(43) S. R. Heller and D. A. Koniver, "Computer Generation of WLN. II. Polyfused, Perifused and Chained Ring Systems", *J. Chem. Doc.*, **12**, 55-59 (1972).

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

(45) E. E. Townsley and W. A. Warr, "Chemical and Biological Data—an Integrated On-Line Approach" in "Retrieval of Medicinal Chemical Information", W. J. Howe, M. M. Milne, and A. F. Pennell, Eds., American Chemical Society, Washington, DC, 1978, ACS Symp. Ser. No. 84.

(46) E. V. Krishnamurthy, P. V. Sankar, and S. Krishnan, "ALWIN Algorithmic Wiswesser Notation System for Organic Compounds", *J. Chem. Doc.*, **14**, 130-141 (1974).

(47) P. V. Sankar, E. V. Krishnamurthy, and S. Krishnan, "Representation of Stereoisomers in ALWIN", *J. Chem. Doc.*, **14**, 141-146 (1974).

(48) K. Subramanian, S. Krishnan, and E. V. Krishnamurthy, "Huffman Binary Coding of WLN Symbols for File-Compression", *J. Chem. Doc.*, **14**, 146-149 (1974).

Graphics Challenge WLN. Can WLN Hold Fast?[†]

DIANE R. EAKIN

Fraser Williams (Scientific Systems) Ltd., Glendower House, London Road South, Poynton,
Cheshire SK12 1NJ, England

Received August 24, 1981

Chemist-oriented graphics systems are now available, allowing the research chemist to manipulate structures by using computer techniques. The presentation looks at the role these graphics systems play in research and compares it with the role which standard WLN-based systems play. It discusses the techniques necessary to support chemical information retrieval and the development required for WLN to meet the challenge of chemist-oriented graphics. A graphics-oriented enhancement of CROSSBOW is discussed. A machine aligned connection table (MACT) is proposed, as is an interface between WLN and Chemical Abstracts Connectivity Tables.

INTRODUCTION

The chemist in research has seen many changes in the last decade; one such change has been the use of computerized technique to support his main synthetic effort. These include the following:

On-line reference retrieval systems now supplement his conventional library publications.

Substructure and structure retrieval systems available on company data banks replace previous manual card indexes.

Drug design techniques are now in use with computerized structure–activity correlation.

Structure elucidation usually involves some form of computerized interpretation following spectroscopic analysis.

Many chemists see molecular modeling as a vital aid to interpreting their research results.

Other chemists now use computerized methods to aid synthetic pathway design.

Computers are thus very vital to chemical research, and the chemist's involvement with computers is changing. The advent of interactive graphics systems means the chemist can have meaningful dialogues with computer systems in the chemist's own language—the structure diagram. This is particularly important for applications such as molecular modeling and synthetic pathway design, where interaction with the system is essential.

But some chemists are now putting on pressure to ensure that they can have direct access to computer files for other purposes—such as substructure retrieval. Will any problems be solved by giving chemists direct access to company files? Where does this leave the existing WLN-based systems? Can they move forward and meet the challenge? Or is it time to move away from WLN? Will graphics just be an expensive toy? These are questions being asked by many people in many organizations in America, Europe, and Japan, and it is important that the people making these decisions evaluate carefully the options before them.

STRUCTURE REPRESENTATION

Firstly, however, perspective is important. WLN is not an alternative to graphics—WLN is merely a method of structure representation. Computerized chemical handling systems principally use one or more of the following ways of describing the chemical to the computer:

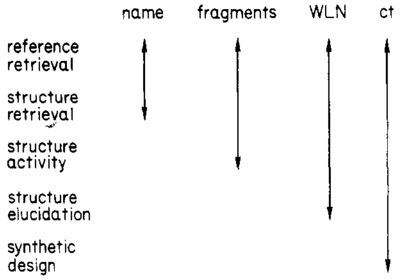
Chemical Nomenclature. This covers everything from simple common and trade names to complex systematic nomenclature such as that used by Chemical Abstracts.

Fragment Codes. This covers from the simple manual fragment code typified by the Derwent Ring Code to the complex algorithmically generated systems used by the BASIC or IDC groups.

Notations. Almost exclusively WLN, but some companies and organizations have developed other notational systems.

Connection Tables. These vary from simple atom-bond connectivity matrices to the more complex, involving stereochemistry, etc. Graphics-based systems rely almost solely on connection tables.

The use of a particular structural representation in a given application obviously varies from organization to organization. In general terms the following diagramatic representation is seen to apply:



Connection tables are used extensively where detailed structure manipulation is required. But the picture is much

[†] Presented before the Division of Chemical Information, 180th National Meeting of The American Chemical Society, Las Vegas, NV, Aug 27, 1980.

0095-2338/82/1622-0101$01.25/0 © 1982 American Chemical Society

more complicated than this; users can effectively convert from WLN to connection tables and to fragment codes. Since this is so, why is WLN not more widely used? For example, WLN is ideally suited for use in text-based retrieval systems. Yet its use here is very limited, the only major system being the ICRS data base from ISI.

There are obviously many external forces at work, regardless of the technical capability of WLN. The influence of Chemical Abstracts in the text area may be one. WLN is most successful in those areas where such external factors are minimized—within individual industrial organizations. Its main use is concentrated in industrial compound registers, as typified by the speakers at the 180th National Meeting of the American Chemical Society meeting from Searle, ICI, DOW, and Diamond Shamrock.

REQUIREMENTS FOR CHANGE

Within most of these organizations WLN has become the province of the chemical information specialist. Today, there are external pressures on them from chemists—chemists wanting direct access, wanting a full structure graphics interface with the company data bank. These chemists go to meetings such as the 180th National Meeting of the American Chemical Society, see graphics systems working, become enthusiastic, and perhaps forget the implications in their own environment. Chemical information specialists must be able to respond to these pressures from a position of knowledge. They must impress on chemists the total environment and be able to understand the broader issues associated with such changes. An examination of these includes the following:

Size of a Data Base. The size and content of the data base is important to the effectiveness with which a chemist can use that data base. As computers become more powerful, response time becomes less of an issue, but it is particularly relevant today if one is considering a large data bank on a small machine.

Modeling. Chemists often do not distinguish adequately between the need for modeling on small data sets, isolated on demand from the main data file, and the need for interactive access to the large data base itself. Many companies would do well to consider an interim step where existing search systems are used to feed subsets of data in the appropriate format to local modeling systems.

Extension to Substructure Search. Many chemists do have a need to perform substructure searching on a central data bank, and there are certainly advantages to their performing such searches themselves. But one still has to develop skills to use even the best graphics-based system. One has to learn the physical techniques of handling the graphics equipment, the system itself, and the way it works. Those of us who have carried out substructure searches on behalf of chemists know the importance of question definition, of knowing the content of the data base, which classes of compounds to avoid, and so on. Conceptualizing the impact of a particular substructure on a diverse file is difficult, and an effective and manageable subset can often only be obtained through a pragmatic approach. How often will the research chemist have to use the system to achieve this ease of use? Are such needs the requirements of many chemists within an organization or only a few?

Integrated Research Data Bases. Most chemists do not want access to chemical structural data alone—there is a requirement for integrated research data banks containing chemical structure, chemical property, sample availability, and biological activity data. The integration of chemical structural data into these data banks is very important.

Stereochemistry. Research chemistry is developing—the current emphasis is on stereochemical factors and the significance of this on biological activity. In many cases, two-dimensional representation is no longer sufficient, and the implication of this on the handling of a large file previously devoted to two-dimensional information must be considered.

These are just some of the issues which have to be evaluated; others I could have mentioned include data integrity, registration policy, and so on. These issues obviously vary from one company to another, and each will need to develop a strategy to meet the technical and financial targets of its environment.

DEVELOPMENT OF CROSSBOW AND CROSSFIRE

The CROSSBOW[1] system (*C*omputer *R*etrieval of *O*rganic *S*ub*s*tructures *B*ased *O*n *W*iswesser) was, as the name suggests, developed to provide chemical structure retrieval on a data base where compound information was encoded in WLN. CROSSBOW was designed to overcome some of the problems identified with WLN-based systems:

(1) Limitations on Substructure Searching: CROSSBOW ensures that *all* types of substructure searches can be answered with accuracy and precision via multilevel searching. WLN is processed automatically to give fragment codes and connectivity tables. Users of the CROSSBOW system thus have fragment searching, WLN string searching, and atom connectivity searching available to them.

(2) Communication with End Users: Chemists generally do not wish to learn WLN. CROSSBOW converts WLN into a two-dimensional structure diagram suitable for interpretation by the chemist himself.

(3) Integration with Other Data: CROSSBOW is designed as a series of structure-handling tools and can be integrated into a total research information system involving spectroscopic data, physical property, biological data, etc.

CROSSBOW was designed for two-dimensional structure handling, economic batch-oriented processing, and use by a trained chemical information specialist acting as intermediary between the chemist and the system. It is successfully used in this context by some 16 companies throughout the world.

Some thought has obviously been given to the need for change to meet the challenge of today, as outlined above,

to handle stereochemistry and all its implications
to make use of interactive computing techniques
to facilitate direct chemist's use via the graphics terminal

To this end we have been developing CROSSFIRE (*C*omputerized *R*epresentation *o*f *S*tructures *f*or *I*nteractive *Re*trieval). CROSSFIRE is designed to make the best use of two technologies: (1) the graphics techniques used in molecular modeling and (2) the substructure search techniques evolved during many years of practical experience with CROSSBOW. It is designed to cover all aspects of

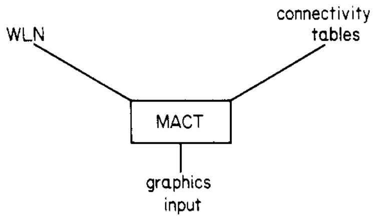
compound registration
structure and substructure search
molecular modeling
report production
data base maintenance

using graphics interfaces.

The system is designed to be integrated into an existing R & D situation and to allow users to build upon developments which have taken place within that environment. Hence the system handles the chemistry of the R & D environment in such a way that it can be used in conjunction with other parameters which are more specific to a particular environment—biological test results, physical properties, etc.

MACHINE ALIGNED CONNECTION TABLE, MACT

One of the first considerations in CROSSFIRE was the structural language to be used as the basis of the system. The requirement was for a structural representation which would 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 180000 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 © 1982 American Chemical Society