A General Chemical Compound Code Sheet Format*

By JULIUS FROME** and PAUL T. O'DAY**
U. S. Department of Commerce, Patent Office, Washington, D. C.

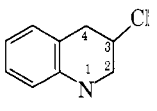
Received May 23, 1963

This paper describes a composite-punch card system for the mechanized information retrieval of organic chemical compounds with definite structures. The system is designed to retrieve both fragmented units and their connections. Its salient feature is a departure from the art-oriented restrictions of previous chemical composite-punch card retrieval aids. The General Chemical Compound Code Sheet Format is designed to record all types of definable organic chemicals. The code format was not intended to apply to inorganic structures, but there is no reason why it would not be useful in this area.

The Composite-Punch Card Format Approach.—The composite-punch card approach is commonly used to code and mechanize the retrieval of chemical information, especially when the size of file is small, when cost is an important factor, and when a computer is not readily available.

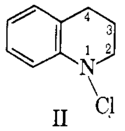
Compositing is the coding of subject matter (in this case, chemical structures) in a given document, into one code format without the use of "links," "interfixes," or other devices for keeping the information separated. This results in a mixing of the coded data and in the creation of new relationships that do not always reflect the parent document. Thus, when chemical structures are composited, a new structure is generated within the code sheet that contains all of the characteristics of individual compounds in the document.

For example, assume a document to contain two separate compounds



I

and



II

A composite system would record on one format both of the characteristics 1-Cl and 3-Cl without keeping the information separate; *i.e.*, if one observes the code sheet without reference to the document he will not know whether there is one compound with both a 1-Cl and a 3-Cl, or two separate compounds as shown.

A searcher wishing to retrieval all documents containing 1,3-dichloroquinoline would receive the above document in a search using a composited system although it is not pertinent. This is generally termed a "false-drop."

Many times, additional descriptors are added to composite systems to reduce the frequency of false-drops; *e.g.*, in the above case the availability of a descriptor for "dichloro compounds" would have eliminated the nonpertinent reference. The use of these types of descriptors is limited by the number of locations available in the coding format after the necessary general subject matter codes have been assigned. Their effect is also often minimized by the complexity of the art coded and the need for a high percentage of descriptors per document.

The composite method being described here employs a coding format that is based on the standard 80-column, 12-row punch card. This allows a total of 960 locations. A number of these must be allocated to document identification so there are approximately 900 locations remaining for recording subject matter. The usual composited system assigns one punch card to each document in the file, but occasionally two or more are used to allow for more selective retrieval of the documents.

It is readily apparent that the selectivity of the search procedure depends on the judicious assignment of subject matter to the available punch card locations. Redundant or unimportant descriptors, if present in sufficient number, can seriously hamper the effectiveness of this type of system as an efficient search tool.

Despite the restrictions of the punch card format, and the inefficiencies caused by the ever-present false-drop, the composite-punch card approach has important advantages that recommend its use for the information retrieval of chemical structures.

A major advantage of composite-punch card systems is the relatively low cost of coding, processing, and searching the information. Since an entire document is generally treated as a unit, the coding costs are not affected by the complexities of dividing the document into separate structures or other divisions to maintain lines of distinction between overlapping subject matters. Redundant coding of information is held to the minimum resulting in a high degree of coding efficiency.

In addition, the composite-punch card systems are also less expensive to process since each document is represented by only one punch card. Noncomposite systems may have up to several hundreds of cards for more complex documents.

The ability to search the composite-punch card system with a multicolumn sorter makes this operation inexpensive and convenient. Updating the file is no problem as the new cards need only be added to the deck.

No system is a good search tool unless the information in its file is reliable and consistent. Generally speaking, composite-punch card systems contain inherent characteristics that contribute to this end. Since the coder cannot code correctly unless he knows what the system can handle

* Presented before the Division of Chemical Literature, 142nd ACS National Meeting, Atlantic City, N. J., September 9-14, 1962.

** Defense Documentation Center for Scientific & Technical Information (formerly ASTIA), Cameron Station, Alexandria, Virginia.

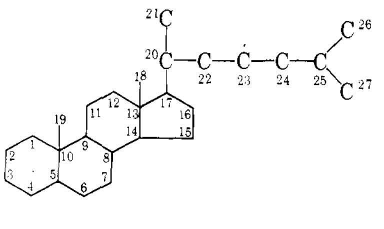
he must be continually aware of the available descriptors. The usual composite code sheet coding format has a limited number of controlled terms that are constantly before the coder.

Since much information appears time and again in patents, the probability of errors of omission decreases with each occurrence. Codes omitted inadvertantly have a good chance of being picked up as the coding progresses through the document. The limited number of descriptors in a compact format also allows convenient checking of the code document by a second analyst.

. Due to the compositing of information in these systems references are often retrieved that are related to the search question but are not directly in point. However, this can be an advantage to the searcher who is interested in searching "around" a given concept. The patent attorney and the examiner are painfully aware of the worth of combinations of references with closely related teachings. Although the composite feature produces a greater number of search answers, this is often compensated by the retrieval of an important reference that would not have answered the search question in a non-composite file.

Evolution of the General Code Sheet.—The general code sheet format had two major forerunners in the U. S. Patent Office that contributed significantly to its conception and development. The earlier of the two predecessors involved the coding of steroids (1). Its selective retrieval powers are predominantly based on a series of relationship sections that generically record the existence of a chemical grouping at one of the numbered steroidal positions. The success of this method was immediately apparent, and further research along the composite-punch card line produced a novel approach to the storage of information on organo-phosphorus compounds. (2) This method relies heavily on a group of three "nodes," or matrices, that record the generic relationship between a given type of phosphorus nucleus and the fragments directly attached, once removed, and farthest removed from it. The organo-phosphorus system is the immediate precursor of the general code sheet approach. The general code sheet is an extension of the logic used in the two previous systems.

The Steroid System.—The retrieval logic of the steroid system[1] is primarily based on the unique numbering sequence assigned to the tetracyclic ring system



The types of chemical groups that are usually substituted on the steroid nucleus are then divided into 24 categories, plus one additional category for "miscellaneous" which includes groups not classified into the other 24.

The code sheet format contains these 25 categories each associated with a listing of 24 possible steroid positions. For example, the coding of a keto group on the 3-position of the steroid nucleus would appear on the code sheet as

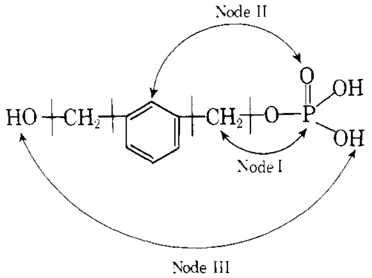
=O	
Ex	21
23+	20
22	10
1	11
2	12
③	13
4	14
5	15
6	16
7	17
8	18
9	19

Since two punch card columns are used for each chemical category there is a total of 50 columns used in this section. The remaining 30 columns are allocated to specific description of groups within the categories, to patent identification, plus a few locations for some descriptors highly specific to the steroid art.

The general acceptance and apparent success of the system spurred activity on devising a composite-punch card system of broader scope. This gave birth to the organo-phosphorus approach.

The Organo-Phosphorus System.—The major need in an extension of the steroid approach is an effective substitution for the highly specific relationship section based on the steroid numbering system. The initial step chosen to loosen the requirement of a large unique nucleus bases the relationship retrieval on a smaller, more common, unique substitutent—the element phosphorus.[2]

The condition that each compound in the system contain at least one phosphorus atom provides a starting point for logical relationship retrieval. This is accomplished by the use of three sets of matrices, which have as coordinates different types of phosphorus, phosphorus-oxygen, and phosphorus-sulfur nuclei along one axis and the different generic areas of chemical fragments along the other. The first matrix records the relationship between the central phosphorus-containing nucleus and those chemical fragments directly attached to it. The second matrix records the relationship between the central phosphorus nucleus and those fragments once removed from it. The third matrix performs a similar function for the fragments farthest removed (terminal fragments) from the phosphorus nuclei. The matrices are termed the first node, second node, and terminal node, respectively. The following is a graphic illustration of the types of relationships recorded in the coding of a typical compound



In addition to the relationship retrieval matrices the system incorporates a section for the specific retrieval of the chemical fragments in the organo-phosphorus compounds without regard to their neighboring substitutents.

An attempt is made in this system to reduce the number of false-drops due to compositing by dividing the subject matter of the documents into two separate categories and coding a single code sheet for each. One code sheet is used for all compounds containing phosphate or thiophosphate nuclei, *i.e.*, the structure $X=P(X)_3$, where X = oxygen or sulfur. The second code sheet is used for compounds with all other types of phosphorus nuclei. Thus, each document coded will have one or two code sheets assigned to it depending on whether it contains one or both classes of phosphorus compounds.

Since the approach is inherently limited to those arts that require a unique atom in each compound, the apparent success of this method only solves a part of the retrieval problem.

Additionally, both the steroid and phosphorus approaches can be held to be only conclusively effective on a relatively small file. There is no way of predicting what effect a much larger file would have on these systems. The search time would be proportionally longer but the number of drops per search would depend on the direction of the arts as they produce new documents for the file. If they develop in a highly concentrated, overlapping manner, the effectiveness of either system would be minimized accordingly. On the other hand, art development that does not require constant heavy reliance on a few of the code sheet areas might render the systems effective for files many times their present size.

The next effort in this area was an attempt to devise and test a code sheet format that is applicable to all of organic chemistry. The remainder of this paper is devoted to our attempt to solve this problem.

The General Chemical Compound Code Sheet.—The previous systems have depended upon the inherent logic of a particular type of chemical structure for relationship retrieval. The General Code Sheet departs from this limitation in an attempt to apply a punch card composite system to a broad area with no reliance on unique, built-in logic for retrieval of connections and relationships between chemical groups. The system has been initially applied to 3500 U. S. patents in the organometallic art, so some of the aspects of the code sheet will be described within that context. The code sheet is shown in Table I.

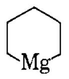
The major need in bridging the gap between the art-oriented predecessors and the universal approach is the construction of a coding format that will retrieve relationships between chemical fragments. To be truly general in scope such a format cannot be bound by any requirement for the existence of unique units in the compounds to be coded. The general code sheet attempts to solve this problem by using a matrix with coordinates that contain all of the possible chemical fragments within their definitions.

The code sheet is divided horizontally into two parts. The upper half uses 25 punch card columns and contains relationship descriptors that indicate the connections of the fragments of chemical structures. A classification of chemical fragments into 24 generic categories is used for descriptor assignments. Linkages of fragments defined by these categories are coded by recording the appropriate column and row designation.

Some specific fragments within these generic categories have been assigned descriptors in the lower half of the code sheet. Those without a specific descriptor assignment are recorded in the open-ended listing found in the "Miscellaneous Dictionary" section. The system also has provision for the specific and generic coding of rings and ring systems.

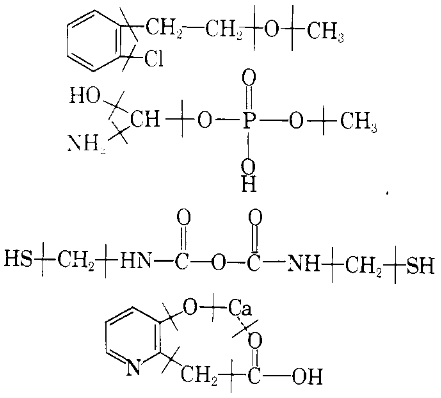
Fragmenting.—Before a compound is coded it must be fragmented according to the following rules:

1. Rings consisting solely of carbon, nitrogen, oxygen, or sulfur, *e.g.*, benzene, thiophene, are coded as a unit. Rings containing other components are broken up and the parts are coded separately, *e.g.*



is considered as Mg connected to a 5-membered alkylene.
2. Combinations of nitrogen, sulfur, oxygen, carbonyl carbon, and $N-\overset{|}{C}=N$ are considered as units.
3. All hydrocarbon groups irrespective of unsaturation or branching are kept intact.
4. Combinations of phosphorus, oxygen, and sulfur are grouped together.
5. Hydrogen is disregarded in fragmenting.

The following examples illustrate the fragmenting step:



Fragment Relationships.—After a compound being coded is fragmented, relationship codes are recorded for each place of separation between the fragments. The upper section of the code sheet, comprising columns 1–25, is used for this purpose. This relationship section is based on a matrix that uses as coordinates 24 generic categories which include all of organic chemistry. The basic form of the matrix is shown in Fig. 1.

The matrix has been listed vertically on the code sheet to reduce coding errors. One set of coordinates is used as headings for the columns and the appropriate coordinates from the other axis are listed beneath. Both sets of coordinates are identical, so the parent matrix is triangular in shape. Thus each succeeding vertical list diminishes in size by one term. The lists beneath the first four headings are identical in size because the relationships—alkyl, alkylene, alkinyl, alkenyl—are treated as a single fragment according to fragmenting rule no. 3, and are therefore omitted from the matrix.

To illustrate, the relationships for 1-chloro-2-aminoethane are coded as shown in Fig. 2.

	Halo	Aryl	Cycloalk	COOR	S-COOR	Metal	S Hetero	O Hetero	N Hetero	N,C,S	N,C,O	S,O	C,N	—C=O	N,O	Amine	-O-(S)	Phos.	OH(SH)	Misc.	
Alkyl																					Alkyl
Alkylene																					Alkylene
Alkenyl																					Alkenyl
Alkinyl																					Alkinyl
Halo																					Halo
Aryl																					Aryl
Cycloalk																					Cycloalk
COOR																					COOR
S-COOR																					S-COOR
Metal																					Metal
S Hetero																					S Hetero
O Hetero																					O Hetero
N Hetero																					N Hetero
N,C,S,																					N,C.S
N,C,O																					N.C.O
S,O																					S.O
C,N																					C,N
—C=O																					—C =O
N,O																					N,O
Amine																					Amine
-O-(S)																					-O-(S)
Phosphorus																					Phos.
OH(SH)																					OH(SH)
Miscellaneous																					Misc.

Figure 1.

Fragment Descriptions.—The lower half of the code sheet, columns 25–51, is devoted to recording specific information about the fragments coded generically in the relationship section. In *coding* fragment descriptors, all codes that apply, both specific and generic, are coded for each fragment. In *searching*, the best practice is to ask only for the most specific code applicable, omitting redundant generic descriptors.

Miscellaneous Connection Codes.—The body of art first coded on the General Code Sheet comprised about 3500 patents classified in organometallic chemistry. Since it was not clear that the general approach to punch card compositing was going to be effective, a number of miscellaneous sections, some art-directed, were incorporated into the sheet. Three of these sections are supplements to the "Fragment Connection" half of the code sheet and are labeled "Number of Metal Connections," "Number of Identical Metal Connections," and "2-Identical Connections." These sections are optional in any given application of the General Code Sheet and, if desired, may be replaced with other art-oriented categories or with general descriptors.

Metal Connections.—Column 52, rows 0–5, are used to indicate the total number of metal connections for each individual metal in the compound. The total is independent of actual valence as in Werner complexes where

$Cl|CH_2\text{-}CH_2|NH_2$

ALKYLENE

	8	Halo
	9	Aryl
	11	Cycloalk
2	12	COOR
3	0	S-COOR
	1	Metal
	2	S Hetero
	3	O Hetero
	4	N Hetero
	5	N,C,S
	6	N,C,O
	7	S,O
	8	C,N
	9	Carbonyl
	11	N,O
3	12	Amine
4	0	O(S)
	1	Phosphorus
	2	OH(SH)
	3	Misc.

Figure 2.

Note that the relationship codes may be recorded only one way. If reference is had to the "Halo" and "Amine" relationship columns, no descriptors will be found listed there for "Alkylene."

Table I. Left Side

CONNECTIONS

		CYCLOALK
	11	Cycloalk
10	12	COOR
11	0	S-COOR
	1	Metal
	2	S-Hetero
	3	O-Hetero
	4	N-Hetero
	5	N,C,S
	6	N,C,O
	7	S,O
	8	C,N
	9	Carbonyl
	11	N,O
11	12	Amine
12	0	O,(S)
	1	Phosphorus
	2	OH,(SH)
	3	Miscell.

		COOR
	4	S-COOR
	5	Metal
	6	S-Hetero
	7	O-Hetero
	8	N-Hetero
	9	N,C,S
	11	N,C,O
12	12	S,O
13	0	C,N
	1	Carbonyl
	2	N,O
	3	Amine
	4	O,(S)
	5	Phosphorus
	6	OH,(SH)
	7	Miscell.

		S-COOR
	8	S-COOR
	9	Metal
	11	S-Hetero
13	12	O-Hetero
14	0	N-Hetero
	1	N,C,S
	2	N,C,O
	3	S,O
	4	C,N
	5	Carbonyl
	6	N,O
	7	Amine
	8	O,(S)
	9	Phosphorus
	11	OH,(SH)
14	12	Miscell.

		METAL
	0	Metal
	1	S-Hetero
	2	O-Hetero
	3	N-Hetero
	4	N,C,S
	5	N,C,O
	6	S,O
	7	C,N
	8	Carbonyl
	9	N,O
	11	Amine
15	12	O,(S)
16	0	Phosphorus
	1	OH,(SH)
	2	Miscell.

		S-HETERO
	3	S-Hetero
	4	O-Hetero
	5	N-Hetero
	6	N,C,S
	7	N,C,O
	8	S,O
	9	C,N
	11	Carbonyl
16	12	N,O
17	0	Amine
	1	O,(S)
	2	Phosphorus
	3	OH,(SH)
	4	Miscell.

		O-HETERO
	5	O-Hetero
	6	N-Hetero
	7	N,C,S
	8	N,C,O
	9	S,O
	11	C,N
17	12	Carbonyl
18	0	N,O
	1	Amine
	2	O,(S)
	3	Phosphorus
	4	OH,(SH)
	5	Miscell.

		CARBONYL
	8	Carbonyl
	9	N,O
	11	Amine
22	12	O,(S)
23	0	Phosphorus
	1	OH,(SH)
	2	Miscell.

		C,N
	0	C,N
	1	Carbonyl
	2	N,O
	3	Amine
	4	O,(S)
	5	Phosphorus
	6	OH,(SH)
	7	Miscell.

		S,O
	3	S,O
	4	C,N
	5	Carbonyl
	6	N,O
	7	Amine
	8	O,(S)
	9	Phosphorus
	11	OH,(SH)
21	12	Miscell.

		N,C,O
	5	N,C,O
	6	S,O
	7	C,N
	8	Carbonyl
	9	N,O
	11	Amine
20	12	O,(S)
21	0	Phosphorus
	1	OH,(SH)
	2	Miscell.

		N,C,S
	6	N,C,S
	7	N,C,O
	8	S,O
	9	C,N
	11	Carbonyl
19	12	N,O
20	0	Amine
	1	O,(S)
	2	Phosphorus
	3	OH,(SH)
	4	Miscell.

		N-HETERO
	6	N-Hetero
	7	N,C,S
	8	N,C,O
	9	S,O
	11	C,N
18	12	Carbonyl
19	0	N,O
	1	Amine
	2	O,(S)
	3	Phosphorus
	4	OH,(SH)
	5	Miscell.

METALS

	0	I-A
	1	Li
	2	Na
	3	K
	4	Rb
	5	Cs
	6	II-A
	7	Be
	8	Mg
	9	Ca
	11	Sr
44	12	Ba
45	0	III-A
	1	B
	2	Al
	3	Ga
	4	In
	5	Tl
	6	IV-A
	7	Si
	8	Ge
	9	Sn
	11	Pb
45	12	V-A
46	0	As
	1	Sb
	2	Bi
	3	VI-A
	4	Se
	5	Te
	6	Po
	7	I-B
	8	Cu
	9	Ag
	11	Au
46	12	II-B
47	0	Zn
	1	Cd
	2	Hg
	3	III-B
	4	Sc
	5	Y
	6	La
	7	IV-B
	8	Ti
	9	Zr
	11	Hf

47	12	V-B
48	0	V
	1	Nb
	2	Ta
	3	VI-B
	4	Cr
	5	Mo
	6	W
	7	VII-B
	8	Mn
	9	Tc
	11	Re
48	12	VIII-B
49	0	Fe
	1	Ru
	2	Os
	3	Co
	4	Rh
	5	Ir
	6	Ni
	7	Pd
	8	Pt
	9	Lanthanide
	11	Ce
49	12	Other
50	0	Actinide
	1	Th
	2	U
	3	Other
	4	Hydride
	5	Bis Alkyl
	6	Tris Alkyl
	7	Tetra Alk
	8	Unkn. Conn
	9	Cyc Metal
	11	Chelate
50	12	
51	0	Mono Met
	1	Poly Met
	2	2 M
	3	3 M
	4	4 M
	5	5+M
	6	Mixed Met
	7	2 M
	8	3 M
	9	4+M
	11	
	12	

		# of Metal Connections
	0	1 Conn.
	1	2 Conn.
	2	3 Conn.
	3	4 Conn.
	4	5 Conn.
	5	6+Conn.

		# of Identical Metal Conn.
	6	1 I.C.
	7	2 I.C.
	8	3 I.C.
	9	4 I.C.
	11	5 I.C.
52	12	6+I.C.
53		

		2+ Identical Frag.Connections
	0	Alkyl
	1	Alkylene
	2	Alkinyl
	3	Alkenyl
	4	Halo
	5	Aryl
	6	Cycloalk
	7	COOR
	8	S-COOR
	9	Metal
	11	S-Hetero
53	12	O-Hetero
54	0	N-Hetero
	1	N,C,S
	2	N,C,O
	3	S,O
	4	C,N
	5	Carbonyl
	6	N,O
	7	Amine
	8	O,(S)
	9	Phosphorus
	11	OH,(SH)
	12	Miscell.

MISCELLANEOUS DICTIONARIES

ANION		P-OTHER		Misc N,C,S,O		MISCELL.	
55	56	57	58	59	60	61	62
0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9
11	11						
12	12						

RING DESCRIPTORS

		Ring System (General)
	0	2+R.S.
	1	2 Rings
	2	3 Rings
	3	4 Rings
	4	5 Rings
	5	6 Rings
	6	1 N-Ring
	7	2 N-Rings
	8	3 N-Rings
	9	1 S-Ring
	11	2 S-Rings
63	12	1 O-Ring
64	0	2 O-Rings
	1	1 Mis-Het R
	2	2 Mis-Het R
		Fused Face
	3	M
	4	N
	5	P
	6	Q
	7	R
	8	S
	9	T
	11	U
	12	V

		RING POSITION #
	0	#1
	1	#2
	2	#3
	3	#4
	4	#5
	5	#6
	6	#7
	7	#8
	8	#9
	9	#10
	11	#11
65	12	#12
66	0	#13
	1	#14
	2	#15
	3	#16
	4	#17
	5	#18
	6	#19
	7	#20-24
	8	#25-29
	9	#30
	11	Mono R
	12	Poly R

RING INDEX #

67	68	69	70
0	0	0	0
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9

U.S. DEPT OF COMMERCE
Patent Office

R&D J.Frome & P.O'Day
December 1, 1961

DOCUMENT IDENTIFICATION

71	72	73	74	75	76	77

Analyst__________________ Checked By__________________

Table I. Right Side

FRAGMENT

		ALKYL
	0	Halo
	1	Aryl
	2	Cycloalk
	3	COOR
	4	S-COOR
	5	Metal
	6	S-Hetero
	7	O-Hetero
	8	N-Hetero
	9	N,C,S
	11	N,C,O
1	12	S,O
2	0	O,N
	1	Carbonyl
	2	N,O
	3	Amine
	4	O,(S)
	5	Phosphorus
	6	OH,(SH)
	7	Miscell.

		ALKYLENE
	8	Halo
	9	Aryl
	11	Cycloalk
2	12	COOR
3	0	S-COOR
	1	Metal
	2	S-Hetero
	3	O-Hetero
	4	N-Hetero
	5	N,C,S
	6	N,C,O
	7	S,O
	8	C,N
	9	Carbonyl
	11	N,O
3	12	Amine
4	0	O,(S)
	1	Phosphorus
	2	OH,(SH)
	3	Miscell.

		ALKINYL
	4	Halo
	5	Aryl
	6	Cycloalk
	7	COOR
	8	S-COOR
	9	Metal
	11	S-Hetero
4	12	O-Hetero
5	0	N-Hetero
	1	N,C,S
	2	N,C,O
	3	S,O
	4	C,N
	5	Carbonyl
	6	N,O
	7	Amine
	8	O,S
	9	Phosphorus
	11	OH,(SH)
5	12	Miscell.

		ALKENYL
	0	Halo
	1	Aryl
	2	Cycloalk
	3	COOR
	4	S-COOR
	5	Metal
	6	S-Hetero
	7	O-Hetero
	8	N-Hetero
	9	N,C,S
	11	N,C,O
6	12	S,O
7	0	C,N
	1	Carbonyl
	2	N,O
	3	Amine
	4	O,(S)
	5	Phosphorus
	6	OH,(SH)
	7	Miscell.

		HALOGEN
	8	Aryl
	9	Cycloalk
	11	COOR
7	12	S-COOR
8	0	Metal
	1	S-Hetero
	2	O-Hetero
	3	N-Hetero
	4	N,C,S
	5	N,C,O
	6	S,O
	7	C,N
	8	Carbonyl
	9	N,O
	11	Amine
8	12	O,(S)
9	0	Phosphorus
	1	OH,(SH)
	2	Miscell.

		ARYL
	3	Aryl
	4	Cycloalk
	5	COOR
	6	S-COOR
	7	Metal
	8	S-Hetero
	9	O-Hetero
	11	N-Hetero
9	12	N,C,S
10	0	N,C,O
	1	S,O
	2	C,N
	3	Carbonyl
	4	N,O
	5	Amine
	6	O,(S)
	7	Phosphorus
	8	OH,(SH)
	9	Miscell.

		MISCELLANEOUS
24	11	Miscell.

		OH,(SH)
24	9	Miscell.

		PHOSPHORUS
	6	Phosphorus
	7	OH,(SH)
24	8	Miscell.

		O,(S)
	2	O,(S)
	3	Phosphorus
	4	OH,(SH)
24	5	Miscell.

		AMINE
	9	Amine
	11	O,(S)
23	12	Phosphorus
24	0	OH,(SH)
	1	Miscell.

		N,O
	3	N,O
	4	Amine
	5	O,(S)
	6	Phosphorus
	7	OH,(SH)
23	8	Miscell.

FRAGMENT DESCRIPTIONS

		ALKYL
	0	Alkyl
	1	Lo (1-7C)
	2	1C
	3	2C
	4	3C
	5	4C
	6	5-7C
	7	Hi (8+C)
	8	8-12C
	9	13+C

		ALKYLENE
	11	Alkylene
25	12	Lo (1-7C)
26	0	1C
	1	2C
	2	3C
	3	4C
	4	5-7C
	5	Hi (8+C)
	6	8-12C
	7	13+C

		ALKINYL
	8	Alkinyl
	9	Lo (1-7C)
	11	2C
26	12	Hi (8+C)
27	0	Mono ≡
	1	Poly ≡

		ALKENYL
	2	Alkenyl
	3	Lo (1-7C)
	4	=CH$_2$
	5	2C
	6	3C
	7	4C
	8	5-7C
	9	Hi (8+C)
	11	8-12
27	12	13+C
28	0	Mono =
	1	Poly =

		HALOGEN
	2	Halo
	3	F
	4	Cl
	5	Br
	6	I
	7	1 Halo
	8	2 Halo
	9	3 Halo
	11	4 Halo
28	12	5+Halo
29		

		ARYL
	0	Aryl
	1	2+
	2	1 Sub
	3	2 Sub
	4	D (O)
	5	E (M)
	6	F (P)
	7	3 Sub
	8	G (O,M)
	9	H (O,M,P)
	11	J (M)
29	12	4 Sub
30	0	5 Sub
	1	6 Sub
	2	Ind
	3	Spiro
	4	Fused

		CYCLOALKYL
	5	Cycloalk
	6	2
	7	3+
	8	Saturated
	9	Unsat.
	11	3,4M
30	12	5M
31	0	diene
	1	6M
	2	7+M
	3	Bicyc (1C)
	4	Bicyc (2C)
	5	Ind
	6	Spiro
	7	Fused

		COOR
	8	COOR
	9	2+
	11	R is H
31	12	R is not H
32	0	COHalo
	1	OCOO
	2	Anhydride
	3	Miscell.

		S-COOR
	4	S-COOR
	5	2+
	6	CSSH
	7	CSSR
	8	CSOH
	9	CSOR
	11	COSH
32	12	COSR
33	0	CSHalo
	1	XCXX
	2	Anhydride
	3	Miscell.

		S-HETERO
	4	S-Hetero
	5	3,4M
	6	5M
	7	6M
	8	7+M
	9	O-Cont
	11	N-Cont
33	12	Other Cont
34	0	1S
	1	2+S
	2	Ind
	3	Spiro
	4	Fused

		N,O
	5	NO
	6	NO$_2$
	7	NOH
	8	Miscell.

		O-HETERO
	9	O-Hetero
	11	3,4M
34	12	5M
35	0	6M
	1	7+M
	2	N-Cont
	3	S-Cont
	4	Other Cont
	5	1-O
	6	2+O
	7	Ind
	8	Spiro
	9	Fused

		N-HETERO
	11	N-Hetero
35	12	3,4M
36	0	5M
	1	6M
	2	7+M
	3	O-Cont
	4	S-Cont
	5	Other Cont
	6	1N
	7	2N
	8	3+N
	9	Salt
	11	Ind
36	12	Spiro
37	0	Fused

		CARBONYL
	1	C=O
	2	CHO
	3	C=S
	4	CHS
	5	=O (ring)
	6	=S (ring)
	7	Metal Carb.
	8	1 CX
	9	2 CX
	11	3 CX
37	12	4 CX
38	0	5 CX
	1	6+CX

		N,C,O
	2	$N-\overset{O}{C}-$
	3	$N-\overset{O}{C}-O-$
	4	$N-\overset{O}{C}-N$
	5	CNO
	6	NCO
	7	Miscell.

		S,O
	8	SO
	9	SO$_2$
	11	SO$_3$H
38	12	SO$_4$
39	0	Miscell.

		C,N
	1	CN
	2	NC
	3	$N-\overset{N}{C}-$
	4	$N-\overset{N}{C}-N$
	5	Miscell.

		N,C,S
	6	$N-\overset{S}{C}-$
	7	$N-\overset{S}{C}-S-$
	8	$N-\overset{S}{C}-N$
	9	SCN
	11	CNS
39	12	Miscell.
40		

		OH,(SH)
	0	OH
	1	SH
	2	1 XH
	3	2 XH
	4	3 XH
	5	4 XH
	6	5+XH

		AMINE
	7	2+
	8	Primary
	9	Secondary
	11	Tertiary
40	12	Quaternary
41	0	Diazo
	1	Azide
	2	Nitride
	3	=NH
	4	=NR
	5	Hydrazine
	6	Salt
	7	Miscell.

		O,(S)
	8	-O-
	9	-O-O-
	11	-S-
41	12	-S-S-
42	0	-(S)$_{3+}$-
	1	2 X
	2	3 X
	3	4+X

		PHOSPHORUS
	4	X=PX$_3$
	5	1+S
	6	4S
	7	=S,2S,O
	8	=S,S,2O
	9	=S,3O,
	11	=O,3S
42	12	=O,3O
43	0	=O,2S,O
	1	=O,S,2O
	2	P-Other
	3	P-X-P
	4	Poly-P
	5	Cyclic-P
	6	P+3
	7	P+5

		MISCELLANEOUS
	8	Ferrocene
	9	
	11	
	12	Miscell.

the coordination number of the metal is the significant figure for this category. All metal bonds are considered whether they are predominantly ionic, coordinate, or covalent. A double bonded attachment is considered as one connection.

Identical Metal Connections.—Column 52, rows 6–12, are used to code the number of identical groups attached to a given metal. For the purpose of this category, groups are identical if they are defined by the same specific descriptor in the fragment identification codes (columns 25–50). If a metal is bonded to two identical fragments by two different bond types, it is coded as 2-identical connections. The nature of the bond, covalent, coordinate, ionic, etc., is irrelevant in determining the number of identical connections for a given metal.

Other Identical Connections.—Columns 53 and 54 are used to record the existence of two or more identical connections between fragments in a given compound. Connections are identical within the meaning of this cateogry if the fragments involved are defined by the same descriptors in the fragment identification section (columns 25–50). Coding in this category is done in pairs of descriptors as each connection recorded must involve two fragments.

One fragment may have two identical connections (*e.g.*, the N in a dimethyl amine group), or the identical connections may be the product of two separate pairs of fragments (*e.g.*, the two halo-alkylene links in dichlorodimethyl ether).

The identical connections are recorded by coding each appropriate generic term involved in the particular repeated connection.

Miscellaneous Dictionaries.—Columns 55–62 are used to record fragments that do not appear specifically on the code sheet. Each dictionary is an "open-ended" listing of these miscellaneous fragments that is expanded as new fragments appear in the coding operation.

When a new fragment occurs in coding, a number is assigned to it in the appropriate dictionary. The number is recorded by coding the corresponding numbers in the code sheet miscellaneous dictionary section. For example, the number assigned to anion "X" is found in the anion dictionary. Assuming the number assigned is 21 the anion is coded as shown.

ANION	
55	56
0	0
1	[✓ 1]
[✓ 2]	2
3	3
4	4

Each dictionary has one or more "entrance" codes in the "Fragment Description" section. One of these must be coded whenever the dictionaries are used. Whenever more than one code is entered in a dictionary on the same code sheet, the fragments themselves are drawn on the code sheet next to the dictionary. This allows efficient manipulation of the codes or revision, if necessary, and also facilitates later reference to the code sheets to check for errors.

The anion dictionary (col. 55, 56) includes various anions and cations that are not on the code sheet. It is also used to code the rarer metals that have not been assigned specific codes in the "Metal" section. The entrance code for the Anion dictionary is found at column 43, row 11.

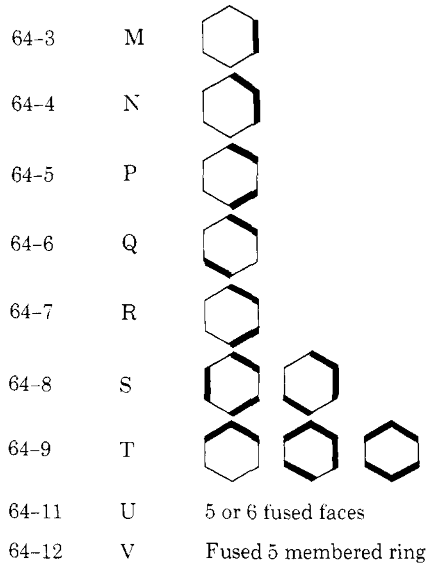
Another dictionary (col. 57, 58) is used to record phosphorus, phosphorus–oxygen, and phosphorus-sulfur fragments that do not have specific descriptions in column 42, rows 7–12, and column 43, rows 0 and 1. The entrance code is the "P-Other" descriptor at column 43, row 2. Polyphosphates are treated as a unit and are coded in the P-Other dictionary.

The Miscellaneous N,C,S,O dictionary (col. 59, 60) is used to record all combinations of nitrogen, carbonyl carbon, sulfur, oxygen, and the group $N-\overset{|}{C}=N$ that do not have specific places elsewhere on the code sheet. "Miscellaneous" descriptors that are entrance codes for this dictionary are found in categories COOR (32–3); S-COOR (33–3); N,O (34–8); N,C,O (38–7); S,O (39–0); C,N (39–5); N,C,S (39–12); and Amine (41–7). Each fragment coded in this dictionary should receive one of these codes.

A final dictionary (col. 61, 62) acts as a catch-all for terms that have no place within any of the generic categories on the code sheet. Its entrance code is found at column 43, row 12. This dictionary is used to record unusually complex or indefinite chemical compositions that cannot be depicted, fragmented, and coded in the specific descriptor section of the code sheet.

Ring Descriptors.—Columns 63–70 on the code sheet are used for coding specific and generic information about rings, ring systems, and their points of substitution. It is expected that this section will be one of the most useful and valuable portions of the coding format for selective retrieval. An attempt has been made to provide wide generic search potential as well as specific retrieval of rings and ring systems.

Fused Face.—The fused face codes indicate the fused sides of 6-membered rings in fused ring systems. The fused faces for all 6-membered rings are recorded. Code 64–12 is recorded for all fused 5-membered rings.

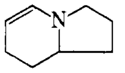


Ring Position Number.—The Ring Position Number codes (col. 65, 66) indicate the point of attachment of subtitutions on ring systems and hetero single rings. The preferred numbering systems given in the "Ring Index"[3] are used.

These codes are used for all ring systems (fused or spiro linkage of two or more rings) and all hetero single rings. Benzene and independent cycloalkyl ring position substitution numbers are not coded.

Ring Index Number.—Columns 67–70 are used to record the Ring Index Number[3] of rings and ring systems. As in the case of the Ring Position Number codes these codes are only applied to ring systems and to independent hetero rings. The Ring Index Numbers for benzene and for independent cycloalkyl rings are not coded. Adequate retrieval of these rings is provided in the fragment description codes. Whenever more than one Ring Index Number is coded on the same code sheet, the separate numbers, along with the names of the ring systems are written in the margin next to the Ring Index section. This allows efficient revision and checking of the codes.

Coding Ring Systems.—Ring systems are coded in the Ring Index Number codes. The individual rings that make up the system are also coded in the appropriate fragment descriptor categories in columns 25–50. When a hetero atom is a bridge in a ring system it is considered part of all the rings fused at that point.



The Ring Index Number (#1276) is coded for the system followed by coding of:

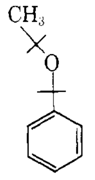
	N-hetero	(35–11)
	6-member	(36–1)
	1N	(36–6)
	fused	(37–0)
and,	N-hetero	(35–11)
	5-member	(36–0)
	1N	(36–6)
	fused	(37–0)

When substitutents are attached to ring systems the relationship codes are recorded for the particular ring to which it is attached. The nature of the remainder of the ring system does not affect this code. If a substituent is attached to a bridge it is considered substituted on all rings fused at that point.

Document Identification.—Columns 71–77 are used to record the document number. In the case of U. S. patents the entire number is entered. For other documents an identification code is devised and recorded in these columns.

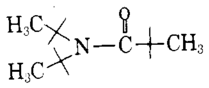
The following examples illustrate the use of the system.

2–4
10–6
25–0, 1, 2
29–0, 2
30–2
41–8



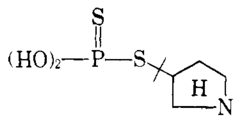
(1)

1–11
25–0, 1, 2
38–2
53–0
54–2



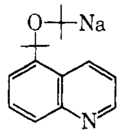
(2)

19–3
35–11
36–0, 6, 11
42–4, 5, 8
43–7
65–2
66–11
67–0
68–1
69–4
70–2



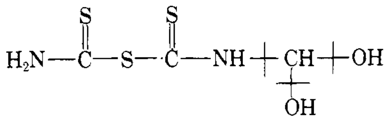
(3)

9–11
10–6
15–12
29–0
30–4
35–11
36–1, 6
37–0
41–8
44–0, 2
51–0
52–0, 6
63–1, 6
64–3
65–4
66–12
67–1
68–7
69–0
70–7



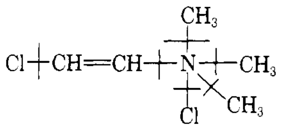
(4)

3–5
4–2
25–11, 12
26–0
39–12
40–0, 3
53–1
54–11
59–7
60–5



(5)

2–3
6–0
7–3
8–11
25–0, 1, 2
27–2, 3, 5
28–0, 2, 4, 8
40–12
53–0
54–7



(6)

Organometals.—Once the general format had been conceived and crystallized, a body of art was selected for the initial application of the system. Attempt was made to select an art that would provide a severe yet comprehensive test of the new coding format. The art also had to be highly active so a useful file would be produced. The area chosen was organometals comprising about 3200 U. S. patents.

Since it was not altogether certain that the code sheet was selective enough for a file of this size, a portion of the format, columns 44–52, was set aside for art-directed descriptors. This is a variable section that can change with the art or, after experience with the system, be assigned generally applicable descriptors.

For the application to organometallics, columns 44-52 are primarily used to specifically identify about 60 metals, their frequencies in a given compound, and the number and similarity of their connections. Tests can be run to determine the effectiveness of the system with and without this type of selective assistance and the sheet may be modified accordingly. The remaining parts of the code format, columns 1-43, and 55-70 are intended to be identical for all arts.

In addition to assigning about 10% of the code sheet to art-specific descriptors an attempt was made to reduce the effect compositing has on the incidence of false-drops. It was noted above in the description of the organophosphorus approach that work was done on the dividing of each coded document into two separate subject matter areas. In the application of the general code sheet to its initial art an extension of this thinking was incorporated into the coding procedure. Each patent was trisected into claims, examples, and disclosure, and one code sheet was used for each of the areas. This provides data on the effect of this type of subdivision on the number of false-drops received when searching with the system.

Although the system has had only a limited application to a single art, a few indications of useful modifications of the format have been obtained. It is apparent that the upper section contains a number of relationship descriptors that are impossible according to the fragmenting definitions. For example a fragment connection for a N,C,O group to a "carbonyl" group cannot exist as the whole is merged into one large N,C,O fragment. These useless locations may be ferreted out and put to better use.

It is also apparent that there is a need for expansion of the "Miscellaneous N,C,S,O dictionary." This section received a heavy concentration of coding and should be split into two or three separate dictionaries, possibly using the codes gleaned from the relationship section just mentioned.

Initial Tests and Evaluation.—To date, some testing data are available about the use of the general code sheet for searching in the organometallic art. The first group analyzed were 459 patents classified into class 260, subclass 439, of the Official U. S. Patent Office Classification (iron, nickel, and cobalt organometal compounds). From these 459 patents, using a table of random numbers, a random sample of 50 patents was selected. Since the claims of the patents were once real search questions, a representaative claim was selected from each of the 50 patents. Questions were formed from the claims and a mechanized search was carried out for each question.

A three-part analysis procedure was used in the coding operation, so answers were received from four search decks: claims, examples, disclosures, and a totally composited file. The latter deck was produced by combining all of the codes for the examples, claims, and disclosure sheets for each patent into one card.

If the document generating the question was retrieved in the search it was considered correctly coded. If it was not retrieved it was counted as an error. In the first test the questions were asked by inexperienced searches and at a later date the same selected claims were searched by an experienced machine-system patent searcher.

The inexperienced searchers retrieved the patent in 92% of the searches using the composite deck, while the experienced searcher retrieved the document in 96% of the searches. The retrieval results for the inexperienced searchers are given in Table II.

Table II. Accuracy of Retrieval Based on 50 Patents Searched
(Inexperienced Searchers)

	Documents retrieved		Encoding accuracy confidence interval at
Section of patent	Number	Percentage	$\alpha = 0.05$, %
Composite	46	92	81-98
Disclosure	44	88	76-95
Examples	41	82	69-91
Claims	39	78	64-88

The confidence intervals expressed in the tables is obtained from the "Tables of the Cumulative Binomial Probability Distribution" prepared by the Staff of the Computation Laboratory, Harvard University.[4] The confidence intervals are found at the 0.05 level of significance. That is, if the experiment were conducted repeatedly, 95% of the confidence intervals would contain the true value for the entire file. The table from which the confidence intervals were taken assumes a finite population correction factor of 1.

In the questions by the inexperienced searchers a total of 390 descriptors were queried. Table III shows the number of descriptors retrieved and the descriptor accuracy confidence intervals for each deck.

Table III. Accuracy of Encoding Based on all 390 Descriptors Used
(Inexperienced Searchers)

	Retrieved descriptors		Descriptor accuracy confidence Interval at
Section of patent	Number	Percentage	$\alpha = 0.05$, %
Composite	379	97	95-98
Disclosure	377	97	95-98
Examples	368	94	92-96
Claims	363	93	90-95

In a similar study with an experienced mechanized patent searcher a total of 246 descriptors were queried for the 50 questions. The composite deck retrieved 239 of these for an accuracy of 97%. The confidence interval at $\alpha = 0.05$ for descriptor accuracy is 94-99%.

The inexperienced searchers were directed to ask all of the pertinent matrix and metal codes plus an additional five fragment descriptors in each search. Table IV shows the document retrieval accuracy for searching only the matrix and metal sections of the code sheet. Table V presents the same data for the use of the fragment descriptor section of the code format.

Table IV. Accuracy of Retrieval Using the Matrix and Metal Descriptors Based on 50 Patents Searched
(Inexperienced Searchers)

	Documents retrieved		Encoding accuracy confidence interval at
Section of patent	Number	Percentage	at $\alpha = 0.05$, %
Composite	47	94	83-99
Disclosure	46	92	81-98
Examples	44	88	76-95
Claims	44	88	76-95

Table V. Accuracy of Retrieval Using Fragment Descriptors Based on 50 Patents Searched
(Inexperienced Searchers)

	Documents retrieved		Encoding accuracy confidence interval at
Section of Patent	Number	Percentage	$\alpha = 0.05$, %
Composite	49	98	89-100
Disclosure	48	96	86-100
Examples	47	94	83-99
Claims	45	90	78-97

Tables VI and VII show the data from Table III broken down into the matrix and metal codes (Table VI) and the fragment descriptor codes (Table VII). Note that the percentage accuracy in Tables VI and VII show little variance indicating equal reliability for the relationship and fragment descriptor sections.

Table VI. Accuracy of Encoding Based on 183 Matrix and Metal Descriptors
(Inexperienced Searchers)

	Retrieved Descriptors		Descriptor accuracy confidence interval at
Section of patent	Number	Percentage	$\alpha = 0.05$, %
Composite	178	97	94-99
Disclosure	178	97	94-99
Examples	172	94	90-97
Claims	172	94	90-97

Table VII. Accuracy of Encoding Based on 207 Fragment Descriptors
(Inexperienced Searchers)

	Descriptors retrieved		Descriptor accuracy confidence interval at
Section of patent	Number	Percentage	$\alpha = 0.05$, %
Composite	201	97	94-99
Disclosure	199	96	93-98
Examples	196	95	92-98
Claims	191	92	88-96

The analysis procedure involved the use of two analysts for each patent, the second analyst checking the work of the first. It was noticed during the tests that on occasion the checker had deleted the codes that produced the retrieval error. A thorough check was made of the nonretrieved documents that should have answered the question and the data are given in Table VIII. If similar results are received from a test of the entire field, future checkers will be instructed not to delete codes that they think are incorrect to ensure the reliability of the file.

Table VIII. Errors Due to Checkers' Deleting Correct Codes Based on 50 Patents Searched

	Patents not retrieved due to checking errors	
Section of patent	Number	Percentage
Composite	1	2
Disclosure	2	4
Examples	3	6
Claims	2	4

It is not certain how much of the data reflects the true condition of the entire file. The sample was entirely drawn from a single area of the total file: iron, cobalt, and nickel organo compounds. These patents were the first to be analyzed and most of the coders were not experienced with the system, so it is likely that a comprehensive test of the entire 3500 patent file will give higher reliability percentages. There are also a few machine "clean-up" operations that may be performed on the file to correct some of the errors.

BIBLIOGRAPHY

(1) J. Frome, "Revised Steroid Search System Coding Manual," U. S. Patent Office Research and Development Report No. 19. U. S. Department of Commerce, Washington, D. C.

(2) J. Frome, P. T. O'Day, F. S. Sikora, and M. S. Gannon, "Manual For A Punch Card Retrieval System for Organic Phosphorus Compounds," U. S. Patent Office Research and Development Report No. 22, U. S. Department of Commerce, Washington 25, D. C.

(3) Patterson, Capell, and Walker, "The Ring Index," 2nd Ed., American Chemical Society, Washington 6, D. C., 1960.

(4) "Tables of the Cumulative Binomial Probability Distribution," Staff of the Computational Laboratory, Harvard University Press, Cambridge, Massachusetts, 1955, Table I.

(5) J. Frome and P. T. O'Day, *J. Chem. Doc.*, **2**, 249 (1962).

(6) J. Frome, J. F. Caponio, P. H. Klingbiel, and P. T. O'Day, Abstracts, 143rd National Meeting of the American Chemical Society, Cincinnati, Ohio, January, 1963.

(7) J. Frome, P. T. O'Day, and E. Lewis, Abstracts, 143rd National Meeting of the American Chemical Society, Cincinnati, Ohio, January, 1963.