The Chemical Abstracts Service Chemical Registry System. V. Structure Input and Editing

ANTONIO ZAMORA* and DAVID L. DAYTON

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received April 22, 1976

The Chemical Abstracts Service Chemical Registry System is a computer-based system that uniquely identifies chemical substances on the basis of their structural features. The Registry System currently contains records for more than 3.4 million different substances. Although there are several ways of entering data into the CAS Chemical Registry System, the majority of the transactions for storage or retrieval of data are entered using chemical typewriters. This paper describes the conventions used for typing structure diagrams, the computer programs which decode the typed structure into a connection table, and the edits which are performed during decoding.

INTRODUCTION

The Chemical Abstracts Service (CAS) Chemical Registry System[1] uses structure typewriters as its primary means of input. Although interactive graphic devices provide alternative methods of entering data[2,3] to the Registry System, the economy and the flexibility of the structure typewriter system are responsible for its predominant use.

When the CAS Chemical Registry System was established in 1964, structures were input as redundant connection tables.[4] This type of input was slow and tedious since for each atom of a structure all the attached atoms and the corresponding bonds had to be described. In 1965 CAS started experimenting with structure typewriters which produced punched paper tape. Computer programs were developed to generate connection tables from the typed structure data, but it was not until after the typewriters were replaced with similar machines that generated magnetic tape in 1967 that structure typewriters were routinely used to input data to the Chemical Registry System.[5] The programs used to support the structure typewriters in this early system had limited capabilities which required a substantial number of structures to be entered in connection table mode.

In 1972 a key-to-disk system was installed to replace the magnetic tape recording units. At this time all the programs supporting the chemical typewriters were changed to unify all sources of input for the Chemical Registry System. This was done to make it possible for structure typewriters to be used for polymers and incompletely defined substances, which had previously required connection table input, and to increase productivity by simplifying structure typing conventions and by improving the diagnostic facilities. This paper describes the input procedures and edits introduced in 1972.

OVERVIEW OF CURRENT SYSTEM

The CAS chemical typewriter system consists of a Varian 620i minicomputer supporting twelve chemical typewriters in a standard key-to-disk operation. The typewriters are specially modified IBM 735 Selectric typewriters with reverse index and a special type element. The keyboard arrangement is that described by Mullen for the Shell Chemical Typerwriter.[6] Figure 1 shows one of the chemical typewriters in use; the minicomputer, a magnetic tape unit, and a disk drive can be seen in the background. Figure 2 illustrates the characters available on the typewriter. One of the most frequently used characters is the large dot, which represents a carbon atom with an appropriate complement of hydrogens. Thus, when a single bond is attached to the large dot, the large dot represents a carbon atom with three hydrogens. The use of some of the other characters will be explained in more detail below.

In the CAS Chemical Registry System, data typed on the chemical typewriter are preceded by mnemonic data identifiers; for example, REG identifies Registry Numbers, STR identifies structures, etc. These data identifiers are used by the minicomputer to assure that individual data items are entered in the proper sequence. The minicomputer also uses the data identifiers to monitor data elements which must meet special constraints. Registry Numbers, for instance, are checked to verify that they have correct check digits. Check digits are error-detecting codes derived algorithmically from the preceding digits, which must agree with the data entered. Using the minicomputer in this way ensures that the input data are reasonable and prevents more costly recycles because the data can be corrected during initial input. The speed and the memory capacity of the minicomputer impose limitations on the types of edits which can be applied without degrading the performance of the typewriters as input devices. Therefore, the editing tasks requiring a large amount of memory or computer time take place in the main computer (an IBM 370/168).

At the end of each day a magnetic tape, generated by the key-to-disk system and containing all the transactions for the Registry System, is routed to the main computer. Programs at the main computer construct connection tables from the data recorded by the typewriters and edit the tables before sending them to the Registry System.

PROGRAM INTERPRETATION OF TYPED STRUCTURES

The chemical typewriter input program (TIP) creates a connection table from the string of characters generated by the typewriter. The chemical typewriter stores characters in the sequence in which the keys are activated. The resulting character string is translated into a two-dimensional image by a program which uses the space, backspace, index, and reverse index characters to determine the coordinates of other characters. A set of input conventions makes it possible to overcome the limitations in the character set of the chemical typewriter. In addition, the mnemonic data identifiers allow the typewriter operator to enter whole structures or portions of structures in connection table format for structures which are cumbersome to type. "Shortcut" notations for commonly occurring chemical groups and a "ditto" feature also help to increase the rate at which structures are typed. Chemical shortcuts will be discussed more extensively below.

Since no constraints are placed on the typewriter operators regarding the size of the typed structure, TIP performs an initial scan of the string to determine the amount of core storage required to build a two-dimensional image of the structure. The program allocates the necessary storage, creates the two-dimensional image, and then proceeds to analyze the stored image. The stored image is examined using a formal

* To whom correspondence should be addressed.

Figure 1. The Chemical Abstracts Service structure typewriter system.

a b c d e f g h i j k l m n o p q r s t u v w x y z

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

1 2 3 4 5 6 7 8 9 0 ⫽ / ‖ ● | – ≈ \ ⑊

( ) + / – o < : > • ◦ , * . ↗ ® ↖ \

Figure 2. Characters available on the chemical typewriter.

description which specifies the valid relationships possible between the bonds, elements, chemical shortcuts, abnormal masses, charges, and valences in the stored image.

An initial attempt to derive a formal description using Backus Naur Form[7] (BNF) was made, but it proved inadequate owing to the two-dimensional spatial relationship of the basic entities. A special formal description for the stored two-dimensional image had to be developed. This formal description is a grammar that is applicable to any structure which can be typed with the chemical typewriter and defines the valid relationships of atoms, bonds, and other structural entities in a two-dimensional space. The formal description defines "nodes" as alphanumeric strings which may contain embedded hyphens when the hyphens are preceded and followed by alphabetic characters. The "environment" of a node is defined to be the spaces immediately surrounding a node. Charges, masses, and valences in the environment of a node are assigned to the node if they are not ambiguous and if they are appropriate for the node. For instance, a mass is valid in the environment of a node if the node is an element symbol but not if the node is a chemical shortcut. In addition, the mass may not be in the environment of more than one node. The formal description also specifies the conditions under which a bond is attached to a node. Since some chemical shortcuts may have up to three distinct attachments, it is necessary for TIP to record the character of the shortcut to which a bond is attached. This is particularly important for chemical shortcuts which represent asymmetrical groups of atoms. Figure 3 illustrates the distinctions which are made for the amino acid shortcuts *Ala* and *Gly* depending on how the bonds are attached.

Approximately 12 000–15 000 structures are input through the CAS chemical typewriter system every week. Input rates vary substantially depending on the skill and experience of the operators. An operator with four weeks of experience may type approximately 20 structures per hour, whereas a well-trained operator may consistently type from 80 to 100 structures per hour. In 1975 an average rate of 51.4 structures per hour was recorded for all the structure input operators. Table I summarizes structure input rates from 1972 to 1976. Interpretation of Table I is complicated because the rates are affected by staff turnover and by a latency period which occurred between the time new typing conventions were introduced and the time benefits from the new conventions were observed. The general upward trend of the rates is clear, however. The input rate reported for 1976 will probably be somewhat lower when averaged over the entire year, but it reflects the recent introduction of simplified typing conventions for structures with charged atoms as well as a greater percentage of experienced operators. While the input rate has increased, the error rate of the input transactions has decreased from 9.3% in 1971 to 3.9% in 1975.

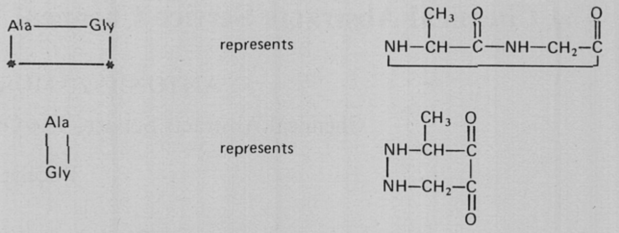


Figure 3. Chemical shortcut representations.

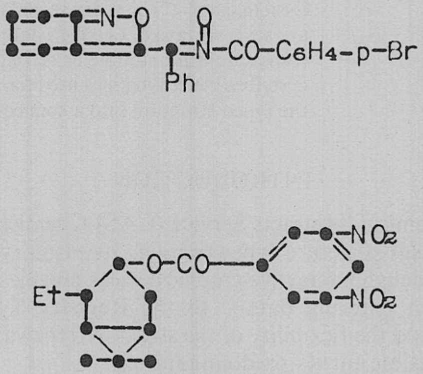


Figure 4. Examples of flattened and conventional rings.

Table I. Structure Input Rates from 1972 to 1976

Period	Structures Per Hour
Jan.-March 1976	67.3
1975	51.4
1974	36.1
1973	38.4
1972	20.4

STRUCTURE TYPING CONVENTIONS

The basic purpose of typing conventions is to speed up typing of a structure, to extend the capabilities of the structure typewriter, and to represent structures in a form easily recognized by chemists. The readability of the typed structure is important for applications where the typed structure image is to be routed directly to a graphic display device. For applications where the graphic representation of the typed structure is not important, it is possible to type structures faster by flattening the rings of the structure. This technique reduces the number of index and reverse index operations (up and down movements of the platen) and results in greater usage of the horizontal bonds, which are close together in the keyboard. Figure 4 illustrates flattened and conventional rings.

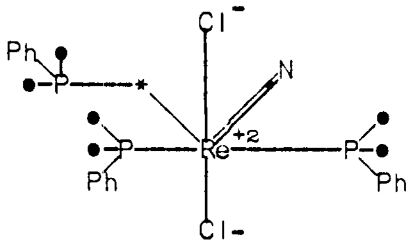


Figure 5. Examples of extended bonds, bent bonds, and charges.

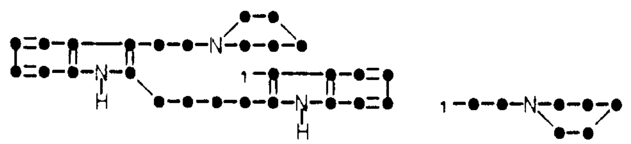


Figure 6. Use of linking nodes.

Display quality is not necessarily sacrificed by flattening the rings to achieve higher input rates because CAS has recently developed a Structure Display System which generates structure diagrams from the connection table data stored in the Chemical Registry System. Substances are displayed with the standard ring shapes and orientations even when the structures are input with distorted rings. This is possible because a reference file of approximately 18 000 ring shapes provides the templates for displaying structures containing rings.

Overtyping Conventions. Overtyping consists of typing one character on top of a previously typed character. A character may be overtyped by backspacing and striking a different character. However, any other combination of the space, backspace, index, or reverse index keys may be used to reposition the structure for the overstrike. Stereochemical representations for display on graphic devices may be obtained by overstriking single bonds with "a" or "b" to indicate alpha or beta bonds. Structural repeating units are delimited by single or double bonds overtyped by parentheses as illustrated in Figure 9. Triple bonds are represented by overtyping double bonds with "center dots" (not to be confused with "carbon dots" or periods; see Figure 2). Any character is replaced by a blank when overtyped with a "delete circle" (not to be confused with zero or lower case o; again see Figure 2). All other overstrikes will result in replacement of the original character except that the characters representing space, backspace, index, and reverse index do not overlay any other character.

Bonding Conventions. The main function of the overtyping conventions is to extend the character set of the chemical typewriter. The function of the bonding conventions, however, is to simplify the work of the operator by reducing the constraints imposed by the chemical typewriter. Bonds can be extended by joining them end-to-end with other bonds of the same type. Overtyping any one of the component bonds of an extended bond by a valid overstrike is equivalent to overtyping all the bonds. Bonds may be bent by attaching them to an asterisk (see Figure 5). When four bonds are attached to an asterisk the bonds are assumed to cross. Atoms which require many attachments can be extended by using any number of "greater than" signs (>) to increase the number of positions to which a bond may be attached. For instance, Re>>>> could have been typed in place of the Re in Figure 5 to facilitate bonding since any bond to a "greater than" sign is interpreted to be attached to the element preceding it. Complex structures may be typed in separate pieces and connected by linking nodes. Linking nodes (see Figure 6) are numerical, logical connectors which occur in pairs and which are deleted by the program after the connection table is built from the typed structure.

-Ph	-CN	O_2N-
-CO-	-NO_2	-Bu
-NH-	-o-C_6H_4-	TMS-
-Et	H_2N-	t-Bu-
-OH	-Pr-i	HO_2C-
-NH_2	-m-C_6H_4-	-HN-
HO-	-Pr	i-Pr-
HCl	-Bu-t	-SO_3H
-CO_2H	-SO_2-	-Gly-
-Ac	OXALATE	-Pro-

Figure 7. Most commonly used shortcuts.

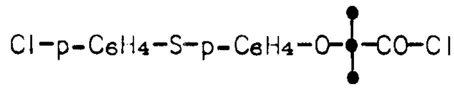


Figure 8. Extensive use of shortcuts creates almost linear structures.

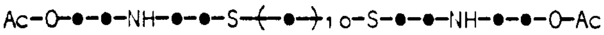


Figure 9. Example of a structural repeating unit representing a chain of ten carbon atoms.

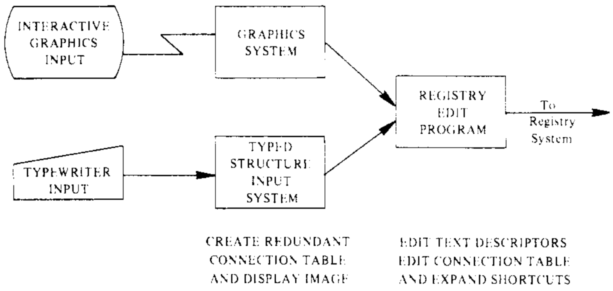

INTERACTIVE GRAPHICS INPUT
GRAPHICS SYSTEM
REGISTRY EDIT PROGRAM
To Registry System
TYPEWRITER INPUT
TYPED STRUCTURE INPUT SYSTEM
CREATE REDUNDANT CONNECTION TABLE AND DISPLAY IMAGE
EDIT TEXT DESCRIPTORS EDIT CONNECTION TABLE AND EXPAND SHORTCUTS

Figure 10. Flowchart of boxes.

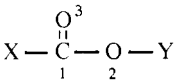
Other Conventions. Overtyping and bonding conventions are supplemented by various other conventions which aid the operator of the chemical typewriter as well as the chemists which prepare the input documents containing the chemical structures. A "ditto" feature makes it possible to save a partial, commonly occurring substructure and to reuse it later by recalling it. Approximately 200 chemical shortcuts are available to speed up typing and to allow display of chemical structures in a form more suitable for chemists. Shortcuts may represent common groups of atoms or complete compounds. Atoms and shortcuts are required to be bonded explicitly to each other. Figure 7 lists the most commonly used chemical shortcuts in descending frequency of use from left to right. Extensive use of shortcuts increases input rates by creating highly linear structures (see Figure 8). Long chains of atoms may be represented as repeating groups (see Figure 9).

EDITING

The major function of the Chemical Structure Input program is to create a connection table from the data generated by the chemical typewriter. In addition, edits are performed to assure that the graphic representation of the typed structure is correct. This means that diagnostics are generated when atoms are attached to themselves, when it is not possible to determine to which atom a mass or a charge belongs, etc. The Structure Input program, however, does not edit the structure for chemical sense. This function is per-

S67 SHCT CO_2,((-,1,1),(-,3,2))

S68 SHCT O_2C,((-,1,2),(-,3,1)),SYN=S67



S67 ATOM 1,C,(-X,=3,-2)

ATOM 2,O,(-Y,-1)

ATOM 3,O,(=1)

Figure 11. Shortcut table format and the corresponding shortcut expansion.

formed by the Registry Edit program.

Figure 10 illustrates that the Registry Edit program is the central point where the structures are edited regardless of whether the structures are input using chemical typewriters or interactive graphics terminals. The basic purpose of all editing functions is to prevent invalid data from entering the data base.

The Registry Edit program has editing functions which encompass all aspects of the structure data. Text descriptors describe special features of chemical substances which cannot be deduced from their topology, such as stereochemistry. These descriptors are rigorously edited. A BNF description of the format and content of text descriptors provides the basis for detecting format errors for isolated substances. In addition, before a substance record is added to the Registry File, text descriptors are checked for consistency with those already on file. Conflicts have to be resolved by a chemist. Text descriptors have been described in greater detail by Blackwood et al.[8]

The Registry Edit program requires a molecular formula to accompany every chemical structure. The standard format of a molecular formula is also defined by a BNF description. This makes it possible to compare the molecular formula recorded by a chemist against the standard format and against the formula derived by the computer from the typed structure. Whenever the formulas disagree, the structure is rejected by the edit program. If manual analysis determines that the error is due to a transcription error, the structure is retyped; otherwise, it is routed to a chemist for resolution.

Various edits which are performed on individual atoms and groups of atoms of a structure have been described by Dittmar et al.[1] These edits apply to atoms with nonstandard valences, charged atoms, and coordination atoms.

One function of the edit program is to expand chemical shortcuts and repeating groups into the atom-by-atom representation required by the CAS Chemical Registry System. Although this function could have been performed during the analysis of the typed structure, it would have forced other input alternatives (such as graphics input) to duplicate this task.

All the valid shortcuts are defined by a set of internal tables in the edit program. The tables describe each shortcut completely and can easily be modified to accommodate additional shortcuts. Figure 11 illustrates the internal table format for the shortcuts $-CO_2-$ and $-O_2C-$. Each chemical shortcut has an identifier (e.g., S67 for CO_2), the shortcut itself, and a set of attachment entries. The number of attachment entries corresponds to the number of bonds attached to the shortcut. Each attachment entry has three components: bond type, character of the shortcut to which the bond is attached, and the atom number to which that attachment corresponds. Thus, the entry (-,3,2) indicates that a single bond attached to the third character of the shortcut is equivalent to attaching the single bond to atom two of the shortcut expansion. Shortcut expansions are in the form of redundant connection tables. The same shortcut expansion may be used for two separate shortcuts by identifying them as synonyms. It should be noted that the direction of the bonds attached to the shortcuts is unimportant, but the character to which the bond is attached makes a significant difference and is carefully checked by the program.

The shortcuts -Ac and -Pr require a deviation from the normal shortcut expansion procedure since these symbols may also represent the elements actinium and praseodymium. Whenever the molecular formula associated with a structure contains the symbols Ac or Pr, any occurrence of the symbol in the structure is assumed to be the element rather than the shortcut. Thus, the shortcuts -Ac and -Pr may not be used in actinium or praseodymium compounds. This limitation is insignificant since such compounds are rare.

CONCLUSION

This paper has presented an overview of some of the most fundamental activities of the CAS Chemical Registry System, namely, data input and verification. These processes are responsible for safeguarding the integrity of the data base. The initial objectives of unifying the sources of input and making better use of human resources have been achieved. This has been demonstrated by a significant reduction in the error rate, and by structure input rates which have more than doubled since the system was installed. In addition, the greater flexibility allowed by the system for data management transactions provides convenience which is difficult to quantify.

ACKNOWLEDGMENT

The authors would like to recognize the efforts of Gary A. Davis and Ronald T. Khouw in programming the Structure Input program and the Registry Edit program, respectively. The authors also wish to thank Anne W. Ryan for assistance with the chemical shortcut definitions and for providing input and error rate statistics. Thanks are also due to the National Science Foundation for providing substantial assistance in the development of these programs under Contract C656.

LITERATURE CITED

(1) P. G. Dittmar, R. E. Stobaugh, and C. E. Watson, "The Chemical Abstracts Service Chemical Registry System. I. General Design", *J. Chem. Inf. Comput. Sci.*, **16**, 111 (1976).

(2) N. A. Farmer, F. A. Tate, C. E. Watson, and G. A. Wilson, "Extension and Use of the CAS Chemical Registry System", CAS Report No. 2, April 1973, pp 3–10.

(3) N. A. Farmer and J. C. Schehr, "A Computer-Based System for Input, Storage, and Photocomposition of Graphical Data", Proceedings of the ACM, Vol. 2, Nov 1974, pp 563–570.

(4) D. P. Leiter, H. L. Morgan, and R. E. Stobaugh, "Installation and Operation of a Registry for Chemical Compounds", *J. Chem. Doc.*, **5**, 238–242 (1965).

(5) R. G. Hefner, P. M. Keesecker, and D. F. Rule, "Keyboarding Chemical Information", *J. Chem. Doc.*, **7**, 232–236 (1967).

(6) J. M. Mullen, "Atom-by-Atom Typewriter Input for Computerized Storage and Retrieval of Chemical Structures", *J. Chem. Doc.*, **7**, 88–93 (1967).

(7) P. Naur, Ed., "Revised Report on the Algorithmic Language ALGOL 60", *Commun. ACM*, **6**, 1–17 (1963).

(8) J. E. Blackwood, P. M. Elliott, R. E. Stobaugh, and C. E. Watson, "The Chemical Abstracts Service Chemical Registry System. III. Stereochemistry", presented to the Division of Chemical Information, 170th National Meeting of the American Chemical Society, Chicago, Ill., Aug 24, 1975.