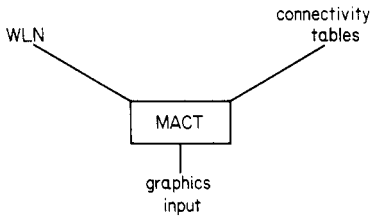
provide an interface between existing methods.



The three methods we had to consider were

(1) the existing WLN data banks
(2) the existing Connectivity Table data banks
(3) two- or three-dimensional graphics input

The *M*achine *A*ligned *C*onnection *T*able (MACT) is designed to be this interface and to provide an appropriate search language. MACT is a highly structured record allowing

atom detail, including stereochemistry, to be identified and stored
ring information to be isolated and searched effectively
a component hierarchy to be established and used for parent/salt identification

WLN becomes an optional input language in CROSSFIRE since its potential as an effective compound descriptor is recognized. It cannot be the main structure language of CROSSFIRE because of

inadequacies in assigning and processing stereochemical descriptors
problems in generation from connection tables or three dimensional graphics input

Do we consider this to be the end of WLN?

CONCLUSIONS

WLN is an effective structural language for systems where a trained information specialist acts as intermediary. In addition, it can act as a powerful descriptor in an integrated research data base. It conveys the structure of the molecule in its own right and yet can be expanded to an atom connectivity record or structure display for further processing.

However, its use is limited once chemists require to interface directly with the system via graphics terminals. In this context its value may continue as an additional search descriptor, e.g., to identify complex ring systems or as an inexpensive input technique. Given present trends, it is difficult to see WLN as having a long future in the U.S.

But it is important to remember that there is an escalation of cost when a company goes to graphics-based systems, and that connectivity tables are in themselves difficult to interpret without translation to the structure diagram. A connection table system needs a graphics interface to make it useable whereas a WLN-based system requires no special equipment.

The current economic climate might help WLN to prolong its field of influence, and the intermediary may indeed be an economic necessity in the short term.

REFERENCES AND NOTES

(1) Eakin, D. R. "The ICI CROSSBOW System". In "Chemical Information Systems"; Ash, Janet E., Hyde, E., Eds.; Wiley: London, 1975; pp 227-242.

Applications of the Wiswesser Line Notation at the Dow Chemical Company†

V. B. BOND, C. M. BOWMAN, L. C. DAVISON, P. F. ROUSH,* and L. F. YOUNG
Information Systems Development, Dow Chemical Company, Midland, Michigan 48640

Received August 24, 1981

The Wiswesser Line Notation has been used at the Dow Chemical Company since the early 1960s to provide machine representation of the approximately 180 000 defined structures in its compound data base. Substructure fragments and connection tables are derived from the notation for structure searching, pattern recognition, structure drawing, and other purposes. The notation itself has been used as a basis for clustering structures to form prototype groups for biological screening purposes. The inconsistencies, incompleteness, and "one dimensionality" of the WLN have presented a number of problems in developing computer algorithms for its analysis and interpretation. The use of computer graphics as a means of entering and storing structure data is being investigated as a replacement for the WLN.

A well-functioning chemical information system is vitally important to an organization's success in the highly competitive chemical industry. Such a system should provide a mechanism for storage, verification, retrieval, and analysis of those chemical compositions involved in the company's business. At the very basic level, the system should

(1) record a compound in such a way as to determine its uniqueness from other compounds (registration and verification);
(2) provide information about which compounds in the system possess certain substructural requirements (substructure search);
(3) retrieve descriptive and other associated information about each compound (e.g., names, molecular formulas, physical properties, screening data, etc.).

This is common knowledge which has been well recorded in the literature. Recently, more sophisticated analysis techniques have expanded the potential usefulness of compound data bases. Research in the area of pattern recognition, structure elucidation, molecular design, organic synthesis, and reaction indexing has suggested ways in which existing data can be used to design structures, predict activity–structure correlation, and suggest new and perhaps optimal reaction mechanisms.

The greatest problem still facing the chemical information area today is, however, how to successfully manage increasingly larger volumes of chemically oriented data. Dow first addressed this problem in the early 1960s. At that time, it had become apparent that a fragmentation code[1] was not adequate to handle the need for more comprehensive structure information from the company's rapidly growing compound file. What was needed was a means of structural identification

† Presented at the Symposium on the Use of the Wiswesser Line Notation, 180th National Meeting of the American Chemical Society, Las Vegas, NV, Aug 27, 1980.

0095-2338/82/1622-0103$01.25/0

which would "completely describe the relationship of the atoms in the molecule",[2] in other words, some form of notation system. Fortunately, at this same time, Dow had established a research computing facility within the company's corporate research department. The use of the computer for storage and retrieval of large volumes of chemical structure information seemed like a logical solution to the data management problem. Thus, the ultimate requirement for a notation system was that it be "complete, unique, unambiguous, concise, readable, economical and capable of being manipulated by existing computer equipment".[2] The Wiswesser Line Notation "was the best overall choice that could be made from those notations and topological representations available".[2]

Since the time of this decision, the WLN has been used by Dow to encode about 180 000 chemical substances into machine-readable form. A number of computer programs were developed in connection with the notation.

(1) A "checker" program was developed for verifying notation accuracy by calculating a molecular formula from the notation and comparing it with one entered by the encoder.[2] This program served as a pattern for those developed by ICI Ltd. (CROSSBOW),[3] Pomona College (WLKEN),[4,5] and others. Currently, an adaptation of the Pomona College program provides for on-line interactive encoding and verification.[6]

(2) A "pathfinder" program was used for the generation of WLN's from the nonconsecutive locants of complex polycyclic ring systems.[7] This program was used to generate WLN's for the structures in the Ring Index. It has been enhanced and modified by Chemical Abstracts Service for generating WLN's for the Parent Compound Handbook.[8-10]

(3) A highly sophisticated numeric fragmentation code, which used a system of numbers to form fragments that represented various structural features, was also developed. These were stored and searched with an inverted index.[11] The user was able to construct meaningful fragments to represent his query and coordinate these fragments with Boolean logic. This system has been replaced by an online interactive substructure search which operates at two levels of specificity: a bit screen fragment code and an atom-by-atom search based on a connection table.[12]

Dow activities with the WLN have been well recorded in the literature.[1,2,6,7,11,12] It is a well-known fact that over the past 10–15 years there has been a great deal of activity in the area of chemical information handling. For this reason, it would be desirable at this time to evaluate the company's experience with the notation and determine whether it is still the best structure identification tool for a large compound data base.

First, we list the advantages.

(1) The WLN is concise. Over 180 000 notations can be stored on 25 cylinders of IBM 3350 disk.

(2) It is highly descriptive. It was useful for permuted and other types of printed indexes when these were in use at Dow. Trained users may still use the WLN for screening search results. Substructure fragment screens and connection tables can be generated programmatically for most notations. Ring backbones are easily identified. This feature has been used to cluster compounds for prototype selection in drug design.

(3) There is a mechanism (MANTRA suffixes) for linking multistructure compounds. Each component in such substances, however, retains its individual identity for searching purposes.

(4) The WLN has been selected for identification of structures in a number of chemical information tools in the public domain (e.g., Index Chemicus Registry System, Parent Compound Handbook, various suppliers' catalogs, etc.). This provides the potential for matching structures in the Dow registry system with those available elsewhere [however, see (1) under disadvantages below].

In connection with these advantages, it should be noted that the encoding, verification, and storage of 180 000 notations represent a significant investment of time and manpower by Dow. This "inertia" factor must be weighed heavily in considering alternate methods of structure identification. Furthermore, the information system which has evolved around the WLN has, to date, been reasonably adequate to serve the company's needs. These factors make the evaluation process for Dow quite different than that of a company which is just establishing its first machine-readable compound data base.

Nevertheless, the use of the WLN as a storage medium has presented a number of problems. These, coupled with new advances in computer technology and demands for more sophisticated chemical structure data by Dow researchers, require an objective evaluation of the adequacy of the existing system. Generally, the problems with the WLN may be itemized as follows:

(1) The WLN, although conceived as a canonical notation, has not remained so in actual use. Not only have different users adapted the notation to local conventions, but the notation itself has changed over the years. While some change and improvement is desirable and necessary, it poses considerable difficulty for those concerned with data integrity. In many areas, the rules are ambiguous and subject to a great deal of interpretation. Registration based on a "unique" WLN is ineffective, and interinstallation file comparison is cumbersome and often incomplete. Retrospective conversion to accommodate rule changes is costly and error prone. Dow does not use canonical WLN, and the molecular formula is still used as a manual registration tool.

(2) The notation is not comprehensive. The development of WLN rules for polymers, inorganics, and other such "unusual" compounds has been extremely slow. About 2500 polymers and 2000 inorganics (2% of file total) are presently in the Dow data base. No attempt has been made to construct an algorithm for these classes of compounds because of the rule ambiguities.

(3) Computer-based algorithms for interpreting the WLN are complex and, if not carefully conceived, can be error prone. The biggest problem here is that atom connectivity is often obscured at branch points, particularly in the case of multivalent atoms. Dow currently encodes about 150 compounds per week, of which about 97% are analyzed correctly.

(4) Users of the WLN—both encoders and chemists—must be specially trained. Chemists seem much more receptive to drawn structures than to the WLN. Encoders must have some background in chemistry to produce consistently accurate notations.

(5) Substructure searching with the WLN is cumbersome. Searches involving branching atoms are difficult, if not impossible, to construct and lead to many "false drops".

Generally speaking, the problems surrounding the use of the WLN seem to center around its inconsistencies, its incompleteness, its one dimensionality, and, for the sake of computer applications, its context-dependent constructs.

It is interesting to note that a common denominator among all chemical structure systems of any significant size seems to be the desirability of connection tables as the ultimate storage and/or searching medium. This is true whether the source of input is a line notation (such as the WLN), a structure typed on a chemical typewriter, or a sophisticated interactive graphics system. The use of a connection table as a structure record in the 5 million compound Chemical Abstracts Registry System lends some credence to this statement. The reason for this choice is that the connection table provides a simple, complete, unambiguous description of the topology

of a molecule. The identity of each atom in the structure and its interrelationships with other atoms can easily be determined by examining the table, either visually or programmatically. Such a basic description of the molecule is necessary for precise and efficient substructure searching as well as pattern recognition, structure–activity correlations, physical property predictions, and other such calculations. These latter techniques are fast becoming an essential component of the chemical information systems of the 1980s. The economics of producing chemicals in a highly competitive marketplace has required industrial organizations to derive much more meaningful data from their chemical data bases. The ability to predict activity, establish prototype structures, propose reaction mechanisms, and suggest alternate synthesis routes can provide a company with a valuable, cost-effective tool for better utilizing its data resources. Computer technology has made all of this possible. The connection table plays a vital role in all of these developments. The argument that the connection table requires too much storage has been obviated by the advent of low-cost, high-density storage devices. Thus, the argument for the WLN or other notations as a concise method of structure identification is no longer a compelling one.

Another interesting development in computer technology is the increasing use of graphics as an input/output medium. This provides an exciting challenge to the information chemist as a possible replacement for more conventional data entry and retrieval techniques. Indeed, a number of chemical structure graphics-based systems are already operating (e.g., Chemical Abstracts Service,[13] Molecular Design Ltd.,[14] The Upjohn Company[15]) with considerable success.

Graphics software and hardware are still relatively expensive. Increased interest in this area, however, should result in the development of commercially available software tools, the cost of which would then be distributed over a wide variety of users. Already, the related hardware has followed the industry trend toward lower cost. In any case, the chief advantages presented by graphics are twofold:

(1) Data entry is a simple matter of "drawing a picture" on a screen or tablet with a light pen or some other device. This requires relatively unskilled, less expensive labor than a highly structured notation system.

(2) The topology of the molecule can be more easily determined for storage as a connection table. Display coordinates may also be stored for retrieval of structure drawings.

What conclusion, then, can be drawn from the above discussion? It is Dow's belief that graphics-based systems with dedicated processors and large areas of low-cost storage are the way of the future for large chemical structure data bases. The bottom line for such systems is actually the input costs, which, with increasing manpower expenditures, are becoming prohibitive for rule-bound notation systems. No system will adequately describe all categories of compounds. But a system where input is a simple description of the topology of a molecule is bound to be more comprehensive and accurate than one that is subject to analysis and interpretation. Notation systems may still be highly useful for small files where printed tools are still the primary search method. But for data bases of any size and for applications involving sophisticated data analysis techniques, their usefulness has become obsolete.

REFERENCES AND NOTES

(1) Opler, A.; Norton, T. R. "A Manual for Programming Computers for Use With a Mechanized System for Searching Organic Compounds"; The Dow Chemical Company, Western Division: Pittsburg, CA, 1956.

(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*, 437.

(3) "The CROSSBOW Handbook: A Guide for Users and Potential Users of the CROSSBOW System"; available from Fraser Williams (Scientific Systems Ltd.) Cheshire SK12 INJ, England.

(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. II. Computerized Decoding and Manipulation of Ring Structures Coded in WLN". *J. Chem. Doc.* **1974**, *14*, 65–69.

(6) Bowman, C. M.; Davison, L. C.; Roush, P. F. "On-Line Storage and Retrieval of Chemical Information. I. Structure Entry". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 228–230.

(7) Bowman, C. M.; Landee, F. A.; Lee, N. W.; Reslock, M. H. "A Chemically Oriented Information Storage and Retrieval System. II. Computer Generation of the Wiswesser Notations of Complex Polycyclic Structures". *J. Chem. Doc.* **1968**, *8*, 133–138.

(8) Ebe, T.; Zamora, A. "Wiswesser Line Notation Processing at Chemical Abstracts Service". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 33–35.

(9) Zamora, A.; Ebe, T. "Pathfinder II. A Computer Program That Generates Wiswesser Line Notations for Complex Polycyclic Structures". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 36–39.

(10) Zamora, A. "An Algorithm for Finding the Smallest Set of Smallest Rings". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40–43.

(11) 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.

(12) Bond, V. B.; Bowman, C. M.; Davison, L. C.; Roush, P. F.; McGrew, R. D.; Williams, D. G. "On-Line Storage and Retrieval of Chemical Information. II. Substructure and Biological Activity Searching". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 231–234.

(13) Blake, J. E.; Turner, N. A.; Haines, R. C. "An Interactive Computer Graphics System for Processing Chemical Structure Diagrams". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 223–238.

(14) "Company Offers Computer-Assisted Chemistry". *Chem. Eng. News* **1979**, *57*, June 18.

(15) Howe, W. J.; Hagadone, T. R. "Progress Toward an On-Line Chemical and Biological Information System at the Upjohn Company". In "Retrieval of Medicinal Chemical Information"; American Chemical Society: Washington, DC, 1978; ACS Symp. Ser. No. 84, p 107.