

# Handling Chemical Information in the Du Pont Central Report Index<sup>†</sup>

JOHN L. SCHULTZ

Central Report Index, Information Systems Department, E. I. du Pont de Nemours and Co., Inc., Wilmington, Delaware 19898

Received September 5, 1974

**The Central Report Index operates a system of files for storing and retrieving document references to, and structural information about, chemical compounds. Compounds are identified by registry numbers. Correct assignment of registry numbers is aided by molecular formula/structure diagram, name, and topology files. Descriptors, similar to chemical fragments, are generated for each compound by computer from its topology record. The registry numbers, document references and abstracts, names, and molecular formulas for classes of compounds are retrieved by computer searching of the descriptor file, the topology file if necessary, the document file, abstract file, name file, and molecular formula files.**

## INTRODUCTION

This paper describes files and procedures used for handling chemical information in the Du Pont Central Report Index. The emphasis is on the chemical-type files used for document indexing and document reference retrieval. In both operations the two functions of identification of chemicals and of grouping chemicals into classes are involved.

The Central Report Index was established in 1964<sup>5</sup> to index and retrieve information contained in proprietary Du Pont documents, mostly research and development reports. Indexing and retrieval are carried out by information chemists. Currently about 71,000 documents are covered, with 3,500 new ones being indexed each year.

Each document is assigned an accession number and indexed under chemical and general terms. Link letters are used with all terms, and role indicators are used with chemical terms. The document-term-link(-role) posting so created are keyboarded for computer storage and retrieval in the Information Flow System.<sup>3</sup> Also stored for retrieval by computer for each document are the title, author, and an abstract written by the indexer.

General terms denote indexing concepts other than chemical compounds. Each general term is a code of up to 35 alphanumeric characters that spell out a name for the concept, *e.g.*, OXIDATION. There are currently about 9,000 general terms; additional ones are added as needed.

Chemical terms denote chemical compounds, which for our purposes include not only individual nonpolymers but also polymers and materials that are conveniently assigned chemical terms even though not strictly compounds. There are currently about 110,000 chemical terms; about 7,000 new ones are added each year.

Each chemical term is a seven-character code, for compact storage, called a registry number (also known as a "compound number" or "C-number"). The first six characters comprise a six-digit number; the seventh is a letter, called the check digit, calculated by an algorithm from the six-digit number. The check digit helps prevent miscopied registry numbers from entering the system: at updating the computer rejects any registry number in which the check digit and the six-digit number do not correspond. There are separate registry number ranges for organic nonpolymers (numbering about 83,000), inorganic nonpolymers (numbering about 5,000), and polymers (numbering about 22,000). Within each range registry numbers are assigned to new compounds in sequence.

The Central Report Index supplies document references and abstracts by retrospective searching in reply to inquiries for information from Du Pont technical personnel, mostly research chemists and engineers. About 2,350 inquiries are answered each year.

## OVERVIEW OF FILES FOR HANDLING CHEMICAL INFORMATION

As shown schematically in Figure 1, chemical information is handled through a series of files interlinked by one common feature: the registry number. A given compound is denoted by the same registry number in all these files. All the files except the Molecular Formula Card File are IBM 360 computer-based and are therefore mechanically as well as intellectually interlinked through the registry numbers.

**Document Reference Files (First-Level Files).** In document reference files, sometimes called first-level files, the document is the indexed item and the registry number (chemical term) is an attribute of the document. The chemical information in the document is indexed by means of registry numbers as described in the Introduction. The document reference, and other document references containing the same information, is retrieved by specifying one or more of these registry numbers at search.

The Central Report Index document reference file is the Master Accessions File of the Information Flow System.<sup>3</sup> In it are stored the document-term-link-role postings described in the Introduction.

The Computer Abstract File contains the title, author, and abstract of each indexed document.

**Files for Identification and Classification of Compounds (Second-Level Files).** In these files, sometimes called second-level files, the compound (denoted by the registry number) is the indexable item, and it is indexed under attributes such as molecular formula, structure diagram, name, structural features, and properties.

Files for identification are used to determine whether any given compound exists in the system, to find its registry number if it does exist, and to assign it a new registry number if it does not exist.

Files for classification are used to group compounds into classes on the basis of structural or other attributes. The registry numbers for compounds of any desired class are retrieved from these files by specifying the appropriate attributes at search.

The current system of second-level files went into operation in early 1973, having been developed from topology, fragmentation, and other systems that had been set up in 1964.<sup>2,5</sup>

<sup>†</sup> Presented before the Division of Chemical Literature, 168th National Meeting of the American Chemical Society, Atlantic City, N.J., Sept 9, 1974.

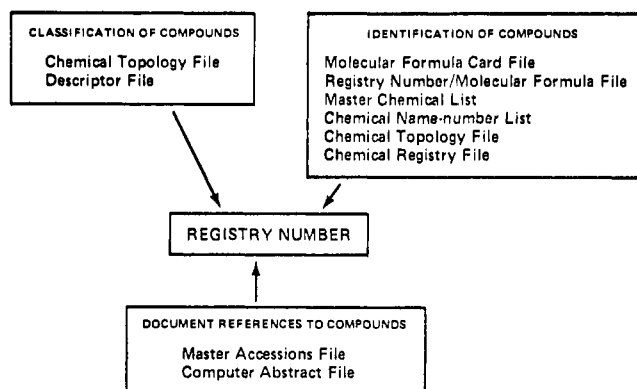


Figure 1. Files for handling chemical information.

## FILES FOR IDENTIFICATION OF COMPOUNDS

**Importance of Identification.** Since each compound is denoted by its registry number, one must find the appropriate registry number in order to index a document dealing with any given compound or to retrieve document references to that compound. This process of finding registry numbers, and the converse process of finding what compounds are denoted by any given registry numbers, is a process of identification. Unless the same compound is associated with the same registry number every time, information will be scattered and lost, and subsequent retrieval will be impossible. This is true also of chemical terms other than registry numbers: names, line notations, etc. Moreover, compound identification is a high-volume and therefore high-total-cost operation. In Central Report Index experience, for instance, the volume of repetitive identification of compounds already in the system is five times as high as that of the one-time entry and classification of compounds new to the system. At the current indexing rate of about 3,500 documents per year at an average depth of about 12 compounds per document, there are 42,000 times when one must answer the questions:

Is this compound in the system?

If so, what is its registry number?

In about 7,000 instances the compound turns out to be new; in the other 35,000 instances, or five times as many, the compound is old and the existing registry number must be assigned to the document reference.

There are about 800 instances a year in which information on individual compounds is retrieved by retrospective searching in answer to inquiries from clients; here also, the identification process is used to find the registry numbers.

**Types of Files Used for Identification.** By itself a registry number conveys no intelligible information;\* there is nothing in it to suggest the compound it denotes. To find the registry number of a compound to be indexed or searched, therefore, it is necessary to match, either manually or by computer, some intelligible representation of the compound with a file containing the same kind of representation for existing compounds, together with the registry number of each compound, arranged in a known order. Conversely, to find what compound is denoted by a given registry number, one consults the appropriate file arranged in registry number order.

The types of representation used in the Central Report Index for finding registry numbers are

- Molecular formulas (with structure diagrams)
- Names
- Topology records

\* A name or a line notation does of course convey intelligible information if one knows the principles by which it was derived, but the necessity of correct identification remains; i.e., the same name or notation must be assigned to the same compound every time.

Matching of molecular formulas and structure diagrams to find registry numbers for indexing is usually done by nontechnical personnel. Matching of names created by technical personnel is done by a computer program or by technical or nontechnical personnel. If none of these routes turns up a registry number, (a) if the compound is a polymer or inorganic nonpolymer, a new registry number is assigned to it by an information chemist; (b) if the compound is an organic nonpolymer, a topology record of the compound is created for matching by a computer program against the file of existing topology records. If the total topology record is matched, the existing registry number is printed out. If it is not matched, the computer program adds it to the file and assigns it the next available registry number.

**Molecular Formula Files.** The Molecular Formula Card File contains, for each nonpolymer, a card bearing its molecular formula, structure diagram, and registry number. Cards are filed in molecular formula order. This file is the only non-computer-based chemical file in the system.

The Registry Number/Molecular Formula File contains the registry number and molecular formula of every nonpolymer. It is printed in both registry number and molecular formula order. For compounds with topology records, entries are made by computer program from stored element count. For other compounds entries are made by keyboarding of manually produced input.

**Name Files.** The Master Chemical List is primarily a tool for display or manual matching of systematic names. It contains the registry numbers, *Chemical Abstracts* names, and molecular formulas for about 53,000 nonpolymers (currently those whose names are required for frequent repetitive lookup or for display in computer-produced search files on selected areas of technology), and registry numbers and Central Report Index names for all 22,000 polymers. Each entry is also manually assigned an alphabetizing code, a seven-digit number that puts it into its alphabetical position; the alternative would be a highly complex computer sorting program. Registry numbers, names, and alphabetizing codes are entered by keyboarding; molecular formulas are picked up by computer from the Registry Number/Molecular Formula File. The Master Chemical List is printed in name (alphabetizing code) order and registry number order.

The Chemical Name-Number List is primarily a tool for computer matching of names and for recording of nonsystematic names. It contains about 31,000 registry number/name entries for about 15,000 compounds. Names, modified where necessary to start with letters, are alphabetized character-by-character by computer and are limited to 48 characters or less. These names are chiefly short *Chemical Abstracts* names, trivial or non-*Chemical Abstracts* names, tradenames, mnemonics, inorganic line formulas, jargon, and internal Du Pont code designations.

An important feature of the Chemical Name-Number List is "autocoding." Autocoding permits the indexer to write the name of a compound at indexing without finding its registry number. If this name matches a name already on the list, a computer updating program automatically substitutes the registry number for the name. Once the indexer is familiar with the kinds of names found on the list, he can index thousands of frequently encountered compounds by autocoding. About a third of the 42,000 yearly registry number indexings are done with autocoding and therefore require no human matching to find registry numbers.

The Chemical-Name-Number List is printed in name and registry number order.

**The Chemical Topology File.** The Chemical Topology File is a computer file of topology records for organic nonpolymers, each with the registry number and molecular formula of the compound represented by the record. Each to-

polymers record consists of a compact connection table showing the atoms in the molecule, the bonds that join the atoms, and auxiliary information ("parameters") about the atoms.

The Chemical Topology File functions as an identification tool through the computer matching of topology records. If molecular formula and name file matching does not turn up a registry number for an organic nonpolymer, a topology record is submitted for the compound. In most cases this record does not match any of the records already in the Chemical Topology File; the computer program therefore adds the record to the file and assigns the next available registry number to the compound. If the record does match a record already in the file, the computer program prints out the existing registry number.

**Input of Topology Records.** The procedure described here for input of topology records is an adaptation of the CAS Registry System developed by Chemical Abstracts Service<sup>4,6</sup> from work done initially at Du Pont.<sup>1</sup>

Input of a topology record begins with a two-dimensional structure drawing, made from an indexed document by an information chemist or copied from the document by a nontechnical person with checking by the information chemist. This structure drawing is simply an atom-bond diagram of the kind familiar to all chemists, and generally it represents the only technical effort required for topology input (beyond the initial operation of indexing the document, of course); nontechnical personnel perform the rest of the input operations. In most cases the structure diagram is typed with an "Invac" chemical typewriter\*\* whose typing strokes are captured on magnetic tape and converted to a compact connection table. In other cases the connection table is manually coded\*\* for computer input via keyboarding (see Figure 2); the non-hydrogen atoms are numbered sequentially and the following parameters are entered for each number as appropriate.

**ELEM**—element symbol (may be left blank for carbon)

**GROUP NO.**—used in certain cases as a shortcut for coding groups of atoms that are repeated sequentially in the structure

**BOND**—type of bond connecting the given atom to the atom whose number appears in the following ATOM ATT column. Bond input types are

- 1 single bond
- 2 double bond
- 4 triple bond

(Alternating "aromatic" bonds are drawn and entered as alternating single and double bonds. These are converted by computer program to a special type of alternating bond, to which is assigned the type number 8.)

**ATOM ATT**—the number of the atom attached by the bond cited in the BOND column. (Up to 15 such attachments are permitted.)

**CHARGE**—sign and value of charge, if any.

**AB VAL**—abnormal connection valence; entered when the sum, bond lines<sup>†</sup> to the atom + absolute value of charge + hydrogen count, does not equal the sum, normal connection valence + 2n, where normal connection valence is a value stored in the computer for each element (e.g., 4 for carbon, 3 for nitrogen, 2 for sulfur) and n = 0, 1, 2, . . . . Abnormal valence values equal to the sum, normal connection valence + 2n, where n = 1, 2, . . . , are automatically recognized and stored by the computer program.

**H COUNT**—number of attached hydrogen atoms; entered only for noncarbon atoms. For carbon atoms the program automatically supplies the hydrogen count required to make the sum, number of bond lines + absolute value of

\*\* We are investigating the use of a "Datapoint" data entry station to replace both the "Invac" and manual coding of connection tables.

† Single bond = one line; double bond = two lines; triple bond = three lines; two alternating "aromatic" bonds to an atom = three total lines, three = four total lines.

MF - C<sub>33</sub>N<sub>4</sub>O<sub>1</sub>S<sub>2</sub>H<sub>56</sub>

C-

TID- C 35083

M

FOR COMPLETE  
DETAILED STRUCTURES  
CONTACT ALAN L. B.  
INSTITUTE FOR COLLS

CONNECTION TABLE CODING

ATOM	ELEM	GROUP NO.	BOND	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT	CHARGE	ATOM ATT
------	------	-----------	------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------	--------	----------

Figure 2. Coding of topological connection table.

charge + hydrogen count, equal 4 or the value entered under AB VAL.

Special structure-drawing conventions are used for structures other than completely defined covalent ones: salts, complexes, incompletely defined structures, boron cages, etc.

Each incoming topology record is assigned a temporary identification number (TID) to identify it during input.<sup>2</sup> Entries for the Master Accessions File, manually assigned entries for the Descriptor File (see below), and entries for the Chemical Name-Number List are also input under the TID. At updating these are converted to permanent entries by automatic substitution of the new or existing registry number for the TID.

For purposes of compound identification the most significant feature of the topology input program is unique table generation.<sup>6</sup> The program numbers the non-hydrogen atoms in a unique, or canonical, order so that a given two-dimensional structure diagram is converted to and recognized as the same compact connection table every time. It is this feature that permits identification of compounds by matching of topology records.

**The Chemical Registry File.** This is the directory file for all registry numbers that exist in the Chemical Information System. For each registry number it shows whether that number appears in the Chemical Topology File, Descriptor File, and/or Master Accessions File.

## FILES FOR CLASSIFICATION OF COMPOUNDS

**Definition.** Classification here means the grouping of compounds into classes on the basis of their structural and other attributes. Mechanistically, files are created in which each registry number is associated with codes denoting the appropriate attributes. The registry numbers for compounds belonging to any class may then be retrieved, as shown later, by specifying the appropriate codes in a retrospective search. The two files used for classification are the Chemical Topology File and the Descriptor File.

**The Chemical Topology File.** As described above, the Chemical Topology File contains a complete structure record for each organic nonpolymer, and it functions as an identification file through matching of these complete records. It also functions separately as a classification file through matching of partial records: by the search program described later, any desired structural attribute may be coded as an atom-bond network and matched against the Chemical Topology File to retrieve the registry numbers whose topology records contain this network. Since the structural attribute is defined and coded only as needed, the Chemical Topology File is a postclassification file.

**The Descriptor File.** Descriptors are terms, up to 23 characters long, each denoting some individual structural or other attribute of a compound. There are about 600 non-polymer descriptors. (The current system of polymer descriptors is not covered here. It is about to be replaced with a new system based on polymer class and constituent monomers.) In the Descriptor File each registry number is associated with the appropriate descriptor terms. For specified descriptors each registry number/descriptor posting includes also a count, *i.e.*, the number of times the structural feature denoted by the descriptor occurs in the molecule. Since descriptors are predefined terms, the Descriptor File is a preclassification file.

Descriptors for structural attributes are equivalent to what are often called fragments. Descriptors for non-structural attributes, *e.g.*, physical properties, can be used for data retrieval, although this feature has not been implemented in the Central Report Index.

**Computer Generation of Descriptors.** Most descriptors for organic nonpolymers are generated by computer program from the topology records the first time they enter the Chemical Topology File. This requires no human effort or judgment beyond the initial programming and the correct input of the topology records. The computer is programmed to post a new registry number to preselected descriptors whenever it finds the prescribed atoms, bonds, and/or other parameters present in its topology record. For those descriptors that take a count, it also generates a count of the number of times the structural feature is found in the topology record. Descriptors are generated by computer for the following types of features:

**Elements**—The descriptors are the element symbols. In addition there is a collective descriptor M for all metals.

**Bonds**—Seven descriptors of the form BOND-*t-l*, where *t* = bond type (1 = single, 2 = double, 4 = triple, 8 = alternating "aromatic") and *l* = bond location (1 = open chain, 2 = ring). Example: BOND-1-2 denotes a single ring bond.

**Element-bond-element triplets**—Each denotes two elements and the bond joining them. Descriptors are of the form el-e2-*t-l*, where el = element symbol first in alphabetical order, e2 = element symbol second in alphabetical order, and *t* and *l* denote the bond type and location, respectively, as for the bond descriptors. Example: O-S-1-2 denotes an oxygen-sulfur single ring bond. There are about 280 element-bond-element triplets covering all the non-inert-gas nonmetals and the collective metal "atom" M. These descriptors afford detailed and versatile access to structural features, especially in conjunction with topological searching, and avoid the proliferation of descriptors for seldom-encountered functional groups and rings.

**Abnormal connection valence**—Ten descriptors of the form VALENCE-ABNORMAL-*n*, where *n* = ZERO, 1, . . . , 9-UP and denotes a numerical value of abnormal connection valence as defined under Input of Topology Records.<sup>6</sup>

**Ring parameters**—Descriptor names are as follows:

RING	Number of individual rings
RING-el	Number* of ring atoms of element whose atomic

symbol is el, where el = As, B, I, N, O, P, S, Sb, Se, Si, Te, or the collective metal symbol M  
 Number\* of carbon bridgehead atoms  
 Number\* of non-carbon bridgehead atoms  
 Number\* of non-carbon ring atoms  
 Number\* of carbon spiro atoms  
 Number\* of non-carbon spiro atoms

RING-BRIDGEHEAD-C

RING-BRIDGEHEAD-NON-C

RING-NON-C

RING-SPIRO-C

RING-SPIRO-NON-C

**Functional groups**—Descriptors for about 70 common functional groups. To generate each descriptor, the computer is programmed to examine ingoing topology records with the appropriate prerequisite descriptors for elements, bonds, element-bond-element triplets, etc., then to look for the prescribed network of atoms and bonds that define the functional group. Functional group definitions are programmed in terms of any bond types and environments, any element symbols (including hydrogen), and the following special symbols:

CM	- carbon bonded to a non-carbon atom by a nonaromatic multiple bond
CZ	- carbon not bonded to a non-carbon by a nonaromatic multiple bond
HET	- any non-carbon-non-hydrogen atom
M	- any metal atom
REAL	- any non-hydrogen atom
X	- any halogen atom

For certain functional groups there are separate descriptors for cyclic and acyclic occurrences.

The following are examples of functional group descriptor prerequisites and definitions:

Descriptor	Definition	Prerequisites
CARBOXAMIDE-3-ACYCLIC	(C, H)-CM-N-CZ  (open chain bonds)	At least 3 C-N-1-1, at least 1 C-O-2-1
ETHER-CYCLIC	CZ-O-CZ (ring bonds)	At least 2 C-O-1-2

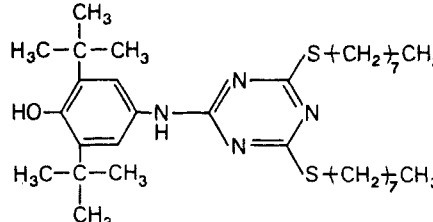
**Classes of compounds containing carbon-carbon double bonds**—Four descriptors for classes C-C double bond compounds that are of interest in polymer chemistry. The computer examines each C-C double bond in the topology record and generates one of the descriptors, if appropriate, in the following descending order of precedence: ENE-CONJUGATED (nonaromatic conjugated dienes, trienes, etc.), ACRYLIC, ALLYL, VINYL.

**General structural features**—The more commonly used of these descriptors are ATOM-NON-H (number of non-hydrogen atoms), FREE-RADICAL (presence of one or more free radical sites, defined as the occurrence of specified lower-than-normal connection valences with specified elements), HYDROCARBON, ION-NEG (presence of one or more negative charges), and ION-POS (presence of one or more positive charges).

An example of computer-generated descriptors is shown in Figure 3.

**Manual Assignment of Descriptors.** Computer generation of descriptors applies only to compounds having topology records, *i.e.*, organic nonpolymers. For polymers and inorganic nonpolymers all descriptors are assigned manually by an information chemist who examines each new compound when it first enters the system. In some cases these descriptors are the same as those generated by computer for organic nonpolymers. The resulting input is keyboarded and added to the Descriptor File.

\* In the entire structure, not in an individual ring system.



Descriptor	Count
ATOM-NON-H	40
C	33
N	4
O	1
S	2
BOND-1-1	29
BOND-8-2	12
C-C-1-1	22
C-C-8-2	6
C-N-1-1	2
C-N-8-2	6
C-O-1-1	1
C-S-1-1	4
RING	2
RING-N	3
RING-NON-C	3
AMINE	1
AMINE-2	1
AMINE-2-ACYCLIC	1
HYDROXY	1
HYDROXY-ARYL	1
SULFIDE	2

Figure 3. Computer-generated descriptors.

## RETRIEVAL OF CHEMICAL INFORMATION

**Retrieval for Indexing.** The procedures for identification of compounds described above are actually procedures for retrieval of registry numbers, for use in indexing, from the Molecular Formula Card File, Master Chemical List, Chemical Name-Number List, or Chemical Topology File. This may be called "full-code" matching since one matches an entire molecular formula/structure diagram, name, or topology record code against a file of existing codes. As already shown, this is a high-volume operation, requiring about 42,000 lookups a year. It is essential for correct storage of information and therefore for correct retrieval of information for retrospective searching.

**Retrieval for Retrospective Searching.** Figure 4 shows schematically how the files are interrogated to retrieve information on an individual compound, *e.g.*, 3-chloro-1-propanol, or on a class of compounds, *e.g.*, 3-chloro alcohols.

The first step is to find the registry number(s) of the compound(s). For an individual compound this is done by a procedure identical with that already described for finding the registry number for indexing: by matching the molecular formula/structure diagram, name, or topology record against the Molecular Formula Card File, Master Chemical List, Chemical Name-Number List, or if necessary Chemical Topology File. For a class of compounds the registry numbers are found by interrogating the Descriptor File and subsequently, if necessary, the Chemical Topology File. From the Descriptor File are obtained the registry numbers that have been posted to the appropriate descriptors for that class. If the available descriptors do not fully define the class, the answer from the Descriptor File search is further refined by a Chemical Topology File search.

The registry numbers found or retrieved above are then passed through subsequent steps to obtain all or any of the following types of information:

Accession numbers of documents containing information on the compound(s) denoted by the registry number(s)

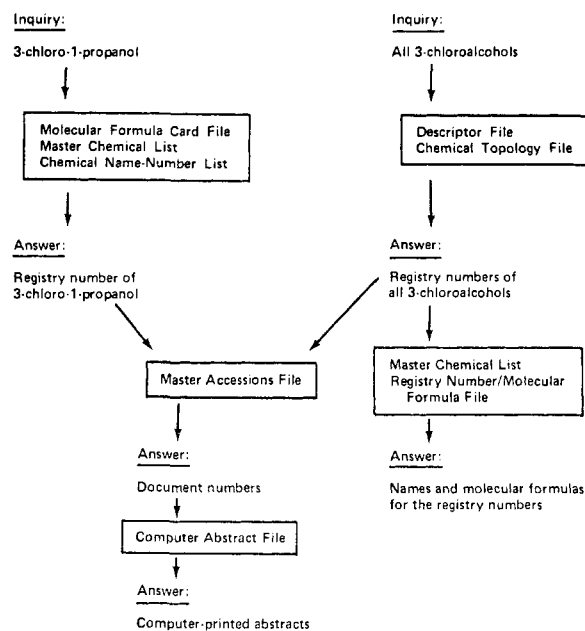


Figure 4. Retrieval of chemical information by retrospective searching.

(from the Master Accessions File).

Abstracts of these documents (from the Computer Abstract File).

Names and/or molecular formulas for the registry numbers for a class of compounds (from the Master Chemical List and/or Registry Number/Molecular Formula File).

The step of finding the registry number of an individual compound for retrospective searching is a manual one. All the other steps are carried out by computer programs, and the retrieval sequence may be programmed to run from beginning to end or to stop at any point to allow human intervention before the next stage. This is made possible by the fact that the files are interlinked through the registry numbers as already shown in Figure 1.

The two succeeding subheadings show how registry numbers for a class of compounds are retrieved from the Descriptor File and Chemical Topology File and carried through the system to retrieve information from the Master Chemical List, Registry Number/Molecular Formula File, Master Accessions File, and Computer Abstract File. Figures 5-9 show the input and answers for an inquiry on "properties of hindered monophenols containing (a) an ether or secondary amine group para to the hydroxyl, (b) heterocyclic nitrogen but no other hetero ring elements, and (c) no chlorine atoms."

**Retrieval from the Descriptor File.** As shown in Figure 5, a Descriptor File search consists of up to four sections: the Identification Section (optional) allows input (for printout with the search answer) of identifying information such as a title showing the subject of the search; the Logic Section specifies the search terms and logic to be used in searching the Descriptor File; the Substructure Section (optional) specifies whether the Chemical Topology File is also to be searched; and the Output Section specifies the format and further processing, if any, of the search answer. All except the Substructure Section are manipulated in a manner similar to the Identification, Logic, and Output Sections in Master Accessions File searching as described in an earlier publication.<sup>3</sup>

The Logic Section contains one or more statements that specify the search terms to be used and the Boolean logic to be performed on them. Each statement is in the form of an

"Hindered monophenols containing (a) ether or sec-amine para to OH, (b) heterocyclic N but no other heterocyclic elements, (c) no Cl".

001	CHEM. SEARCH	PHEN/JS	(16 characters maximum)
002	IDENTIFICATION SECTION:	Cross out Identification Section if not used.	
003	SEARCH TITLE	HINDER MONOPHENOLS-P-2-AMINE,	
004	(50 characters maximum)	ETHER-N-HETERO-NO CL.	
009	END IS		
010	LOGIC SECTION:	Answer names are 16 characters maximum.	
100	\$ PHEN (*) * HYDROXY-ARYL * 1 * (AMINE-2 > 0 +		
110	ETHER > 0) * RING-NON-C * RING - N - CL > 0.		
799	END LS		
800	SUBSTRUCTURE SECTION:	Cross out Substructure Section if not used.	
810	SS( PHEN )	= 28383.	
899	END-SS	If unspecified: STRUC_DESC_MAX (answer name) = 10,000	
900	OUTPUT SECTION:	Either PASSROLES or ANS FORM = NO PASS must be coded.	
910	PASSROLES:	PHEN	)*( 0 ): RSS PHEN.
920	ANS FORM = FILE.		
998	END OS	If unspecified: ANS_MAX = 10,000	
999	END CHEM. SEARCH *		

Figure 5. Descriptor file search input.

equation. The left-hand member is an answer name up to 16 alphanumeric characters long; the right-hand member consists of the search terms related by the standard Boolean operators "\*" (and), "+" (or), "¬" (and not), and "." (end of statement). A statement may be a final answer statement, shown by a \$ before the answer name, or a subanswer statement. The answer from a final answer statement is printed for the user. The answer from a subanswer statement merely enters into the logic of some other answer statement through its answer name's appearing as a search term in that answer statement. A final answer name may also appear as a search term in another final answer statement. Answer names are used as search terms to avoid re-coding logic that is common to two or more answer statements.

Search terms can be of several types, the most common of which are the following:

**Descriptor.** The term for a descriptor that does not take a count is the descriptor by itself.

**Descriptor relation count.** One type of term for a descriptor that takes a count consists of the descriptor followed by an operator (the relation) and a number (the count). The relation shows the comparison to be made between the count stored in the Descriptor File and the count specified after the relation; it determines whether the count stored in the Descriptor File meets the requirement of the inquiry. Thus AMINE-1 > 2 means that a compound must have more than two primary amine groups to be an answer to the inquiry. The relations used are = (equal to), > (greater than), < (less than), >= (greater than or equal to), <= (less than or equal to), and ¬ = (not equal to).

**Descriptor relation descriptor.** The other type of term for descriptors that take a count consists of two descriptors separated by one of the relations described in the preceding paragraph. It specifies the numerical relationship that must exist between the counts for the two descriptors. Thus RING-NON-C = RING-N means that, for a compound to be an answer to the inquiry, the number of nitrogen atoms in rings must equal the total number of noncarbon atoms in rings, i.e., no non-nitrogen heterocycles are permitted.

• Subanswer names or final answer names. As shown above, an answer name may appear as a search term in another answer statement.

**Retrieval from the Chemical Topology File.**<sup>2</sup> If the

Descriptor File answer statement in the Logic Section does not provide as precise an answer as required, that answer is further refined by a search of the Chemical Topology File. The topology records for the registry numbers turned up in the Descriptor File search are examined for the presence of the atom-bond network desired; those registry numbers whose topology records contain this atom-bond network constitute the answer. This is a "part-code" search for an atom-bond network contained within any number of topology records, in contrast to the "full-code" search for compound identification by matching of an entire topology record. A Chemical Topology File search is always preceded by a Descriptor File search; topological searching of the file of 83,000 topology records would not be economically feasible without such previous screening. The average size of a Descriptor File search passed on for topological searching is about 1700 registry numbers (2% of the file).

About 47% of the Descriptor File search answers are passed on for further refinement by topological searching. The chief features that require such refinement are specific ring systems, functional groups not covered by the functional group descriptors, and configuration of two or more functional groups within a molecule (e.g., 1,6-alkylene diamines). Experience will show where further descriptors are needed for such features, but the guiding principle is to make special provision only for those features that are requested frequently.

To trigger a search of the Chemical Topology File the searcher writes "\*" after the answer name of the final answer or subanswer that is to be carried through for topological searching, and enters a statement in the Substructure Section showing this answer name and the name to be assigned to the topology search.

An example of a Chemical Topology File search is shown in Figure 6. The first step in coding a topology search is to determine what network(s), or group(s), of atoms must, may, or must not be present. The diagram of these is called the substructure drawing. Each group is assigned a two-digit group number and the group numbers are combined in a Boolean statement called the Group Combination Logic. For example the statement, 01\* (02 + 03) ¬ 04., signifies that group 01 must be present, either group 02 or 03 must be present, and group 04 must not be present. Two strings of atoms with no atoms in common must be assigned different group numbers because the program interprets consecutively cited atoms as being joined.

The atoms within each group are then numbered. Numbering may be in any order except that (a) no numbers may be skipped and (b) for search efficiency that most unusual atoms or sequences should be given the lowest numbers. Each such numbered atom is called a node. Topological searching consists of comparing each node of the query with the node(s) (non-hydrogen atoms) of each candidate topology record for the following:

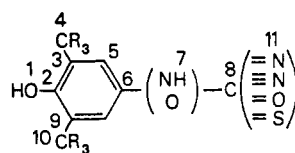
Are the elements the same?

Are the bond connections from the previous node the same?

Are there special parameters, e.g., charge, connection valence, hydrogen count?

If the node being compared fulfills all the test conditions, the next node is tested for the same parameters. If not, the computer is programmed to "walk back" to previous nodes and to test nodes on other paths proceeding from them. This continues until all the nodes in the query structure have been tested or matched with nodes in the topology records for all the registry numbers passed on from the Descriptor File search. If a match on any node fails after all alternatives have been tested, that topology record is rejected. Only registry numbers for topology records that match the query structure are selected as answers.

It is readily apparent that search efficiency demands that there be as little "walking back" as possible. This is accomplished by (a) giving low numbers to the most unusu-



GROUP 01: 1-2-3-4-3-5-6-7-8  
 GROUP 02: 2-9-10  
 GROUP 03: 8-11

Form Q: Group Combination Logic											
1	2	3	4	5	6	7	8	9	10	11	12-79: Group Logic — (Zero-filled Two-Digit Group Numbers)
2	8	3	8	3							Operators: + (Union) * (Intersect) - (Negation) = (Equation ends with period)
0 0 1 0 01 * 02 --- 03.											
Form P: Atom Parameters Substructure Coding											
SEQ. NO. (1-10)	P	GROUP NO. (1-11)	ATOM NO. (1-11)	NEG. SPEC. (N)	ELEMENT RECITED ATOM NO. (1-11)	BOND TYPE (1-8)	LOC. (1-2)	H CNT.	CONN. VAL.	CHARGE SIGN/VALUE	NON-H ATT.
0 0 1	P	0 1	0 1		O			1			
0 0 2			0 2								3
0 0 3			0 3				8	2	0		
0 0 4			0 4		C		1	3	0		4
0 0 5					0 3						
0 0 6			0 5				8	2			
0 0 7			0 6				8	2	0		
0 0 8			0 7		N		1	3	1	3	
ALTERNATE SPECIFICATIONS											
0 0 9			0 8		C						2
0 1 0		0 2			0 2						
0 1 1			0 9					0			3
0 1 2			1 0		C		1	3	0		4
0 1 3		0 3			0 8						
0 1 4			1 1		N		6	3			
ALTERNATE SPECIFICATIONS											
					2 3						
					2 3						

Figure 6. Chemical topology file search input.

al features so that there will be as few alternative starting places as possible and (b) specifying as many parameters as possible for each query node so that the program can decide at the earliest possible node whether it is proceeding down a correct path of atoms.

The query node numbers and parameters of the substructure are then coded for computer input. The following features are coded:

SEQ. NO.—a sequence number to keep the coded lines in the proper order.

GROUP NO.—the group number explained above.

ATOM NO.—the query node number described above, specified for an atom the first time it is cited. (For subsequent citations of the same atom see ELEMENT ... below). In place of such a number the following symbols may be used: UD (the node must be a previously matched one, i.e., coded on a previous line) or UN (the node may not be a previously matched one).

Except for a recited atom (see ELEMENT ... below), the preceding items are mandatory while the following are optional.

NEG. SPEC.—negative specification. An N in this field signifies that the node must not have the combination of parameters specified in the succeeding fields.

ELEMENT, RECITED ATOM NO., SYMB. EL.—an element symbol in this field requires the node to be an atom of that element. A blank means that any element is acceptable. When a previously cited note is recited, its node number is entered in this field and all other fields except SEQ. NO. are left blank. Symbolic element is a rarely used feature that permits two or more nodes, each of which can be a choice of elements, to be specified as the same or different.

BOND—type and location of the bond connecting the node to the preceding node. Type is denoted, as above, by 1 (single), 2 (double), 4 (triple), or 8 (alternating aromatic); and location as 1 (open chain) or 2 (ring). A choice of types

CRI CIS MOLEFORM/NAME SEARCH ANSWER	
PHEN	08 JUL 74 PAGE 1
107716J	C 33 N 3 0 2 5 2 H 55 PHENOL 4 / (4,6-BIS(OCTYLTHIO)-5-TRIAZIN-2-YL) OXY/-2,6-DI-TERT-BUTYL-
6385G	C 33 N 4 0 1 5 2 H 56 PHENOL 4 / (4,6-BIS(OCTYLTHIO)-5-TRIAZIN-2-YL) AMINO/-2,6-DI-TERT-BUTYL-

Figure 7. Name/molecular formula answer.

PHEN	001	SEARCH: PHEN/JS	(16 characters maximum)
	010	LOGIC SECTION:	
	020	\$ PHEN = RGNPROPERTIES * RSSPHEN.	
	800	END_LS	
	901	OUTPUT SECTION: If unspecified: ANS_MAX = 2000, maximum abstracts = 1000.	
	902	ANS_FORM = ABSTRACTS	
	998	END_OS	
	999	END_SEARCH	

Figure 8. Input for master accessions file search utilizing answers from descriptor file/chemical topology file search.

or locations is denoted by the sum of the numbers for the individual ones. Thus 6 (= 4 + 2) in the TYPE field denotes a double or triple bond. Specification of bond type/location is optional because consecutive citation of two nodes in the same group already requires that the two nodes be connected. The BOND fields are left blank if any bond type/location is acceptable, if only one is chemically possible or likely (e.g., a carbon to chlorine bond would be a single open-chain bond), or if the parameters H CNT., CONN. VAL., and/or NON-H ATT. leave only one as possible or likely. These parameters are used instead of bond type/location whenever possible because they give the computer information one node earlier than do the bond parameters.

H CNT.—the number of hydrogen atoms (including zero) attached to the node. A blank means that any value is acceptable.

CONN. VAL.—connection valence (whether normal or abnormal) as defined above. A blank means that any value is acceptable.

CHARGE—sign and/or absolute value of a charge on the node. A blank means that any value is acceptable.

NON-H ATT.—number of nodes (non-hydrogen atoms) attached to the node. A blank means that any value is acceptable.

ALTERNATE SPECIFICATIONS—for any node, up to 12 alternative combinations of parameters may be coded.

**Retrieval of Names, Molecular Formulas, Document References, and Abstracts.** The registry numbers that satisfy the requirements of the inquiry are printed out as a search answer. Registry numbers are seldom of value by themselves, so various options are coded in the Output Section of the Descriptor File search (Figure 5) to instruct the computer to retrieve intelligible information.

Coding of ANS-FORM = FILE plus a Molform/Name Print Request adds to this answer the Master Chemical List name, if any, the molecular formula from the Registry Number/Molecular Formula File, or both. An example is shown in Figure 7. This feature is of use chiefly to show the inquirer what compounds the inquiry turned up or to permit human screening (seldom required) of registry numbers before they are passed on for document retrieval.

The most important option is one that triggers the retrieval of document references to the registry numbers turned up in the search. This is a statement of the form, PASSROLES (final answer name) = (role indicator(s): subanswer name for document search). It instructs the



**CRI INQUIRY REFERENCE**

CRI REPORT NO. UO 420359  
 DEPT. REPORT NOS. NPR-2-B  
 REPORT DATE 2/66 Report available from Energy and Materials Department  
 TITLE UV, Absorbers, Antioxidants, and Fluorescent Brighteners - Geigy Industrial Chemicals  
 AUTHOR McKinney, J.W.

ABSTRACT Melting point, availability and price, and selected physical properties were reported for the UV absorbers "Tinuvin" P, 326, and 327 (all hydroxyphenyl benzotriazoles). Similar data were reported for "Irganox" 565, 858, 1010, and 1076 hindered phenol type antioxidants. The price of "Tinopal" fluorescent brightener was reported.

Figure 9. Computer-printed abstract.

computer to take the registry numbers of the search answer for the specified final answer name, pass them on to the Master Accessions File to retrieve the document accession numbers indexed under them with the specified role indicators, and manipulate this body of document numbers, collected as a search term with the specified subanswer name, to obtain a document number answer. An example of the Master Accessions File input for such a search is shown in Figure 8. A further option, ANS-FORM = ABSTRACTS., in the Master Accessions File search causes the computer to print abstracts for the document numbers retrieved (Figure 9).

**Volume and Types of Search Inquiries.** Table I shows the number of inquiries per year answered by the central Report Index, with emphasis on the various types of inquiries requiring retrospective searches of the chemical files.

#### FLEXIBILITY AND UTILITY OF THE SYSTEM

The Central Report Index procedures for handling chemical information bring together input and retrieval capabilities which evolved through more than 20 years of operating experience with several predecessor systems throughout Du Pont. During this time various features were added or modified as they appeared to be necessary or advantageous. As part of this cumulative experience, several summary comments can be made on the flexibility and use of the current system.

(a) Effective retrieval is impossible without correct input, no matter what the sophistication of the searching system may be. Though not always obvious or accepted, this recognition has led to the creation of a variety of files—molecular formula/structure diagram, name, and topology—for the repetitive identification of compounds and the correct assignment of their associated registry numbers. Procedures have been set up to keep these files as accurate and complete as possible so that document references and other records may be posted to the proper registry numbers and therefore positioned in the proper file locations.

(b) The basic task in serving a large clientele of research chemists and engineers is to provide the source references to information on chemical compounds. The emphasis has always been on guiding the client to appropriate Du Pont documents containing information relevant to his inquiry. Thus the manipulation of registry numbers by the system is not an end in itself, but only a means for retrieving the information associated with them; the end product of a search is therefore not registry numbers of chemicals but citations and abstracts of documents.

(c) Computers are useful aids for the fast, accurate manipulation of large collections of records in several different kinds of files. Computerization of most of the files relating to chemicals and interlinking of them through the registry numbers allow flexibility in using them for a variety of purposes. Thus, for example, the Chemical Topology File is both an input and a retrieval tool; the Master Chemical

Table I. Volume and Types of Search Inquiries

Type of search	No./year <sup>a</sup>	% of total <sup>a</sup>
No compounds specified	1200	51
Compounds specified	1150	49
Individual compounds	800	34
Classes of nonpolymers	200 <sup>a</sup>	8 <sup>a</sup>
By functional groups	160	7
By element-bond-element triplets	35	1.5
By specific ring systems	30	1.3
By elements	30	1.3
By presence or absence of rings	20	0.9
By classes of ring systems	15	0.6
Classes of polymers	150	7
Total	2350	100

<sup>a</sup> Numbers of subclasses total more than 200, and percentages more than 8, because more than one type of feature is often specified.

List is a name display tool both for indexing and for enhancing the interpretation and utility of search answers. Moreover the search system, starting with the structural specifications for a given class of compounds, is programmed to collect, carry through, and tabulate registry numbers, names, molecular formulas, document references, and document abstracts with no human intervention beyond the initial coding of the search logic and options.

(d) Atom-bond topology records provide a powerful and versatile tool for specifying chemical classes *via* precise substructures. The reserve capabilities of the Chemical Topology File can be applied when needed to supplement the requests for common structural classes, such as functional groups, which are predefined in the Descriptor File. In addition, this flexibility restrains the potential ballooning of a preclassified structural file into an unmanageable file size and plethora of descriptors.

(e) A balance of features, capabilities, and file manipulations is needed. For example, a combination of sequential and inverted file structures is beneficial, depending on the density of records posted to individual terms;<sup>3</sup> a variety of files for the identification of chemicals is necessary, as noted earlier; an optional capability for preclassification (descriptor terms) and postclassification (topology records) of chemicals by structural characteristics is highly useful. Frequently, in our experience differences in viewpoint about the strengths and benefits of competing arrangements center not on their need but on where the balance point should be set. Thus, no single procedure or technique provides the one best arrangement for indexing and retrieval of chemical information in every situation.

#### FUTURE DIRECTIONS

Most of the current descriptors denote structural features of compounds. A logical extension would be the use of property descriptors for data retrieval. Some applications of this kind are now under consideration.

A new descriptor system for polymers is planned to go into operation by early 1975. The descriptors will denote polymer classes and the monomers (actual or prescribed) of which the polymers are made. It will be possible to retrieve monomer registry numbers by Descriptor File and, if required, Chemical Topology File searching, then retrieve the registry numbers of the polymers associated with these monomers.

The only Chemical Information System File not yet computer-based is the Molecular Formula Card File. The cards are still produced, filed, and accessed manually. A desired goal is to mechanize, digitally or optically, the storage and retrieval of structure images by their molecular formulas and registry numbers.



## ACKNOWLEDGMENTS

The system described here developed over several years through the efforts of many people in Du Pont, especially the Central Report Index and Information Retrieval Systems groups. The author acknowledges the advice and help given by Dr. Melvin L. Huber of the Central Report Index in the preparation of this paper.

## LITERATURE CITED

(1) Gluck, D. J., "A Chemical Structure Storage and Search System Devel-

- oped at Du Pont," *J. Chem. Doc.*, **5**, 43 (1965).  
 (2) Hoffman, W. S., "An Integrated Chemical Structure Storage and Search System Operating at Du Pont," *J. Chem. Doc.*, **8**, 3 (1968).  
 (3) Hoffman, W. S., "Du Pont Information Flow System," *J. Chem. Doc.*, **12**, 116 (1972).  
 (4) Leiter, D. P., Morgan, H. L., and Stobaugh, R. E., "Installation and Operation of a Registry for Chemical Compounds," *J. Chem. Doc.*, **5**, 238 (1965).  
 (5) Montague, B. A. and Schirmer, R. F., "Du Pont Central Report Index: System Design, Operation, and Performance," *J. Chem. Doc.*, **8**, 33 (1968).  
 (6) 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).

## Hoffmann-La Roche's On-Line/Batch Interactive Chemical Information System<sup>†</sup>

A. SHENG,\* L. LUPI, M. RONAYNE, A. SPRULES, and S. ZORNETZER

Hoffmann-La Roche Inc., Nutley, New Jersey 07110

Received September 30, 1974

**This paper presents the current view of Roche's integrated Chemical Information System which is user-oriented and modularly designed for easy expansion. The essential elements of the data base for on-line interrogation and the various on-line search procedures—Ro number search and Wiswesser notation search—for different applications are described.**

## SYSTEM OVERVIEW

This paper provides an overview of Hoffmann-La Roche's Chemical Information System. It explains the main features of the system, the advantages and limitations of its various modules, and the means by which the subsystems are interrelated. The design characteristics of the system were modular, versatile, user-oriented, and open-ended to allow individual subsystems to be added to the overall integrated system for easy expansion and modification.

The system was jointly developed by the Research Systems Section of Management Information Services Department and the Research Records Office of Research Services Department. It is implemented on a Honeywell 600/6000 time-sharing system with CRTs and teletypewriters to form a network of terminal-to-computer communication on a data base of over 100,000 compounds. In addition, "Vistas," which are visual display devices monitored by a special software package, are placed in various locations to provide management and technical personnel the momentary status and utilization of the operating system. Each job is assigned an identifying name or number which is displayed on the Vista, thereby allowing the user to follow the step-by-step execution of his program (Figure 1).

The chemical information system provides internal information services for the scientists of the Research Division. Inquiry, search, and retrieval are performed on request through the Research Records Office (RRO) which serves as a focal point for storage and retrieval of technical information.

Figure 2 is a system flow chart showing the various technical components and the processing of the interrelated data elements within the chemical information system. The

data bank consists of the following subsets: chemical name, chemical structure, chemist name, chemist number, molecular formula, and Wiswesser notation<sup>1</sup> of the registered compounds. Each compound is identified by a nine-digit number called the Ro number (Roche registry number). The three major files which form the essential components of the data base are the chemical name file, the chemical structure file, and the Wiswesser notation file. These are randomly structured files stored on magnetic disk and are processed by a series of programs written in Fortran and Assembly Language for man-machine interaction in entry, update, search, and retrieval.

## TERMINAL DEVICES AND INPUT PROCEDURE

**A. On-Line Terminals.** Two remote terminal devices, the teletypewriter and the CRT, have been used for communication with the central processor both to enter and retrieve information.

The teletype, Model 37, with paper tape punch and reader module, is commonly used for processing various subsystems. It is equipped with half forward and reverse line space, chemical bond symbols, and upper and lower case characters. A special feature of the character set is the inclusion of "Octobliques,"<sup>2</sup> an extension of the existing dot-bond notation, for depicting the spatial arrangement in complex structures. The set consists of eight oblique lines: two slants each of slope +2.0, -2.0, +0.5, and -0.5. Each pair of like slopes is situated in opposite halves of the character matrix.

Chemical data are handled by the following procedure. After a compound has been assigned an Ro number, the TTY 37 is used for recording the chemical data including the structure on a special form called a data sheet. The reverse line space feature allows the operator to type the structure along any convenient path, for example, around a

<sup>†</sup> Presented before the Division of Chemical Literature, 168th National Meeting of the American Chemical Society, Atlantic City, N.J., Sept 10, 1974.

\* To whom correspondence should be addressed.