ECONOMICS AND KNOWHOW

To generate the necessary algorithms and computer programs for converting from one system to another, one needs both money and knowhow. Few organizations have both in the area of chemical structure representation.

The Institute for Scientific Information is one organization that has the necessary knowhow for creating such programs and has already undertaken the conversion of WLN to the Ring Code[5] fragment scheme. This was done for an economic reason, namely, to sell more subscriptions to the *Index Chemicus Registry System* (ICRS).[7] However, in developing these programs, another project became feasible—i.e., conversion of WLN to still other systems.

ISI believes that the Wiswesser Line Notation is the most economical form of structural input for a chemical retrieval system and that WLN has many additional advantages as well, which make it the system of choice for many organizations. However, because it is inexpensive, unique, and unambiguous it also serves as an ideal starting place for developing programs for the automatic generation of other codes used by individual organizations. In fact, ISI plans to introduce in 1973, a new service called CHEMTRAN for just this purpose. CHEMTRAN is a practical way of bringing money and knowhow together to improve both communication and cooperation in the chemical structure handling field.

CHEMTRAN

Simply stated, CHEMTRAN is a service that will develop programs to be used in converting (or translating) from one structure representation to another. At present, programs have been developed for converting WLNs to a unique connectivity table. Programs for converting from WLN to Ring Code fragmentation code are nearly complete. The programs developed for these projects, as well as the intermediate records that are generated, can successfully serve as the foundation for all subsequent interconversion programs. Consequently, an organization does not have to reinvest the effort ISI has already expended in reaching this level.

For example, any organization wishing to have its fragment code computer generated (with all the quality control this permits) could switch to WLN, take advantage of the great flexibility WLN offers, and still maintain the fragment code its users have become accustomed to for searching. Or it could use WLN as a less expensive form of input to its present system. In either case, the company investment made in programs developed for processing and searching a particular fragment code can be preserved. Furthermore, users can move towards what we feel is the ideal system—one which includes a fragment code, WLN, and a connectivity table. One CT is already deliverable under CHEMTRAN. Others can be. This planned new service is announced in the belief that it will expedite the development of programs for the interconversion of chemical structure systems.

LITERATURE CITED

(1) Zipf, G. K., "Human Behavior and the Principle of Least Effort," Hafner, New York, 1949.
(2) Steidle, W., "Possibilities of Mechanical Documentation in Organic Chemistry," *Pharm. Ind.* **19**, 88-93 (1957).
(3) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw Hill, New York, 1968.
(4) Morgan, H. L., "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service," *J. Chem. Doc.* **5**, 107-13 (1965).
(5) Granito, C. E., Roberts, S., and Gibson, G. W., "Wiswesser Line Notations to Ring Codes. Part I," *J. Chem. Doc.* **12**, 190-6 (1972).
(6) Fugman, R., Braun, W., and Vaupel, W., "GREMAS—A New Method of Classification and Documentation in Organic Chemistry," *Nachrichten fur Dokumentation* **14**, 179-90 (1963).
(7) Garfield, E., Revesz, G. S., Granito, C. E., Dorr, H. A., Calderon, M. M., and Warner, A., "Index Chemicus Registry System: Pragmatic Approach to Substructure Chemical Retrieval" *J. Chem. Doc.* **10**, 54-8 (1970).

Some Chemical Notation Cooperative Activities*

WILLIAM J. WISWESSER,** CHARLES L. CRUM, KURT J. WINDLINX and RICHARD A. CREAGER
Fort Detrick, Frederick, Md. 21701

Received January 18, 1973

Cooperative efforts between Fort Detrick and various organizations over the past eight years and the consequences of these efforts are discussed.

Fort Detrick's first cooperative activity in chemical information management—with the J. T. Baker Chemical Co. in 1964—was one of the most stimulating and productive of a continuing series of such ventures. J. T. Baker's representative, C. T. Kleppinger, had asked Raymond R. Myers, then at Lehigh University, for suggestions having innovative value in their corporate plan to introduce a large new line of organic laboratory chemicals. Myers recommended management of the chemical structure information with a chemical notation, in cooperation with Wiswesser at Fort Detrick. A. J. Barnard, Jr., in charge of chemical information management at J. T. Baker, approved the plan and a punched-card deck was started with the mutual agreement that the information also would become experimental material for eventual development of a Fort Detrick chemical-biological data base. All agreed that this was a premium opportunity to

* Presented before the Division of Chemical Literature, 164th Meeting, ACS, New York, N. Y., Aug. 30, 1972.
** To whom correspondence should be addressed.

mechanize a chemical file with items of high interest, and, while learning how to do this, to share complementing skills and test the operating concepts with others who were specialists in chemical catalog work. This was recognized as an important first step to prepare machine records for input to a "mini-data base," but the card file that resulted was not given the dignity of being called a "data base," because at that time no computer services with the presently familiar "third-generation" equipment were available at either place; all that both groups had was the earlier-generation tabulating equipment, and a united determination to do the very best with what was available.

One of the first serendipic benefits came from the limitation that all structure-searching and correlating work during the initial development had to be done with card sorters. This also was the case with the chemical-biological file being started at Fort Detrick on herbicides and related plant-growth regulators. Kleppinger and Barnard at J. T. Baker quickly confirmed the Fort Detrick experience, that the simplest of all obvious file-managing measures were the five sets of classifying digits described as "BATCH" digits in 1964—**B** for the "Basic or Benzene" ring measure, **A** for the Atomic class, **T** for the Total number of hetero-atoms, **C** for the units part of the carbon-atom count, and **H** for a similar measure from the hydrogen-atom count, in this case the sum of the unit and tens digits. (Inclusion of the tens-digit broke the odd-even limitations of H-atom counts and gave full usage of all ten digits, regardless of the presence or absence of other odd-valent atoms like nitrogen and the halogens.) The J. T. Baker executives confirmed that their technical assistants could master the B, A, T-definitions and reliably file the accumulating chemical literature by these "BAT" measures, after just a few days of training.

The second serendipic benefit came from the J. T. Baker requirement for a "tabulating" chemical name that would resemble the catalog name closely enough to be used reliably by customers and purchasing agents. Here was a golden opportunity with Baker's nomenclature skills to implement the suggestions for "computer-oriented chemical names" that also were overlooked "sleepers," described with the BATCH digits and first reported at the ACS Meeting in Chicago more than ten years earlier. The J. T. Baker data sheets showed, in addition to the structure diagrams and standardized molecular formulas, typewritten names that were to be put onto the IBM cards; but these names contained lower-case letters, lower-case italic prefix marks, prime marks, and other punctuation marks that were not available with the tabulating equipment at that time. Nor are most of these marks available as international standards in computer centers today. A logical analysis of the prefix marks showed that they had no information content to stockroom managers and purchasing agents, other than "A, B, C" tags assigned to distinguish positional isomers. Thus, the entire Greek alphabet was denoted by punctuating the corresponding tabulating letter with the ampersand mark: A& clearly denoted "alpha," B& denoted "beta," and G& denoted "gamma." An asterisk denoted the prime mark. The hyphen-slash combination denoted the left parenthesis, and the slash-hyphen combination denoted the right one; and so on for a "full service bank" of other tabulating translations.

These two serendipic benefits alone were so encouraging that they led to a feature publication of J. T. Baker "*BATCH*" *Directories* introducing the new line of organic laboratory compounds in 1965 and 1966. At the same time, the mechanical simplicity and adaptability of the accompanying chemical notation for these high-interest chemicals was so clearly demonstrated that the cooperative effort also generated a series of publications—three in 1966[1,3] and in 1967;[4] two in 1968,[5,6] and a number of joint presentations to groups of chemists and information managers. This cooperation also provided an ideal proving ground for the inorganic chemical notations, because no group could claim to be more concerned with practical descriptions of inorganic compounds than the J. T. Baker Chemical Co., American pioneers in this line of products.

The simplest half of this J. T. Baker card file also helped provide balanced guidance—emphasis on frequently met types of structures—in a more spectacular development a few years later. Permission was obtained to use this card deck in response to a request for such testing material from Deena Koniver and her associates in the Division of Computer Research and Technology at NIH. Shortly thereafter, Detrick was rewarded with a list of discrepancies in this manually created WLN file, and a pictorial structure display for the others. This long listing of diagrams was generated from a program "warm-up" exercise by Richard Feldmann, a new member of the NIH team that programmed their PDP-10 computer to flash out WLN descriptions from hand-drawn diagrams, input with a Rand writing tablet. These and related program developments, stimulated by cooperating discussions with the Detrick group, led to a series of impressive NIH publications.[7-10] A fifth paper[11] discusses computer techniques for a nested or structured file of compounds that closely resembles the WLN grouping of rings, heteroatoms, and branches.

Enthusiastic cooperation between Fort Detrick and the Industry Liaison Office at Edgewood Arsenal around 1963–1964 led to friendly rivalries on possible extensions of the WLN in the "third-generation" computer age. One of the first outstanding benefits of this interaction was the so-called WLN Permuting Program developed by Sorter, Gelberg, and Granito at ILO, described in an early series of papers.[12-14]

Just after Peter F. Sorter left the Industry Liaison Office at Edgewood Arsenal to head up new chemical information services at Hoffmann-LaRoche in Nutley, N. J., he demonstrated the ease of use of the WLN by encoding all of the organic structures described in the 1960 edition of the *Merck Index*. He cooperated with Detrick by providing an IBM card file of these WLN records, and after the corresponding molecular formulas were added at Detrick, these became welcome file-building computer input for pilot studies in the U. S. Army's Chemical Information and Data System (CIDS Program) at Edgewood Arsenal. In October 1966, he and other WLN users cooperated with the Army by coming to a two-day CIDS Conference at Edgewood and reporting on their user experiences.[15]

This 1966 conference highlighted the fact that the most significant of all chemical notation cooperative activities were those going on since 1960, when E. G. Smith at Mills College volunteered to write a comprehensive WLN manual with the collective guidance of users like Bonnett and his associates at G. D. Searle in Chicago, Addelston at Winthrop Laboratories in New York, Gelberg and Granito at Diamond Alkali, Bowman and his associates at Dow, Horner at Stanford Research Institute, Renard and his associates at ILO, Hyde and his associates in ICI, Sorter and his associates at Hoffmann-LaRoche, and Wiswesser and his associates at Detrick. The ten pioneering users also became charter members of the Chemical Notation Association, an organization dedicated to provide educational effort, standardizing stability, and open-ended extension of the "official" rules of whatever chemical notation appears to be providing maximum benefits to the majority of users. Thus far, none of these serious users of the WLN have "surrendered" to any other alternative.

The *Chemical Notation Association Newsletter* in April of this year reported that it then had 89 members and associates in United States, 27 in England, 4 in Japan, 3 in

Holland, and 2 in France. "Since the Association was organized less than seven years ago with a membership of ten, these data give some notion of the growth of interest in the Wiswesser notation since that time." This growth in seven years has been from 10 to 125.

The Aldrich Chemical Co. generously provided Fort Detrick with a deck of some 8000 IBM cards showing WLN descriptions of their catalog, and a corresponding deck of molecular formula cards to be merged with them, for potential use in the CIDS file-building experiments and in the planned Fort Detrick computer searches for new or untested screening compounds. At that time (1966–8) Fort Detrick's "computer chemistry" was limited to an incomplete BATCH-generating program for its second-generation UNIVAC (using 90-column cards). The primitive program required a fixed format for the input molecular formula, and Aldrich soon had much more sophisticated substructure-searching programs[16] before Detrick's database programs were completed. However, this early difficulty with a different kind of molecular formula format stimulated the development of a fast yet powerful "checker" program that would accept such format variations and reject input records showing disagreements between the WLN record and the molecular formula, regardless of what format was used for the latter. This experience later proved valuable in a cooperative effort between Fort Detrick and the National Cancer Institute, which involved the processing of some 6000 published descriptions of "Compounds Screened for Carcinogenic Activity."[17] The input molecular formulas for these compounds followed a "N, O, P, S" citing sequence that was a complete departure from the international Hill-CAS sequence of "C, H, then all others in alphabetic order."

Imperial Chemical Industries in 1966 authorized Ernest Hyde and Lucille Thomson to "set up shop" near Montreal and make a feasibility study on the possible "computerization" of ICI files with WLN input records. A "dot-plot" extension of WLN symbols for high-speed structure display with Detrick's old UNIVAC computer, implemented in 1964 and reported in 1966,[18] inspired Hyde and Thomson to create a serendipic benefit of first magnitude at that time—the computer generation of structure diagrams from WLN input.[19,20]

Cooperation among several groups of chemical notation users was a basic part of the "Common Data Base" created by FDA's Science Information Center and the National Library of Medicine, with data contributions from Chemical Abstracts Service, and WLN descriptions from Dow's Computation Research Laboratory. An important related effort was the volunteered file of WLN descriptions from E. G. Smith for all the structures in the Ring Index, processed by the Dow group—with their "Pathfinder" computer program checking the most complex ring systems—and published by Chemical Abstracts Service.[21]

The Institute for Scientific Information, like ICI, carried the WLN "dot-plot" implementation (with A, L, D, T symbols for alkyl groups) far beyond its Detrick beginnings with a powerful open-ended set of programs that generate "Ring Code" substructure-searching cards or bit screens from input WLN descriptions of the corresponding complete structures.[22] While Detrick did not participate in this development, ISI had cooperated with the Detrick group in demonstrating a basically related development—the first large-volume generation of "bit screens" or substructure-searching signals from WLN input records by a relatively simple and fast computer program.[23] These WLN-generated bit screens provided a fiftyfold increase over "string searching" in central-processing speed, by avoiding the tedious string examination for all those records that lacked the necessary "bits and pieces." The Detrick experiences with these bit screens showed that the "exact matching" requirement for the bits alone in searches for specific structures generally reduced a file of 30,000 WLN descriptions to the two, one, or none that contained only those bits and pieces. This very high discriminating power of the combinations of "assortments" of WLN marks again reflects the carefully deliberated selection of WLN symbols.

The 1950 symbol set of course is not as perfect as the extension that Wiswesser devised at Detrick in 1964, with the "dot-plot" letters for separate display of every x-ray reflecting atomic group—e.g., A, L, D, and T for distinct segments of the alkyl chains. Eugene S. Domalski at the National Bureau of Standards acknowledged several years ago that these added symbols, giving a "freedom from chemical bondage" or from undesirable unsaturation marks, were vitally necessary for satisfactory descriptions of free radicals. Meanwhile he demonstrated the generally satisfactory nature of the present WLN descriptions for his compilation of biologically related stable compounds that have reported thermodynamic data. Detrick cooperated with him in helping to provide WLN descriptions for each item in his name index.[24]

Breen and Breen in another division of the National Bureau of Standards provided some rather unusual cooperation with NIH, FDA, and Detrick by initiating corrections on some "confused" inorganic WLN descriptions in the "Common Data Base" previously mentioned. These descriptions were confused by a conflict between CAS rules and WLN rules; or rather, by trying to superimpose the CAS rules over the WLN rules, in violation of "Equal Opportunity" for chemical notations. Koniver's group at NIH provided a card file of all the inorganic records (correct and incorrect) to Detrick, and later these cards were given to the Breens as elementary training material. The younger Breen, Barry, put the "correct" inorganic cards in one pile, and marked the others with what he thought were the correct WLN's. He naturally was not 100% perfect in this training exercise because he was a *junior*-high school student at that time. He had learned about the notation through his sister Bettijoyce Breen, who uses a computer terminal in the NBS Office of Standard Reference Data.

A very gratifying three-way cooperative effort by NIMH, NIH, and Fort Detrick culminated in the *Science* report of June 30th.[25] Earl Usdin wanted to include WLN descriptions with other data in his third publication of "Psychotropic Drugs and Related Compounds," so a first set of notations were retrieved by their CAS Registry numbers from the FDA-NLM Common Data Base. Then Deena Koniver in NIH trained her assistant, a summer pharmacy student, to draw diagrams at the Rand writing tablet for computer-generation of the other needed WLN descriptions. Those that were not accepted by the PDP-10 program were encoded at Detrick. Then the molecular formulas were provided in one set of cards, the WLN descriptions on another, and both were merged at Detrick to be run through its "checker" program. This triple participation slowed down the final production considerably, but very few errors should have escaped the attention of the multiple checkups.

The Flavor and Extract Manufacturers' Association maintain data on a very high interest group of "GRAS Substances,"[26] food additives that are "generally regarded as safe." Representatives of this Food Additives Committee of FEMA, Richard L. Hall and William H. Stahl, provided data sheets to Detrick for "computerizing" this premium group of compounds with WLN descriptions from the diagrams, and names and molecular formulas from the data sheets. An additional deck of cards was prepared to show the value of the Detrick substructure bit screens in the last 20 columns of the cards, in place of molecular

formulas. This FEMA-GRAS collection is less than 2000 cards, so the IBM sorter has become the ultimate in economy for substructure searching at the McCormick laboratory and several other places where the FEMA-GRAS "Educator" deck was copied.

William H. Stahl also initiated another cooperative effort with ASTM and Fort Detrick on a closely related small catalog of high-interest compounds—those having threshold odor or taste data. Again the cards were generated with notations, names, and molecular formulas at Detrick, and ASTM is publishing the computer-compiled results[26] prepared through a contracted center. All agreed that this card deck should have the same layout as the FEMA-GRAS deck with molecular formulas, because both would be closely associated "micro data banks" for information managers who are just beginning to hear about the WLN and how many ways it can help them in the processing of chemical structure information.

The Thermodynamics Research Center at Texas A&M University decided about five years ago to add WLN descriptions to their data bank of some 10,000 compounds with spectroscopic and thermochemical data. Detrick cooperated by offering to help generate the WLN records, but after the first hundred or so were put on cards and mailed to Texas, all the rest were done by the Research Center. At the present time four members of the Chemical Notation Association are from this Center at Texas A&M.

Walter A. Bowles provided an opportunity for the WLN to penetrate the Rocky Mountains at the Denver Wildlife Research Center of the USDI. In January 1971, he offered to provide diagrams on some 4000 repellent-screened compounds and interested the sponsor at U.S. Army Natick Labs in the desirability of preparing an addendum to their forthcoming report that would cite the WLN for each compound. Most of the coding was done by our associates in the Industry Liaison Office at Edgewood Arsenal, and the resulting deck of 4000 cards were shipped to Denver this summer.

In all of these cooperative efforts, we have tried to stress the great need to *promote voluntary standardizations* in chemical information management through user experiences which begin with modest IBM-card catalog investments.

LITERATURE CITED

(1) Barnard, A. J., Jr., Kleppinger, C. T., and Wiswesser, W. J., "Retrieval of Organic Structures from Small-to-Medium Sized Collections," *J. Chem. Doc.* **6,** 41-8 (1966).

(2) Barnard, A. J., Jr., Kleppinger, C. T., and Wiswesser, W. J., "Computer-oriented Chemical Names," *Ibid.,* **6,** 48-57 (1966).

(3) Barnard, A. J. and Wiswesser, W. J., "Some Innovations in Chemical Information Management," NIH Seminar, Bethesda, Md., *Inform. Retrieval Lett.* **2**(6), 1-3 (1966).

(4) Barnard, A. J., Jr. and Wiswesser, W. J., "Computer-serviced Management of Chemical Structure Information," *Lab. Management* **5,** 34-6, 38, 40-4 (1967).

(5) Barnard, A. J., Jr., Broad, W. C., Kleppinger, C. T., and Wiswesser, W. J., "Some Techniques for the Machine Management of Small Chemical Data Systems," in Proceedings of the Wiswesser Line Notation Meeting of the Army Chemical Information and Data Systems Program, pp. 85-101, (EASP 400-408, edited by J. P. Mitchell), Edgewood Arsenal Special Publication, Edgewood Arsenal, Md. (1968).

(6) Wiswesser, W. J. and Barnard, A. J., Jr., "The Retrieval of Chemical Structure Information," *Brit. Soc. Rheol. Bull.* **2,** 3-15 (1968).

(7) Farrell, C. D., Chauvenet, A. R., and Koniver, D. A., "Computer Generation of Wiswesser Line Notation," *J. Chem. Doc.* **11,** 52-9 (1971).

(8) Feldmann, R. J., and Koniver, D. A., "Interactive Searching of Chemical Files and Structural Diagram Generation from Wiswesser Line Notation," *Ibid.,* **11,** 154-9 (1971).

(9) Heller, S. R. and Koniver, D. A., "Computer Generation of WLN. II. Polyfused, Perifused, and Chained Ring Systems," *Ibid.,* **12,** 55-9 (1972).

(10) Miller, G. A., "Encoding and Decoding WLN," *Ibid.,* **12,** 60-7 (1972).

(11) Feldmann, R. J. and Heller, S. R., "An Application of Interactive Graphics," *Ibid.,* **12,** 48-54 (1972).

(12) Sorter, P. F., Granito, C. E., Gilmer, J. C., Gelberg, A., and Metcalf, E. A., "Rapid Structure Searches via Permuted Chemical Line Notations," *Ibid.,* **4,** 56-60 (1964).

(13) Granito, C. E., Gelberg, A., Schultz, J. E., Gibson, G. W., and Metcalf, E. A., "Rapid Structure Searches. II. A Keypunch Procedure for the Generation of an Index for a Small File," *Ibid.,* **5,** 52-5 (1965).

(14) Granito, C. E., Schultz, J. E., Gibson, G. W., Gelberg, A., Williams, R. J., and Metcalf, E. A., "Rapid Structure Searches *via* Permuted Chemical Line-Notations. III. A Computer-produced Index," *Ibid.,* **5,** 229-33 (1965).

(15) Mitchell, J. P., editor, "Proceedings of the Wiswesser Line Notation Meeting of the Army Chemical Information and Data Systems (CIDS) Program, 6th-7th October, 1966." Edgewood Arsenal Special Publication EASP 400-408, 1968. (AD-665,397 from National Technical Information Service, Springfield, Va. 22151)

(16) Buth, W. F., "Fragment Information Retrieval of Structures," *Aldrichimica Acta* **1,** 3-5 (1968).

(17) National Cancer Institute, "Survey of Compounds Which Have Been Tested for Carcinogenic Activity," Public Health Service Publication No. 149 and five supplements (1951, 1957, 1969, 1971, 1972, and 1973).

(18) Wiswesser, W. J., "The 'Dot-Plot' Computer Program," *Abstracts of Papers,* 152nd Meeting, ACS, New York, 1966; and reference 15.

(19) Thomson, L. H., Hyde, E., and Matthews, F. W., "Organic Search and Display Using a Connectivity Matrix Derived from Wiswesser Notation," *J. Chem. Doc.* **7,** 204-9 (1967).

(20) Hyde, E., and Thomson, L. H., "Structure Display," *Ibid.,* **8,** 138-46 (1968).

(21) Chemical Abstracts Service, "Wiswesser Line Notations Corresponding to Ring Index Structures," American Chemical Society, 1968. (PB 180,901 from National Technical Information Service, Springfield, Va. 22151)

(22) Granito, C. E., Roberts, S., and Gibson, G. W., "The Conversion of Wiswesser Line Notations to Ring Codes. I. The Conversion of Ring Systems," *J. Chem. Doc.* **12,** 190-6 (1972).

(23) Granito, C. E., Becker, G. T., Roberts, S., Wiswesser, W. J., and Windlinx, K. J., "Computer-generated Substructure Codes (Bit Screens)," *Ibid.,* **11,** 106-10 (1971).

(24) Domalski, E. S., "Selected Values of Heats of Combustion and Heats of Formation of Organic Compounds Containing the Elements, C, H, N, O, P, and S," *J. Phys. Chem. Ref. Data* **1,** 221-77 (1972).

(25) Koniver, D. A., Wiswesser, W. J., and Usdin, E., "Wiswesser Line Notation: Simplified Techniques for Converting Chemical Structures to WLN," *Science* **176,** 1437-9 (1972).

(26) American Society for Testing and Materials, Committee E-18, "Compilation of Odor and Taste Threshold Values," Philadelphia, Pa., 1973.

(End of Symposium)