Multi-Level Retrieval Systems

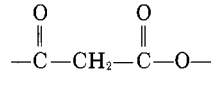
III. A Generic Chemical Search System Using Optical Coincidence Cards*

LEE N. STARKER, JOHN A. KISH, and FRANCES H. ARENDELL
Warner-Lambert Research Institute, Morris Plains, N. J.

Received November 29, 1969

The ability to enter data into Termatrex cards from previously prepared IBM cards has led to development of a generic search system that currently provides a highly flexible search capability. The availability of the IBM card file suggests a ready conversion to computer searching at some future date. Examples of coding techniques and search strategies are discussed.

Many systems have been devised over the past 10 years that have had as their major aim the simplification of the work required to carry out generic chemical searches. In this respect, we define a generic search as one that involves the selection of all chemical compounds having one or more structural characteristics in common. Examples would be all pyridine compounds bearing an amine group on the ring; or those compounds in which the grouping



may be found, etc. Obviously, these kinds of problems cannot be readily solved by use of the conventional name, subject, or formula indexes, and other methods must be devised. One of the first attempts was that made by Fletcher,[1] who employed a file of cards arranged in a molecular formula order that placed the least common elements before the more common ones. These cards were mounted in Acme Kardex files so that manual searches could be conducted at a relatively high rate of speed. The system operated satisfactorily until the file passed the 20,000-compound mark and was then replaced by a punched-card system.[2]

The evolution of chemical structure search systems should also include a mention of the classification system devised by Wiselogle,[3,4] and its adaptation to edge-notched cards by Arendell.[5]

As data processing equipment became more available to workers in scientific areas, the number and sophistication of the search techniques that were employed increased significantly. As a result, we have seen many punch-card oriented systems,[6] and a smaller number of computer-oriented ones,[7] come into use. The attempts of Chemical Abstracts Service to provide generic searches of all chemical structures are still in the developmental process.[8]

*Presented in part before the Third Middle Atlantic Regional Meeting, ACS, Philadelphia, Pa., February 1968, and in part before the Fourth Middle Atlantic Regional Meeting, ACS, Washington, D. C., February 1969.

Also of interest has been the development of machine permuted line-notations,[9] formula indexes,[10] and a combination of codes specifying molecular formulas and generic characteristics.[11]

Despite the nominal speed of computer searching, however, many information groups find that their access to the computer is so limited that they must seek other solutions to their problems. For those in this situation, we can recommend consideration of optical coincidence techniques. This approach permits the user to search readily, and with a great deal of flexibility, without having to be scheduled into a computer. He can carry out search requests at will and often at his own desk.

The information group at the Cancer Chemotherapy National Service Center (CCNSC) developed an optical coincidence generic search system,[12] which was based on the 18,000-document card originated by the National Bureau of Standards.[13] Although it is a compact and versatile system, it suffers from the fact that neither the drill, the reader, nor the cards are readily available on a commercial basis.

We have over the past years developed a significant body of experience with the optical coincidence equipment developed by the Jonker Corporation.[14] Their Termatrex machines have been used by us in a number of information retrieval applications,[15] and after a study of other possible search methods, we decided to adapt a generic search system to this equipment.

A particularly attractive advantage of the peek-a-boo technique for generic chemical searches is the ability to browse through the deck and, where necessary, to get the best approximate answer. Thus, if a question that asked for all pyridine compounds containing alkyl and nitro groups gave no answer, it would immediately be possible to eliminate the alkyl cards and determine the nitropyridines in the collection, or eliminate the nitro cards and determine the alkylpyridines. It is, of course, possible to do the same thing with real-time access to computers, but this is still a relatively rare possibility for most information groups.

Thought also had to be given to the fact that as our collection grew in size, and as computer handling became simpler and more economical, conversion of the system would have to be considered. Ideally, this conversion would be possible without a need to re-keyboard the data. The use of the J-400 Termatrex drill, which is driven by IBM punch cards, makes this possible and has been detailed earlier.[15]

Our considerations then involved the construction of a fragment code that could be entered into IBM punch cards, searched through a Termatrex deck, and eventually placed on tape for computer searching.

SYSTEM AND CODE DEVELOPMENT

The many fragment-oriented generic structure search systems that have been reported have most of their basic fragments in common. In general, the fragments are those which are recognized by the synthetic organic chemist as being the basic building blocks of the compounds with which he deals. Variations in the different systems have depended to a large extent on the types of compounds to be encountered in the collection under consideration. Very few of these systems were meant to deal with the total world of organic chemistry, but they could be readily expanded to include new fragments arising out of new research programs or interests.

Accordingly, we chose to base our system on those functional groups or fragments, rings and fused ring systems, that occurred most frequently in our own collection of compounds.

Molecular Formula Card (Card 1). After an initial analysis of the problem, it seemed that we would be able to encompass the information needed with four IBM card formats (three of which were identical, except for column headings) for input to the Termatrex. It appears that these formats will also prove to be satisfactory for eventual computer input. Card 1 was designed to carry certain basic information, and was laid out as follows:

cc	1–12	Compound Number
	13–37	Molecular Formula
	38–74	Unassigned
	75–78	Generic Codes
	79–80	Controls

The compound number field, cc 1–12, was designed to accommodate our relatively complex compound numbers, which include designation of the source of the material, its number, whether or not it is a salt, and for multiple preparations, the lot number or other differentiation.

The inverted molecular formula layout is illustrated in Figure 1 and shows prepunched atom codes in cc 13, 15, 17, 25, and 28. Up to nine atoms can be recorded for N, S, and Ø, while 99 can be shown for C and H. Up to two kinds of halogens are recorded in cc 19–24, using cc 19–20 and 22–23 for the atom designations; other elements are entered into cc 31–37 as required. In addition, for an amine salt we enter the actual acid in this field—right-justified; while for acid salts, we code the free acid and enter the base portion into cc 31–37, but on a left justified basis (Figure 2).

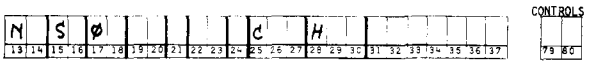


Figure 1. Molecular formula card

The generic code field (cc 75–78) is used to indicate, on a direct punch basis, the fact that the compound under consideration falls within one of a group of broad, predetermined classes of particular interest (steroid, protein, etc.) or possesses a quality or property that cannot be readily demonstrated in another way (*d*-isomer). The codes that have thus far been assigned are shown in Figure 3.

The absence of a zone punch in cc 80 denotes Card 1 of the system. If more than one card of this type is required, it is numbered sequentially in cc 80. The last card of the sequence also carries a dash (-) in cc 79. This occurs most frequently where a material is encountered that has no definite structure and is recorded by each of its names (Figure 4). In this case, the information is punched directly into cc 13–37 without regard for the molecular formula elements. The presence of the dash in cc 79 provides the ability to determine the presence of all cards in a sequence. If a set of cards shows an unbroken series of card numbers, the last of which has a dash, we then know that the set is complete.

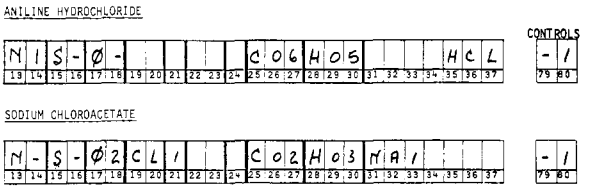


Figure 2. Molecular formulas of salts

cc	75	76	77	78
12				ORG-ORG SALT RIGHT-HAND
11				
0				
1	STEROID	CIS	POLYMER	ORG-ORG SALT ACID PART
2	YOHIMBANE	TRANS	RADIOACTIVE	ORG-ORG SALT BASE PART
3	BIPHENYL	INDEFINITE ISOMER		SALT OF ORG ACID INORG BASE
4		D ISOMER	MIXTURE	SALT OF ORG BASE INORG ACID
5		L ISOMER		
6	SUGAR	DL, MESO ISOMER		CHELATE
7	DIPEPTIDE		METALLO-ORGANIC	HYDRATE
8	TRIPEPTIDE	SPIRO	INORGANIC	SOLVATE
9	POLYPEPTIDE	1, 1 DISUBSTITUTED	MISC	

Figure 3. Generic codes

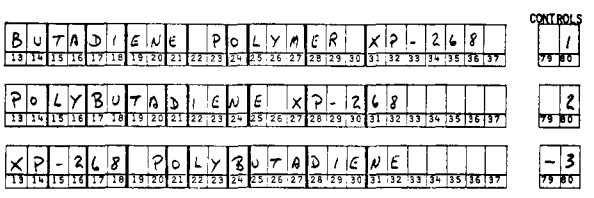


Figure 4. Indefinite structures

The card numbering commences again for each new card type, which is not defined by the presence of a zone punch. Card columns 38–74 are unassigned and are available for some future designation.

Although there are three different types of codes for the fragments, they are all punched in one general type of card format. These cards are differentiated, according to the type of code fragment they contain, by a zone punch in cc 80.

Thus, Card 2 contains functional group fragments and is marked by the 12 punch in cc 80. The monocyclic rings are recorded on Card 3, and show a zero punch in cc 80, while the fused ring systems are recorded on Card 4 and are differentiated by an 11 punch in cc 80.

These cards show the following layouts:

cc	1–12	Compound Number
	13–14	Attachment A
	15	Cyclic Code
	16–22	Fragment or Ring Code
	40	Ring Size
	41–46	Attachments B, C, and D
	47	Number of Groups being Coded
	48	Number of Substituents on the Group
	75–78	Generic Codes (repeated from Card 1)
	79–80	Controls

Fragment Codes (Card 2). Fragment codes relating to the functional groups that are generally found in the Warner-Lambert compound collection are entered, one code per card, in cc 16–21 of Card 2 and are designed to offer increasing levels of either genericity or specificity, as may be desired in a given search. The six-digit code is constructed as follows:

cc	16	A letter denoting a broad group of organic compounds (Figure 5)
	17	The number of nitrogen atoms in the fragment
	18	The number of sulfur atoms in the fragment
	19	The number of oxygen atoms in the fragment
	20	The number of carbon atoms in the fragment
	21	A serial number to differentiate between codes which are identical in the first five digits. Thus, a primary alcohol is H00111, a secondary alcohol is H00112, a primary thioalcohol is H01011, etc.

This approach to code construction makes it possible to retrieve all alcohols and their thioanalogs by use of the 'H' code only. A requirement for all compounds containing fragments with two oxygen atoms (—COO—, NHCOO, etc.) is readily satisfied by searching out all compounds with a 19/2 code. On the other hand, application of a complete six-digit code will result in the retrieval of compounds bearing a specific group—tertiary hydroxyls, aldehydes, etc.

H = ALCOHOLS, ETHERS, THIOALCOHOLS, SULFIDES

Ø = CARBONYLS

M = AMINES, NITRILES, ISONITRILES, IMINES

Y = AMIDES, UREAS, IMIDES, CYANATES

L = HALOGENS

Figure 5. Selected first digit fragment codes

ATTACHMENTS. To more fully define each fragment, up to four groups that may be attached to it can be recorded. These attachments are coded into the four attachment fields as follows:

cc	13–14	Attachment A
cc	41–42	Attachment B
cc	43–44	Attachment C
cc	45–46	Attachment D

The net result, of course, is that each fragment is coded twice—once as a major group, and secondly as an attachment to another group.

Each such field is a two-column field in which the first column denotes the generic level of the attachment (Figure 6), while the second column contains a more specific definition of the attachment.

GENERIC LEVEL	SPECIFIC LEVEL
H	(NONE)
A = ALKYL	SPECIFY NO. OF C's (2, 3, 4, ETC.)
L = ALICYCLIC	SPECIFY RING SIZE (3, 4, 5, 6, ETC.)
P = PHENYL	(NONE)
T = HETERO	N = NITROGEN Y = OTHER HETERO S = SULFUR P = NØ Ø = OXYGEN V = NS, ETC.
X = HALOGEN	F = FLUORINE B = BROMINE K = CHLORINE I = IODINE
F = OTHER FUNCTIONS	N = $-NR_2$, $-NO$, $-NO_2$, ETC. O = -OH, -OR, -O-O-, ETC. S = -SH, -SR, $-SO_2NH_2$, ETC. G = METAL-ORGANIC C = -CN, -COOR, ETC.

Figure 6. Attachment codes

Thus, an 'L' in the first column of any attachment field represents an alicyclic ring, which is modified by a number in the second column to show the number of carbons in the ring (L6 = cyclohexyl).

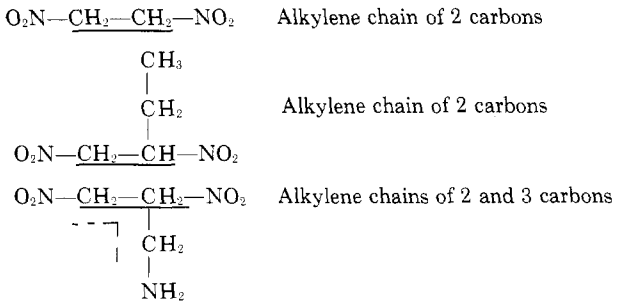
For nonsymmetrical fragments, the 'A,' 'B,' etc. attachments are carefully defined. As an example, when coding esters, the 'A' attachment is always the group on carbon, and the 'B' attachment is always the group on oxygen (A—COO—B). This makes it possible to search for, and differentiate between, ethyl acetate ($CH_3COOC_2H_5$, attachment A = A1, attachment B = A2) and methyl propionate ($C_2H_5COOCH_3$, attachment A = A2, attachment B = A1).

For symmetrical groups such as the amines, the attachments are assigned by certain simple rules of precedence. For methylethylaniline, the sequence of groups on the amine fragment would be:

Attachment A = A1 (for methyl)
Attachment B = A2 (for ethyl)
Attachment C = P (phenyl)

ALKYL, ALKYLENE CHAINS. Although these fragments are coded in the same format as are the functional group fragments, the basic code assigned for them is developed somewhat differently.

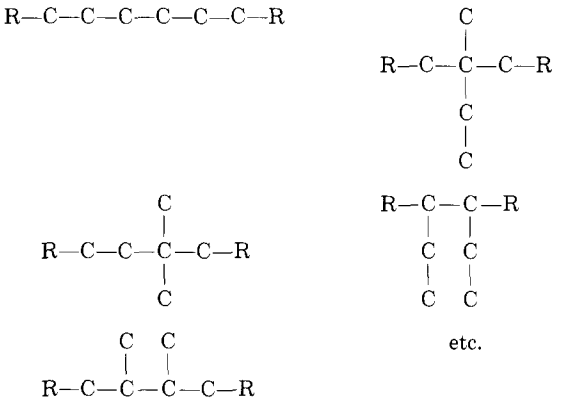
Alkylene groups are defined as carbon chains linking two functional groups and/or rings. The chain is considered to include only those carbon atoms directly between the functional groups and rings, and does not include any branching carbons. Thus, in the following examples:



The codes for these groups are developed as follows:

cc 16 Generic Letter A
17 Number of Carbons between Fragments
18 Number of Double Bonds
19 Number of Triple Bonds
20 Number of Conjugated Bonds
21 Total Number of Carbons in chain (including branching)

The last digit of this code is interesting since it permits the retrieval of any alkylene chain, regardless of the groups or rings it connects, and regardless of how the side chains are constructed. Thus, a request for all carbon skeletons containing six carbons would be carried out by searching for '6' in cc 21, and would yield (among others) the following:



ALKYL GROUPS. These groups are handled similarly to the alkylene chains. Alkyl groups are defined as carbon chains attached to only one functional group or ring. Their codes are constructed as are the codes for the alkylenes, except that the first letter is 'C' and the last digit (cc 21) represents the fact that the group is straight chain (21/0) or branched (21/1).

Thus, $CH_3CH_2CH_2$—Br is coded C30000
While, CH_3—CH(—CH_3)—Br is coded C30001

Thus, all propyl groups can be retrieved and differentiated into *normal* and *iso* chains.

CYCLIC CODES. Cyclic variations of normally linear fragments (esters and lactones, amides and lactams) have been satisfactorily handled by a number of systems[2, 6d] in the past, and we have modified this to the extent of indicating directly the size of the ring involved.

Thus, an ester is given the code B00212. If a lactone is under consideration, we prefix the ester code with a 'C' (in cc 15) to indicate a cyclic group, and a numeral (usually 5 or 6) in cc 40, to denote the size of the ring that carries the lactone. This permits the retrieval of lactones with esters, if desired, or the retrieval of each of these groupings alone. Only attachments that are exo to the lactone portion of the ring are coded for these groups.

Finally, for each group so coded, we record the number of such groups that are contained in the compound and the number of substituents attached to each group.

Monocyclic Codes (Card 3). A separate deck of cards is punched with the codes for any monocycles that occur in the compound. The general card layout is the same as for the fragment cards (Card 2), except that cc 15 and 40 (used to denote cyclic forms of fragments) are not employed.

The seven-character ring number, which is punched in cc 16–22, is again built up so as to move from a highly generic level to a highly specific one. The codes are developed as follows:

cc 16 Ring Size
17 Saturation or Number of Double Bonds
18 Number of N Atoms
19 Number of S Atoms
20 Number of O Atoms
21 Arrangement of Hetero Atoms
22 Classification Number

Thus, for piperazine the code would read 602004:

6 = Six-membered ring
0 = No double bonds
2 = Two nitrogen atoms
0 = No sulfur
0 = No oxygen atoms
4 = 1,4 Arrangement of the hetero atoms

Attachments to the rings are defined as noted previously, but no more than four such groups can be listed. Orientation of these groups on the ring is not coded, since our experience has been that in generic searching this property is not usually one of prime importance and other factors can be used to reduce the file to a manageable size.

Fused Ring Codes (Card 4). The fourth card format is reserved for the fused ring components of organic compounds. The fused ring number is constructed in an analogous manner to the ring and fragment numbers and is built up as follows:

cc 16 Number of rings in the system
17 An alpha code designating the hetero atoms in the system (Figure 7)
18 Number of 5-membered rings
19 Number of 6-membered rings
20 Number of rings that have <5 and >6 members
21 Saturation or number of double bonds in the system
22 Classification number

NO HET ATOM	- 0	N_3O+S+	- R
N	- E	N_4	- S
NO	- F	N_4O+S+	- T
NO_2+	- G	N_5+	- U
NS	- H	N_5O+S+	- V
NOS	- I	N_6O+S+	- W
N_2	- J	N_7O+S+	- X
N_2O	- K	Ø	- A
N_2O_2+	- L	O_2+	- B
N_2S	- M	S+	- C
N_2S_2+	- N	OS	- D
N_3	- Ø	F.E.	- Y
N_3O+	- P	Z	- PARTLY DEFINED
N_3S+	- Q		

Figure 7. Hetero atom code

Thus, indole is coded as 2E11040, and isoindole is coded as 2E11041, where the 'E' refers to the presence of one nitrogen.

Attachments are not shown for fused rings, but the number of substituents on the rings is coded in cc 48.

TERMATREX INPUT

Although the initial conversion of our fragment codes to machine language is *via* IBM punch cards, our search mode at present is wholly *via* the Termatrex optical-coincidence system.

The code, as designed and used at the input level, contains a much higher level of specificity than we feel is needed with the current collection of compounds. Accordingly, we do not input to the Termatrex system all of the information that has been punched into the IBM cards. Thus, when we consider attachments to the various fragments, we input only the generic level of the two-column field (i.e. 'T' for a hetero ring), but not the specific level ('N' for hetero N ring). This means that our level of retrieval is somewhat restricted, and our false drop is increased, but this will not become an annoyance until the number of compounds in the system increases substantially.

The IBM cards are sorted initially on cc 80 to separate the different kinds of cards and/or codes, and the resultant decks (Figure 8) are handled separately for entry to the Termatrex.

Each group of cards is then sorted on a column-by-column basis, and the compound numbers contained in each sort are drilled into the Termatrex cards on a unit-digit basis. It is true that this procedure makes possible a substantial false drop, but it has been our experience that this false drop has not been onerous. It has run at the average level of 10–15%, and in many instances has been considered to be an advantage, since it gives a degree of browseability which is often desirable.

Several examples of completely coded compounds will give a better insight into the use of the system.

A very simple compound, methyl propionate, is coded as shown in Figure 9. The ester grouping is coded on line 1. Here, B002120 in cc 16–22 represents the —COO— moiety. The 'A' attachment, coded as A2 (cc 13–14), specifies the two-carbon chain attached to the ester carbon, while the 'B' attachment, coded as A1 (cc 41–42), shows

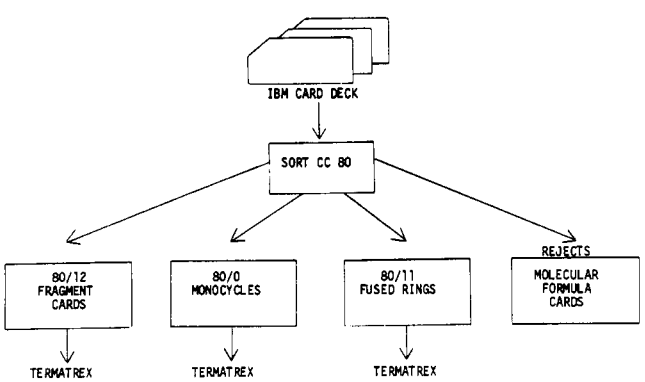


Figure 8. Termatrex sort sequence

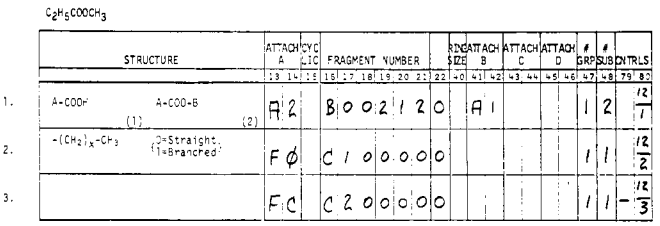


Figure 9. Examples of complete codes

the one-carbon alkyl attached to the ester oxygen. Line 2 details the presence of a methyl group (C100000 in cc 16–22) attached to an oxygen function (FØ in cc 13–14). Finally, line 3 records an ethyl group (C200000 in cc 16–22) attached to a carbon function (FC in cc 13–14).

A more complex compound is shown in Figure 10. In this instance, line 1 shows a primary amino group attached to phenyl; line 2 codes a nitrile group attached to phenyl; and line 3 records the methyl group, also attached to phenyl. The phenyl group, itself, is coded on line 4 (6300000 in cc 16–22) and is shown to bear three attachments, a one-carbon chain (A1 in cc 13–14), a nitrile (FC in cc 41–42), and the amino group (FN in cc 43–44). The card for this code carries the 80/0 punch to distinguish a monocyclic code from the fragment codes which all carry the 80/12 punch.

Finally, we show the coding for a fused ring compound in Figure 11. The tertiary nitrogen is coded on line 1, and the alkyl groups on lines 2 and 3. The fused ring

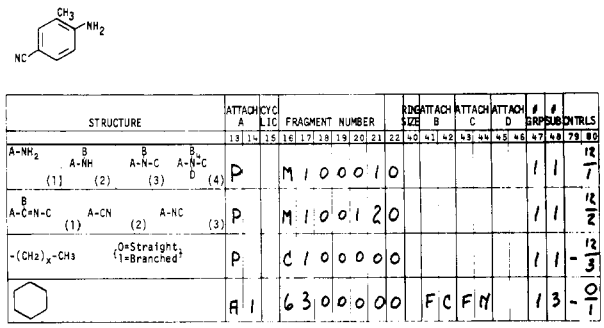


Figure 10. Examples of complete codes

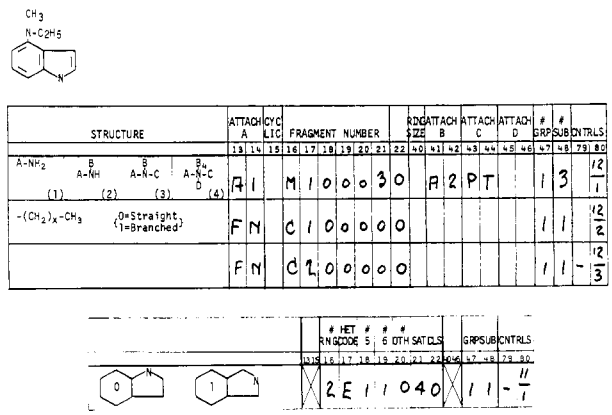


Figure 11. Examples of complete codes

code is shown on line 4, where 2E11040 (cc 16–22) defines the indole nucleus. For these rings, only the total number of substituents is shown in cc 48. The 80/11 punch identifies this card as a fused ring code.

SEARCHES

This system has been in full operation for the past 18 months and has resulted in a significant saving in the time required to carry out generic searches. In general, the results we have obtained can be summarized as follows:

Average number of compounds per search	75
Average false drop per search	13%
Average search time	18 min.

The searches are generally reported by preparing Xerox copies of the structures that have been obtained. A microfilm copy of this file is being prepared and should substantially reduce the time required for this procedure. Where the search requestor has his own file of structure cards, the results of the search may be reported simply as a list of compound numbers. In this instance, the information center does not normally evaluate the search results and weed out false drops.

As we have mentioned previously, the theoretical capacity for false drops in the system as we have thus far implemented it is quite high. To this point, it has remained well within practical limits. We plan, however, to eventually increase the depth of the Termatrex decks by adding the second character of the attachment codes; and most importantly, by recording the more common groups and rings on single cards. This means that instead of retrieving all nitro groups by reading out a combination of six Termatrex cards equivalent to the nitro group code (R10201), the answer would be obtained from a single Termatrex card devoted only to those compounds in which this group was present.

The current approach does, however, permit maximum flexibility in retrieval. This should become clearer from the following example—showing a search for fused ring systems involving one nitrogen.

A combination of five Termatrex cards specifying the following information

2 fused rings
1 nitrogen in a fused ring system
0 five-membered rings
2 six-membered rings
0 rings of other sizes

will yield all two-ring systems containing only six-membered rings, and with one nitrogen (quinolines and isoquinolines). If we now add the '5 double bonds' card, we will eliminate all reduced forms of this compound, while the final addition of the classification number '0' will yield only quinolines. Substitution of this '0' by the classification number '1' card, will present only isoquinolines. Further manipulation of these cards will give a number of other answers of varying levels of genericity.

SUMMARY

We have developed a generic search system that is extremely versatile in output and is based on a comparatively simple input. Coding is not a tedious operation, and can be carried out by relatively inexperienced personnel, with a minimum amount of training. Average coding times of three minutes per compound are readily achieved.

Search strategies, similarly, are simple to devise and easy to carry out. Since the system will readily permit browsing, it can serve as an idea-generator.

REFERENCES

(1) Fletcher, J. H., and D. S. Dubbs, *Chem. Eng. News* **5888** (1956).
(2) Starker, L. N., and J. A. Cordero, J. CHEM. DOC. **2**, 12 (1962).
(3) Wiselogle, F. Y., "A Survey of Antimalarial Drugs 1941–1945," J. W. Edwards, Ann Arbor, Mich., 1946.
(4) Bukle, E. H., E. D. Hartnell, A. M. Moore, L. R. Wiselogle, and F. Y. Wiselogle, *J. Chem. Ed.* **23**, 375 (1946).
(5) Arendell, F. H., J. CHEM. DOC. **1**, 47 (1961).
(6) a. Frome, J., J. CHEM. DOC. **1**, 1 (1961).
b. Geer, H. A., A. M. Moore, C. C. Howard, and C. A. Eady, *Ibid.*, **2**, 110 (1962).
c. Willard, J. R., and E. J. Malkiewich, *Ibid.*, **4**, 211 (1964).
d. Wheeler, K. W., E. R. Andrews, F. Fallon G. L. Krueger, F. P. Palopoli, and E. L. Schumann, *Amer. Doc.* **9**, 198 (1958).
(7) a. Oatfield, H., J. CHEM. DOC. **7**, 37 (1967).
b. Leary, P. T., J. M. Cattley, J. E. Moore, and D. G. Banks, *Ibid.*, **5**, 233 (1965).
c. Frome, J., *Ibid.*, **1**, 76 (1961).
(8) Anonymous, *Chem. Eng. News* **47**, 48 (1969).
(9) a. Sorter, P. F., C. E. Granito, J. C. Gilmer, A. Gelberg, and E. A. Metcalf, J. CHEM. DOC. **4**, 56 (1964).
b. Gelberg, A., W. Nelson, G. S. Yee, and E. A. Metcalf, *Ibid.*, **2**, 7 (1962).
c. Bonnett, H. T., and D. W. Calhoun, *Ibid.*, **2**, 2 (1962).
(10) Garfield E., *Ibid.*, **3**, 97 (1963).
(11) Sher, I. H., J. O'Connor, and E. Garfield, *Ibid.*, **4**, 49 (1964).
(12) Ihndris, R. W., J. CHEM. DOC. **4**, 274 (1964).
(13) Wiedbach, W. A., and J. Stern in "Punched Cards—Their Applications to Science and Industry," R. S. Casey, J. W. Perry, A. Kent, and M. M. Berry, Ed., Reinhold, New York, 1958, Chapter 6.
(14) Jonker Corporation, Gaithersburg, Md.
(15) Starker, L. N., K. C. Owen, and B. C. Batson, J. CHEM. DOC. **9**, 161 (1969).