An Integrated Chemical Structure Storage and Search System Operating at Du Pont*

WARREN S. HOFFMAN
E. I. du Pont de Nemours & Company, Inc., Secretary's Department, Wilmington, Del.

A topological chemical structure storage and search system has been installed for the use of four centralized information groups within Du Pont. Input and search subsystems have been integrated with computerized information retrieval systems for technical reports and patents. Updating and searching of chemical structure and document oriented files may be performed together. Computer programs are modified and extended versions of the Chemical Abstracts Service registry and experimental substructure search systems. Polymers and other chemical classes have been added. Future development of the system will result in full integration of chemical structure and document system files and programs and improved input and searching methods.

Several centralized information groups at Du Pont which index technical reports and patents have converted to mechanized information retrieval systems within the last four years. Computers are employed in both updating and searching of files.

The Chemical Structure Storage and Search System (CS^4) discussed in this paper is the current implementation of the system described by D. J. Gluck of Du Pont in 1964 (*1*). The earlier version developed and tested most of the concepts in use today. Algorithms and programs were given and described to the Chemical Abstracts Service. CAS produced production versions of the input registration programs (*2, 3*) and an experimental substructure search system (*4, 5*).

During a continuing period of cooperative effort, the CAS programs have been installed for use in the production environment at Du Pont. Operation of the chemical structure system, integration of the system with the primary document systems, modifications to the CAS programs, and expected future developments are described in succeeding sections.

DESCRIPTION OF THE CHEMICAL STRUCTURE SYSTEM

The Chemical Structure Storage and Search System is used as a second-level index to the primary document system files. CS^4 stores the topology of chemical structures in the form of connection tables. A registry number is assigned to each input structure identified by a temporary identification number (TID). Inquiries are submitted in the form of atom-bond sequences called substructures. Registry numbers of compounds which contain the specified substructure are retrieved. The input programs are called the Registry System. The inquiry programs are called the Substructure Search System.

The CS^4 System is used by four centralized Du Pont information groups which index and search collections of internal technical reports and U. S. and foreign patents. A Central Chemical Registry group coordinates input and searching activities related to the chemical structure system. CS^4 is integrated with both update and search cycles of the computer systems used by each group to store and search information indexed from documents.

Document Systems Organization. The configuration of a typical Document-CS^4 system is shown in Figure 1. Index terms selected for each document are stored in one of two inverted files. The General Term File uses alphanumeric fields to identify nonchemical terms with associated document accession numbers. The Compound File is similar in organization but uses numeric compound numbers to identify terms. Both files provide for link and role indicators. A separate list of accession numbers is associated with each term-role combination.

Supporting these files are several other computer-maintained second level index files. The Thesaurus File stores each valid term in the General Term File, relationships among terms, and other term descriptions. The Fragment File stores chemical compound descriptions in predetermined functional groups and auxiliary descriptors. The Fragment File is an inverted file with compound numbers posted to fragment term codes. The CS^4 Registry Files store structures of chemical compounds in the compound files of the four operating groups.

CS^4-Document System Update Relationship. Both general and chemical compound terms indexed for a typical document accession are written on an indexing sheet. A TID number is assigned to compound terms for which the registry number cannot be determined by lookup on chemical name-number lists or in other manual references. A structure diagram and TID number are written on a form which initiates processing by the CS^4 Registry System.

Input to the General Term and Compound Files is in the form of records containing an index term, roles, accession number, and link. Each file is updated separately, although input may be submitted together. After sorting by term, input records are edited and merged with the most recent master file.

Linkage of the CS^4 Registry and Document Systems is accomplished by temporarily storing Document System compound file input records identified by a TID number until the proper registry number may be determined.

[*] Presented before the Division of Chemical Literature, 154th Meeting, ACS, Chicago, Ill., September 1967.

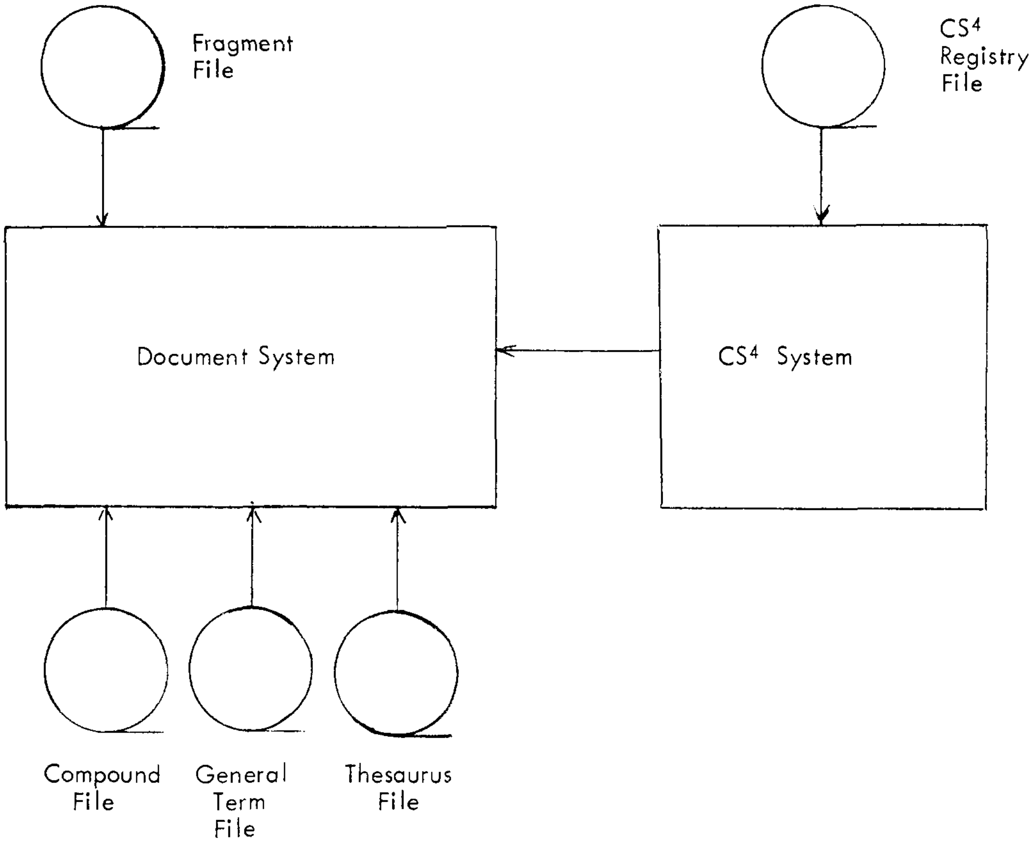


Figure 1. Typical Document–CS^4 System configuration

The operation of the linkage is shown in Figure 2. A tape file produced by the final registry output edit contains a TID number-registry number pair for each input compound. The registry number is substituted for the TID number in each Document System input record. Output records are identical to Document System input records for which the registry number was determined by manual lookup.

Connection table input is submitted outside the input stream of the Document System. The independent development of the CS^4 and Document System processing programs at Chemical Abstracts Service and Du Pont, respectively, prompted the development of the linkage approach rather than direct integration of programs. Users generally are not aware of the degree of independence between the two systems.

CS^4-Document System Search Relationship. Configuration of the combined CS^4-Document System for searching is shown in Figure 3. Unlike updating, no special linkage program is required. Input to the document system consists of complete questions composed of registry numbers retrieved by substructure searching and other complete questions containing compound and general terms. Records on the two tapes are of the same format.

Document System inquiries use Boolean logic in combining postings to terms specified on the inquiry cards. One type of Document System question produces no final output but can be referenced by one or more other questions. The inquirer who prepares the CS^4 inquiry may direct that answers be submitted in that form to the Document System. The collection of accession numbers which refers to one or more registry numbers may then be combined logically with accession numbers which reference general terms or other compound terms. A typical question of this sort combines specifications at the generic level of substructure and at the more specific level of compounds and general terms to retrieve document accession numbers as answers.

Registry Input Preparation. A chemical structure must be known to be registered. Some ambiguous word descriptions of chemicals in documents can be interpreted by more than one structure. Because each unique structure receives a registry number, it is possible to have more than one registry number for a chemical.

It is beyond the scope of this paper to discuss conventions and procedures for drawing unambiguous and reproducible structures. Input is assumed to be a chemical structure diagram, without any judgment of how well the structure represents the chemical.

Input to the CS^4 Registry System is in the form of connection tables which describe the structure of a compound in terms of atom-bond connections, element symbols, and bond type codes. Connection tables are prepared on a form by clerks using chemist prepared structure diagrams. The form consists of a molecular formula card and a connection table sheet with a piece of interleaved carbon paper. A sample form is shown in Figure 4. The clerk numbers atoms of the compound in any order. Each atom number row corresponds to an atom. The atom number and bond type are entered for each attached atom. The number of hydrogen atoms on non-carbon atoms is entered in the HYG column. Use of the UOXST

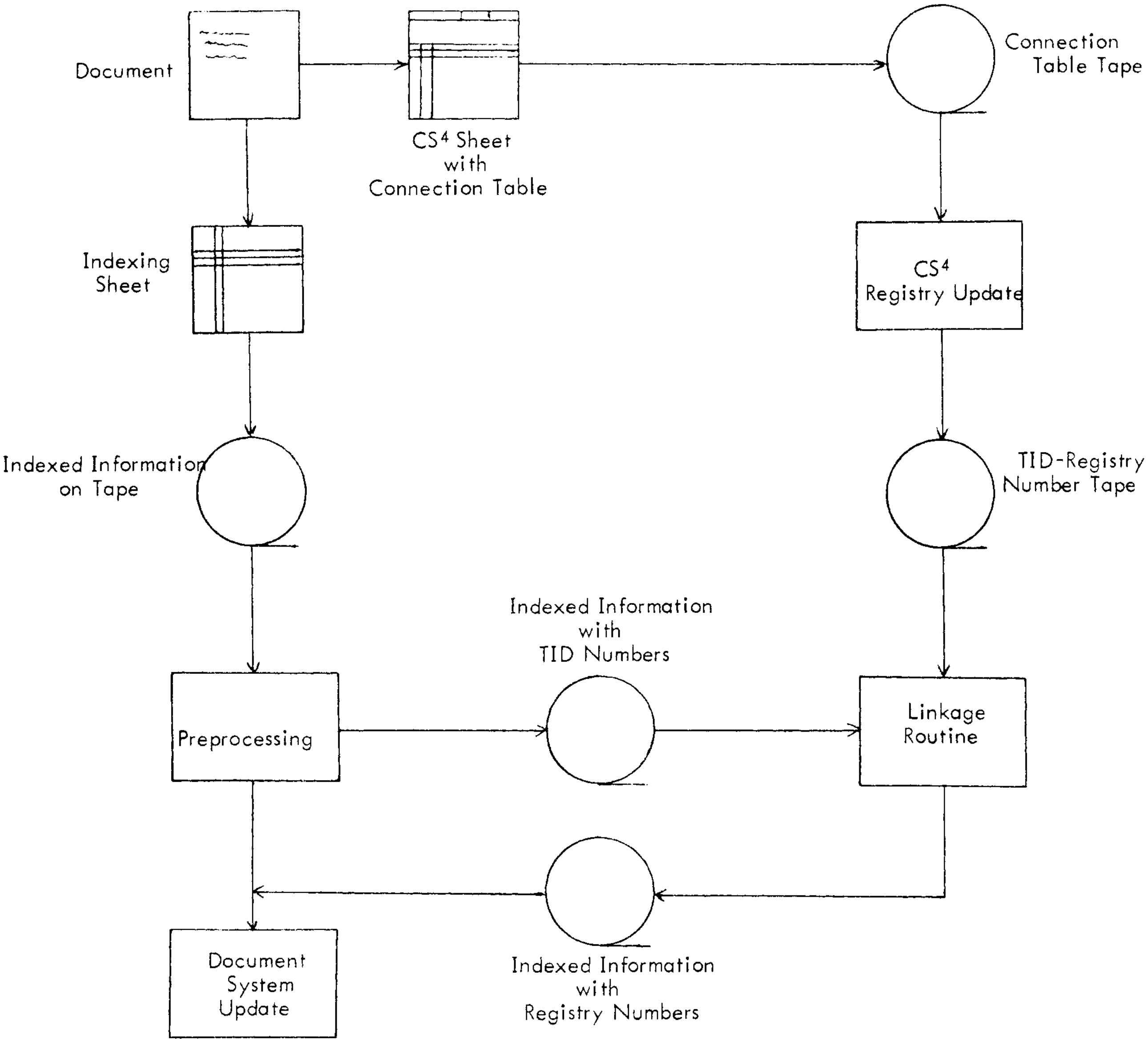


Figure 2. CS^4 Registry–Document System update relationship

and UCONN columns is discussed later. The CHARGE and UMASS columns are used to enter charges and unusual isotopic mass values, respectively.

The card is separated from the connection table sheet and filed by TID number. Later, the registry number is written on the card which is kept in a file arranged by molecular formula. The registry number is written on the connection table sheet, which is filed by registry number.

CS^4 Registry System. A general flow chart of the CS^4 Registry System is shown in Figure 5. Connection tables are keyboarded directly onto magnetic tape and sorted. A line-by-line edit is used to correct errors detected during the previous update. Input structures are then checked for errors. Error messages and the input connection tables are printed. Bonds in rings are detected and marked with a special set of codes.

A unique (canonical) form (numbering) is computed for each input compound. The structure is renumbered using an algorithm developed by Morgan (*6*). The algorithm ensures that the same numbering will be computed regardless of the order the atoms were numbered by the clerk. The input connection table is reformatted to a compact form requiring six characters per atom plus seven characters per ring exclusive of record identification characters. The compact table stores atom connections, bond codes, and element symbols in separate lists.

Separate non-polymer and polymer registry files are updated in a two-step process. If an input structure is found to be already in the file, the registry number of the existing record is extracted. Otherwise, a new number is assigned. Status records containing the TID number, new or existing registry number, and molecular formula are prepared for printing. A tape with TID and registry numbers for each input compound is produced for use by the Document System update.

A screen file containing information extracted from the connection table record for each compound is updated as the last step in each CS^4 Registry System update. Use of this file is discussed later.

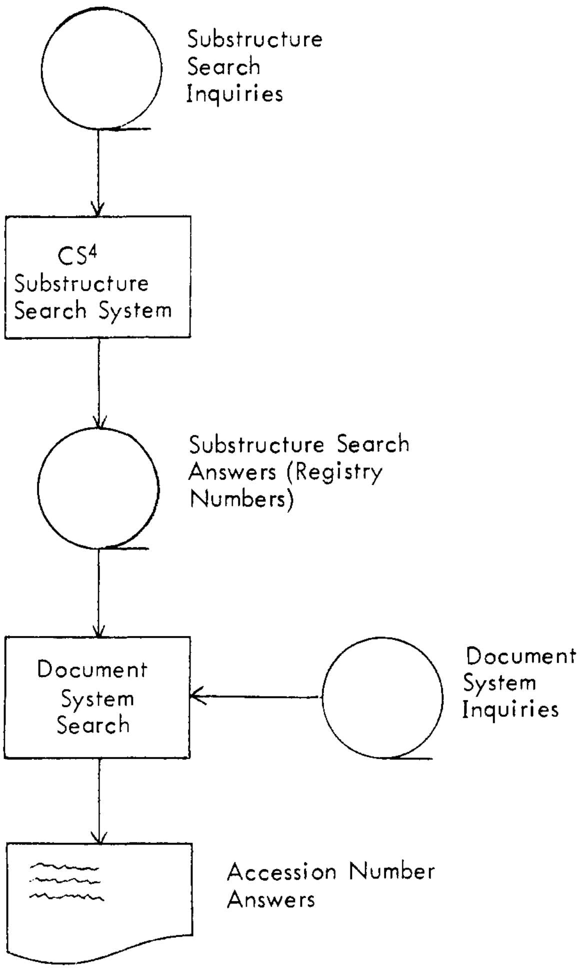

Substructure Search Inquiries
CS4 Substructure Search System
Substructure Search Answers (Registry Numbers)
Document System Search
Document System Inquiries
Accession Number Answers

Figure 3. CS^4–Document System search configuration

Search Input Preparation. Information chemists prepare inquiries for the CS^4 Substructure Search System. Each inquiry includes a substructure drawing, identification information, screens, and substructure coding. A typical drawing is shown in Figure 6. This will find all 2,2′-dichlorinated ketones.

Each atom of the inquiry is assigned an atom number. Execution of the inquiry is not dependent upon the order in which numbers are assigned. The substructure is composed of one or more groups to which a logical operator—AND, OR, or NOT—is assigned. The combination of groups defines the substructure inquiry.

Identification information tells the search system how to dispose of registry number answers. The inquirer may request that answers be printed, passed on to the Document System search, or both.

Screens prepared by the inquirer specify minimum features of a registry compound to justify attempting to iteratively map the substructure onto the compound. For example, if chlorine and nitrogen atoms are specified in the substructure, compounds not containing both elements should be rapidly eliminated from consideration. An inquiry may consist solely of screen specifications. Selection of a Screenout Option instructs the system to regard compounds meeting screen criteria as final answers.

Five types of screens are used. A sixth, employing machine generated fragment codes, is being developed but is not discussed in this paper. Screens permit specification of atom and ring counts, bond types and counts, molecular formulae, and element-bond-element triplets.

All screens are of the form shown in Figure 7. An example of construction of the molecular formula screen for an inquiry requiring the presence of at least two chlorine atoms is given. Boolean logic is employed in screening. Except for atom and ring counts, each screen item may have an AND, OR, or NOT relationship to other screen items. Items may be organized into groups which have an AND or OR relationship to other groups.

Substructure coding is a special form of connection table. Consecutive specification of inquiry atom numbers implies that the atoms are connected. A series of atom numbers traces a path which will be followed by the iterative search. The logical operator assigned to each group is specified on the first atom of each group. For the inquiry shown in Figure 6, a single path commencing at atom number one (chlorine) and terminating at atom number six (also chlorine) will trace the substructure. The inquiry would be coded as a single "AND" group. The path would go from atom number four back to atom number three before proceeding to atom number five.

CS^4 Substructure Search System. A general flowchart of the CS^4 Substructure Search System is shown in Figure 8. Inquiry forms are keyboarded directly onto magnetic tape. An edit program reprints each question and diagnostic or error messages. The Screenout program compares the screens prepared by the inquirer with information stored on the screen file. Only compounds which pass the screens will be iteratively searched.

The Substructure Search program consists of two main phases. In the first, each substructure inquiry is compiled into a series of machine language instructions which will actually perform the search. For example, if the inquiry specifies that a particular atom must be chlorine, the compiler will generate an alphabetic constant of "CL," an instruction to compare the contents of an area in the compound with the constant, and a branch to the next instruction of the inquiry if the comparison is made successfully.

Phase Two performs the substructure search. The compact connection table for compounds which passed the screens for one or more questions is expanded to a one atom per row connection table. Only items required by a question are expanded. For example, if a question did not specify bonds, the bond list would not be expanded. Actual machine addresses substituted for atom numbers are referenced directly by the compiled questions. After each compound is read and expanded, control is transferred to the compiled questions. If the iterative search is successful in mapping the inquiry onto the compound structure, the registry number and inquiry number are written on an output tape. An editing routine either prints registry number answers or formats the answers for the Document System search.

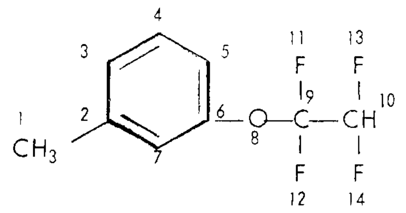
Computer Hardware and Software. With a few minor exceptions, all programs in the CS^4 Registry and Substructure Search Systems are run on an IBM 7010 Computer under the PR-155 Operating System. The computer has 100,000 characters of core storage, an IBM 1301 Disk Drive, and seven magnetic tape drives.

A serious problem encountered during installation of the CS[4] System was that the entire Document System operates under a special Du Pont operating system for the 7010. The main differences between the two operating systems are in input/output macro instructions, label routines, and supported file formats. Programs written for one operating system require a very large number of minor changes to be made compatible with the other operating system.

AUTOCODER, the IBM 7010 assembly language, is used for all major processing programs. COBOL is used for all output edits and the substructure search input

M.F.— $C_9 H_8 F_4 O_1$ | C— 41638 | TID— 123456

STRUCTURE—



NAME--

PAGE ___OF___

Atom	Elem.	Group	BOND	Att.	BOND	Att.	BOND	Att.	BOND	Att.	BOND	Att.	BOND	Att.	BOND	Att.	CHARGE	U-OXST	U-MASS	U-CONN	HYG	Atom
No.			1	1	2	2	3	3	4	4	5	5	6	6	7	7						No.
1			1	2																		1
2			1	1	1	3	2	7														2
3			1	2	2	4																3
4			2	3	1	5																4
5			1	4	2	6																5
6			2	5	1	7	1	8														6
7			1	6	2	2																7
8	O		1	6	1	9																8
9			1	8	1	10	1	11	1	12												9
10			1	9	1	13	1	14														10
11	F		1	9																		11
12	F		1	9																		12
13	F		1	10																		13
14	F		1	10																		14
15																						15
16																						16
17																						17
18																						18
19																						19
20																						20
21																						21
22																						22
23																						23
24																						24
25																						25

Figure 4. Connection table

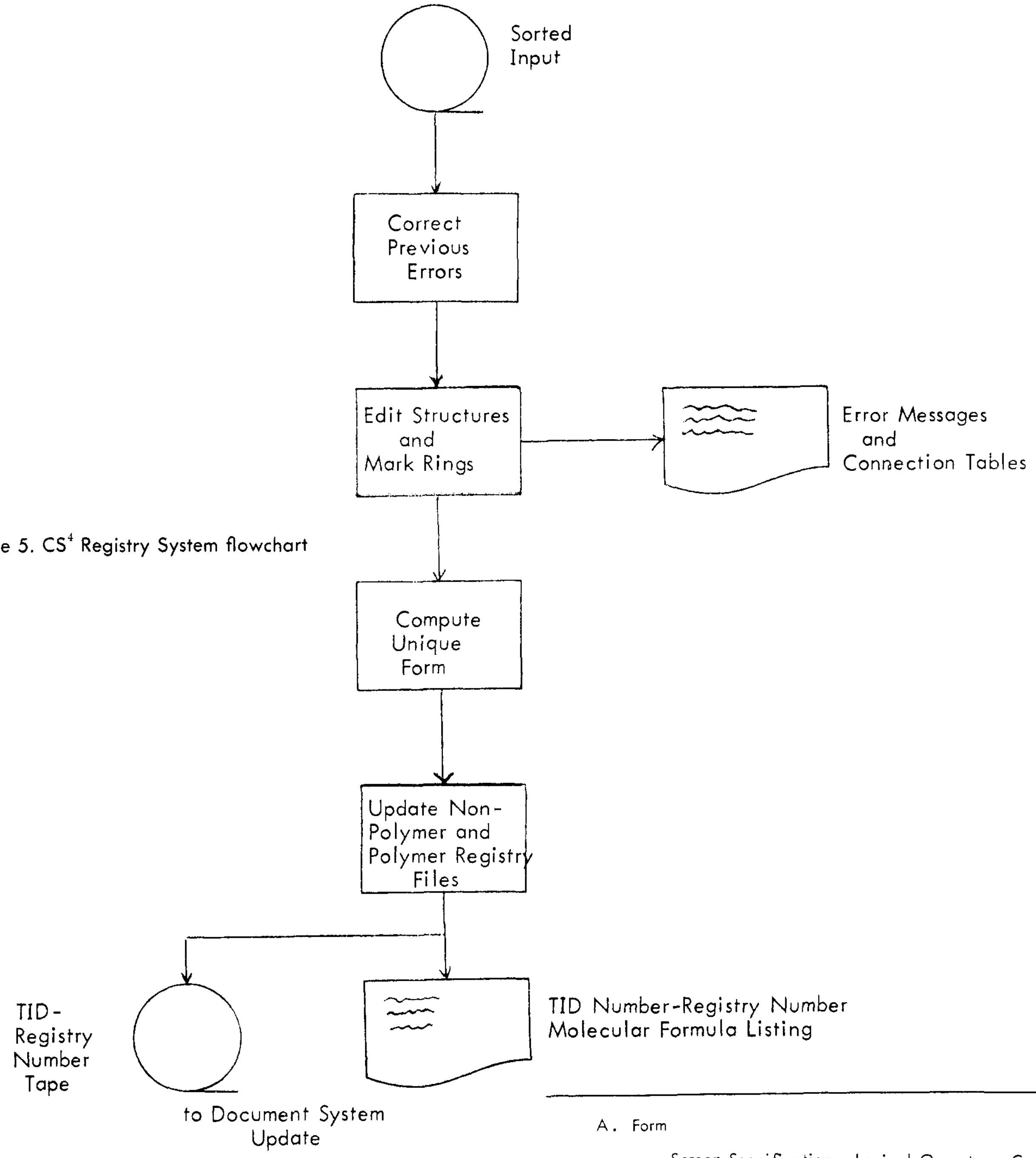

Sorted Input
Correct Previous Errors
Edit Structures and Mark Rings
Error Messages and Connection Tables
Compute Unique Form
Update Non-Polymer and Polymer Registry Files
TID-Registry Number Tape
TID Number-Registry Number Molecular Formula Listing
to Document System Update

Figure 5. CS⁴ Registry System flowchart

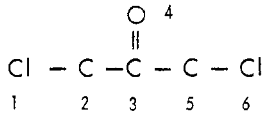


Figure 6. Substructure drawing

A. Form

Screen Specification - Logical Operator - Count

B. Example

Chlorine - Greater Than or Equal to - 2

Figure 7. Screen item descriptor

edit. When the system is run on IBM System/360 computers emulating the 7010, selected COBOL programs may be converted to the considerably more efficient 360 COBOL.

Operating Performance. The level of operating performance is a function of many factors, including people, money, and time. The CS⁴ System as it now operates represents only a starting point. The statistics discussed here will change as the system evolves.

The registry files currently contain approximately 55,000 non-polymers and 14,000 polymers, for a total of 69,000 compounds. The collection is growing at the rate of 18,000 annually.

Connection table preparation has averaged 5 minutes or $.183 per compound, based on $2.20 per hour of clerical time. The input error rate has averaged 25.6% since January, 1966. Approximately 5 minutes of clerical time are required to correct each error detected during the previous

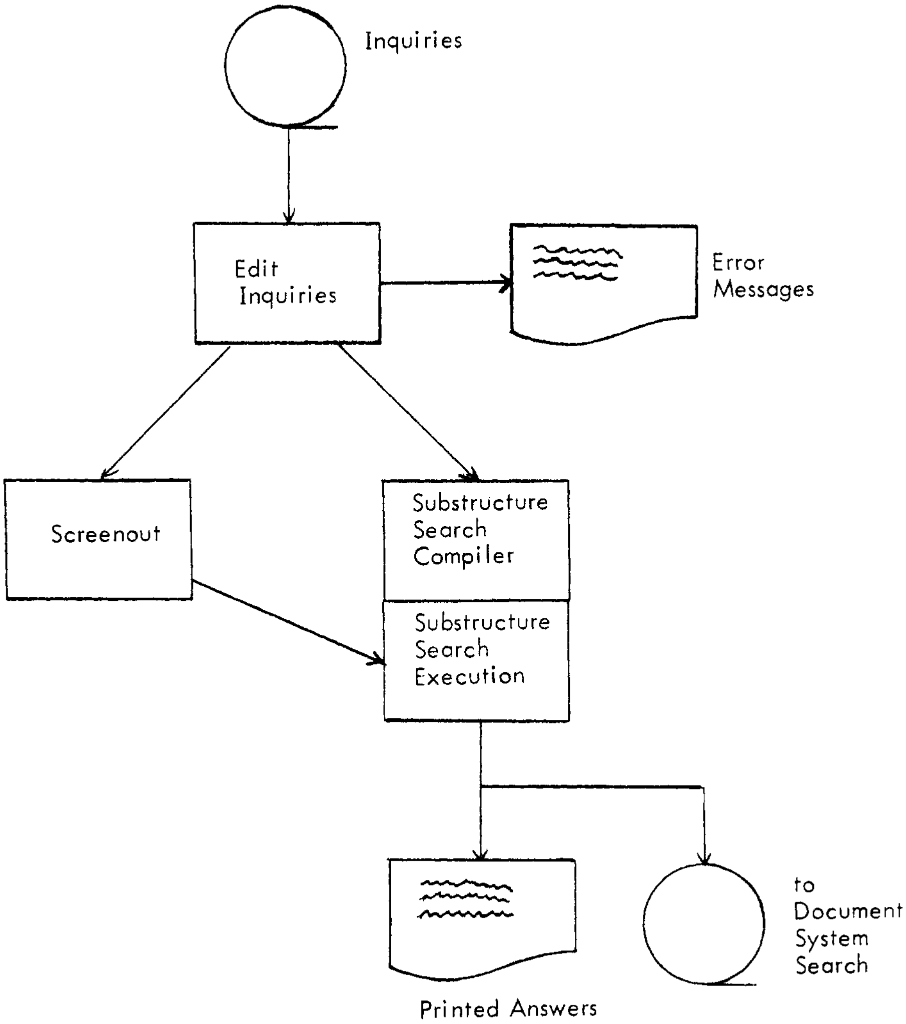

Inquiries
Edit Inquiries
Error Messages
Screenout
Substructure Search Compiler
Substructure Search Execution
to Document System Search
Printed Answers

Figure 8. CS[4] substructure search system

update. Error correction cost is equivalent to 1.25 minutes, or $.046 per input structure. Keyboarding directly to magnetic tape, without verification, costs $.275 per compound. All computer processing costs $.378 per compound, including update of the screen files. The latter costs approximately $.05 per compound. Total input cost per compound is $.882. Search computer costs are a function of the number of questions run in a batch. Searches of the non-polymer registry file cost $54. per search. Searches of the polymer registry file cost somewhat less. The above cost figures do not include technical time to prepare structures and searches and clerical efforts such as filing of molecular formula cards and connection table sheets.

Both registration and search costs reflect a very large amount of systems overhead. All work and master files use tape, and operator intervention is frequently required. Incremental cost per search question is only $28. Software and hardware limitations require each system to be divided into many more programs than would be necessary using more modern equipment.

EXTENSIONS OF THE CHEMICAL STRUCTURE SYSTEM

The original registry and experimental substructure search programs received from Chemical Abstracts Service handled only simple covalent compounds. Major modifications and extensions have been incorporated to permit registration and substructure searching of coordination compounds, complexes, and polymers. Several routines have been added and files modified to give each group using the CS[4] System the appearance of operating its own system.

Coordination Compound and Complex Registration. Changes to permit registration and substructure searching of compounds containing atoms with six or less non-hydrogen attachments have been made. The six-atom figure is a limitation of the current computer programs, not the technique.

Two quantities, called connection number and oxidation state, are defined for each atom. Connection number is the sum of bond values, including hydrogen, attached to an atom. Oxidation state is an attribute which describes the number of valence electrons donated to or accepted from a ligand. Connection number is essentially the same as "Valence" in the Chemical Abstracts Service Registry System. A standard connection number value has been chosen for each element. The standard oxidation state value is equal to the standard connection number value. Unusual connection number and oxidation state values for an atom, defined as different from the standard value, are entered in the UCONN and UOXST columns on the connection table sheet (Figure 4). Both values are entered if either is unusual, although only unusual values are stored in the compact connection table on the registry master file.

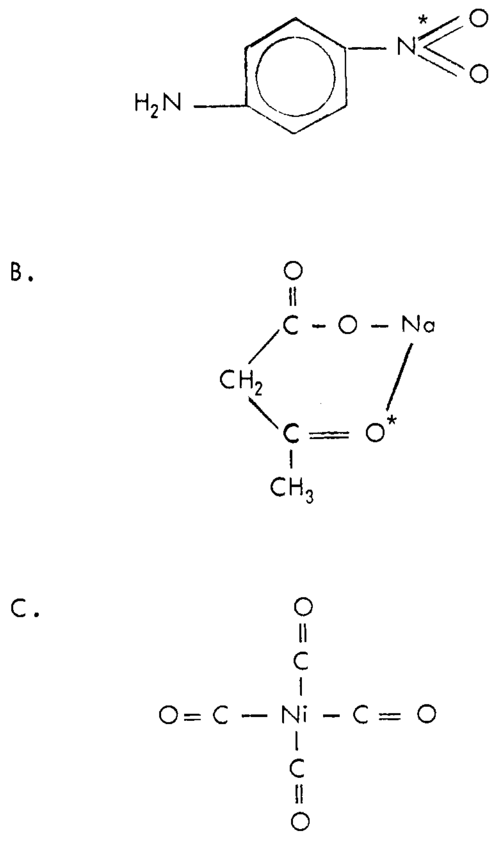
Three compounds structured using the connection number/oxidation state conventions are shown in Figure 9. Because nitrogen has a standard value of 3, the "N" of the nitro group of Compound A is shown with an unusual connection number of 5. The oxidation state of this nitrogen is considered to be standard. Compound B is a typical coordination compound. The sodium and oxygen atoms, attached by a single connection, each have an unusual connection number but standard oxidation state. Compound C is a complex. The carbon of each carbonyl group and the central nickel atom have unusual connection numbers and oxidation states.

In a substructure search, specification of unusual connection numbers or oxidation states or unusual configurations permit retrieval or exclusion of atoms in coordination compounds. The inquirer may also specify that only standard values are acceptable. Free radicals, carbenes, onium compounds, and amine salts are among other chemical classes structured using the connection number/oxidation state conventions.

Polymer Structuring. Linear polymers are structured for the CS^4 System using coded representations of one or more significant repeating units (SRU's) and end groups. An SRU is the smallest set of two or more atoms which, reproduced sequentially, represents the significant topological features of the polymer backbone. An end group is the group of atoms which terminates a polymer chain. The end group is usually different from a significant repeating unit. Additional conventions, not yet implemented, have been developed for grafted, cross-linked, and post-reacted polymers.

Each SRU is drawn as a series of atoms attached to a dummy atom. This dummy atom symbolizes that the attached significant repeating unit appears many times in the polymer chain. The symbol "A" was chosen for the dummy atom. A polymer may have any number of SRU's. A typical polymer appears in Figure 10. The ends of the SRU, —CH$_2$— and —CHCl— are connected to the dummy atom. The lower case letters "a" and "b" in Figure 10 are called bond descriptors. All bonds from the terminal atoms of an SRU to a dummy atom must be descripted. Bond descriptors indicate the ways in which the SRU may be linked to other SRU's, as described below. Bond descriptors are used only to indicate unlikeness. Assignment of either an "a" or "b" to any bond is permitted.

A copolymer is a polymer with more than one significant repeating unit. An example is shown in Figure 11. The SRU's derived from styrene and vinyl chloride each are



*Nitrogen: Conn. No. = 5
Ox. State = 3

Sodium: Conn. No. = 2
Ox. State = 1

* Oxygen: Conn. No. = 3
Ox. State = 2

All Carbons: Conn. No. = 3
Ox. State = 2

Nickel: Conn. No. = 4
Ox. State = 0

Figure 9. Connection number/oxidation state convention examples

Unusual values

Polyvinyl Chloride

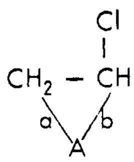


Figure 10. Typical polymer representation

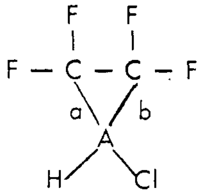


Figure 12. Polymer with end groups representation

Vinyl Chloride - Styrene Copolymer

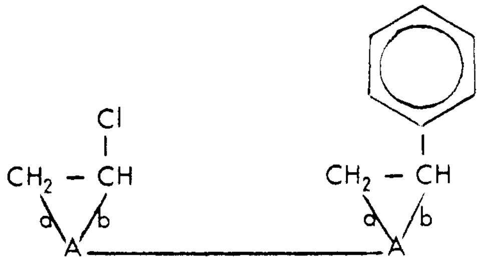


Figure 11. Typical copolymer representation

attached to a dummy atom and the dummy atoms are linked by a single connection. The substituted carbon atoms are each connected to a dummy atom by a single bond with a "b" bond descriptor.

No specific differentiation is made between "addition" and "condensation" polymers. In each case the structure is drawn as a series of significant repeating units. Permitted and non-permitted connections are indicated by bond descriptors. The relative abundance of different repeating units, molecular weight, or starting materials are not included in input information prepared for each polymer. These items, if known and indexed, are stored in the Document System files. The number of SRU's is not necessarily equal to the number of starting materials represented in the polymer backbone. Nylon 66 would have only one SRU made up of adipoyl and diaminohexamethylene segments.

The rules on tracing paths between SRU's are:

> The descriptors on bonds upon entering and leaving dummy atoms must be different.
>
> Unlimited jumping from dummy atom to dummy atom is permitted, provided that the rule above is observed.

In the copolymer illustrated in Figure 11, a path from the chlorine substituted carbon to the dummy atom could continue to the unsubstituted carbon of the same SRU or, after jumping to the dummy atom of the styrene derived SRU, to the phenyl substituted carbon. $—CH_2—$ to $—CH_2—$ and —CHCl— to —CH—Phenyl— connections would not be permitted.

End groups are drawn connected to dummy atoms by bonds without bond descriptors. Any number of end groups may be indicated. In copolymers, end groups may be attached to any dummy atom, although that location will not be maintained in the registry file. A polymer with end groups is shown in Figure 12. The chlorine and hydrogen atoms are attached to the dummy atom. Hydrogen is indicated on a coded connection table by a hydrogen count of 1 on the dummy atom.

Each end group is considered to be independent of all other end groups and significant repeating units. Path tracing is not permitted between end groups and SRU's or between end groups and other end groups.

Polymer Registration. Polymer structures are drawn on the CS[4] input forms. Connection tables are prepared and keyboarded. Connection table records are processed by the computer programs used to process non-polymers. The two types of input are intermingled in the input stream. After computation of the unique form, polymer structures are added to a separate master file tape.

Changes to the input edit and unique form generation programs were necessary for polymers. Besides checking each structure for errors, the edit program reformats and changes each polymer structure. Because of the changes, the structure representation stored on the master file differs from the form presented at input. No necessary information has been lost. All changes either facilitate generation of the unique form or simplify substructure search processing.

The first special polymer function performed during the edit phase is to associate each significant repeating unit with a dummy atom. Chemists are permitted to draw the vinyl chloride–styrene copolymer discussed earlier as shown at the top of Figure 13. This form would be converted to the form shown in the middle of Figure 13.

Each significant repeating unit is isolated by erasing all connections between dummy atoms. The effect of this operation is to place all SRU's at an equal level, removing the bias introduced by the order in which the SRU's are listed. For similar reasons, each end group is attached to a dummy atom J0 (J-zero). The SRU dummy atom symbol is translated from "A" to "Q0" (Q-zero). "Q" and "J" do not appear as the first character of any element symbols. The final form of the vinyl chloride–styrene copolymer is shown at the bottom of Figure 13.

The group which developed polymer coding conventions at Du Pont specified that the location of the dummy atom within the SRU of polymers containing a single SRU was not critical. Each form represented the same polymer. To permit generation of a unique form for such polymers, the Q0 dummy atom is isolated from the SRU. All possible locations for its insertion, defined as non-ring single bonds in the minimum length path between ends of the SRU, are marked as "g" bonds. The SRU of poly 1,4-butadiene, would appear as shown in Figure 14. The Q0 dummy atom is reinserted at the location of the highest "g" bond after computation of the unique

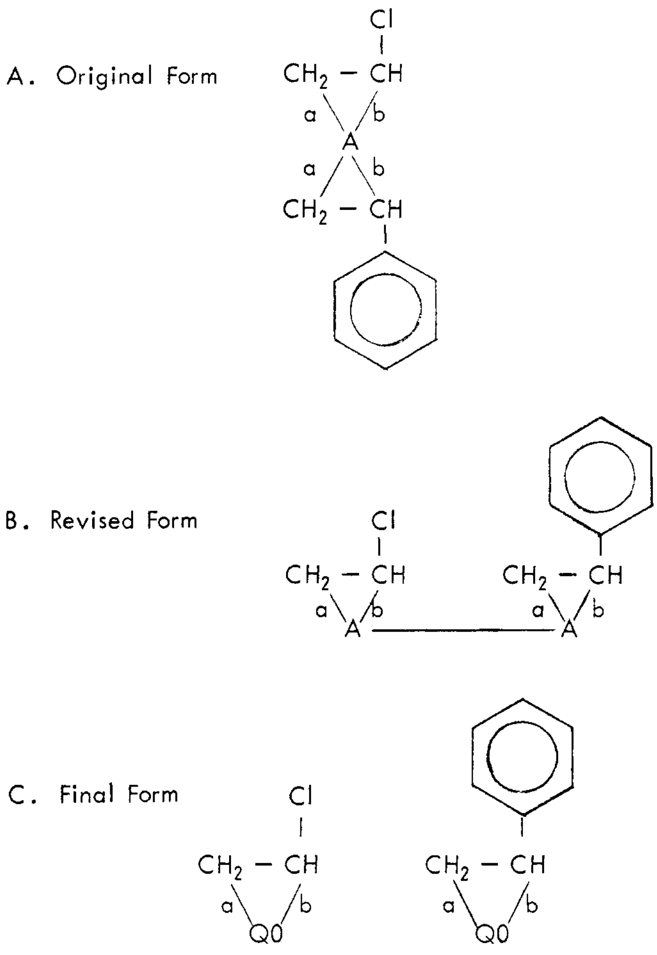


Figure 13. SRU isolation

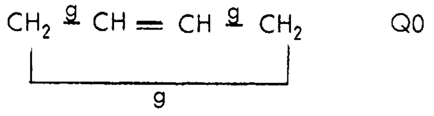


Figure 14. Single SRU polymer

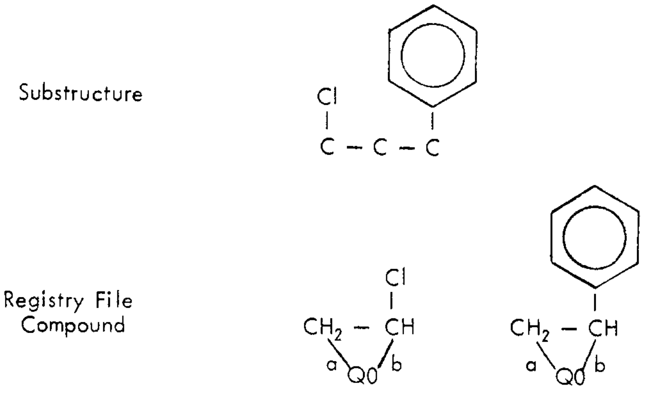


Figure 15. Polymer search example

form. Remaining "g" bonds are then changed to single bonds. The routine which detects and marks bonds in rings was modified so that bonds within significant repeating units, other than those which would be considered ring bonds in the corresponding non-polymer compound, remain noted as chain bonds.

Relatively few changes were required to the program which computes the unique form for each input compound. To compute a unique numbering without regard to bond descriptors, two copies of the bond list must be maintained. The second version, which has single bonds substituted for descripted bonds, is used in the computations. To ensure uniqueness including consideration of bond descriptors, the last descripted bond in the bond list is made a "b," and all other descripted bonds are reversed if necessary. A special procedure invoked if one or more SRU's are symmetrical is not discussed in this paper.

In all other respects, polymer structures are registered in the same manner as non-polymers. The final printout of TID and registry numbers and molecular formulae summarizes non-polymer and polymer registrations in separate sections of the listing.

Polymer Substructure Searching. The polymer version of the CS^4 Substructure Search System maps a sequence of atoms and bonds onto a structure, even if the complete sequence is not found within a single significant repeating unit. For example, the sequence shown at the top of Figure 15 will be found in the polymer at the bottom. During the mapping operation, whenever a descripted bond to a dummy atom is encountered, the program stores the bond descriptor letter, jumps if necessary to another Q0 atom, and continues mapping with the atom attached to the Q0 by the opposite type of descripted bond to that saved. Specification of a substructure as a string of atoms will be mapped into the polymer "across SRU's." The effect is the same as if all possible SRU attachment combinations were stored on the file.

The inquirer may restrict the search to atoms of a single SRU. For this option, it is not necessary to specify the location of the dummy atom within the SRU.

End group searches are coded identically to searches of SRU's. The inquirer must indicate that a substructure search is intended for mapping only against end group atoms. The search edit and substructure search connection table expansion routines translate the element symbol of each end group atom to another symbol. The search is then performed in a normal manner. None of the translated element symbols corresponds to normal element symbols, preventing undesired mapping.

Multiple File Usage. Four character positions of the text area of the registry record for each compound are used to indicate whether the Document System of each group which uses the CS^4 System has information concerning the compound. The contents of each position indicates one of three conditions: the group definitely has information concerning the compound, definitely does not have information concerning the compound, or the status of the compound is undefined.

The information stored in the text area was obtained by dumping a file of valid registry file references from the several compound number-document files. Several sorting procedures were required to merge references with the registry file.

On a going-forward basis, when a compound is first registered, the text area is set to signify that the group which submitted the compound has related information. The identity of the group is obtained by examination of the TID number. If the same compound is later submitted by a different group, the text area is appropriately updated.

Contents of the text area appear on miscellaneous file printout records. Major use is made during the substructure search screenout program. The group which originated each question is identified by a character on the inquiry identification record. The compound is rejected if the text position corresponding to the group indicates no related information is stored in the Document System File.

FUTURE PLANS

Implementation of the CS^4 System at Du Pont required a series of compromises to permit use of the available programs. It was desired to make the system operational as soon as possible and to gain experience with use of a topological chemical structure system. Anticipated changes to the chemical structure and related systems will occur over a period of years, until essentially all procedures and routines have changed.

We plan to experiment with keyboard-to-magnetic tape or keyboard-to-computer devices to capture chemical structure diagrams to replace the costly and error-prone clerical connection table coding process. The keyboard will be connected to a typewriter-like device or a cathode ray tube. Future substructure searches will be prepared and submitted in a similar manner. The CS^4 System, both programs and files, will be completely integrated with other machine systems in operation or to be started in the future.

SUMMARY

A chemical structure storage and search system operating at Du Pont has been described. The system integrates a topological file of connection tables with other computer systems which store information indexed from documents.

Du Pont registration and search programs are extended and modified versions of the Chemical Abstracts Service registry and experimental substructure search systems. Registration and searching of coordination compounds and complexes is performed using conventions which define a connection number and oxidation state for each atom. Polymers are described in terms of significant repeating units and end groups. Characters in each registry record indicate the presence or absence of information relating to the compound in document system files.

Future development of the system will result in full integration of chemical structure and document system files and programs and improved input and searching methods.

ACKNOWLEDGMENT

I would like to acknowledge the considerable contributions of Carolyn C. Benton, Rodney T. Jones, Jr., and Diane M. Trimble in the programming, installation, and testing of the systems described in this paper.

LITERATURE CITED

(1) Gluck, D. J., "A Chemical Structure Storage and Search System Developed at Du Pont," J. Chem. Doc. **5**, 43 (1965).

(2) Tate, F. A., H. L. Morgan, D. P. Leiter, and R. E. Stobaugh, "A Mechanized Registry of Chemical Compounds," presented at the 1965 Congress of the International Federation for Documentation, Washington, D. C., October 12, 1965.

(3) Leiter, D. P., Jr., and H. L. Morgan, "Installation and Operation of a Registry for Chemical Compounds," J. Chem. Doc. **5**, 238 (1965).

(4) Cossum, W. E., M. L. Krakiwsky, and M. F. Lynch, "Advances in Automatic Chemical Substructure Searching Techniques," *ibid.*, p. 33.

(5) Krakiwsky, M. L., R. W. White, and W. C. Davenport, "Searching for Subsets in Machine Records of Chemical Structures at the Chemical Abstracts Service," presented at the 1965 Congress of the International Federation for Documentation, Washington, D. C., October 12, 1965.

(6) Morgan, H. L., "The Generation of a Unique Machine Description for Chemical Structures," J. Chem. Doc. **5**, 107 (1965).