

FIG. 9

REASONS FOR SEARCH REQUESTS

COMPOUNDS RELATED TO AN ACTIVE COMPOUND (COMMERCIAL).  
 MODEL COMPOUNDS FOR IR AND UV COMPARISONS.  
 MODEL COMPOUNDS FOR NMR COMPARISONS.  
 CHECK NEW ANALYTIC PROCEDURE.  
 FOLLOW-UP TESTING LEAD.  
 LOCATE SYNTHETIC INTERMEDIATES.  
 PRE-PROJECT SURVEY.

FIG. 11

1960 GENERIC MACHINE SEARCHES

	41 SEARCHES		
	RANGE	AVERAGE	MEDIAN
DECK SIZE	5,000-32,000	15,000	14,000
NO. OF SORTS	1-9	4	4
SEARCH TIME	0.2 min/1000- 2.8 min/1000	1.38 min/1000	1.60
DROP	0 - 484	91	32
FALSE DROP	0 - 100%	13%	3%

FIG. 10

The need for an improved and more rapid search has been graphically demonstrated by the increased number of search requests which have been received. These requests are now being made at the rate of about 200 per year. Prior to the installation of the system, an average of 15-20 searches per year was carried out at Pearl River.

Another area in which we hope to make extended use of the system is in the study of structure-activity correlations. This activity has not yet begun, but it is hoped that preliminary studies can begin in the near future.

## Generic Mechanized Search System\*

By JULIUS FROME

Office of Research and Development, Patent Office, U. S. Department of Commerce, Washington, D. C.

Received August 24, 1961

During the last three and one half years, in answering actual questions requested, the U. S. Patent Office has made well over ten thousand mechanized searches. This includes searches on the steroids (2, 3), resins (4, 9, 10), and phosphorus compounds (9). Mechanized searches were made on various machines such as computers, *i.e.*, SEAC, Bendix G-15, RAMAC 305, and punched card machines, *i.e.*, single column sorters, ILAS, and IBM-101. Although most of these mechanized searches were for patent examiners, a great many were conducted on an experimental basis for research workers in industry.

From an analysis of the questions submitted, it was quite apparent that a majority of the searches desired were generic rather than specific. In further studying mechanization from the viewpoint of the questioner, the following facts became apparent: (1) generic as well as specific searches are a necessity; (2) system should be compatible

to various machines; (3) system should be segmentable in accordance with need. The user must be able to make generic as well as specific searches. The system must be compatible with several machines since some users have access to punched card machines and some to computers. Finally, for some users a great depth of information and detail is necessary, whereas others are satisfied with less. Therefore, from the questioner's viewpoint, the system should be able to be segmented so as to give him as much or as little information as he desires.


From the viewpoint of an information scientist, a system also should have these capabilities: (4) should not have limited subject application; (5) should be capable of storing and retrieving compounds, processes, compositions, biological data; (6) should have an open-end dictionary; (7) Information extracted should be useful in systems organized for: (a) random access searching, (b) serial searching. A system has been developed which accomplishes many of the above objectives within practical limits.

\*Presented before Division of Chemical Literature, American Chemical Society, St. Louis, Missouri, March, 1961.

The preparation of an intermediate punched card suitable for use either as input for a punched card system using the serial approach or as input to a computer using the random access method of search will be described.

While the system is not limited to any machine, this paper will describe the new system as applied to the IBM 101 with a Row-by-Row attachment (1).

*Features of New System.* The proposed system has several features of the VS<sub>3</sub> (5) which used the ILAS (6) to search. The following are important aspects of the new system: (1) use of semiautomatic techniques in preparing

file; (2) use of building blocks (NO<sub>2</sub>, COOH, );

(3) open-end dictionary; (4) no limit to the number of punched cards per document; (5) searches generically as well as specifically; (6) relationship by signals and interfixes; (7) can easily be segmented; (8) easily convertible to other machines.

*Format.* The basic word unit of the system is an 80 column horizontal row of a standard punched card. Figure 1 shows the format.

Signals	Modulant	Ring System Position	Ring Conn	Chain Conn	Subject Matter		Interfix
Col. 1-7	Col. 8-13	Col. 14-18	19	20	Col. 21-40	41-68	69-80

Fig. 1.—Card format.

*Signals.* The placing of codes in groups is accomplished by signals. Thus, for example, all the codes pertaining to a compound can be correlated by signals. All the compounds in a document or in a given mixture within a document can be indicated by means of signals. The properties of a compound or its biological effects can also be correlated by means of signals. There is no limit to the number of codes that can be included in a signal and therefore no limit to the amount of information that can be recorded for a document. The present system makes provision for seven signals, but more or less signals may be used in accordance with the need. To date we have used these signals in the system: Signal 5, end of compound; Signal 6, end of mixture; Signal 7, end of document.

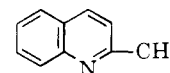
*Modulants.* The modulant is a device for using the same codes to indicate different things. Thus, for example, modulant 2 stands for the generic term alkyl and indicates that all the codes in the same row refer to alkyl codes. Modulant 10 is alkenyl and indicates that the codes which follow refer to alkenyl. In other words the modulant determines the meaning of the codes which follow it. Provisions have been made for at least a thousand different modulants (four additional columns would be required to utilize this many modulants). However, this number can be enlarged greatly if needed. To date we have used only about forty modulants. These modulants are usually generic, e.g., modulant 2 is the generic term alkyl. The species or specifics are indicated by the codes further associated with the modulant.

16	8	4	2	1
----	---	---	---	---

Cols. 14-18

Fig. 2.

*Ring System Position.* A ring system is defined as a single ring or a collection of fused rings, the fused rings taking precedence over the single ring. The ring system position refers to the position to which the substituent identified by the modulant and codes is attached. Thus, for example, in the formula



the CH<sub>3</sub> group is attached to the 2 position of the quinoline ring and is so indicated in the row which describes the methyl group. The Patterson, Capell & Walker Ring Index (8) is used for the numbering of the ring system position. The five digits, (14-18, Fig. 1), allow for 32 different positions as shown in Fig. 2. More positions can be allocated as needed.

*Ring Connected.* A punch in this position (Col. 19) indicates that the substituent in question is attached to a ring.

*Chain Connected.* A punch in this position (Col. 20) indicates that the substituent is chain connected.

*Subject Matter Codes.* Columns 21-40 further define the substituents or rings when taken in conjunction with the modulant. Details of various modulants and codes are set out in the appendix.

*Columns 41-48.* These are surplus columns and may be used for further modulants or any other codes desired.

*Interfix.* This feature has been described already (5). Briefly it shows the connective relationship between the various fragments of a compound. The interfix will be punched in columns 69-80. A modification of this information can be so manipulated that it may be searched on an IBM 101, with Row-by-Row attachment. This is beyond the scope of the paper, but is available upon request.

*Columns 49-69.* In using the 101 Row-by-Row, these columns are left free. In the formation of intermediate cards, which will be described next, information for a system using the RAMAC computer is put in these columns.

*Intermediate Card.* One of the important features of the system is the use of an intermediate punched card. This intermediate card makes the system usable or compatible with several machines, such as a random access computer, an IBM 101 with Row-by-Row attachment, or an IBM 1401.

The information to be placed on the intermediate card is obtained by semiautomatic means as described in (9).

The intermediate punched card is then generated from the information fed into a RAMAC 305 which automatically produces the card by a suitable program (9). Of course, the intermediate card may also be prepared by ordinary hand punching but this is a slow and tedious process.

Each intermediate card contains an 80 column row of punches allocated as shown in Fig. 3 and the table; The card contains

Signals	Cols. 1-7	
Modulant	Cols. 8-13	
Ring System No.	Cols. 14-18	
Ring Connected	Col. 19	
Chain Connected	Col. 20	
Subject Matter	Cols. 21-40	
Unused Space	Cols. 41-48	
Document No.	Cols. 49-55	) RAMAC
Sequence in Document	Cols. 56-58	) RAMAC
Accession No.	Cols. 59-62	) RAMAC
Compound No.	Col. 63	) RAMAC
RAMAC address	Cols. 64-68	) RAMAC
Interfix	Cols. 69-80	

Row 9	Signals	Modulant	R.S.	R.C.	C.C.	Subject Matter	Pat. Seq. No.	Pat. No.	Acc. No.	Cpd. No.	RAMAC Address	Inter- fix
	Col. 1-7	8-13	14-18	19	20	21-40					41-68	69-80

Fig. 3.—Format of the intermediate card.

For use in a RAMAC random access system the intermediate cards (Cols. 49-68 of Fig. 1) would be fed directly into the RAMAC and would be stored at the proper address. These intermediate cards also can be used to feed a 1401 computer, the information in a single row being stored between two word marks. The description of such systems is beyond the scope of this paper.

For use in a 101 with Row-by-Row attachment, the first 12 intermediate cards are gang punched into a single card and the other intermediate cards are punched into another card. Thus if a document contained 25 intermediate cards, these cards would be gang punched into 3 punched cards. The system will now be described for a 101 with Row-by-Row attachment.

*Principles of Coding.* (1) A compound is divided into its various building blocks, *i.e.*, benzene ring, COOH, NO<sub>2</sub>, CH<sub>3</sub>, etc.

(2) The relationship of the various building blocks is indicated: (a) positions on ring systems; (b) ring or chain connected; (c) relative positions, ortho, meta, para; (d) how building blocks are connected (interfix).

(3) Its utility or properties are indicated.

(4) Compounds in the same mixture or process are indicated by the assignment of an appropriate Signal.

Each building block is shown by a separate horizontal row in the punched card. If there are more than twelve building blocks in a compound, the next card is used. As many cards as are needed for a compound are used. The several cards may be considered the same as a continuous 80 channel tape.

Each building block is generically identified by its modulant. The other codes in the remainder of the row

more specifically identify the building block, *e.g.*, for CH<sub>3</sub> the modulant is "alkyl," and the other codes in the row identify this building block specifically as methyl.

The relationship of the building block, whether chain or ring connected, its interfix and whether it is ortho, meta or para, is also shown by the codes in the same rows as previously indicated.

The use or other properties of the compound also may be indicated by the use of modulants and codes. A modulant is assigned a generic use, such as bactericidal, and the codes within that modulant are used to indicate the specific bacteria killed.

Two or more compounds may be indicated as in a mixture by means of signals.

*Example.* The compound to be coded is shown in Fig. 4. It is stated to be non-toxic to man. It is used with acetone as a solvent.

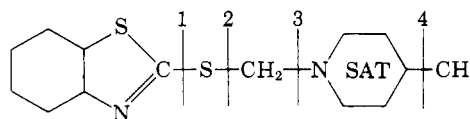


Fig. 4.

The compound is separated into building blocks and coded (Table I).

The system is not art oriented as any compound may be coded that has a definite structural formula.

The following are sample questions that can be answered by this system: (1) A thiazole compound non-toxic to man; (2) A compound containing a piperidine ring with an added methyl group; (3) a benzothiazole substituted on the 2 position with an -S- group which is further attached to a hydrocarbon chain; (4) A 2-methylene-4-methyl-piperidine compound that is non-toxic to man; (5) a piperidine compound, not toxic to man, used with acetone as a solvent.



## SUMMARY

A system has been described which has these features: (1) potential for generic, specific, or combination searching; (2) unlimited dictionary; (3) not art oriented; (4) searches compound, process, mixtures and biological data; (5) compatibility to various machines; (6) can be segmented to user's needs.

## APPENDIX

This appendix is a sample of some of the codes used. It is not intended to be exhaustive but merely exemplary. The title is intended to be the generic modulant, and the subject matter codes are under the modulant. The number beside the title is the modulant number. The numbers next to the specific codes are the column numbers.

TABLE I

Building blocks	Generic modulant	Relationships			Subject matter	Interfix			
		Ring system	Ring conn.	Ch. conn.					
Benzene	Aryl				 fused face				
Thiazole	S-Hetero				Thiazole, unsatd., fused	X			
Ring system word	R. system word				2R, 1N, 1S, 1 Benzene				
—S—	Ether	2	X	X	—S—	X	X		
—CH <sub>2</sub> —	Alkylene	1	X	X	—CH <sub>2</sub> , —St, Low, Para		X	X	
	N-Het 6M				Piperidine, IND, SAT, IN			X	X
CH <sub>3</sub>	Alkyl	4	X		Methyl, St, Low, Para				X
Non-toxic to man	Property				Non-toxic to man				
Signal	End of compound								
CH <sub>3</sub>	Alkyl			X	Methyl, Low, St.	X			
C=O	Carbonyl			X	C=O	X	X		
CH <sub>3</sub>	Alkyl			X	Meth, Low, St		X		
Solvent	Property				Solvent				
Signal	End of compound								
Signal	End of mixture								
Signal	End of document								

Card Position Col.	Generic modulant subjects			
	(1) Alkyl (A)	(7) COOR (H)	(13) O-Hetero (L)	(29) Amine (T)
21	Lo Alk (1-7)	COOH	1-0	NH <sub>2</sub>
22	1 C	COOR	2+—0	NHR
23	2 C	COHAL	N-cont.	NR <sub>2</sub>
24	3 C	COF	other het	
25	4 C	COCI	3M	
26	5 C	COBr	4M	quat.
27	6 C	COI	5M	diazo
28	7 C		furan	azide
29	Hi Alk (8+)		4H-furan	nitride
30	8 C		oxazole	= NH
31	9 C		6M	= NR
32	10 C		pyran	N <sup>5</sup>
33	11, 12 C		morph	
34	13, 14 C		KETAL	
35	15 + C		MISC	
6	Strt.		SAT	
37	Br.		UNSAT	MISC
38	0	O	IND	O
39	M	M	SPIRO	M
40	P	P	FUSED	P

Complete modulant and subject matter code lists are available from the author upon request.

## REFERENCES

- (1) H.P. Luhn, *Row-by-Row Scanning Systems*, I.B.M. Research Center, Yorktown Heights, N. Y., May 8, 1959.
- (2) J. Frome and J. Leibowitz, *A Punched Card System for Searching Steroid Compounds*, Patent Office R. & D. Report No. 7, Washington 25, D. C., 1957 (superseded by report No. 11).
- (3) J. Frome and J. Leibowitz, *A Manual for Coding Steroids*, Patent Office R. & D. Report No. 11, Washington 25, D. C., 1958.
- (4) J. Frome, J. Leibowitz, and D.D. Andrews (O. R. & D.) and Joseph D. Grandine, Steven T. Polyak, and Karl G. Siedschlag, Jr. (du Pont), *A System of Retrieval-Compounds, Compositions, Processes, and Polymers*, Patent Office R. & D. Report No. 13, Washington 25, D. C., 1958.
- (5) J. Leibowitz, J. Frome, and D.D. Andrews, "Variable Scope Search System: VSs." Preprints of papers for the "Proceedings of the International Conference on Scientific Information," Washington, D. C., National Academy of Sciences-National Research Council, 1958, Area V, pp. 291-316.
- (6) Don D. Andrews, *Interrelated Logic Accumulating Scanner (ILAS)*, Patent Office R. & D. Report No. 6, Washington 25, D. C., 1957.
- (7) Julius Frome, H.R. Koller, Jacob Leibowitz, H. Pfeffer, and Don D. Andrews, *Recent Advances in Patent Office Searching: Steroid Compounds and ILAS*, 1957; Reprinted as chapter 25 in *Advances in Documentation and Library Science. Volume II. Information Systems in Documentation*, Interscience Pub., Inc., New York, N. Y., 1957, pp. 447-77.
- (8) Patterson, Capell and Walker, *Ring Index*, 1960 (ACS).
- (9) Julius Frome, *Semi-Automatic Indexing and Encoding*, Patent Office R. & D. Report No. 17., Washington 25, D. C. 1959.
- (10) Jacob Leibowitz, Julius Frome, and Don D. Andrews, *Variable Scope Patent Searching by an Inverted File Technique*, Patent Office R. & D. Report No. 14., Washington 25, D. C., 1958.