The "box" is simply limits, left-hand, top, and bottom, of the structure. The punctuation is the right-hand limit. This makes it easier for the keypunch operator to enter the structure onto a punch card.

Arbitrarily I have limited the number of characters in the horizontal line on cross-hatched paper to 98 since the computer is limited in the length of line it can print. Furthermore, I have limited the number of rows arbitrarily to 13 to accommodate our particular report-writing program.

As the record is being transformed from the punched card format to the output format, the 704 calculates certain control information it needs in order to remember how to reproduce the structure when called upon. It counts the number of horizontal rows in the structure (three or five for the benzene ring); it also counts the number of columns in each row of the structure beginning at the left edge as indicated by the box line and terminating with the punctuation. Editing is then performed by the 704 to ensure that the structure does not contain an incorrect number of rows and that each row does not contain an improper number of characters.

Incorrectness occurs, of course, as a result of human errors. I have noted two empirical coincidences; both of which are used in this editing process: (1) the number of rows in the rewritten structures must of necessity always be odd, 1, 3, . . . , 13; and (2) the number of squares from the box in any row must be odd, since all elements are single-character symbols as are the bonds. However, the last position in every row is signaled with a punctuation mark. Thus the number of characters, including blanks, in a row is always even, two through 98. When errors are detected, the structure is rejected from entrance to the permanent file. If correct, the control numbers are stored as a part of the structure record. After the structure is stored the computer proceeds to calculate the molecular formula and insert it in its proper place in the record.

As the structure is being analyzed, further checking by the machine is made to determine the accuracy of the input data. The rules for rewriting structures are integrated in the program so that the computer is able to take a sophisticated *look* at the chemist's rewritten structure and the keypunch operator's work. It will not allow any atom to have too many or too few bonds, nor is a 7-bond permissible with atoms for which ionic bonds are not "legal." Improper atom and bond symbols and misplaced characters are recognized by the computer.

A chemist proposing a machine search of the chemical structure files must state precisely the elements and bonds he wants, how they should be connected, and what he does *not* want. These search specifications are transformed to IBM cards, and become the set of rules by which the computer will perform the search.

Search questions such as finding all the derivatives of resorcinol are meaningless to the computer. The chemist posing the question must determine for the computer what is meant by a derivative. It can pull out all phenols, all compounds having a benzene ring structure, or the computer can pull out all compounds containing two hydroxyl groups. Control data include the molecular formula requirements and the substructure, rewritten in the same manner as the structure records. Further control information is utilized to control the switching network during the search.

REFERENCES

(1) W.H. Waldo, R.S. Gordon, and J.D. Porter, *Am. Document.*, **9** (1), 28 (1958).

(2) W.H. Waldo and M. DeBacker, Preprints of the International Conference on Scientific Information, Washington, D. C., Nov. 1958, Area 4, pp. 49–68.

Application of a Line Formula Notation in an Index of Chemical Structures*

By. H. T. BONNETT and D. W. CALHOUN
G. D. Searle & Co., Chicago 80, Illinois
Received August 24, 1961

Any laboratory engaged in synthetic organic chemistry finds an index to past efforts essential. Over a period of years the total number of compounds studied often becomes large enough to make inviting the use of machinery in the creation and use of indexes to such files of compounds.

Such was the situation in the Searle Laboratories. Over a period of about twenty-five years the laboratories had made and screened thousands of compounds for biological activity. New compounds were being entered into the file at an increasing pace. This experience, of course, is not unique.

Our initial effort was directed to a chemical structure index. While many types of information could be put in such an index, the information we chose included the following: (a) the structure of the compound; (b) the name of the chemist who submitted the compound; (c) the identifying number of the compound; (d) a code for rough classification; (e) a functional group index. To this list should be added the desire that the format chosen be capable of extension to compounds taken from published literature.

The design of any index must be tailored to the facilities available. Until accounting type equipment for research purposes could be justified by use, it was determined to use the facilities of the accounting department tabulating section. The equipment available included the usual IBM key punch, verifier, sorter, collator, duplicating punch, and tabulator machines widely used in accounting operations. The basic indexing principles adopted, using this equipment, was to prepare one punched card per compound, and to translate questions asked of the index into the corresponding manipulations of the punched cards. But the machinery is located in another building and is not available at all times. These factors obviously invited development of a physical form of index which would not require manipulation of cards for all searches. To meet this need, it was desirable to sort the file into

ordered sequence by structural concepts and then to print the resulting ordered sequence.

Inclusion of the structure on an index card is attractive. Several methods are in use or have been proposed. An obvious method is that of applying the drawn structural formula to the card. Chemists like this; by training and experience they are accustomed to the use of structural formulas. Unfortunately, simple machines can do nothing with pictures and, in view of card wear, reproduction of decks becomes necessary at too frequent intervals. Reproduction of decks employing drawn structural formulas is a chore. Computers present other approaches: *e.g.*, Waldo[1] has programmed the IBM 702 and 704 to print out a formula. Opler[2] has programmed the IBM 704 to display structural formulas on a cathode ray tube face. These machines are costly and, in any event, unavailable to us. We chose to explore the Wiswesser[3] line-formula notation, for which an adaptation to simple unmodified accounting equipment had been published.

Benson[4] and Smith[5] have demonstrated that the Wiswesser notation can be used with the basic accounting equipment manufactured by Remington Rand and IBM, respectively. Smith's work in particular has been extensive in that his file now includes more than 50,000 compounds which have been punched into IBM cards.

Chemists for a long time have used line formulas in describing structures of compounds. The line formula notation is simply an extension and codification of this practice. The basic idea is to employ letter symbols for functional groups and numbers for alkyl chain sizes; these segments then are cited in connecting order from one end of the molecule to the other. By citing first the symbols for terminal functional groups or non-benzene ring structures if present, the notation focuses attention on structural elements that tend to characterize compounds and determine their properties and uses. In alphabetized lists of notations this principle tends to bring together related compounds, and therein lies the ability of the notation to create indexes of considerable power.

Furthermore, the notation can be read at sight like a structural formula. This feature of visual recognizability is a second important utility—the index cards themselves, dropped in a search or tabulated into a list, can be scanned to select desired compounds.

These points are illustrated in Table I containing a list of simple compounds as they would appear in our tabulated list index (but with names added here).

In our lists the notation prefix (discussed later) is used as the means of dividing the deck (and the lists), enabling us to enter directly into any given major section. The classification number offers a means of direct access to finer subdivisions of the file when required. In non-benzene cyclic structures the description of the ring nucleus is enclosed within "L---J" or "T---J" pairs (for carbocyclic or heterocyclic structures, respectively). Numerals within these pairs of symbols denote ring sizes and if followed by "T" denote saturated rings. Heteroatoms are indicated by letters and ring locations (locants) are identified by letters *preceded by a blank space*. Substituents attached to the rings are described in locant sequence, by first giving the locant immediately followed by a recitation of structural elements in connecting order.

Table I
Tabulated List Index

Notation	Notation prefix	Class no.	Name
E1	&	&610	Bromomethane
E2	&	&610	Bromoethane
G1	&	&610	Chloromethane
G2	&	&610	Chloroethane
Q1	&	&010	Methanol
Q2	&	&010	Ethanol
Q2Q	&	&020	Ethylene glycol
Q3	&	&010	Propanol
Q3G	&	&620	Trimethylene chlorohydrin
QV1	&	&020	Acetic acid
QV2	&	&020	Propionic acid
Z2	&	&110	Ethylamine
Z2Q	&	&120	Ethanolamine
Z4	&	&110	Butylamine
QR	0	0010	Phenol
QVR	0	0020	Benzoic acid
L6TJ	1	1000	Cyclohexane
T6NJ	1	1110	Pyridine
T6NJ BG CVQ	1	1740	2-Chloronicotinic acid
T6TM DMJ	1	1220	Piperazine
T6TM DOJ	1	1120	Morpholine
T6TMJ	1	1110	Piperidine
L66J	2	2000	Naphthalene
T56 BMJ	2	2110	Indole
T66 BNJ	2	2110	Quinoline
T66 CNJ	2	2110	Isoquinaline
L B666J	3	3000	Phenanthrene
L C666J	3	3000	Anthracene
T C666 BM ISJ	3	3520	Phenothiazine
LSA FVJ E OQ	4	4020	Estrone
LSAJ E FQ OQ	4	4020	Estradiol
LST FV MU OVJ A E	4	4021	Androstenedione
LST MU OVJ A E FV1	4	4021	Progesterone
LST MU OVJ A E FV1 FQ	4	4031	17-Hydroxyprogesterone
LST MU OVJ A E FV1 GL AT6TNJ	4	5131	16-Piperidylprogesterone

Definitions

Letter preceded by a space = Locant, designating position in a ring system
E = bromine
G = chlorine
M = —NH—
Q = hydroxyl
U = double bond
V = carbonyl—CO—
Z = —NH_2
L---J = Parentheses enclosing carbocyclic ring.
T---J = Parentheses enclosing heterocyclic ring.

Abbreviations

"ST" for "E5T B6T6T6T"[a]
"SA" for "E5T B6T6T6"[a]
A locant standing alone denotes a methyl group, being an abbreviation, *e.g.*, "*A*" for "*A1*."

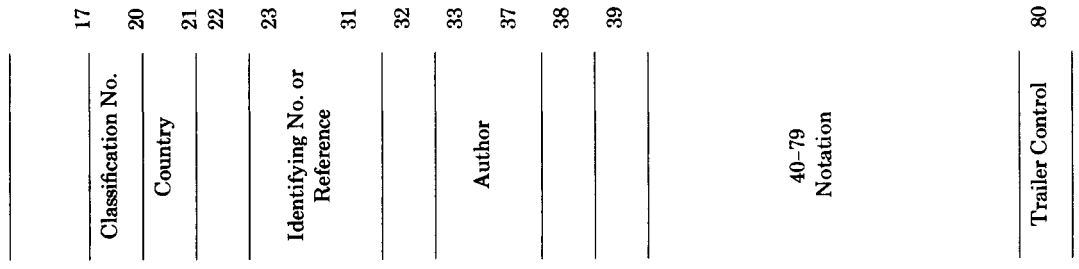
[a] These abbreviations are ours.

The details of the card format that has been developed thus far will be understood from Table II and the discussion which follows.

Each of the reference, chemist's name (author), and notation fields is preceded by a single column code (in cols. 22, 32, and 38, respectively) to increase the versatility of these fields.

Half the space on the card is reserved for the line-formula notation. Our choice of 40 spaces for the notation rests upon several factors: (1) The model 402 IBM tabulator that was initially available to us is limited to 43 alphabetic printing positions per line, of which we use 3 to indicate the chemist. (2) Forty spaces are adequate for the majority of our compounds. A test sample of our compounds indicated: (a) an average notation length of about 22 spaces; (b) that about 96% of our compounds would fit into 40 spaces and, (c) by abbreviating two types of

Table II.
Card Format



ring systems 98.5% would fit. (3) The other information to be recorded requires upwards of half the card. The various fields on the card are discussed later.

THE NOTATION FIELD-COLUMNS 40-79. The alphanumeric notation description of the compound structure is recorded in columns 40-79. For long notations additional cards are required, and column 80 is used for trailer control.

The symbols employed in writing the notation are the blank space, the ampersand, 26 capital letters, and the ten numerals. (In addition, the hyphen may be employed if one has access to a model 407 tabulator.) Additional symbols are available on the 407 tabulator; these, however, require a three-hole punching pattern, and thereby complicate sorting and the ranking sequence of symbols.

Column 39 is used for a prefix preceding the notation. This prefix is shown in Table III, and serves to distinguish inorganic compounds, aliphatic compounds, benzene compounds (in which benzene is the only cyclic structure), and other cyclic structures. A numeral in this column indicates the number of rings in the first described (*i.e.*, highest priority) cyclic system in the compound. While Wiswesser[3] has suggested the use of a prefix symbol, we have departed from his manual by using the prefix to determine priority of a compound or ring system. This means that a bicyclic ring system outranks a monocyclic, after which non-benzene ring compounds divide into carbocyclic or heterocyclic groups on the second symbol (*i.e.*, "L" = carbocyclic, "T" = heterocyclic).

Table III.
Notation Prefix

Blank space – Inorganic compound
& – Aliphatic compound
0 (zero) – Benzene compounds (benzene as the only cyclic structure)
Numerals 1-9 – The number of rings in the highest priority non-benzene ring system in the notation (9 means 9 or more).

Column 38 carries a code to distinguish the type of material recorded in columns 39-79. We use the numeral "1" in column 38 to indicate that the material which follows is the structure of a compound in Wiswesser notation. Any file of compounds, however, is apt to include compounds for which a structure is not available, and it may be necessary to resort to other means of describing the compounds, such as an arbitrary English word name, the empirical formula, etc. (For housekeeping purposes we use a card for each number even if it merely carries the legend "No Compound.") If we use English words in lieu of the structure of a compound, we record the numeral "2" in column 38, and in addition, leave columns 39 and 40 blank. In this way we guarantee that when we do an alphabetic sort on the field represented by columns 39-79, any such alphabetic material will sort out (and print) at the beginning, followed by inorganic compounds, aliphatic compounds, benzene compounds, and finally other cyclic structures arranged in order of complexity from monocyclic to polycyclic.

AUTHOR FIELD. Column 32 carries a code to distinguish the type of name recorded in columns 33-37. If a compound has been made by a chemist in our laboratories, column 32 carries the numeral 1, followed in columns 33-37 by the initials of his first two given names, and the first three letters of his last name. This field is used in the same manner, but with other code numbers, to identify industrial laboratories, universities, assignees of patents, etc.

REFERENCE FIELD. Column 22 carries a code indicating the type of reference recorded in columns 23-31. If the compound in question has been made in the Searle laboratories, column 22 carries the numeral "1," and a five digit number is recorded in columns 27-31, which we call an SC number (for Searle Compound). In cards representing other than internal compounds, column 22 is used to distinguish patents, journal articles, etc., and the 9 spaces from columns 23-31 are adequate to record the year of publication in one column using letters or numerals (numerals for the 1960 decade), 3 columns for journal identification, and 5 columns for page reference.

Column 21 is used to distinguish the country of issue or publication of patents and journals, respectively.

CLASSIFICATION NUMBER. Columns 17-20 are used to record a four digit classification number of which each digit, however, is in itself a code. Column 17 is a ring index, and is shown in Table IV.

Table IV.
Ring Index

Blank – Inorganic compounds
& – Aliphatic compound
0 (zero) – Benzene compounds
1-9 – Other cyclic structures (total number of non-benzene rings in the compound)

The symbols recorded in this column parallel those used in the notation prefix, except that the TOTAL number of non-benzene rings in a compound is recorded in the case of cyclic compounds.

Column 18 is used for a code indicating the types of hetero atoms present in a compound. The code for this is

given in Table V, which requires no comment, except to note that the numeral employed is the highest applicable number. We do not overpunch in this column.

Table V.
Heteroatom Index

0 - C, (H, O)
1 - C, (H, O) + 1N
2 - C, (H, O) + 2N
3 - C, (H, O) + 3 (+)N
4 - C, (H, O) + S
5 - C, (H, O) + N + S
6 - C, (H, O) + halogen
7 - C, (H, O) + halogen + N and/or S
8 - C, (H, O) + Rarer atom (*i.e.* other than C, H, O, N.S. and halogen)
9 - Inorganic cpd.

Column 19 is used to record the total number of hetero atoms. The code used is given in Table VI. The code parallels the IBM sequence of symbols.

Table VI
Heteroatom Count

IBM Symbol	Meaning
0-9	0-9
&	10
A-I	11-19
- (hyphen)	20
J-R	21-29
/ (slash)	30 and 31
S-Z	32-39 or more

The slash is a 2-hole symbol including the zero and "1" punches; thus, no distinction is possible between 30 and 31.

Column 20 is used to indicate the type and amount of non-aromatic unsaturation present in a compound. Both zone and numeric punches are used. Double bonds are indicated by numeric punches. Triple bonds are indicated by zone punches, the 12-punch for one triple bond, 11-punch for two triple bonds and slash mark for three or more triple bonds. Again, the punching pattern parallels IBM symbolism so that the number of triple and double bonds is derived readily from the symbols printed. Thus, if a compound contains one double-bond, the numeral "1" appears in this column, whereas if it contains one triple-bond and one double-bond, the letter "A" will appear in this column. Two triple-bonds and one double-bond would appear as the letter "J," etc.

There is nothing "sacred" about the concept built into this "classification" number. We have found these four codes useful; they remain after discarding others that didn't prove useful.

COLUMNS 1-17. Columns 1-5 are used to record a serial number on compounds taken from literature sources. For such compounds we have found it useful to draw the formula along with precise identification, such as the author's number, example number, and the like, on a serially numbered formula sheet from which we write the punch instructions to keypunch operators.

Columns 6-10 are used to record a serial card number (punched into the cards after the alphabetic sorting on the notation) so that the decks can be returned to alphabetic arrangement by a numeric sort of these columns. Such a device probably would not be needed if the file existed in magnetic tape form, for example.

Columns 11-16 are unused at present.

The design of the system provides a total of 8 indexes of which we are printing 7 with a tabulator into list form. These are: (1) structure sequence, (2) numeric sequence, (3) author sequence, (4) ring index, (5) heteroatom index, (6) heteroatom count, (7) unsaturation index.

The structure sequence is achieved by sorting the file of cards alphabetically beginning at column 79 and moving toward and including column 39. This operation causes the file to be organized rather effectively according to structural concepts because such concepts govern the rules of priority upon which the notation itself is organized.

The sorting sequences we use for the ring index, heteroatom index, heteroatom count and unsaturation index are:

Index	Sorting Sequence
Ring index	Cols. 20, 19, 18, 17
Heteroatom index	Cols. 20, 19, 17, 18
Heteroatom count	Cols. 20, 18, 17, 19
Unsaturation index	Cols. 19, 18, 17, 20

Columns 19 and 20 require alphabetic sorting.

We are not in position to report on our functional group field. We have tried several formats that left something to be desired. A few comments may summarize our experience adequately. We believe a relatively small direct code field should be adequate, and that it should be possible to base this on structural fragments appearing also as part of the notation.

Wiswesser[6] proposed a compact field. Smith[5] concluded Wiswesser's design was inadequate especially in not providing enough information about how the fragments are connected. Our attempts also have shown the desirability of indicating how the fragments are connected, including such fairly detailed information as: (1) attachment to rings, (2) attachment to non-carbonyl carbon atom, (3) attachment to carbonyl carbon atom, (4) attachment to non-carbon atom, (5) multiple occurrence, (6) positional relation of hetero atoms in rings, (7) ring sizes, (8) aromatic rings, (9) hetero rings and the like.

The arrangement of the information in the tabulated lists need not parallel the arrangement on the card, but we have found it useful to print the notation prefix in front of the classification number separated by a space. This device aids in spotting the presence of low-ranking ring systems attached to and hidden behind higher ranking systems because the ring index digit of the classification number is larger than the notation prefix.

The system, as we are currently operating it, includes the seven tabulated lists mentioned previously, supplemented by a cumulative list of new incoming compounds arranged according to the ring index version of the classification number. The notation is used for all searches. The classification number is always used in addition for searching the cumulative supplement.

Answers to search questions are delivered as SC numbers. Members of our staff routinely receive copies of all structures taken into the system.

Which of the indexes is used for a given search depends on the search question. The most common search question contains sufficient structural detail to make the index-by-structure the logical choice. The search is simply a telephone directory type of look-up. The purpose of the classification number is to produce relatively fine subdivisions of the total file. It often is surprisingly effective, *e.g.*, a search of our file of several thousand steroids for

those containing triple bonds—in any position, or in a specific position—is run in the unsaturation index and requires about ten minutes. (Here the unsaturation digit must be a character utilizing a zone punch and the ring digit must be four or more.) A search for sulfonamides would be limited to the heteroatom index groups x54 ⁻ x; x75 ⁻ x and x85 ⁻ x because the occurrence of N and S together is limited to the heteroatom index digits 5, 7 and 8, and there must be at least 4 and 5 heteroatoms respectively present. If ring index structural detail is supplied in addition, the search categories can be still further restricted. On the other hand, a problem that would have yielded easily to a search in the functional group field, had it been available, turned up about 450 compounds and required about 4 hours. In view of the performance of this more versatile index, we have discontinued our molecular formula file.

In our experience the generic searches which must be handled by means of a functional group field are in the minority. With some ingenuity in the use of the various index tools available in the present system, we have been able to provide answers in a reasonable length of time. We do not wish to be understood, however, as denying the desirability of a functional group field—we intend to build one—unless availability of variable field searching equipment that would operate on the notation itself makes this unnecessary.

Some confusion and perhaps misconception exists as to the utility of the Wiswesser notation of itself as an indexing tool. From our favorable experience we would comment: (1) it provides a means of recording structures in unmodified machine language in punch cards and tabulated listings; (2) the notation can be scanned much like a structural formula, and (3) organized alphabetically, it provides a structural index of considerable power which nevertheless has limitations for many generic searches.

The most serious limitations and problems we have encountered in using the line formula notation are: (1) First, it must be said, is in mastering the rules themselves. They are not always clear in application, a difficulty which has been pointed to previously.[7]

(2) A given desired functional group may not appear in the same columns from compound to compound. Fixed field equipment does not deal with such search problems effectively. That is why a functional group field is needed. The story ought to be different, however, with variable field equipment. (In selected searches we have used the tabulator successfully to respond to a symbol or locant within a 10 column field and print out the complete card detail.)

(3) A second inherent limitation, that of subordination (*e.g.*, subordinant independent rings) probably is inevitable in any system of rules. Structural features within the notation may offer search difficulty. A variant of the subordination problem is inversion, which occurs principally in aliphatic and benzene compounds. Here the starting point of the notation is determined in an arbitrary way by the alphabetic rank of the terminal symbols.

(4) Isomerism also may present a problem. We have found it desirable to indicate configuration in steroids by a code appearing at the end. While Wiswesser has suggested symbols to indicate configuration, such symbols, if used within the notation, are likely to modify the sorting sequence.

(5) Confusion of zero and the letter "O." It is unfortunate that IBM initially did not design these two symbols to be visually distinct (they are distinct to machines). This has not been a problem in our experience, but we could see how it might be to anyone working with long aliphatic chains. (Recently, we have been informed that IBM has modified the zero in their newer equipment.)

(6) The problem of updating is due not to the notation, but to our use of tabulated listings which, of course, are a "frozen" file. In practice, this has not been burdensome. We supplement our main listings of compounds with a list which is cumulated quarterly. (We also have used a supplemental duplicate deck of cards.) This list could be sorted on the structure, but we have found listings on the basis of the classification number to be adequate. The cumulative supplement is merged into the main file not more often than annually.

(7) Trailer cards. Long structure notations require more than one card. The cards used for trailer situations have an opposite end cut to that of the main deck. Such cards carry punches in column 80, the first card being punched with the numeral "1," the second with numeral "2," etc.

When the file is being sorted alphabetically on the notation to produce a structurally ordered sequence, the second and succeeding trailer cards are removed before beginning sorting, and later inserted by hand. This is necessary because the symbolism on the trailer cards differs from that on the number "1" card so that the trailer cards do not travel together during the sorting operation.

Punch instructions are recorded on a specially designed form on which the information is written in the sequence it will appear on the card. The notation is the last field and the punch operator need not count spaces, but the coder must write only 41 notation symbols per card. Good handwriting is desirable in coding. If a space is meant the coder writes it as an underline. Punch instructions are checked before punching, and the punched cards are verified against the coding sheets. The coding and code checking operations (not including the functional group field) require a total of about two to three minutes for the usual compounds.

BIBLIOGRAPHY

(1) (a) W.H. Waldo, R.S. Gordon, and J.D. Porter, *Amer. Doc.* **9**, 28 (1958); (b) W.H. Waldo and M. De Backer, Proc. Int'l. Conf. Scientific Information, 1958, Area 4, p. 49.
(2) (a) A. Opler, *Chem. Eng. News*, April 28, 1958, p. 108. (b) A. Opler and N. Baird, *Amer. Doc.*, **10**, 59 (1959).
(3) W.J. Wiswesser, "A Line-Formula Chemical Notation," Thomas Y. Crowell Co., 1954. The adaptation to accounting equipment is discussed on pages 33 and 120. The use of a prefix is mentioned on page 120.
(4) F.R. Benson, "Recording and Recovering Chemical Information with Standard Tabulating Equipment," American Chemical Society National Meeting, Sept., 1953, unpublished.
(5) E.G. Smith, *Science*, **131**, 142 (January 15, 1960).
(6) W.J. Wiswesser, "Advances in Chemistry Series," No. 16, p. 76 (1956).
(7) *Chem. Eng. News*,**33**, 2838 (July 4, 1955).