84 (1975).
(10) Morse, D. L., Baker, E. L., Jr., and Kimbrough, R. O., "Propanil-chloracne and Methomyl Toxicity in Workers of a Pesticide Manufacturing Plant," *Clin. Toxicol.* **15**[1], 13–21 (1979).
(11) Sundstroem, G., Jannson, B., and Renberg, L., "Determination of the Toxic Impurities 3,3',4,4'-Tetrachlorazobenzene and 3.3',4,4'-Tetrachloroazoxybenzene in Commercial Diuron, Linuron and 3,4-Dichloraniline samples," *Chemosphere* **7**[12], 973–9 (1978).