

ACKNOWLEDGMENT

We thank the organic chemists who took the time to use the CAOCI fiche. Additionally, appreciation is due to Mrs. Frances Sanderford for her innovative efforts in preparing this manuscript.

REFERENCES AND NOTES

- (1) Alfa Products, Thiokol/Ventron Division, P.O. Box 299, Danvers, MA 01923.
- (2) Pyrazine Specialties, P. O. Box 6933, Atlanta, GA 30315.
- (3) Eastman Kodak Co., Eastman Organic Chemicals, Rochester, NY 14650.
- (4) Aldrich Chemical Company Inc., 940 W. Saint Paul Ave., Milwaukee, WI 53233.
- (5) Takeda Chemical Industries Ltd., 7 Edobashi Nichome, Nihinbashi, Chuo Ku, Tokyo, Japan
- (6) Hopkin & Williams, P.O. Box 1, Romford, Essex RM1 1HA, England.
- (7) Fluka AG, Chemische Fabrik, CH-9470 Buchs, Switzerland.
- (8) "Chem Sources-U.S.A."; Directories Publishing Company, Inc.: Ormond Beach, FL 32074.
- (9) "Chem Sources-Europe"; Directories Publishing Co., Inc.: Flemington, NJ 08822.
- (10) "Chemical Week Buyer's Guide"; McGraw-Hill, Inc.: New York, NY 10020.
- (11) "OPD Chemical Buyers Directory"; Schnell Publishing Co., Inc.: New York, NY 10007.
- (12) "Chemical Economics Handbook"; SRI International Chemical Information Services: Menlo Park, CA 94025.
- (13) Walker, S. B. "Development of CAOCI and Its Use in the ICI Plant Protection Division". *J. Chem. Inf. Comput. Sci.*, paper included in this issue.
- (14) The Fraser Williams Group, 140 Route 17 North, Paramus, NJ 07652, or Fraser Williams (Scientific Systems) Ltd., Flendower House, London Road South, Poynton, Cheshire SK12-1NJ, England.

Diverse Uses and Future Prospects for Wiswesser Line-Formula Notation

WENDY A. WARR

ICI Pharmaceuticals Division, Alderley Park, Macclesfield, Cheshire SK10 4TG, England

Received August 24, 1981

The many and varied uses found for Wiswesser Line Notation (WLN) in the 1970s are reviewed. The advantages of WLN-based systems for storage and retrieval of chemical structures have been proved over several years of increasingly wide usage. Future prospects for the notation are briefly considered in the light of these advantages and certain shortcomings.

Wiswesser Line Notation (WLN)¹ is a compact way of uniquely and unambiguously representing the complete topology of a chemical molecule by a linear sequence of characters. It is used by around 150 organizations worldwide in industrial, governmental, and educational fields. In addition, between 30 and 40 educational institutions teach WLN even though they do not maintain files of notations. The Chemical Notation Association (CNA) mails information to Canada, U.S., United Kingdom, France, Italy, Sweden, Japan, Denmark, Netherlands, Israel, Switzerland, India, Australia, West Germany, Romania, Austria, Belgium, Norway, and Yugoslavia. The rules of WLN are available in English, French, German, and Japanese translations.²

Numerous programs have been written to handle WLN, some part of in-house systems and some, such as the CROSSBOW³ programs, commercially available. The CROSSBOW suite comprises programs for registration with WLN/molecular formula check, fragment generation, bit and string search, connection table generation, atom-by-atom search, structure display, and KWIC listing. Sixteen different organizations, to date, are using the CROSSBOW system.

Most organizations handling WLN need a checker program which generally validates notation syntax and then generates a molecular formula from the notation for comparison with an input molecular formula. ICI and other companies have used such programs for years. The CROSSBOW one is commercially available.³ Other well-known checker programs are the ones used by the Institute for Scientific Information in connection with Index Chemicus Registry System (ICRS)⁴ and the program WISCT (commercially available⁵) written at Pomona College to verify WLN and generate connection tables. WISCT is also used to generate connection tables for the MACCS system.⁶

There are two WLN-string-search programs on the market, RADIICAL⁷ from ISI (Retrieval and Automatic Dissemination of Information from the Index Chemicus and Line Notation) and CROSSBOW. Reckitt and Colman in England have purchased certain CROSSBOW programs but use the IBM package INQUIRE to search notations, molecular formulas, and CROSSBOW fragments.

Many American companies, and Pfizer in England, have in-house software for checking and searching WLN and for generating connection tables (from WLN) to be used in atom-by-atom search and structure display. On the connection table front, CROSSBOW,³ WISCT,⁵ and MACCS⁶ are for sale. Two software houses, CIMI⁸ and Fraser-Williams (Scientific Systems) Ltd.,³ specialize in writing WLN systems for customers. Granito of CIMI has written a program to generate ring codes from WLN^{9,10} (i.e., the code of the Pharma Documentation Ring).

Finally, this brief survey exemplifying (but not comprehensively covering) WLN software would not be complete without mentioning the PATHFINDER program originally written by the Dow Chemical Company but now available from Chemical Abstracts Service¹¹ (or in England via the United Kingdom Chemical Information Service—UKCIS). The program is used to check locant paths in ring systems with bridges and/or crossed bonds. It could, of course, be used for any ring system but it is of most use for "tricky" notations where the encoder is not sure whether he has picked the optimum path for lettering ring atoms.

From software we move to examples of the various data bases that have been created by using WLN. The biggest is ICRS.⁴ ISI supply very up-to-date tapes, on a monthly basis, of novel compounds appearing in the literature, together with bibliographic data and various flags and subject terms. they

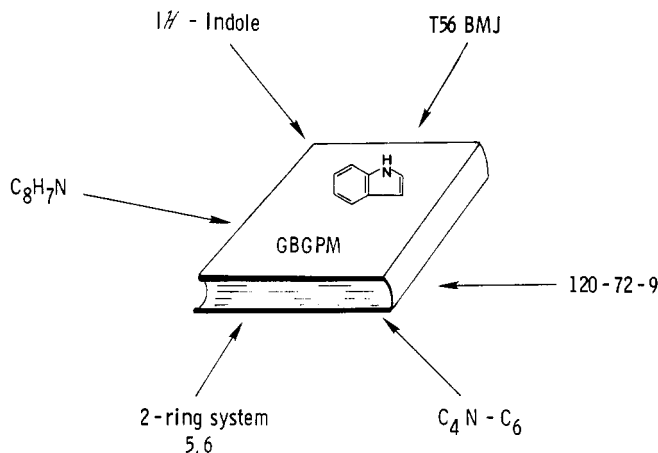


Figure 1.

have thus created a data base, since the early 1960s, of about two and a half million compounds. Cumulated indexes on microfiche and KWIC listings of WLN on microfilm are also available from ISI.⁴

Another major collection of notations is "The Parent Compound Handbook",¹² available from Chemical Abstracts Service (or UKCIS in England). The handbook gives structural formulas for the parents of all published ring systems. Each parent is given a five-letter accession code. There are six indexes so that the structures can be accessed by parent name, WLN, molecular formula, Chemical Abstracts Registry Number, ring analysis, or ring substructure (see Figure 1). Some specialized data bases are available in machine-readable form from CIMI—for example, CIMI-DRUGFILE and the WLN for compounds in the Merck Index.¹³ The present edition of this reference book contains no notations, but it does carry an explanatory chapter on WLN.

Notations for psychotropic drugs may be looked up in a U.S. National Institute of Mental Health publication.¹⁴ This will be mentioned again later. The Pesticide Manual published by the British Crop Protection Council¹⁵ has an alphanumeric listing of WLN, and the Pesticide Index published by the Entomological Society of America¹⁶ also contains notations.

The Commercially Available Organic Chemicals Index (CAOCI) was originally produced by a consortium of chemical companies in the U.K., but it is now updated and marketed by Fraser-Williams (Scientific Systems) Ltd.³ The original CAOCI indexed WLN, molecular formula, chemical name, and supplier data from more than 50 catalogs, and the file had about 120 000 entries with about 52 000 unique WLN. Fraser-Williams have rationalized and reduced the number of catalogs, but they index more compounds from each catalog, since they now include inorganics, biochemicals, etc. The structural data are available in machine-readable form, and there are various indexes on microfiche. Incidentally the Aldrich Chemical Company was using WLN before the advent of CAOCI.

An incomplete list of WLN of toxic chemicals used to be available on tape with the Registry of Toxic Effects of Chemical Substances (RTECS).¹⁷ The notations were later withdrawn, but in 1979 it was announced that WLN would be added for all structures in the near future. About 27 000 "prime names" are on file. Notations for all RTECS compounds have been coded for the Environmental Chemicals Data and Information Network (ECDIN)¹⁸ of the European Economic Community. Compounds in the "List of Organic Chemicals and their Customs Classification", the so-called Brussels classification, have also been coded for ECDIN.

There are also files of physical chemical data with associated Wiswesser Line Notations. The Thermodynamics Research Center in Texas and Pedley's work at the University of Sussex,

England, are notable in the thermodynamic field. Pedley has used his system to publish tables of thermochemical data, but the WLN are not yet available in machine-readable form.¹⁹

Pollack has worked on a substructure index for correlating spectral data,²⁰ and the CRC Press has published on "Atlas of Spectral Data" wherein spectral characteristics and physical constants of organic compounds can be accessed by WLN²¹ (Pollack having provided the WLN expertise).

In Australia, Needleman is using WLN as a tool in a computer-aided learning program,²² and Vaughan has published on the handling of WLN and related physical data.²³ Hansch and Leo of Pomona College, Claremont, CA, have published extensively in the field of structure-property relationships.^{5,24-26} Their data base of physical constants, such as log *P* values, plus WLN is commercially available.⁵ Adamson and Bawden have also published on structure-property correlation²⁷⁻²⁹ and so has Saggers.³⁰⁻³²

Several teams have independently used WLN in reaction indexing research. Osinga,³³⁻³⁵ Lynch,³⁶⁻³⁸ Eakin and Warr,³⁹ and Gelernter^{40,41} are cited here. In connection with his SYNCHEM and SYNCHEM2 programs Gelernter has worked on the computer generation of WLN.^{40,41} Teams at NIH^{42,43} and the Dow Chemical Company⁴⁴ have also worked in this field. In fact the NIH team was able to generate canonical WLN for 81% of the compounds in one of their data bases, and their program was used to produce WLN for the psychotropic drugs publication mentioned earlier.¹⁴

Having exemplified work done using WLN in the past, we shall now briefly consider the use of WLN in ICI Pharmaceuticals Division at present. A recent publication described the online integrated chemical and biological data base at ICI and placed particular emphasis on the sophisticated biological control functions of the system.⁴⁵ Chemical and biological searches (independent or integrated) as well as registration and WLN checking are carried out online. Atom-by-atom searching and structure display are done overnight on hit files which have been suitably merged online.

The CROSSBOW system is also used for batchwise searching of CAOCI, ICRS, and Hansch data bases. It is hoped that in 1982 CAOCI data will be incorporated into the online data base.

Various other enhancements are planned for 1982, the most obvious to the outsider being online structure display. Terminals linked to the online data base are now available for use by research chemists, and it will be a great improvement when chemists inputting molecular formulas or reference numbers can receive structural formulas in response to their single enquiries for chemical data, sample data, and biological test results. Chemists cannot at present run substructure or biological searches. Longer term plans are to provide extra, more user-friendly facilities. The future system could still involve WLN since it is not intended to take chemical registration out of the hands of information scientists. The system will revolve around a canonical connection table, and WLN, while not essential for registration, would still be an input option.

The heavy investment put into WLN by ICI and many organizations makes it unlikely that WLN usage will cease for many years to come. Nevertheless the exponents of WLN are going to have to work seriously on some of its disadvantages if WLN is to have a future. Tautomerism, Markush structures, and stereochemistry are three fields requiring research.

The WLN rules are such that any precisely defined structure can be encoded, and tautomers, e.g., 2-hydroxypyridine and 2-pyridone, have different WLN. This "old-fashioned bond fixation" has both its advantages and its disadvantages in the field of tautomeric and certain mesomeric structures (e.g., chelates).

Krishnamurthi feels that Algorithmic Wiswesser Notation⁴⁶⁻⁴⁸ could be of use in a Markush structure project, and Lynch's team at Sheffield have been considering this in ongoing work on Markush structure representation. As regards stereochemistry, a special subcommittee of the Chemical Notation Association made great progress in stereochemical suffixes for WLN's up until 1979, but no work is being carried out at present. CNA (UK) had high hopes of organizing a research project in WLN, stereochemistry suffixes, structure display, and graphics, but no suitable student could be found for the academic year 1980-1981.

The principle disadvantage of WLN is that it is not user friendly. This can only be overcome by programs which will derive a canonical WLN from something else (but no one has yet produced a cost-effective program to do this for over 90% of compounds), by writing programs to generate canonical connection tables from noncanonical WLN's, or by accepting the intervention of a skilled "middle man".

However, one must recognize that WLN has some great advantages. Compared with most chemical names and connection tables, it is a very compact way of completely representing the topology of a molecule. It has a proven, past performance with large files and is in worldwide usage. Obviously one has to balance the cost of hardware plus software against the costs of extra WLN-skilled personnel, but in many organizations WLN is a cost-effective tool for registration and search. WLN's are readable, listable, and permutable. They are easily printed and transmitted on tape. Chemically meaningful entities are easily detectable in them and in fragment codes generated from them. Systems such as CROSSBOW require little prior processing of files (e.g., tree structuring or inversion) before searches can be run. Relatively cheap hardware and software are used. Moreover, there is only the one line notation with worldwide usage, whereas connection tables being used are many and varied. WLN is therefore useful for interchange of data the world over. For these reasons WLN must have a future in the 1980's.

ACKNOWLEDGMENT

I thank Graham Palmer and George Adamson for helpful discussions.

REFERENCES AND NOTES

- E. G. Smith and P. A. Baker "The Wiswesser Line-Formula Chemical Notation (WLN)", 3rd ed., Chemical Information Management Inc., Cherry Hill, NJ, 1975.
- Contact Wendy A. Warr, ICI Pharmaceuticals Division.
- The marketing rights for the basic CROSSBOW programs are held by Fraser Williams (Scientific Systems) Ltd., Glendower House, Poynton, Cheshire, England SK12 1NJ.
- ICRS tapes and C&C are available from the Institute for Scientific Information, University City Science Center, Philadelphia, PA.
- Corwin Hansch and Albert Leo, Pomona College Medicinal Chemistry Project, Seaver Chemistry Laboratory, Claremont, CA. 1980 Newsletter details WISCT.
- The molecular access system of Molecular Design Ltd., Hayward, CA.
- RADIICAL III is described in "RADIICAL III User's Manual for ICRS", Institute for Scientific Information, University City Science Center, Philadelphia, PA, 1976.
- Charles E. Granito, Chemical Information Management Inc., Cherry Hill, NJ.
- C. E. Granito, S. Roberts, and G. W. Gibson, "The Conversion of Wiswesser Line Notations to Ring Codes I. The Conversion of Ring Systems", *J. Chem. Doc.*, **12**, 190-196 (1972).
- C. E. Granito, S. Roberts, and G. W. Gibson, "Conversion of Wiswesser Line Notations to Ring Codes—Completed System", paper presented at the 168th National Meeting of the American Chemical Society, Atlantic City, NJ, 1974.
- Chemical Abstracts Service, Columbus, OH.
- "Parent Compound Handbook", Chemical Abstracts Service, Columbus, OH, 1977.
- "The Merck Index" (an encyclopaedia of chemicals and drugs), 9th ed., Merck & Co., Inc., Rahway, NJ, 1976.
- E. Usdin and D. H. Efron, "Psychotropic Drugs and Related Compounds", 2nd ed., U.S. National Institute of Mental Health, Washington, DC, 1972.
- "The Pesticide Manual", 6th ed. Charles R. Worthing, Ed., British Crop Protection Council, London, 1979.
- "The Pesticide Index", 5th ed. William J. Wiswesser, Ed., Entomological Society of America, College Park, MD, 1976.
- RTECS is prepared annually by the U.S. National Institute for Occupational Safety and Health (NIOSH) of the Department of Health, Education and Welfare.
- O. Norager, W. G. Town, and J. Howard Petrie, "Analysis of the RTECS Files and Conversion of the Data in these Files for Input to ECDIN", *J. Chem. Inf. Comput. Sci.*, **18**, 134-40 (1978).
- J. Brian Pedley, "Tables of Thermochemical Data", University of Sussex, The School of Molecular Sciences, Falmer, Brighton, England, 1977.
- N. Pollack "A Substructure Index for Correlating Spectral Data", paper presented at the 168th National Meeting of the American Chemical Society, Atlantic City, NJ, Sept 1974.
- "Atlas of Spectral Data and Physical Constants for Organic Compounds", 2nd edition, J. G. Grasselli, Ed., CRC Press, Cleveland, OH 1975.
- M. Needleman, and A. H. Chai, "A Drug Information File Incorporating WLN", *Proc. R. Aust. Chem. Inst. Symp. Chem. Data*, 125-134 (1977).
- G. N. Vaughan, "Australian Applications of WLN in Handling Physical Data", *Proc. R. Aust. Chem. Inst. Symp. Chem. Data*, 84-88 (1977).
- C. Hansch, A. Leo, and D. Elkins, "Computerised Management of Structure-Activity Data. I. Multivariate Analysis of Biological Data", *J. Chem. Doc.*, **14**, 57-61 (1974).
- A. Leo, D. Elkins, and C. Hansch, "Computerised Management of Structure-Activity Data. II. Decoding and Searching Branching Chains and Multiplied Groups Coded in WLN", *J. Chem. Doc.*, **14**, 61-65 (1974).
- D. Elkins, A. Leo, and C. Hansch "Computerised Management of Structure-Activity Data. III. Computerised Decoding and Manipulation of Ring Structures Coded in WLN", *J. Chem. Doc.*, **14**, 65-69 (1974).
- G. Adamson and D. Bawden, "A Method of Structure-Activity Correlation Using WLN", *J. Chem. Inf. Comput. Sci.*, **15**, 215-220 (1975).
- G. Adamson and D. Bawden, "An Empirical Method of Structure-Activity Correlation for Polysubstituted Cyclic Compounds Using Wiswesser Line Notation", *J. Chem. Inf. Comput. Sci.*, **16**, 161-164 (1976).
- G. Adamson and D. Bawden, "A Substructural Analysis Method for Structure-Activity Correlation of Heterocyclic Compounds Using Wiswesser Line Notation", *J. Chem. Inf. Comput. Sci.*, **17**, 164-171 (1977).
- D. T. Siggers, "Uses of the Computer in a Research Program", *World Rev. Pest Control*, **10**, 75 (1971).
- D. T. Siggers "Application of the Computer to a Pesticide Screening Program", *Pestic. Sci.*, **5**, 341-352 (1974).
- D. T. Siggers "The Search for New Herbicides" in "Herbicides", L. J. Audus, Ed., Academic Press, London, 1976.
- M. Osinga and A. A. V. Stuart, "Documentation of Chemical Reactions. II. Analysis of the Wiswesser Line Notation", *J. Chem. Doc.*, **14**, 194-198 (1974).
- M. Osinga and A. A. V. Stuart, "Documentation of Chemical Reactions. III. Encoding of the Facets", *J. Chem. Inf. Comput. Sci.*, **16**, 165-171 (1976).
- M. Osinga and A. A. V. Stuart, "Documentation of Chemical Reactions. IV. Further Application of WLN Analysis Programs: A System for Automatic Generation and Retrieval of Information on Chemical Compounds (AGRICC)", *J. Chem. Inf. Comput. Sci.*, **18**, 26-32 (1978).
- M. F. Lynch and P. Willett, "The Production of Machine-Readable Descriptions of Chemical Reactions Using Wiswesser Line Notations", *J. Chem. Inf. Comput. Sci.*, **18**, 149-154 (1978).
- M. F. Lynch and P. Willett, "The Automatic Detection of Chemical Reaction Sites", *J. Chem. Inf. Comput. Sci.*, **18**, 154-159 (1978).
- M. F. Lynch, P. R. Nunn, and J. Radcliffe, "Production of Printed Indexes of Chemical Reactions Using Wiswesser Line Notations", *J. Chem. Inf. Comput. Sci.*, **18**, 94-96 (1978).
- D. R. Eakin and W. A. Warr "Computerised Aids to Organic Synthesis in a Pharmaceutical Research Company" in "Computer-Assisted Organic Synthesis" W. T. Wipke and W. J. Howe, Eds., American Chemical Society, Washington, DC, 1977, ACS Symp. Ser. No. 61.
- H. L. Gelernter, N. S. Sridharan, A. J. Hart, S. C. Yen, F. W. Fowler, and H. Shou, "The Discovery of Organic Synthetic Routes by Computer", *Top. Curr. Chem.*, **41**, 114 (1973).
- H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spritzer, and J. E. Searleman, "Empirical Explorations of SYNCHM", *Science (Washington, D.C.)*, **197**, 1041 (1977).
- C. D. Farrell, A. R. Chauvenet, and D. A. Koniver, "Computer Generation of Wiswesser Line Notation", *J. Chem. Doc.*, **11**, 52-59 (1971).

- (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:

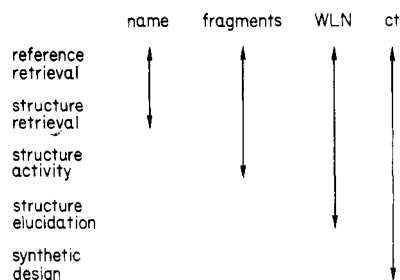
Chemical Nomenclature. This covers everything from simple common and trade names to complex systematic nomenclature such as that used by Chemical Abstracts.

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 diagrammatic 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.