

## On-Line Storage and Retrieval of Chemical Information. I. Structure Entry<sup>†</sup>

CARLOS M. BOWMAN, LINDA C. DAVISON, and PATRICIA F. ROUSH\*

Systems Research, The Dow Chemical Company, Midland, Michigan 48640

Received February 20, 1979

An interactive program has been developed for entering, verifying, and storing chemical structures encoded in the Wiswesser Line Notation. The program calculates a molecular formula from the WLN for checking and then generates a bit string fragment code and connection table of the atoms in the structure. The encoding and entry of the WLN using a CRT has significantly improved the speed of the total compound registration process.

The Wiswesser Line Notation<sup>1</sup> provides a unique, comprehensive, and compact storage medium for chemical structures. From the WLN, other descriptive structural information, such as fragment codes and connection tables, can be generated by computer. These facts have been well recorded in the literature. Notation analysis is a complicated procedure requiring large programs and a good size machine. At Dow, notation verification and generation of structure data have been carried out in batch mode by several different programs.<sup>2,3</sup> Structures were encoded in the WLN from 4" × 6" cards on which was recorded other descriptive information about the compound, such as name, molecular formula, source, and structural drawing. These other data (except the structure drawing) had been entered into a separate part of the compound data base through another series of programs some 6 weeks before reaching the encoder. After notations were verified by the computer, another delay was involved as the erroneous notations were corrected and reentered. Since the master WLN file was updated every quarter, this was a delay of yet another 3 months. In all, a period of 2–5 months could elapse between the time a compound was received for registration and was subsequently available for structure searching.

To rectify the above situation, the entry procedure described in this paper was developed. It was decided to incorporate the encoding process into the compound registration process. Structures would be encoded into the WLN as they were received at the Central Report Index. An on-line WLN analysis program would allow for immediate verification of the notation, and would simultaneously generate a fragment and connection table record for substructure searching and other applications. In this way, all data describing a compound would be entered at once, with no significant time lag among the various steps in the registration process.

### WLN ANALYSIS PROGRAM

A WLN analysis program was developed to scan the notation character by character and generate the following data:

- (A) A molecular formula, to be compared with a formula entered by the encoder. Matching formulas are commonly considered a reasonable check of notation accuracy.
- (B) A bit string fragment code and a connection table to be used for substructure search and other applications. Some notation errors are also detected while determining atom connections for the table.

The final version of the program operates in interactive mode on a Hazeltine 2000 cathode ray tube interfaced with an IBM 370/158 (TSO, MVS operating system) (Figure 1). An interim batch version preceded the on-line one and was

used for debugging purposes. The WLN, fragment, and connection table records are stored in temporary disk files and then updated biweekly onto master tapes. The master tapes are subsequently loaded to disk for searching. In the near future, the program will interact directly with the master files on disk. Connection tables are stored on a removable disk pack, which can be mounted upon request. About 200 compounds per week are processed. The current data base contains about 150 000 compounds.

The program itself is written in PL/I and is derived from a basic checking program developed by Leo et al.<sup>4,5</sup> at Pomona College and purchased by Dow. The original program performs a detailed analysis of the notation, including the generation of a connection table for ring systems. It also calculates a molecular formula. The Dow modification of this program retains the ring analysis procedures as well as those which reformat the input and output molecular formulas. The remainder of the program, which analyzes the backbone of the notation, was completely rewritten to accommodate the inclusion of fragment code and connection table generation algorithms.

As the program scans each character of the notation, it adds the atoms present to the molecular formula count, and records a connection table atom value and one or more fragmentation codes for that character. For the more common WLN symbols (such as W, N, G, Q, V) a section of in-line code is assigned to derive these data for that symbol. The data for less common symbols is derived completely from a table look-up (e.g., scanning an array) of the symbol value, connectivity, and fragment code number. Numeric charge citations are also analyzed for connectivity. Data for non-MANTRAP addends are included in the data for the principal compound, although no connection among these species is recorded unless so indicated (e.g., ionic charge). Analysis of some multi-valent elements is handled by the program (primarily sulfur and phosphorus), but most are not successfully analyzed. Other categories of compounds not handled are polymers, inorganic and coordination compounds, and multiplied notations (methyl contractions, however, are still used at Dow and are correctly interpreted by the program). For these compounds, the notation is proofread and stored, but no other data are generated (about 3% of the file). It is hoped that these compounds will eventually be incorporated into the system through refinement of the algorithm and development of a technique for manually entering the data.

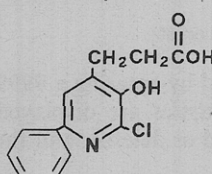
MANTRAP suffixes pose a special problem to the program because they actually represent two or more distinct chemical entities (e.g., Mixtures, Alternate structures, Tautomers, Reaction products, Addends, Polymers) linked together as one compound. It was decided that each component of a suffixed notation would be analyzed separately and that separate fragment and connection table records would be created for them. If any one of these compounds fails to pass the checking process, the entire notation is rejected. Individual WLN's

<sup>†</sup> Presented before the Division of Chemical Information, 175th National Meeting of the American Chemical Society, Anaheim, Calif., March 15, 1978.



Figure 1. Data entry on the Hazeltine 2000 CRT.

C<sub>14</sub>H<sub>12</sub>ClNO<sub>3</sub> DR-0000-00011  
 4-pyridinepropanoic acid: 2-chloro-  
 3-hydroxy-6-phenyl-



Doe J  
 Aldrich  
 7-7-77  
 1 g, Midland

T6NJ BG CQ D2VQ FR

Figure 2. Master Registry Card.

which do pass checking are stored in a temporary file, and if all pass, a suffix checking routine is invoked. This verifies that the entire combination of suffixes used is in conformation with the WLN rules. The suffix checking routine formats the correct notations into a single character string and writes it into the notation update file. The fragment and connection table records are also entered into their respective files.

#### DATA ENTRY

Compounds entering the Dow registry system are recorded on registration forms and verified for uniqueness via a molform comparison in a master compound card file. The verified compound is assigned an accession (or registry) number (an eight-digit number including check digit). The name, molform, source information, and any cross reference numbers are entered through a separate system. The WLN is assigned and entered on-line via the WLN analysis program. A sample compound and its WLN are illustrated in Figure 2.

The interactive nature of the WLN analysis program allows for immediate correction of notation, molecular formula, and other data errors at the time of input. For this purpose, the cursor control and foreground (unprotected)/background (protected) field display features of the Hazeltine 2000 are employed. These functions are programmatically controlled. In this way, the descriptive field identifiers can be formatted and displayed in background or protected mode, which ensures that they will not be erased by the data entry activities. The actual data are entered in foreground or unprotected mode, which allows for the deletion, addition, or insertion of characters in any of the foreground fields before they are transmitted to the computer. This procedure is also referred to as "batch" mode operation, as the data are "batched" before they are sent; i.e., the whole screen is transmitted at once. The program also controls the positioning of the cursor at the beginning of data fields and the use of the tab key to move it from one field to the next. Lines of data may be erased, new lines inserted, or the entire screen cleared for the start of a new compound. A picture of the screen format is shown in Figure 3. In addition to field identifiers, all fields are delimited by quote marks.

DR#:00000011 #MANTRAP:0 ACTION CODE:

MF:"C14-H12-CL-N-03  
 WLN:"T6NJ BG CQ D2VQ FR

Figure 3. Screen format for entering WLN: italics, unprotected field; bold print, protected field.

DR#:00000011 #MANTRAP:0 ACTION CODE:

MF:"C14-H12-CL-N-03  
 WLN:"T6NJ BG CQ D2VQ FR

COMPOUND OK

Figure 4. WLN checked as correct: italics, unprotected field; bold print, protected field.

The Dow registry number is entered at the top of the screen, along with the number of MANTRAP-suffixed entries for this structure (where applicable) and an action code ([A] for addition, [D] for deletion). Since the program does not interact directly with the master file, it creates records in update files bearing the appropriate action to be taken when the actual master file update occurs. Another action code [B] will flag the program to bypass WLN processing (used for those categories of compounds not handled by the algorithm). These three data items are transmitted and verified by the program. They are then erased and displayed in protected mode for reference while the remaining data are entered. The cursor is then positioned at the beginning of the molecular formula field in the middle of the screen. The WLN field follows. The bottom of the screen is used for the display of error messages.

The encoder enters the molecular formula (up to 44 characters) and then tabs to the WLN field and enters the notation (up to 200 characters). Corrections may be made to either of these fields at this point. The encoder then tabs to the asterisk at the end of the WLN field and transmits the data. (In the case of a deletion, only the molform is entered and no checking is done.) If the notation passes the checking process, the message COMPOUND OK (Figure 4) is printed, the screen is cleared, and the field identifiers are redisplayed for the next compound. If a notation error is detected, a message identifying the error (where possible) is displayed at the bottom of the screen, along with the calculated and encoder-generated molforms and the last five entries in the connection table (Figure 5). A correction code field is also displayed at the end of the notation field, and the cursor is positioned at the beginning of the molform field. When the encoder determines his error, he may correct the data and tab down to the correction code field. He then enters a code to indicate that he has corrected the data (C), or that he wishes to "hold" the data in a temporary file for further checking (H), or that he wishes to bypass WLN analysis altogether (B). If a code of "C" is entered, the corrected data are reanalyzed and processed as is the initial entry. Entering a code of "S" in the correction code field or the word START in the WLN

DR#:00000011 #MANTRAP:0 ACTION CODE:

MF:"C14-H12-CL-N-03  
WLN:"T6NJ BE CQ D2VQ FR

CODE: \*

```

ERROR
 9  02      410 0 0 OT V
10  V      911 0 0 002Q
11  Q      10 0 0 0 OV
12  R      6 0 0 0 OT
CMF:C 14H 12BR 1N 10 3
SMF:C 14H 12CL 1N 10 3

```

Figure 5. Error found in WLN: italics, unprotected field; bold print, protected field.

DR#:00000022 #MANTRAP:2 ACTION CODE:

MF:"C14-H12-CL-N-03  
WLN:"T6NJ BG CQ D2VQ FR 5122

\*

Figure 6. Entry of first suffixed compound (with error): italics, unprotected field; bold print, protected field.

DR#:00000022 #MANTRAP:2 ACTION CODE:

MF:"C6-H6-0  
WLN:"QR 5122

\*

Figure 7. Entry of second suffixed compound: italics, unprotected field; bold print, protected field.

field in the initial entry will clear the screen and allow all the data for a compound to be reentered (i.e., restart).

Suffixed notations are analyzed in the same manner as their unsuffixed counterparts (Figures 6 and 7). As each notation in the set is verified, the screen is cleared of all information except the registry number, MANTRAP number, and action code at the top. The suffix checking routine is invoked when the indicated number of compounds has been entered. If a suffix error is detected, each notation is displayed in protected (background) up to the MANTRAP suffix, which is displayed in unprotected (foreground) (Figure 8). The erroneous suffix(es) may then be corrected, retransmitted, and verified, and the message SUFFIXES OK (Figure 9) is printed if they are correct. The screen is then cleared for a new compound.

DR#:00000022 #MANTRAP:2 ACTION CODE:

SUFFIXES OUT OF ORDER

T6NJ BG CQ D2VQ FR 5122  
QR 5122  
CODE: \*

Figure 8. Error found in suffix: italics, unprotected field; bold print, protected field.

DR#:00000022 #MANTRAP:2 ACTION CODE:

SUFFIXES OK\_

Figure 9. Suffixes checked as correct.

The program is terminated by entering a 99999999 in the registry number field. Statistics are displayed giving the number of compounds added or deleted, and the number of MANTRAPS.

Compounds are entered into this system at about 30–50/h. At this rate, the entire encoding process requires about 1 to 2 min per compound. Each notation is examined only once, and at the end of the checking process, all structure data required for that compound is complete.

## CONCLUSION

An interactive program has been developed for entering, verifying, and storing chemical structures encoded in the Wiswesser Line Notation. The program significantly reduces the overhead involved in the encoding process, and improves the timeliness of the structure data base. This program is a part of a series of programs used to record information about the chemical structures in the Dow registry system.

## ACKNOWLEDGMENT

The authors wish to thank Mr. Albert Leo and his colleagues at Pomona College for their assistance in the use of their checking program, as well as A. A. Asadorian, V. B. Bond, J. N. Paige, F. K. Voci, and L. F. Young for their contribution to this effort.

## LITERATURE CITED

- (1) Smith, E. G.; Baker, P. "The Wiswesser Line-Formula Chemical Notation (WLN); 3rd ed.; Chemical Information Management, Inc.: Cherry Hill, N.J., 1976.
- (2) Bowman, C. M.; Landee, F. A.; Reslock, M. H. "A Chemically Oriented Information Storage and Retrieval System. I. Storage and Verification of Structural Information", *J. Chem. Doc.*, **1967**, *7*, 43–7.
- (3) Bowman, C. M.; Landee, F. A.; Lee, N. W.; Reslock, M. H.; Smith, B. P. "A Chemically Oriented Information Storage and Retrieval System. III. Searching a Wiswesser Line Notation File", *J. Chem. Doc.*, **1970**, *10*, 50–54.
- (4) Leo, A.; Elkins, D.; Hansch, C. "Computerized Management of Structure-Activity Data. II. Decoding and Searching Branching Chains and Multiplied Groups Coded in WLN", *J. Chem. Doc.*, **1974**, *14*, 61–65.
- (5) Elkins, D.; Leo, A.; Hansch, C. "Computerized Management of Structure-Activity Data. III. Computerized Decoding and Manipulation of Ring Structures Coded in WLN", *J. Chem. Doc.*, **1974**, *14*, 65–69.