At present, two problems must be resolved before official approval of the "perfluoro" system can be sought. The first of these is the term itself: whether to retain "perfluoro" or to invent a new term to express the same intent. For the reasons given earlier the writer favors a new term. The second problem is the inclusion or exclusion, under the meaning of "perfluoro," of hydrogen attached to atoms other than carbon. If "perfluoro" itself were retained and extended unrestrictedly to hetero atoms, names for many organic compounds of oxygen, sulfur, and nitrogen would be completely unacceptable to organic chemists—for instance, perfluoroethanol for CF_3CF_2OF. In order to retain the traditional functional names which are so universally used in organic chemistry (alcohol, amine, mercaptan) fluorine attached to hetero atoms would have to be excluded or severely restricted. The use of "perfluoro" to cover fluorine attached to *all* atoms is mandatory in the inorganic realm, however, if lengthy repetition of locants is to be avoided. It would, therefore, be necessary to maintain two meanings of "perfluoro," each depending on type of compound or system of nomenclature, and such a situation would be sure to cause confusion and ambiguity. Furthermore, a new system for organic compounds which was not equally applicable to inorganic compounds would have at best only temporary utility, as the distinction between organic and inorganic chemistry continues to blur with passing time. A new term for "perfluoro" would carry no traditional connotations and might, by virtue of its strangeness, alert the organic chemist to unconventional properties of the functional groups named subsequently. The best course seems to be the employment of a new syllable which would include, in the sense of "perfluoro," fluorine attached to all atoms, not only carbon, with special provision made to protect the integrity of a very few groups, possibly —CHO, —COOH, and those in which hydrogen is attached to oxygen.

This account is not meant to be a comprehensive description of the anticipated "perfluoro" system, but merely presents its main principles and intentions. We fully recognize that many familiar names are too deeply entrenched to disappear overnight, even in the event of official adoption of the system, and that some of these names, particularly with very simple molecules, are completely acceptable. The system should be most valuable in furnishing a standardized, consistent nomenclature and in naming highly fluorinated molecules of unusual types; its application to simple molecules and to those which contain relatively large numbers of hydrogen atoms is neither necessary nor desirable.

Low-Cost Storage and Retrieval of Organic Structures By Permuted Line Notations: Small Collections*

J. K. HORNER
Stanford Research Institute, Menlo Park, California 94025

Received October 20, 1966

A low-cost storage and retrieval system is described for a small collection of chemical structures. About 3000 structures from the Stanford Research Institute files were coded according to the Wiswesser Line Notation, a computer program was written, and a permuted line-notation index was generated, all for approximately $700.

Earlier reports have shown that permuted chemical line notations in tabulated lists can be used to rapidly locate specific compounds, classes of compounds having similar ring systems, and compounds having the same functional group (*1-4*). Time and cost data on computer-generated indexes are available for large collections of chemical structures (*3*) and time data on key punch-generated indexes for small collections are available (*2*). The object of this communication is to report the experiences of the Life Sciences Research Area of Stanford Research Institute in collecting and storing at a low cost a small collection of chemical structures.

BACKGROUND

In 1963 it became evident to scientists at Stanford Research Institute (SRI) that a central file of data on organic structures was needed to replace the antiquated molecular formula files kept in each section. An organic chemist with neither background nor training in data storage and retrieval was assigned the task of assessing the current methods of chemical data storage and of adapting one to fit the needs of SRI.

Definitive information in this field proved scanty until the publication in 1964 of the "Survey of Chemical Notation Systems," (*5*) a report of the Committee on Modern Methods of Handling Chemical Information. This comprehensive report plus personal contacts with industrial

* Presented at the Wiswesser Line Notational Meeting of the Army Chemical Information and Data Systems (CIDS) Program at US Army Edgewood Arsenal, Edgewood Arsenal, Md., October 6 and 7, 1966.

and governmental personnel indicated that there existed a lack of sophistication in chemical data handling in a majority of governmental agencies and industrial concerns. Those groups having well-conceived chemical data systems usually used punched or notched cards with fragmentation codes, topological codes, or chemical line notations. The codes were usually searched by hand or by means of a computer.

Certain requisite goals were set as a basis for evaluating the existing data programs while searching for the ideal structure retrieval system for SRI. The input procedure was not to require special or expensive equipment normally not found in an automatic data processing section, and the coding, punching, and preliminary manipulations were to be handled by technicians. With this ideal system one could retrieve from the files one specific compound, all compounds having specific ring systems in common, all compounds related by specific functional groups, and all compounds having a specified relation between groups. Most important, no machine should separate the literature chemist from the data files. This latter requirement was important because the computer section and the chemistry department are physically distant and also because a machine cannot browse and select interesting structures which do not fit the search requirements. This browsing can be quite useful, as anyone who has ever looked up a compound in a formula or name index knows, and a literature chemist cannot browse if he is separated from his chemical data.

The topological codes were rejected because they required that the files be searched by computer for any semblance of sophistication. Fragmentation codes are inherently deficient in that they require a decision by the indexer as to what functional groups, bond types, and molecular fragments will be of interest to the subsequent user. The line notation codes could not be encoded by technicians and some notations used symbols not found on automatic data processing machines. Most line codes were searched by computer and none could search for all spatial relationships. The Wiswesser Line Notation (*6*), however, did use symbols familiar to chemists and rejected symbols not found on automatic processing machines. The most appealing use of this notation was described in 1964 by Sorter and coworkers (*1*) at Edgewood Arsenal. This group permuted the notation lists, then printed the permutations and bound them into volumes. The Wiswesser encoding still could not be done by a technician and all spatial relationships could not be found (*7*), but this system now utilized the computer only to prepare the dictionarylike lists which then allowed the chemist to make structure searches at his desk and also permitted him to browse.

A visit with the Edgewood group indicated that over 60,000 compounds were in their system, and that functional group, molecular fragment, and bond-type searches were being successfully completed with a minimum of effort. This group used a revised version of Wiswesser's Line Notation (*8*) and offered to supply their computer program at no cost. SRI decided to use the Wiswesser Line Notation with permuted notational lists.

The structural line coding was done by the author, who learned the Wiswesser Line Notation from the revised manual. An encoding rate of 125 to 150 structures per hour was possible after two to three months of part-time study. Of the 5000 structures ultimately encoded, approximately 10 were sufficiently complex to necessitate encoding assistance from Prof. E. G. Smith of Mills College. The line notations were entered into the Accession Number Index directly under the structures. The notations were checked and most errors were corrected; but because only one person at SRI knew the code, it was not always possible to correct the errors resulting from a misunderstanding of the rules. Some 54 encoding mistakes were discovered after the line codes were permuted and printed.

AUTOMATIC DATA PROCESSING

A computer program for permuting the notations was obtained from Edgewood Arsenal. This program had been written originally in COBAL language for the Univac II, then modified for the Honeywell 400 in EASY II language. The Honeywell language and the ALGOL 60 language of the SRI Burroughs B-5500 were not compatible and a direct translation, though perhaps possible, would not fully utilize the capabilities of the more sophisticated Burroughs machine. Because the computer problem was well defined, a new program was written here at SRI for the Burroughs B-5500 in the source language of extended ALGOL 60.

The combined card punching and verification times were 100 per hour when the SRI punch operators had to turn the pages of the Accession Number Index to read each new line code. When the codes were written on punch card coding forms, the operators then punched and verified 198 cards per hour. The punched card input format was:

Punch Card 1:	Col. 5–10	Identification number and/or letters
	21–79	Wiswesser Notation
	80	Trailer card indication
Punch Card 2:	(Optional)	
	Col. 1–37	Continuation of Card 1

The first punched cards (2808) used as input created 20,294 entries or 7.3 lines per card. The program was written as three sections.

Program A (4.4 minutes): With this main program the 2808 cards were read into the computer and notations were scanned and permuted on the desired elements and functional groups. The Quickscan (*2*) was formed and the lines were edited into a form suitable for sorting. The entries were then written on magnetic tape.

Program B (12.2 minutes): This COBAL program sorted the 20,294 permutations into alphabetical order and wrote them onto a second tape (*9*).

Program C (4.2 minutes): This program did the final editing of the permuted and sorted entries. This final tape was then checked to see if the entry needed one or two lines of output. The page headings and all entries were written onto a printer backup tape.

The total computer processing time was 20.8 minutes, but Program B has now been eliminated and Programs A and C have been combined with the ALGOL sorter, cutting the total processing time to approximately 15 minutes.

The tape from Program C was printed onto paper on a Univac 1004 card processor. There were 132 characters per line and the printout sheet format was:

Col. 1–4	Identification Number
5–6	Blank
7–18	Quickscan area
19, 20	Blank
21–132	Permuted Wiswesser Notation
77	Indexed symbol

TIME-COST DATA

The cost of preparing the permuted index of 2808 cards from the SRI collection was 26 cents per compound. The total cost of $733 (Table I) included the cost of writing the initial program. As this need be done only once, the annual updating should prove rather inexpensive. Our figures show that 400 new compounds could be encoded, the line notations punched into cards, and the new permutations merged with the existing system to generate a new index for about $61. These structures, therefore, could be put into the index for 15 cents each.

Companies, universities, and others with no inhouse data processing facilities would have this phase of the work performed by an outside commercial service bureau. These concerns charge either by "running time" or "lapsed time." The 2808 cards from the initial SRI file contained data requiring 20.8 minutes of actual computer "running time" or 22.0 minutes of "lapsed time"—*i.e.*, time in the computer. A national bank in our local area charges $195 per hour lapsed time for their B-5500. This processing charge would have been $72 instead of $63 as charged by our SRI data section (Table I). Assuming that the card punching and verification charges would be comparable, it appears that groups with small collections of chemical structures and data can file and retrieve information in a sophisticated manner at reasonable cost.

Table I. Time and Cost Data for SRI Compounds

Encode 2808 structures (125/hr.)	22.4 hr.	$112.00
Write program		396.00
Punch 2808 cards (150/hr.)	19 hr.	76.00
Verify 2808 cards (200/hr.)	14 hr.	57.00
Permute 2808 lines	4.4 min.	13.15
Alphabetize 20,294 lines	12.2 min.	36.80
Edit 20,294 lines	4.2 min.	12.55
Print 21,539 lines (500 lines/min.)	43 min.	30.00
Total cost		$733.50

Cost per structure, $733/2808 = 26¢

The four collected volumes of "Organic Syntheses" were encoded next for personal use. The 1850 line notations generated 7885 lines of permuted output or 4.3 lines per card. Table II shows the time and cost data for preparing an index of these structural notations. The card punching and verification times and the costs were less than before because the notations were entered by the encoder onto 35-line coding forms, and the "Organic Synthesis" notations in general were much shorter than the SRI notations. The low cost of 8 cents per compound for generating this permuted index was also possible because the cost of writing the computer program had already been paid.

Table II. Time and Cost Data for "Organic Syntheses" Compounds

Encode 1850 structures (125/hr.)	14.8 hr.	$75.00
Punch 1850 cards (294/hr.)	6.3 hr.	25.40
Verify 1850 cards (617/hr.)	3.0 hr.	12.00
Permute 1850 lines	1.7 min.	5.20
Alphabetize 7885 lines	4.7 min.	14.10
Edit 7885 lines	1.6 min.	4.80
Print 8368 lines (500 lines/min.)	17 min.	11.00
Total cost		$147.50

Cost per structure, $147.50/1850 = 8¢

The combined time and cost data for the SRI and "Organic Syntheses" compounds are shown in Table III. The combined costs equalled $881 for 4658 structures or 19 cents per compound.

Table III. Time and Cost Data for SRI plus "Organic Syntheses" Compounds

Write program		$396.00
Encode 4658 structures	37.2 hr.	187.00
Punch 4658 cards	25.3 hr.	101.40
Verify 4658 cards	17.0 hr.	69.00
Permute 4658 lines	6.1 min.	18.35
Alphabetize 28,179 lines	16.9 min.	50.90
Edit 28,179 lines	5.8 min.	17.35
Print 30,907 lines	60 min.	41.00
Total cost		$881.00

Cost per card, $881/4658 = 19¢

UTILIZATION OF INDEX

The SRI Permuted Index has been used to search generically for ring systems and functional groups and to retrieve specific compounds. During these searches, 54 encoding errors were detected. Most of these errors were due to carelessness. The remainder of the errors were caused by the author's faulty memory of the rules. This failure to retain certain rules is especially serious when one encodes only once a week or perhaps only once a month. Though 1.9% encoding errors did occur, to our knowledge no structures were "lost," perhaps because some encoding errors were serious, in which case the line notation was indexed in a very strange place and thus was quite noticeable. Alternatively, the error was trivial, in which case the notation was indexed either directly before or after the correct location, so again, the error was easily detectable.

A computer program designed to calculate the molecular formula from the Wiswesser Notation has been written (*10*). The calculated formula can then be checked against the correct molecular formula either by hand or by machine. A checker program such as this should do much to allay the encoding problems of part-time chemical data storage and retrieval workers. A loose-leaf notebook containing the structures and notations of the SRI cyclic ring systems is used as a ring system reference book, as an encoding aid, and to determine quickly whether general cyclic structures are in the index.

Over 5000 compounds were encoded and only 10 structures proved sufficiently troublesome to make outside assistance necessary. The relative ease with which one can learn the Wiswesser system is in part due to its use of known chemical symbols and the use of mnemonics. It is also due in part to the efforts of Dr. Smith in improving the original notation. Assisting Dr. Smith in evaluating and improving the Wiswesser Notation are a group of users known as the Chemical Notation Association. These 14 chemical workers all have encoded at least 5000 structures and are constantly using and improving the notation. Newsletters are circulated and formal meetings are held to ensure that the notation will keep pace with chemical advances and that problems which arise will be promptly discussed and solved. At least 16 structure files coded according to the Wiswesser Line Notation are being maintained. These files contain over one-half million compounds.

ACKNOWLEDGMENT

The author is indebted to Mr. D. A. Kerr who prepared the SRI Molecular Formula and Accession Number Indexes and to Mr. Wm. S. Duvall of the SRI computer section who wrote the ALGOL program.

LITERATURE CITED

(1) Sorter, P. F., Granito, C. E., Gilmer, J. C., Gelberg, A., and Metcalf, E. A., J. CHEM. DOC. **4**, 56 (1964).
(2) Granito, C. E., Gelberg, A., Schultz, J. E., Gibson, G. W., and Metcalf, E. A., *ibid.*, **5**, 52 (1965).
(3) Granito, C. E., Schultz, J. E., Gibson, G. W., Gelberg, A., Williams, R. J., and Metcalf, E. A., *ibid.*, **5**, p. 229.
(4) Gelberg, A., *ibid.*, **6**, 60 (1966).
(5) "Survey of Chemical Notation Systems," National Academy of Sciences-National Research Council Publication 1150, Washington, D. C. 1964.
(6) Wiswesser, W. J., "A Line-Formula Chemical Notation," Thomas Y. Crowell Co., New York, N. Y., 1954.
(7) Personal communication with Mr. Ernest Hyde of the Central Research Laboratory of Canadian Industries Ltd., McMasterville, Quebec, has shown that all substructures and spatial relationships can now be searched by computer-generated fragments of the Wiswesser Notation.
(8) Wiswesser, W. J., "A Line-Formula Chemical Notation," revised by E. G. Smith, McGraw-Hill, to be published, 1967.
(9) The very fast ALGOL sort routine was not available when this program was written. The new routine will perform the sorting operations in approximately half the time.
(10) Personal communication with Dr. C. M. Bowman of the Dow Chemical Co., Midland, Mich.

Atom-by-Atom Typewriter Input for Computerized Storage and Retrieval of Chemical Structures*

J. M. MULLEN
Shell Development Company, Emeryville, California 94608

Received November 11, 1966

Novel features have been added to a paper tape typewriter having a removable typing element: A symbol set has been devised which requires only nine characters for typing common chemical structures. The typewriter has an uncoded "INDEX" key which advances the paper without carriage return. A companion key, "BACK INDEX," was provided which directly retracts the paper. Both have been coded. A tape record containing information sufficient for a computer to calculate an atom-bond connection table for a chemical structure is obtained by typing the structure in any order solely from the keyboard or by use of the reader with prepunched tapes containing frequently occurring substructures. Cost was about one-fourth that of earlier paper tape chemical typewriters.

Only A. P. Feldman (*1*) and his colleagues at Walter Reed Army Institute of Research are known to have made successful prior efforts to use a paper tape typewriter as a computer input device for chemical structures. Their current machine, the Army Chemical Typewriter (ACT), uses a three-shift keyboard; one shift contains a set of 39 chemical symbols, modified slightly from those of Miller and Fletcher (*2*) at American Cyanamid. The tape of the ACT also records the X and Y coordinates of characters as they are recorded on paper. The ACT is technically attractive and seems quite appropriate to the Army's large problem, but its price of $18,000 is questionable for an industrial laboratory.

* Based on a paper presented before the Division of Chemical Literature, 150th National Meeting of the American Chemical Society, Atlantic City, N. J., Sept. 13, 1965.