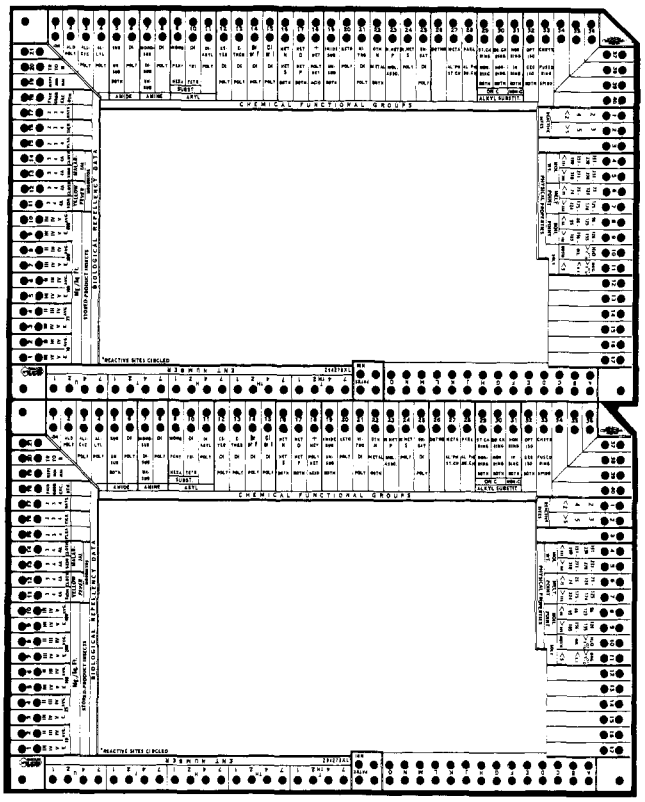


Fig. 2.—Sample copy tandem card with blank center.

to be used. Three copies were made of each tandem original and all were cut in two to give 1,200 original cards and 3,600 copies or four card sets.

The originals were then edge-marked to denote which of three possible positions were to be punched: shallow, intermediate, or deep. Each original was combined with its three copies, and the four were hand-punched simultaneously. (It is important that the intermediate punch cut sharp—otherwise tearing may result.) Machine punching is not practical because the available machines accept only three cards at a time, and punch shallow only. A check was made on all cards to be sure all marked positions were properly punched; improperly punched positions were corrected with McBee card savers. Finally, the cards were separated into the four respective sets, the original set being retained at MRI for subsequent correlation studies. A completed original card is shown in Fig. 1.

The final cost, including the costs of the Xerox machine, cards, and additional time to punch and handle the copies, amounted to approximately 12¢ per card or $144 per extra set. This sum includes approximately 4¢ per special card, 5¢ per card for the Xerox service, and 3¢ for additional labor. Additional expense was incurred in the present program for the purchase of a number of extra original and copy cards to allow for future expansion of the card file.

The only disadvantage to the method is the necessity of having inplant access to a Xerox machine. Going outside for this service would considerably increase the cost.

A Formula Index for Silicon Chemicals*

By CAROL M. LAUER and FRED R. WHALEY
Research Laboratory, Linde Company, Division of Union Carbide Corporation, Tonawanda, N. Y.
Received October 26, 1962

In the deep indexing of literature pertaining to silicon chemistry at Linde Tonawanda Laboratories Information Center,[1,2] the need arose for a mechanized structural formula index which would allow structural configurations and functional groups to serve as a means of identifying complicated molecular species. By suitable coding, formulas are arranged in a linear array based on structure. Such a structural code is most useful in identifying a compound where ambiguity may exist in the nomenclature. A further use when making searches, is to help find compounds having particular structural or functional features.

Because of local equipment limitations, floating field selection was impossible; common structural notations such as those of Dyson[3] and Wiswesser[4] with their codes of varying length could not be used. It was therefore necessary to develop a coding system centered around the silicon atom, and this was done around 1955. The system uses a coördinate index and machine-punched IBM cards. At present over 6,000 chemicals have been coded, and the system has proved valuable in processing retrieval questions in the silicon chemicals field.

It is not the function of this paper to show the complete coding for all the structural elements deemed important by our technical staff, but rather to show the principles involved along with examples for illustrative purposes.

Silicon Configurations.—Silicon compounds are characterized by (1) the number of silicon atoms in the molecule and (2) the number of silicon atoms in each of the possible configurations:

$(R_2SiO_{2/2})_n$	R_4Si
$(RSiO_{3/2})_n$	R_3Si
$(R_3SiO_{1/2})_n$	R_2Si

$SiO_{1/2}$ in the above formulas indicates an oxygen bridge to another silicon atom; R represents any other type of linkage to the silicon atom. This may be an organic, metal-organic, or inorganic radical. Columns 2 through 7 of the Si Formula Index Card (Fig. 1) show the con-

* Presented before the Division of Chemical Literature, ACS National Meeting, September, 11, 1962, Atlantic City, N. J.

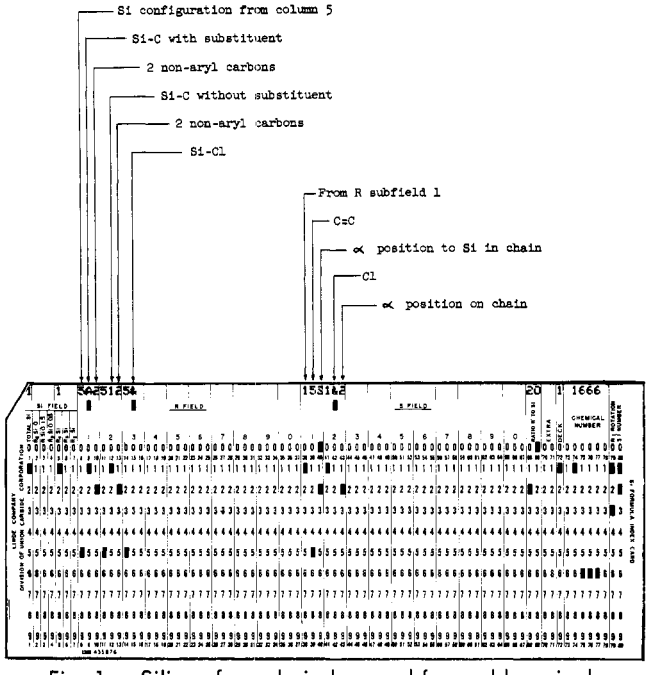
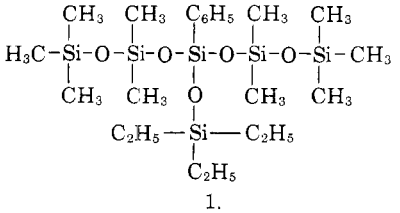
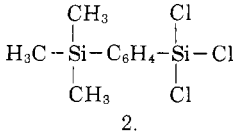


Fig. 1.—Silicon formula index card for α-chlorovinylethyldichlorosilane.

figurations. The numbers are punched in the respective columns (a 9 punch means 9 or more). The first three configurations designate siloxanes, provisions being made in the code for instances where some element other than oxygen acts as a bridge between silicon atoms. The latter three configurations represent the silanes, that is, all configurations that do not contain an Si-O-Si bond. For any given silicon compound there may be silicon atoms in different configurations, as shown in the examples:



1.



2.

In the first example there are three silicon atoms with an $R_3SiO_{1/2}$ configuration, one silicon atom with an $RSiO_{3/2}$ configuration, and two silicon atoms with an $R_2SiO_{2/2}$ configuration. In the latter example there is one R_4Si and one R_3Si. Operating rules within the system call for the assignment of a bridging R group to a particular silicon atom. The code also distinguishes between the cases where different R combinations exist within the same silicon configuration. In example 1, for instance, of the three silicon atoms with an $R_3SiO_{1/2}$ configuration, two have R = methyl and one has R = ethyl. Figure 2 illustrates the coding in the Si field for examples 1 and 2. In example 1, the combination of a 3 punch and an 11 overpunch in column 4 to indicate different R combinations, prints the letter "L."

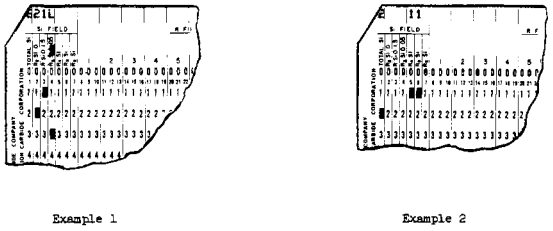


Fig. 2.—Typical coding in the Si field.

In addition to showing the total number of silicon atoms, column 1, by an overpunch, also gives information concerning silicon atoms in cycles, polycycles, or in cycles with carbon atoms.

The Skeletal R Group (The Si-R Linkage).—The next level of analysis of the silicon chemical structure is the identification of the skeletal R group. The R group include Si-C linkages where C is non-aryl or aryl, Si-S, Si-H, Si-OH, Si-X (Cl, Br, I, F), as well as cases where other elements such as oxygen or nitrogen are attached to silicon or bridge silicon to carbon. Each R group is coded in a three-column subfield (columns 8, 9, 10) with ten such subfields available (R field, columns 8 through 37; Figure 1). In the first column of a given subfield the R group is related to the specific silicon configuration by punching the column number of that configuration as shown in the Si field. The linkage to the silicon atom is shown in the second column of the respective subfield. An overpunch is used in this column to indicate the presence of a coded substitutent. The third column shows the number of non-aryl carbons present and indicates (by an overpunch) the valence of R, that is, whether or not the R is divalent and therefore links two silicon atoms or forms a ring with a single silicon atom. This column is left blank if there are no non-aryl carbons and R is monovalent relative to silicon. Using α-chlorovinylethyldichlorosilane as an example, Fig. 1 shows the codes for each R group.

Internal rules of precedence have been established to govern the order of coding of R groups. These rules are particularly useful for coding a chemical with an ambiguous name in order to determine if it has been coded previously.

Substituents on R Groups.—The third level of analysis of the silicon chemical structure is the identification of substituents (S groups) in or on the skeletal R groups. Each substituent is coded in a three-column subfield (columns 38, 39, 40) with 10 subfields available for use (S field, columns 38 through 67; Fig. 1). In the first column of a given subfield, the substituent is connected with the R in the R field on which it is appended by punching the number of the R subfield. An overpunch is also used here to show whether the substituent is on an aryl or non-aryl carbon. The second column of the respective subfield identifies the substituent. All of the important substituents such as alcohols, carbonyls, carboxyls, halides, amines, nitrates, and sulfates are differentiated. All types of non-aromatic unsaturation are also differentiated. The third column of the subfield shows the position of the substituent relative to the nearest Si atom and indicates, by an overpunch, whether or not it is in or on the chain (or ring). In the coding of S groups, arbitrary rules of precedence are followed. Fig. 1 illustrates

the coding in the S field for α-chlorovinylethyldichlorosilane.

Other Parameters Coded.—Provision is also made on the card for coding the R′ to Si ratio (columns 68 and 69, with a decimal point assumed after column 68; Fig. 1). In this case R′ is defined as the total Si-C plus Si-H bonds (the Si linkages that are most stable hydrolytically), and the ratio is the Si-C plus Si-H bonds over the total Si atoms. This will usually differentiate between cases that might be coded the same in the R field.

The chemical number assigned to each compound is shown in columns 73 through 77 (Fig. 1). It is an accession number having no significance with respect to structure or chemical function and serves only as a unique means of identifying the compound. This number is identical with the term number used in the report index to represent this chemical. The structural formula index serves as a satellite file to the report index, and the chemical number is the connecting link between the two indexes.

The Work Sheet.—Each chemical is analyzed for its structural features, which are coded on a work sheet, which in this case is an IBM card. The structural formula is written on the card, a name is given when feasible, and the chemical number assigned. The holes to be punched in each column are indicated until all the necessary fields have been coded. Figure 3 shows the Formula Work Card and the coding for α-chlorovinylethyldichlorosilane.

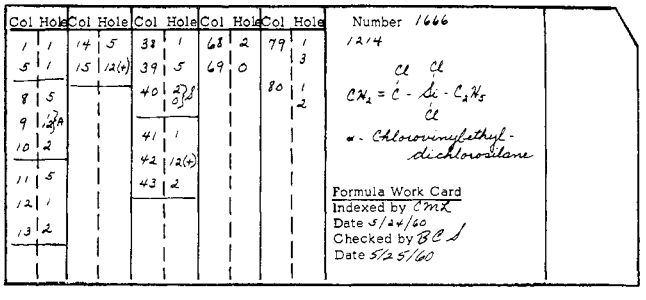


Fig. 3.—Formula Work Card.

Building the Index Files.—From these work cards are punched master cards like the one shown in Fig. 1. The new cards make up the master deck (Deck 1, column 72). Every chemical has its place in the ordered code. From Deck 1, two other decks are punched, using a system of rotation which was adapted from the Chemical-Biological Coordination Center system. The cards from Deck 1 are reproduced exactly and become the R masters which serve as the starting point for rotating R subfields. By suitable manipulation of the reproducing punch, a new card is made for each R group beyond the first one so that each R appears once in the first R subfield (columns 8, 9, 10). The cards, together with the R masters, constitute Deck 2. Figure 4 shows the cards necessary to represent the compound α-chlorovinylethyldichlorosilane in Deck 2.

All cards from Deck 1 which describe substituents are also reproduced and become the S masters. They are rotated so that each substituent required in coding appears once in the first S subfield (columns 38, 39, 40). These cards, together with the S masters, constitute Deck 3. The cards necessary to represent α-chlorovinylethyldichlorosilane in Deck 3 are shown in Fig. 5.

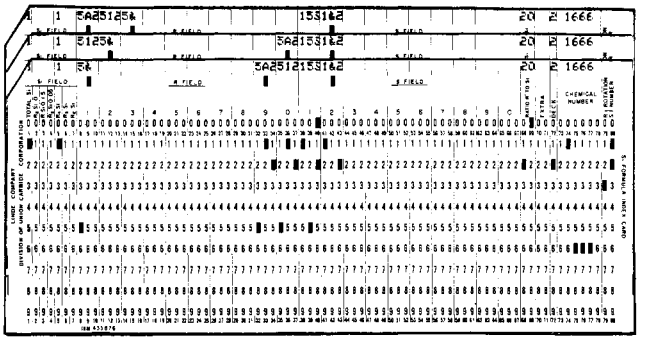


Fig. 4.—Rotation of Si-R linkages in R field.

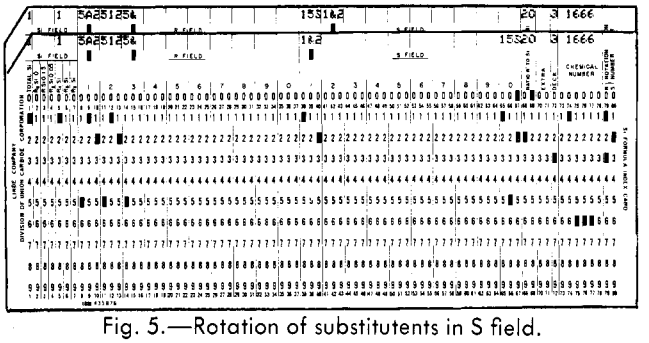


Fig. 5.—Rotation of substitutents in S field.

The deck number for each of the three decks is shown by a punch in column 72. All Deck 1 cards have punched in column 79 the number of R rotations needed to have all R's appear once in the first R subfield. If substituents are present, the number of rotations needed to show all S's once in the first S subfield is punched in column 80. The same information is punched into the cards of Decks 2 and 3. In columns 79 and 80 are punched (with printing suppressed) the rotation number and total number of rotations for the R and S fields, respectively. Instructions for rotations are given on the Formula Work Card (Fig. 3).

Use of the Formula Index.—Each of the three decks serves a different purpose. Deck 1 is maintained in order by the complete code as read across the card. Every chemical has a place in this linear array. Reference to Deck 1 avoids scattering or assignment of several chemical numbers to the same compound because of nomenclature difficulties, and also serves to locate the chemical number of the compound. These points may be illustrated by considering the following compound, which may be named several different ways:

$$(CH_3OCH_2CH_2O)_3SiC_3H_6OCH_2CH\underset{\diagdown\ \diagup}{CH_2}\ O$$

By coding this chemical and checking with the master file (Deck 1), one may determine whether or not the chemical has already been coded. If the chemical has already been coded, its chemical number is found, and duplication is avoided.

Deck 2 is maintained in order by all skeletal R group codes (columns 9 and 10) which have been rotated into the first R subfield. In retrieval it is only necessary to use that part of Deck 2 which meets the structural requirements of any given retrieval question. For example, esters of the type $R_nSi[O\text{-}t\text{-}C(CH_3)_3]_{4-n}$ can be found by using the sub-deck in Deck 2 dealing with the code in columns 9 and 10 for a *t*-butoxy group. If in $R_nSi[O\text{-}t\text{-}C(CH_3)_3]_{4-n}$, R is specified as either phenyl, ethyl, or propyl, the sub-

decks dealing with these codes are sorted together and matched on the collator by chemical number with the *t*-butoxy deck. In this way all *t*-butoxy silicon esters also containing phenyl, ethyl, or propyl groups are identified. This operation will be recognized as an example of combining logical sum with logical product in search strategy.[5]

Deck 3 is maintained in order by all substituent group codes (columns 39 and 40), and is used to search for substituents on R groups. When searching for silicone surfactants containing an $-SO_3H$ or $NH_4^+X^-$, only the sub-decks within Deck 3 for the codes for such groups need be used to obtain the chemical numbers. By use of these numbers in conjunction with the report index, information on chemicals used as surfactants is obtained.

Decks 2 and 3 allow the grouping together of silicon chemicals having in common any specified set of structural elements, and allow almost limitless possibilities in grouping these features. Since Deck 1 groups compounds together based on a single linear relationship and permits selection of individual compounds, the combination of all three decks gives the necessary flexibility from general to specific selectivity required of a formula index.

Physical Properties Index.—To supplement the structural formula index, a physical properties index for silicon chemicals was established. The properties coded in this index are melting point, boiling point, refractive index, and density. Each chemical included in this index is classified by a 4-digit number (class number) which can be determined by examining the formula. The first digit shows the type of linkage to silicon as shown:

0 Si–C
1 Si–Cl
2 Si–F, Si–Br, Si–I
3 Si–O–C (0, 1 or 2 non-aryl C's)
4 Si–O–C (more than 2 non-aryl C's)
5 Si–H
6 Si–OH
7 Si–O–Si
8 Si–N
9 Si–Z (linkage not listed above)

Where there is more than one of the above specified linkages, the highest ranking linkage is coded. The last three digits of the class number are obtained by counting certain structural features in the molecule. The features counted in each case were chosen so as to give a reasonably good spread from 0 to 9 and to be as simple to learn as possible. Balancing off these requirements resulted in the three counts discussed below. The second digit shows the greatest number of non-aryl carbon atoms on any single group, in which one of these carbon atoms is attached to the silicon atom. For example, in methylethylpropylsilane, $CH_3(C_2H_5)(C_3H_7)SiH$, the group with the greatest number of non-aryl carbon atoms is the propyl group, with *three* carbon atoms, one of which is attached to a silicon atom. In triethylsilane, $(C_2H_5)_3SiH$ (Fig. 6), there are no more than *two* non-aryl carbon atoms in any group attached to Si. In dimethylphenylsilane $(CH_3)_2C_6H_5SiH$, there is only *one* carbon atom in any non-aryl group attached to the Si. The third digit shows the total number of aryl groups plus non-aromatic carbon-to-carbon unsaturation points in the molecule. The fourth digit shows the total number of atoms of silicon, oxygen

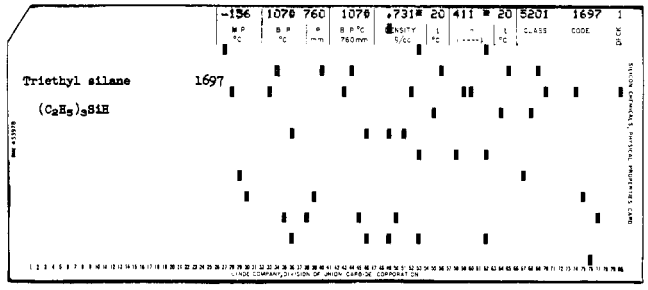


Fig. 6.—Physical properties card.

not attached to silicon, and all other elements except carbon and hydrogen. Since several chemicals may share the same class number, the latter does not identify the chemical uniquely. To accomplish unique identification, its chemical number is given. The chemical number is called the code number in this index. The class number is also indicated on the Formula Work Card (Fig. 3) below and to the left of the assigned chemical number. This index keeps an up-to-date record of all properties reported at our own laboratory and in certain external literature. At intervals the index cards are used to print out a book index with entries in order by class number and within class by chemical number. A typical card used in this print-out is shown in Fig. 6. Standard symbols, following the numerical figures, are sometimes used to code the estimated degree of precision for the properties. By combining this index with the structural formula index, compounds with certain structural features, which also have properties between certain limits, can be identified. It is also possible to process generic questions which delve into structural features and properties by combining the formula index, properties index, and report index. The use of the three indexes is illustrated by a question on the preparation of monosilane derivatives with boiling points between 75° and 125° C. Deck 1 of the structural formula index is searched in column 1 to find all chemicals containing one silicon atom. By collation, all chemicals with the specified boiling point range are found in the properties index. By matching the chemical numbers found from these two indexes, the numbers of the chemicals that meet requirements are obtained. Use of these numbers with the report index yields all references to the preparation of the monosilane derivatives.

REFERENCES

(1) F. R. Whaley, "A Deep Index for Internal Technical Reports," in "Information Systems in Documentation," J. H. Shera, *et al.*, Ed., Interscience Publishers, Inc., New York, N. Y., 1957).
(2) F. R. Whaley, "Operational Experience with Linde's Indexing and Retrieval System" (presented at the Conference on Information Retrieval at Poughkeepsie, N. Y., September, 1959). Published by International Business Machines Corporation (E208040).
(3) G. M. Dyson, "A New Notation and Enumeration System for Organic Compounds," 2nd Edition, Longmans, Green and Co., New York, N. Y., 1949.
(4) W. J. Wiswesser, "A Line-Formula Chemical Notation," T. Y. Crowell Co., New York, N. Y., 1954.
(5) F. R. Whaley, "The Use of a Collator in an Inverted File Index," *Special Libraries*, **53**, No. 2, 65–73, February, 1962.