Retrieval of Organic Structures from Small-to-Medium Sized Collections*

A. J. BARNARD, JR., C. T. KLEPPINGER
J. T. Baker Chemical Company, Phillipsburg, New Jersey

and W. J. WISWESSER
U. S. Army Biological Laboratories, Fort Detrick, Frederick, Maryland

Received October 29, 1965

An IBM punched-card system is described for the storage, retrieval, and printout of chemical structural information. A simple numeric structural-atomic code, the BATCH Number, is introduced. The system is employed in the development of a classified structural directory of organic chemicals.

In the storage and retrieval of structural information from small-to-medium sized collections (that is, with up to possibly 20,000 punched cards), several capabilities exist that are not always realizable with large collections for operational or economic reasons. One capability is the *listing* of some or all of the information stored on punched cards in various classified arrangements and subsequently of resolving many retrieval requests by direct examination of such listings. Indeed, by offset printing, such listings can be made available at moderate cost to a large number of potential users as a classified directory. A second capability is the possibility of saving invested punched-card capacity by providing space in relevant fields for only the shorter entries that account for a large proportion of the total entries. Long, and thereby rare, entries can then be accommodated by the use of secondary cards or even by resort to hand operations.

This paper describes an IBM-card layout that provides effective storage of structural data for compounds of carbon and also facilitates retrieval and listing operations. This punched-card layout should be of special value where small-to-medium sized collections are involved. One salient application is also described, namely, the listing of a portion of the stored information according to an index number having structural and atomic elements, namely, the BATCH Number, and thereby the compilation of an easy-to-use, classified, structural directory. Certain aspects of this approach were suggested by Wiswesser (*1*). The systematic schemes developed for the statement of chemical names using only capital letters, arabic numerals, and four punctuation marks, and for the contraction of the names to fit within a card field of 32 columns, are described in the accompanying paper (*2*).

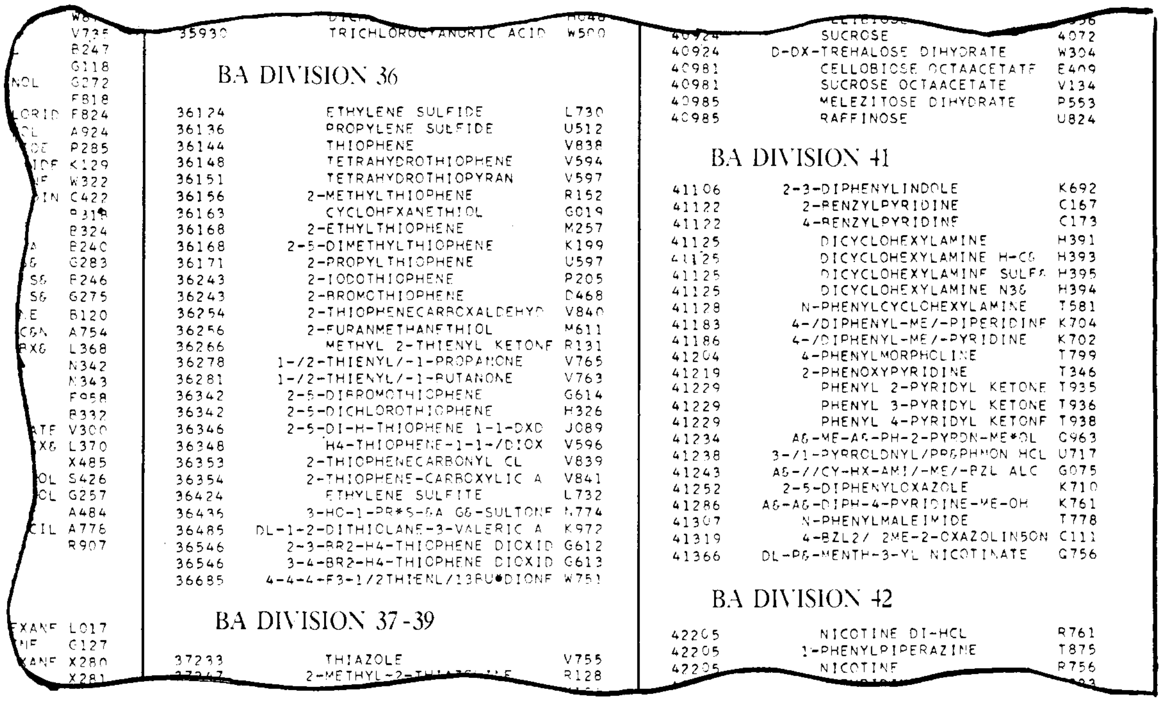
It is believed that the approaches and practices introduced in these papers have widespread application to the effective storage and retrieval of structural, physicochemical, biochemical, analytical, or commercial information.

As of September 1965, the J. T. Baker Chemical Company was offering to the laboratory market over 5000 compounds of carbon intended for research, analysis, and diverse use. The compounds in this line have served as the "commercial deck" employed in the development of the approaches delineated in this paper. Additional experiments were undertaken with files of the U. S. Army Biological Laboratories. Early experience in the growth of the J. T. Baker line revealed the merit of introducing machine storage and retrieval of diverse information. Comments of both technologists and purchasing agents in informal surveys revealed that they experienced difficulty in locating compounds of interest in existing catalogs of suppliers of laboratory chemicals, partly because of limited familiarity with organic nomenclature and partly because of insufficient cross-references. In J. T. Baker laboratory chemicals catalogs, these difficulties are being alleviated by extensive cross-referencing and the inclusion, for carbon compounds, of an empirical formula index. Additional comments, especially from synthetic organic chemists, revealed interest in indexes that grouped similar compounds together. One approach to meeting this need would be the tedious preparation of a conventional functional group index. However, *via* machine listing and offset printing it has been possible to produce a classified structural index, the J. T. Baker BATCH Directory, which can be prepared at moderate cost. This directory serves many of the needs of a conventional functional group index. A portion of a page of this directory is shown in Figure 1.

THE BATCH NUMBER

Collections of knowledge, for many decades, have been classified according to numerical, nonunique codes. The Dewey classification for library collections represents a

* 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.



```
        V73F
        B247
        G118
NOL     G272
        F818
LORID   F824
OL      A924
OE      P285
IDF     K129
NF      W322
IN      C422
        P31B
        B324
A       B240
&       G283
S&      F246
S&      G275
E       B120
C&N     A754
X&      L368
        N342
        N343
        F958
        B332
TE      V300
X&      L370
        X485
OL      S426
OL      G257
        A484
IL      A776
        R907

XANF    L017
NF      G127
ANF     X280
        X281
```

```
35930           TRICHLOROCYANURIC ACID   W500

BA DIVISION 36

36124           ETHYLENE SULFIDE              L730
36136           PROPYLENE SULFIDE             U512
36144           THIOPHENE                     V838
36148           TETRAHYDROTHIOPHENE           V594
36151           TETRAHYDROTHIOPYRAN           V597
36156         2-METHYLTHIOPHENE               R152
36163           CYCLOHEXANETHIOL              G019
36168         2-ETHYLTHIOPHENE                M257
36168       2-5-DIMETHYLTHIOPHENE             K199
36171         2-PROPYLTHIOPHENE               U597
36243         2-IODOTHIOPHENE                 P205
36243         2-BROMOTHIOPHENE                D468
36254         2-THIOPHENECARBOXALDEHYD        V840
36256         2-FURANMETHANETHIOL             M611
36266           METHYL 2-THIENYL KETONF       R131
36278     1-/2-THIENYL/-1-PROPANONE           V765
36281     1-/2-THIENYL/-1-BUTANONE            V763
36342       2-5-DIBROMOTHIOPHENE              G614
36342       2-5-DICHLOROTHIOPHENE             H326
36346       2-5-DI-H-THIOPHENE 1-1-DXD        J089
36348           H4-THIOPHENE-1-1-/DIOX        V596
36353         2-THIOPHENECARBONYL CL          V839
36354         2-THIOPHENE-CARBOXYLIC A        V841
36424           ETHYLENE SULFITE              L732
36435         3-HO-1-PR*S-&A G&-SULTONF       N774
36485   DL-1-2-DITHIOLANE-3-VALERIC A         K972
36546       2-3-BR2-H4-THIOPHENE DIOXID       G612
36546       3-4-BR2-H4-THIOPHENE DIOXID       G613
36685     4-4-4-F3-1/2THIENL/13BU•DIONF       W751

BA DIVISION 37-39

37233           THIAZOLE                      V755
              2-METHYL-2-                     R128
```

```
                SUCROSE                       4072
40924     D-DX-TREHALOSE DIHYDRATE            W304
40981           CELLOBIOSE OCTAACETATF        E409
40981           SUCROSE OCTAACETATE           V134
40985           MELEZITOSE DIHYDRATE          P553
40985           RAFFINOSE                     U824

BA DIVISION 41

41106       2-3-DIPHENYLINDOLE                K692
41122         2-BENZYLPYRIDINE                C167
41122         4-BENZYLPYRIDINF                C173
41125           DICYCLOHEXYLAMINE             H391
41125           DICYCLOHEXYLAMINE H-C&        H393
41125           DICYCLOHEXYLAMINF SULF&       H395
41125           DICYCLOHEXYLAMINE N3&         H394
41128         N-PHENYLCYCLOHEXYLAMINE         T581
41183       4-/DIPHENYL-ME/-PIPERIDINF        K704
41186       4-/DIPHENYL-ME/-PYRIDINE          K702
41204         4-PHENYLMORPHOLINE              T799
41219         2-PHENOXYPYRIDINE               T346
41229           PHENYL 2-PYRIDYL KETONE       T935
41229           PHENYL 3-PYRIDYL KETONE       T936
41229           PHENYL 4-PYRIDYL KETONF       T938
41234       A&-ME-A&-PH-2-PYRDN-ME*OL         C963
41238     3-/1-PYRROLDNYL/PR&PHNON HCL        U717
41243     A&-//CY-HX-AMI/-ME/-BZL ALC         G075
41252       2-5-DIPHENYLOXAZOLE               K710
41286   A&-A&-DIPH-4-PYRIDINE-ME-OH           K761
41307         N-PHENYLMALEIMIDE               T778
41319         4-BZL2/ 2ME-2-OXAZOLIN5ON       C111
41366   DL-P&-MENTH-3-YL NICOTINATE           G756

BA DIVISION 42

42205           NICOTINE DI-HCL               R761
42205         1-PHENYLPIPERAZINE              T875
42205           NICOTINF                      R756
```

Figure 1. Portion of one page of the J. T. Baker BATCH Directory (September 1965) showing BATCH Numbers, commodity names, and commodity numbers.

familiar example. A Formula Index number consisting of five simple numeric measures, each giving possible values ranging from zero to nine, has been proposed (*1*). This structural "classification" code was given the mnemonic designation BATCH Number. An attempt was made to secure a reasonably balanced population distribution among the ten-digit values of each measure for a large collection of compounds.

Statistical confirmation of the original assignments of values is given in the original publication (*1*). In 1964, the assignment of the values of two of the measures, namely, the *B* and *A* digits (see below), was revised in order to reflect better the increasing prominence of monocyclic structures other than benzene and of halogen-containing compounds. This latter change was influenced to some extent by two considerations: (1) the BATCH Number would receive prominent use with small-to-medium size collections of structures, and (2) the BATCH Number in conjunction with a suitable organization of an empirical formula field would provide easy retrieval of the individual halogens. Support for both of these changes is also to be found in the statistical survey of Heumann and Dale (*3*) of more than 59,000 chemicals in the files of the Chemical-Biological Coordination Center of the National Academy of Sciences–National Research Council.

The five measures of the BATCH Number have been selected to relate (1) to the nature (and number) of rings present, if any (*B* digit); (2) to the nature (and number) of atoms present other than carbon, hydrogen, and oxygen (*A* digit); (3) to the number of atoms present other than carbon and hydrogen (*T* digit); (4) to the number of carbon atoms in the empirical formula (*C* digit); and (5) to the number of hydrogen atoms in the empirical formula (*H* digit).

ASSIGNMENT OF A BATCH NUMBER

The scheme for the assignment of a BATCH Number is described below. The "Notes" that follow the description should be examined for an appreciation of some special practices and problems.

The *B* Digit. The *B* implies *B*asis = nucleus or *B*enzene-ring number and the value of the *B* digit is assigned as follows (Notes 1-5):

0 = no rings in structure; acyclic

1 = one benzene ring only in structure

2 = a plurality of (nonfused) benzene rings in structure (Note 3)

3 = one ring *other* than benzene only in structure (benzoquinoidal compounds included)

4 = a plurality of monocyclic rings in structure, at least one of which is *not* benzene

5 = one bicyclic ring system in structure (including a spiran or other mononuclear bicyclic system)

6 = a plurality of bicyclic ring systems or a bicyclic ring system and one or more monocyclic rings

7 = tricyclic ring system(s) in structure

8 = tetracyclic ring system(s) in structure

9 = pentacyclic or higher ring system(s) in structure

The *A* Digit. The *A* implies the *A*tomic class of the empirical formula, and the *A* digit is assigned, from the following scheme, the largest value that fits the empirical formula (Notes 4-8):

0 = hydrocarbons and oxyhydrocarbons (*i.e.*, only carbon and hydrogen present with or without oxygen)

1 = one nitrogen atom

2 = two nitrogen atoms

3 = three or more nitrogen atoms

4 = halogens (any number) present without nitrogen

5 = halogen present with nitrogen (any number of each)

6 = sulfur (any number) present without nitrogen (and with or without halogen)

7 = sulfur present with nitrogen (and with or without halogen)

8 = phosphorus (any number) present (and with or without nitrogen, halogen, or sulfur)

9 = elements other than carbon, hydrogen, oxygen, nitrogen, halogens, sulfur, or phosphorus present (including metal-containing compounds) (Notes 7, 8)

The *T* Digit. The *T* implies *T*otal heteroatomic count. The value of the *T* digit is assigned as follows (Notes 4, 5, 7-9):

Add all atoms in the empirical formula except carbon and hydrogen, and record the total if less than ten. If the total is ten or more, record *zero* for any compound having an *A* digit other than zero (since such a compound must have more than zero heteroatoms) and record *nine* for oxyhydrocarbons ($A = 0$) having more than nine oxygen atoms.

The *C* Digit. The *C* implies *C*arbon-atom count. The value of the *C* digit is assigned as follows (Notes 4, 5, 8):

Record only the *units* count of carbon atoms in the empirical formula. For example, C_2, C_{12}, C_{22}, C_{32}, etc., all have a *C* digit of 2.

The *H* Digit. The *H* implies *H*ydrogen atom count. The value of the *H* digit is assigned as follows (Notes 4, 5, 8, 9):

Record the sum of the units and tens count of hydrogen atoms in the empirical formula or the units part of this sum when it is ten or more. (The *sum* is chosen to assure full use of all digits, regardless of odd/even valence totals.) For example, H_2, H_{20}, H_{39}, H_{48}, H_{57}, H_{66}, H_{75}, H_{84}, H_{93}, H_{102}, etc., all have an *H* digit of 2.

Notes on the Assignment of a BATCH Number

Note 1. Note the following mnemonic devices in the above scheme for the *B* digit: 0 = none; 1, 3, and 5 (all prime numbers) = one of something; 2, 4, and 6 (all divisible by two) = plurality of something; 8 (*i.e.*, two times *four*) = *tetra*cyclic; 0 through 4, no fused rings present; 5 through 9, fused ring present.

Note 2. Chelates present a special problem where uncertainty exists as to the number of coordination bonds actually formed since the number of chelate rings assumed to be present may influence the *B* digit assigned. A least-effort principle is therefore adopted by neglecting *all chelate* rings in establishing the *B* digit. For example, bis(ethylenediamine)copper(II) chloride is assigned a *B* digit of 0 (*i.e.*, no ring other than a chelate ring present) and tris(8-quinolinato)aluminum a *B* digit of 6 (*i.e.*, a plurality of bicyclic (quinoline) rings present).

Note 3. Triphenylmethane dyes are assigned a *B* digit of 2 (*i.e.*, a plurality of benzene rings); in other words, *all three rings are treated equally* as phenyl groups instead of one arbitrarily being considered as quinoidal. A phthalein or sulfonphthalein is treated as having the -COOH or -SO_3H group present as such, that is, as existing in the open form and not in the lactone or sultone form, respectively. In other words, phthaleins and sulfonphthaleins as other triphenylmethane dyes are assigned a *B* digit of 2, unless, of course, a bridge exists between two of the rings to form a tricyclic system as in pyrocatecholsulfonphthalein.

Note 4. In the J. T. Baker practice, polymeric substances are arbitrarily assigned a BATCH Number of 00000. (The absence of anything suggests the presence of "much!")

Note 5. For a monomeric compound where assignment of a value to a digit is impossible because of an incomplete formula, a hyphen is inserted; in the arrangement of a file according to BATCH Number, a hyphen is allowed to precede a zero.

Note 6. The principle of later class applies, that is, assign the largest *A* digit that fits the empirical formula (subject to the restrictions of Notes 7 and 8).

Note 7. In the J. T. Baker use of the Batch Number, salts of organic bases with inorganic acids are treated as the base itself (*e.g.*, aniline hydrochloride, sulfate, and phosphate are all assigned the BATCH Number for aniline). This practice avoids the "scattering" of an organic base and its salts. (Fully quaternized ammonium salts, however, are considered as such.) For a similar reason, metal and unsubstituted ammonium salts of organic acids are treated as the acid itself (*e.g.*, propionic acid, sodium salt, is assigned the BATCH Number of propionic acid).

Note 8. Derivatives of alcohols, phenols, and amines formed with *unipositive* metal ions are listed under the BATCH Number of the parent organic compound (*e.g.*, methanol, sodium derivative has the BATCH Number of methanol and therefore the *A* digit = 0). Derivatives with other metal ions are listed under the BATCH Number of the derivative itself (*e.g.*, methanol, magnesium derivative has the BATCH Number for $(CH_3O)_2Mg$, namely with an *A* digit = 9).

Note 9. In the J. T. Baker practice, compounds are assigned BATCH Numbers without consideration of hydration; in this way, the "scattering" of an anhydrous compound and its hydrate is avoided.

Examples of BATCH Numbers. An understanding of the assignment of a BATCH Number can be facilitated by consideration of a few examples.

Example 1. Acetic acid, CH_3COOH, $C_2H_4O_2$, has a BATCH Number of 00224. $B = 0$ as no ring is present. $A = 0$ as only carbon, hydrogen, and oxygen are present. $T = 2$ as the total of atoms present in the formula other than carbon and hydrogen is two. $C = 2$ and $H = 4$ from the number of carbon and hydrogen atoms in the empirical formula, respectively. (All unsubstituted aliphatic monocarboxylic acids have $BAT = 002$.) (In the J. T. Baker practice, acetic acid, sodium salt would also be assigned the BATCH Number of acetic acid; see Note 7.)

Example 2. Propylamine, $CH_3CH_2CH_2NH_2$, C_3H_9N, has a BATCH Number of 01139. $B = 0$ as no ring is present. $A = 1$ as one nitrogen is present. $T = 1$ as the total of atoms present in the formula other than carbon and hydrogen is one. $C = 3$ and $H = 9$ from the number of carbon and hydrogen atoms in the empirical formula, respectively. (All unsubstituted aliphatic monoamines have $BAT = 011$.)

Example 3. *o*-Aminophenol, $2\text{-}NH_2C_6H_4OH$, C_6H_7NO, has a BATCH Number of 11267. $B = 1$ as one benzene ring is present. $A = 1$ as one nitrogen is present. $T = 2$ as the total of the heteroatoms in the formula is two; that is, one nitrogen and one oxygen are present. $C = 6$ and $H = 7$ from the number of carbon and hydrogen atoms in the empirical formula, respectively.

Example 4. *p*-Aminobenzenethiol, $4\text{-}NH_2C_6H_4SH$, C_6H_7NS, has a BATCH Number of 17267. $B = 1$ as one benzene ring is present. $A = 7$ as both sulfur and nitrogen are present. $T = 2$ as the total of the heteroatoms in the formula is two. $C = 6$ and $H = 7$ from the number of carbon and hydrogen atoms in the empirical formula, respectively.

Example 5. (Thiophene, SCH:CHCH:CH, C_4H_4S, has a BATCH Number of 36144. $B = 3$ since one ring other than benzene is present. $A = 6$ since sulfur is present without nitrogen. $T = 1$ as a single heteroatom is present in the formula. $C = 4$ and $H = 4$ from the number of carbon and hydrogen atoms in the empirical formula, respectively.

Example 6. Triphenyl phosphate, $(C_6H_5O)_3PO$, $C_{18}H_{15}O_4P$, has a BATCH Number of 28596. $B = 2$ as a plurality of benzene rings is present. $A = 8$ as phosphorus is present. $T = 5$ as the total of heteroatoms in the formula is five. $C = 8$ as the digit value of the total of carbon atoms, 18, is eight. $H = 6$ since 15 hydrogen atoms appear in the formula and $1 + 5 = 6$.

Example 7. Cholesterol, $C_{27}H_{46}O$ (tetracyclic), has a BATCH Number of 80170. $B = 8$ as the compound is tetracyclic. $A = 0$ as only carbon, hydrogen, and oxygen are present. $T = 1$ as 1 oxygen atom appears in the formula. $C = 7$ as 27 carbon atoms appear in the formula and 7 is the units value. $H = 0$ since 46 hydrogen atoms appear in the formula, $4 + 6 = 10$, and the units value in this sum is zero.

DISTRIBUTION OF DIGIT VALUES FOR BATCH NUMBERS

It is of interest to record the percentage distribution of digit values for the BATCH Numbers of the commercially more accessible compounds of carbon, as reflected by the J. T. Baker laboratory chemical offerings of September 1965. The relevant percentages are reported in Table I. This "commercial deck" contains 5019 compounds of carbon of which 103 are polymeric and therefore are arbitrarily assigned a BATCH Number of 00000 (see Note 4 on Assignment of a BATCH Number). It will be seen that favorable levelling of the *C*-digit and *H*-digit values is secured. Analysis of the distribution of the *B*-digit and *A*-digit values reveals, as might be expected, that compounds with no rings (39.1%) and with only one benzene ring (24.1%) predominate. Additionally, hydrocarbons and oxyhydrocarbons (37.7%) are prominent.

Table I. Percentage Distribution of Digit Values in BATCH Number

(5019 Commercial Compounds of Carbon, Including 103 Polymeric Compounds Arbitrarily Assigned BATCH No. 00000)

Digit	Percentage distribution of digit values				
value	*B* digit	*A* digit	*T* digit	*C* digit	*H* digit
0	39.1	37.7	10.9	11.2	4.7
1	24.1	16.1	16.6	6.1	9.8
2	7.1	9.2	24.4	9.6	8.7
3	12.0	4.8	17.2	7.2	11.2
4	2.8	12.0	12.7	10.7	10.8
5	6.9	6.0	7.4	8.1	11.9
6	2.7	4.3	5.0	15.5	10.9
7	3.3	6.1	2.9	11.6	11.4
8	1.5	0.7	1.8	13.1	10.7
9	0.5	3.1	1.1	7.0	9.9

On one hand, for this deck, the relative distribution of the *B* digit and *A* digit may be viewed as somewhat unfavorable. However, in any printout associating BATCH Numbers with the corresponding chemical names, most noncyclic and monobenzenoid compounds are readily recognized. On the other hand, the recognition of complex cyclic compounds is facilitated by their less frequent occurence.

IBM-CARD LAYOUT

The punched-card layout summarized in Table II was adopted for structure cards. The interpreted cards shown in Figure 2 delineate the assignment of the fields. This layout was standardized only after some experimentation directed to the optimum size of the fields for the notation and chemical name, and certain aspects of that study have been reported (*2*). The treatment of each field is considered in the following paragraphs.

THE NOTATION FIELD

As can be seen from the summary of the IBM-card layout for a structure card shown in Table II, the Wiswesser Line Notation (*4*), often given the acronym WLN, can be recorded unambiguously in columns 1 through 13. The notation can be extended into the prefix of chemical name field, that is, into columns 14 through 22 to the extent that these are not filled by such a prefix. Where the notation is too long to be accommodated fully in the available prefix area or where "clean" printing of the chemical name is to be expedited, the notation is interrupted at column 12 and an asterisk is placed in column 13 as well as a *zero* in column 80 to serve as control punches. In such a case, the full notation is carried on a second card utilizing the full notation and name fields to the extent necessary. Such a second card carries a control punch of *one* in column 80.

The Wiswesser Line Notation offers the advantage of a unique structure notation employing as characters only capital letters, arabic numerals, and three punctuation marks (with an asterisk as a special signal, notably at the end of the field to indicate a "chopped" notation

Table II. IBM-Card Layout

Columns			
1-9	Wiswesser notation	1 to 9	Notation size (*S*)
10		0	
11-13		A to C	
14-22	Prefix, chemical name	D to L	
23-45	Chemical name		
46-50	BATCH Number		
51	Size, Wiswesser notation (*S*)		
52	Space		
53	C punched, carbon compounds	53-73 Empirical formula (or other data)	
54-56	Number of carbons		
57	H punched, hydrogen compounds		
58-60	Number of hydrogens		
61-73	Remainder, empirical formula		
74-77	Registry and control fields	←J. T. Baker commodity no.	
78-80			

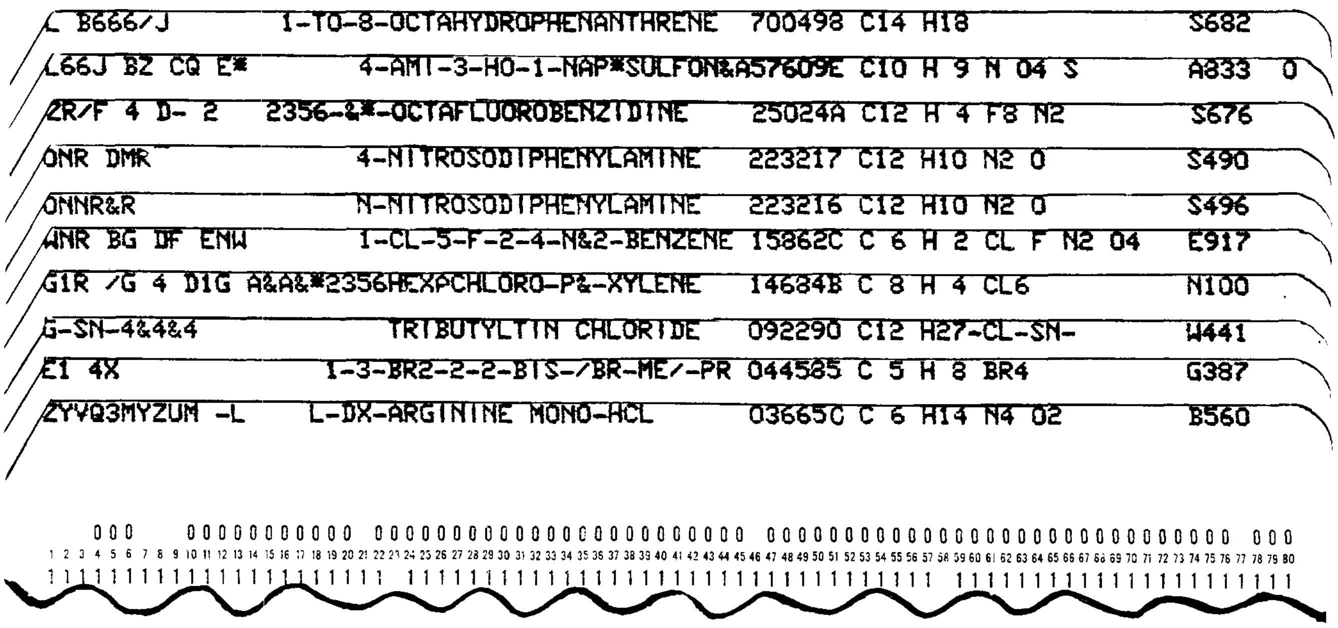
and at the start of the field a cyclic contraction). Consequently, this notation is capable of use with standard, unmodified sorting and listing equipment. Its application to the printout of permuted structural indexes has been described recently by Sorter and co-workers (*11*), and to the computer generation of physical property values from atomic, bond, and group contributions by Brasie and Liou (*5*, *cf*. *6-8*).

The length of the notation is conveniently expressed as 1, 2, . . . , 9, 0, A, B, . . . , Z and is punched in column 51 (see below). The notation size of 4383 compounds of the present commercial deck containing 4920 nonpolymeric compounds is analyzed statistically in Table III. It will be seen that 87.5% of the notations are of size C (=13) or less and thus can be accommodated in the notation field provided in the standard IBM-card layout. Where the notation is allowed to extend three spaces into the nine-space prefix area of the chemical name (size F), 92.8% of the notations can be accommodated.

It is interesting to compare this notation-size distribution with that obtained by Wiswesser (*9*) in the analysis of about 9000 relatively complex compounds in the files of the Chemical-Biological Coordination Center of the National Academy of Sciences–National Research Council: 63.0% had a notation size of C (=13) or less, 73.8% a notation size of F (=16) or less, and 83.9% a size of L (=22). Some years ago, Benson (*10*) assigned line notations to 3360 compounds, then offered by Distillation Products, but did not employ the methyl contraction. Analysis of his deck reveals that 83.4% have a notation size of C or less and 89.5% a notation size of F or less (*9*).

CHEMICAL NAME FIELD

Columns 23 through 45 in the IBM-card layout are reserved for the chemical name and its prefix is allowed to extend back, as required, from column 22 to column 14. To permit the recording of a chemical name, special practices must be adopted since only capital letters, arabic



```
L B666/J        1-TO-8-OCTAHYDROPHENANTHRENE  700498 C14 H18              S682
L66J BZ CQ E*          4-AMI-3-HO-1-NAP*SULFON&A57609E C10 H 9 N 04 S     A833  0
ZR/F 4 D- 2   2356-&*-OCTAFLUOROBENZIDINE     25024A C12 H 4 F8 N2        S676
ONR DMR                4-NITROSODIPHENYLAMINE  223217 C12 H10 N2 0         S490
ONNR&R                 N-NITROSODIPHENYLAMINE  223216 C12 H10 N2 0         S496
WNR BG DF ENW          1-CL-5-F-2-4-N&2-BENZENE 15862C C 6 H 2 CL F N2 04  E917
G1R /G 4 D1G A&A&*2356HEXPCHLORO-P&-XYLENE    14684B C 8 H 4 CL6          N100
G-SN-4&4&4               TRIBUTYLTIN CHLORIDE  092290 C12 H27-CL-SN-       W441
E1 4X                1-3-BR2-2-2-BIS-/BR-ME/-PR 044585 C 5 H 8 BR4         G387
ZYVQ3MYZUM -L     L-DX-ARGININE MONO-HCL      03665C C 6 H14 N4 02        B560
```

Figure 2. Some typical interpreted structure cards.

Table III. Distribution of Notation Size (4383 Compounds of Commercial Deck Assigned Wiswesser Line Notations)

Size			Size		
Designation (*S*)	No.	Cumulative %	Designation (*S*)	No.	Cumulative %
1	1	0.1	D	14	89.6
2	2	1.8	E	15	91.3
3	3	10.4	F	16	92.9
4	4	19.2	G	17	94.1
5	5	29.5	H	18	94.7
6	6	39.6	I	19	95.4
7	7	56.8	J	20	95.9
8	8	60.3	K	21	96.4
9	9	67.8	L	22	96.6
0	10	74.5	M	23	97.0
A	11	80.0	N-Z	24-36	100.0
B	12	84.3			
C	13	87.5			

numerals, and a limited number of punctuation marks are available. In addition, longer names not accommodated by this 32-column field (9 for prefix and 23 for name proper) must be contracted. The systematic approaches adopted in the resolution of these two problems are described in the accompanying paper (*2*).

The use of an offset, variable-length prefix to chemical names has the advantage of facilitating the visual inspection of any printout for a chemical name of interest. For example, see the portion of a page of the J. T. Baker BATCH Directory shown in Figure 1.

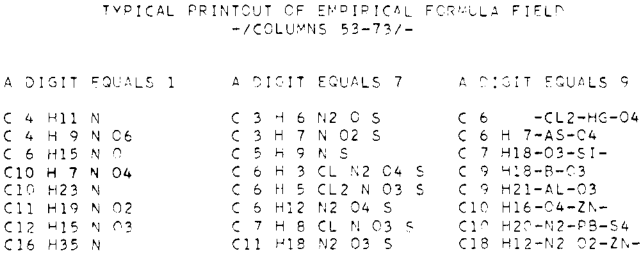
BATCH NUMBER AND NOTATION SIZE FIELD

In the IBM-card layout, the numerical value of the *B*, *A*, *T*, *C*, and *H* digits of the BATCH Numbers are punched in columns 46 through 50. In column 51, the size *S* of the Wiswesser notation (*S* = 1, . . . , 9, 0, A, B, C, . . .) is recorded up to Z for size 36 and larger. This notation size has a number of valuable functions. For example, where the collection is to be aligned according to the BATCH Number, a preliminary sort on column 51 assures that if two compounds, other than simple position isomers, have the same BATCH Number, the more complex one will follow the simpler one. This notation size also can serve as an effective computer control where permutation of the notation and listing would be undertaken following the proposals of Sorter and co-workers (*11*). In some applications, it may be appropriate to allow the notation size to print. In this event, the BATCH Number is followed by a size mark that may be a numeral or a letter, and the mnemonic term "BATCHS Number" is then appropriate.

EMPIRICAL FORMULA FIELD

In the IBM-card layout, the empirical formula field is restricted to columns 53 through 73, that is, to 21 columns. Experience at the U. S. Army Biological Laboratories with an extensive collection of organic compounds, both diverse and complex in structure, revealed that an empirical formula field greater than 25 columns was virtually never required. Overflows, of course, always can be accommodated in a second card that now is provided for notation sizes beyond size C = 13; this formula overflow can be indicated in the first card by an asterisk in the last column of the empirical formula field. With the present collection of over 5000 commercially available compounds of carbon, more than 21 columns were never required and the field was completely filled in only one case.

The printout of some representative formulas in this collection are shown in Figure 3. The treatment of carbon and hydrogen should be especially noted. Space is provided for the extreme of $C_{999}H_{999}$; however, since few, if any, nonpolymeric compounds will exceed $C_{99}H_{99}$, one space is used with each element as a separator. Consequently, all carbon-hydrogen compounds below $C_{100}H_{100}$ will sort "clean" for these elements. A subsequent sort on the separator columns will reveal any carbon and/or hydrogen counts of 100 or more and the few cards, if any, can be dropped into proper sequence by hand.



```
              TYPICAL PRINTOUT OF EMPIRICAL FORMULA FIELD
                          -/COLUMNS 53-73/-

A DIGIT EQUALS 1          A DIGIT EQUALS 7          A DIGIT EQUALS 9

C 4 H11 N                 C 3 H 6 N2 O S            C 6     -CL2-HG-O4
C 4 H 9 N O6              C 3 H 7 N O2 S            C 6 H 7-AS-O4
C 6 H15 N O               C 5 H 9 N S               C 7 H18-O3-SI-
C10 H 7 N O4              C 6 H 3 CL N2 O4 S        C 9 H18-B-O3
C10 H23 N                 C 6 H 5 CL2 N O3 S        C 9 H21-AL-O3
C11 H19 N O2              C 6 H12 N2 O4 S           C10 H16-O4-ZN-
C12 H15 N O3              C 7 H 8 CL N O3 S         C10 H20-N2-PB-S4
C16 H35 N                 C11 H18 N2 O3 S           C18 H12-N2 O2-ZN-
```

Figure 3. Typical printout of empirical formula field: Columns 53–73.

In the remaining columns of the field, the rest of the empirical formula is punched with the elements in alphabetic order (Hill system; *Chemical Abstracts* practice) with a blank space (or hyphen) as a separator from the previous multiplier or letter.

Special recognition is afforded the presence of elements other than carbon, hydrogen, nitrogen, oxygen, phosphorus, sulfur, and halogens, by the insertion of a hyphen in column 60 (unless the hydrogen count is 100 or more) and by enclosing such an element and its multiplier, if any, within hyphens (*e.g.*, -AS-, -CU2-). It should be noted that this situation will only be encountered with a compound having an *A* digit value of 9 in the BATCH Number. All of these less frequently encountered elements are given a two-letter symbol, except boron. Potassium, tungsten, uranium, vanadium, and yttrium therefore are assigned the expanded symbols KA, WO, UR, VA, and YT, respectively.

Studies conducted with a chemical compound file of the U. S. Army Biological Laboratories have established that the value of the empirical formula field is enhanced if the collection is first sorted according to the ten values of the *A* digit of the BATCH Number. In this arrangement, the empirical formulas associated with a single *A* digit show relatively good alignment, either at the top of sensed cards or in a printout. The examples of empirical formulas shown in Figure 3 have been selected from three values of the *A* digit to illustrate this feature. In most uses of the empirical formula field in conjunction with the *A* digit, it is seldom important to sort deeper into the formula field than the element following carbon and hydrogen.

REGISTRY AND CONTROL FIELDS

Columns 74 through 80 of the IBM-card layout are reserved for a registry number and control punches. With the deck of over 5000 commercial compounds of carbon, the J. T. Baker commodity numbers were employed as registry numbers since most are closely assigned to yield an alphabetic arrangement of the commodity names when listed in increasing order. Columns 74–77 were employed for this purpose. In column 80 for the studies conducted so far, only *four* control marks have been utilized. The *absence* of any mark indicates that information is completed on one card. A *one* is punched when the Wiswesser notation is allowed to extend beyond column 13 and, when necessary, as far as column 45. A *zero* is punched when the Wiswesser notation exceeds notation size C (see Table III) and is "chopped" off with insertion of an asterisk in column 13. This latter card is placed in the deck where a printout of the name is sought that is free of the "tail" of the notation. The printout of the BATCH Directory represents such a case. A *nine* is punched in column 80, of any special *heading cards*, spacers, etc. placed in the deck temporarily, for example, to provide Division heads, in the BATCH Directory.

MODIFICATION OF THE IBM-CARD LAYOUT

Modification of the recommended IBM-card layout for the storage and retrieval of physico-chemical, biochemical, analytical, or commercial information and with retention of the capability of some structural correlation appears to present no problems. The punching of a second deck with suppression of the empirical formula field immediately provides a field of 21 columns or of 22 if column 52 is utilized. Where more space is required, a decision is required as to whether the chemical name field with its prefix field or the notation field should be either compressed or eliminated. In any event, retention of the BATCH Number is recommended since its use facilitates simple structural–atomic correlations with a minimum of invested punched-card capacity.

Some economy in the name field can be secured by abandoning the nicety of an offset, variable-length prefix. In this way, the field might be reduced to 24 to 26 columns. Names previously keypunched with an offset prefix can be shifted readily to the smaller field. The shift to a field of 24 columns was undertaken experimentally with the J. T. Baker deck of 5019 compounds. It was found that the names *previously assigned* to 4236 compounds (84.3%) were directly accommodated in the smaller field, 264 names (5.3%) could be made to fit by simple sorting, inspection, and repunching (for example, dropping of terminal E's unnecessary for comprehension), and 472 (9.5%) could be made to fit by additional systematic contraction following the established practices (2). The remaining 47 names (0.9%) could only be accommodated by the dropping of some "bit" of structural information, such as a position-locant.

THE BATCH DIRECTORY

For the preparation of a BATCH Directory, the relevant punched cards are arranged by the BATCH Number and the information of interest is listed. The initial J. T. Baker BATCH Directory consists of the BATCH Number, the assigned commodity name, and commodity number. A portion of one page of this Directory is reproduced in Figure 1. If the information listed per line amounts to 42 to 46 characters, reduction of the listings secured with conventional line printers to 52 to 50% of original size will allow the printing of three columns on standard 8.5 by 11 inch paper. Consequently, more than 300 lines of listings can be carried on one side of such a sheet as well as appropriate page and BATCH Division headings.

If the listing is arranged according to increasing BATCH Numbers, with these treated as conventional five-digit numbers, the nature (and number) of rings present is given precedence over the type and number of atoms present. Actual permutation of the digits of the BATCH Number would yield additional index numbers providing on listing additional discrimination in answering retrieval requests. However, the user would experience difficulty in establishing whether he has, for example, a BATCH, ATBCH, or ABTCH Number at hand. The same goal can be secured by maintaining a single order of the digits in the BATCH Number, but treating various ones as primary, secondary, etc., in additional listings.

Of these other possible arrangements, treatment of the *A* digit as primary, the *T* digit as secondary, the *B* digit as tertiary, etc., has proved the most valuable in practical trials. By this arrangement, the type and number of atoms, other than carbon and hydrogen, present in the formula are given precedence over the nature (and number) of rings. These two arrangements can be delineated by the following two listings of the same eight BATCH Numbers:

Digits									
B	*A*	*T*	*C*	*H*	*B*	*A*	*T*	*C*	*H*
①	2	3	4	5	3	①	2	4	5
0	0	1	1	4	0	0	1	1	4
0	1	2	1	3	1	0	1	6	6
0	6	4	1	4	3	0	1	6	3
1	0	1	6	6	5	0	1	0	8
1	1	2	6	7	0	1	2	1	3
1	6	4	6	6	1	1	2	6	7
3	0	1	6	3	0	6	4	1	4
5	0	1	0	8	1	6	4	6	6

In the left of the above listings, spaces are used to indicate where the *B* digit changes value, and in the right where the *A* digit changes.

In the use of such a two-part BATCH Directory, the part to be consulted in seeking an answer to a particular inquiry will depend largely on whether rings or atoms are the primary consideration. This selection will be tempered obviously by any foreknowledge of the probable number of entries in the relevant class or classes of structures.

ACKNOWLEDGMENT

W. C. Broad is thanked for his assistance and constructive criticism at various stages of the work reported in this paper.

LITERATURE CITED

(1) Wiswesser, W. J., "Literature Sources of Mammalian Toxicity Data, with Special Emphasis on Tabulating Machinery Applications," Advances in Chemistry Series, No. 16, American Chemical Society, Washington, D. C., 1956, p 64.

(2) Barnard, A. J., Jr., Kleppinger, C. T., Wiswesser, W. J., *J. Chem. Doc.*, **6**, 48 (1966); presented at the 150th National Meeting of the American Chemical Society, Atlantic City, N. J., Sept 12, 1965.

(3) Heumann, K. F., Dale, E., "Statistical Survey of Chemical Structures," in "Progress Report in Chemical Literature Retrieval," G. L. Peakes, A. Kent, and J. W. Perry, Ed., Interscience Publishers, Inc., New York, N. Y., 1957, pp 201-214.

(4) Wiswesser, W. J., "A Line-Formula Chemical Notation," Thomas Y. Crowell Co., New York, N. Y., 1954, 149 pp; Smith, E. G., Wiswesser, W. J., *ibid.*, 2nd ed, in preparation.

(5) Brasie, W. C., Liou, D. W., *Chem. Eng. Progr.*, **61**, No. 5, 102 (1965).

(6) Wiswesser, W. J., *ibid.*, **61**, No. 6, 19 (1965).

(7) Gibson, G. W., *ibid.*, **61**, No. 7, 12 (1965).

(8) Brasie, W. C., Liou, D. W., *ibid.*, **61**, No. 7, 16 (1965).

(9) Wiswesser, W. J., unpublished data.

(10) Benson, F. R., Abstracts, 124th National Meeting of the American Chemical Society, Chicago, Ill., Sept 1953, p 5G.

(11) Sorter, P. F., Granito, C. E., Gilmer, J. C., Gelberg, A., Metcalf, E. A., *J. Chem. Doc.*, **4**, 56 (1964); Granito, C. E., Gelberg, A., Schultz, J. E., Gibson, G. W., Metcalf, E. A., *ibid.*, **5**, 52 (1965); Granito, C. E., Schultz, J. E., Gibson, G. W., Gelberg, A., Williams, R. J., Metcalf, E. A., *ibid.*, **5**, 229 (1965).

Computer-Oriented Chemical Names*

A. J. BARNARD, JR., and C. T. KLEPPINGER
J. T. Baker Chemical Company, Phillipsburg, New Jersey

and W. J. WISWESSER
U. S. Army Biological Laboratories, Fort Detrick, Frederick, Maryland

Received October 29, 1965

A system is described for the transcription of chemical names into capital letters, arabic numerals, and four symbols, namely, & - * /. To allow such transcribed names to be accommodated in a fixed punched-card field, regular practices are introduced for the contraction of frequently cited functional groups, prefixes, and suffixes.

In the management of structural information for small-to-medium size collections, the authors have demonstrated (*1*) the effective use of a readily mastered structural-atomic code, namely, the BATCH Number. The mechanics involved in the preparation of classified structure directories based on the use of this code and the IBM-card layout employed are considered in the previous paper (*1*). One problem that required detailed study was the recording of systematic chemical names in the limited typography allowed by conventional electronic-processing equipment and their contraction, where necessary, so that they could be accommodated in a fixed punched-card field. A fixed field is necessary not only to save invested punched-card capacity but to secure economy of space in any extensive listing. The goal is to achieve transcription and contraction of a chemical name without loss of structural information and with the result still capable of being read by the interested technologist after only brief inspection of the practices adopted. The systematic approaches adopted for the "computer-oriented" expression and contraction of chemical names are the subject of this paper and are employed in the J. T. Baker BATCH Directory (2), which considers over 5000 compounds of carbon offered by a single supplier of laboratory chemicals. Certain of the practices adopted in the transcription of chemical names have evolved from the suggestions of Wiswesser presented in 1953 (*3*) and published in 1956 (*4*).

COMPUTER-ORIENTED TRANSCRIPTION OF CHEMICAL NAMES

Basically, all conventional tabulating machinery allows the use of only capital letters and the arabic numerals. It is difficult to envision the economic transcription of systematic organic chemical names with all of their typographical complexities into these 36 characters (and the use of a blank space as a 37th one). Preliminary studies suggested that at least four additional symbols would be required even if a single symbol was assigned different meanings in different positions within a name; that is,

* 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.