computer program, but the number of items on tape is constantly being increased. As the tape file grows, literature searching becomes more valuable.

2. A sufficient number of components of our data is in machinable form to allow for great future flexibility in our output. The rules that form the input typing requirements are detailed so that other information can be extracted at a later date if desired. The following have been considered: (a) Additional statistics as the number of times descriptors have been used, or the average number of descriptors per document. (b) Document usage statistics for "weeding" purposes. (c) Lists of documents by security classification for classified document inventories. (d) Lists by date of document for declassification purposes. (e) The possible replacement of index card files by book indexes. (f) The addition of generic relationships to the "Dictionary of Terms," which would be handled automatically by the computer.

3. The maintenance of the normal library card files and the manual uniterm search methods in conjunction with the computer program is advantageous in that multilevel access to material is available. We feel that the computer prepared output, dissemination of information automatically, and literature searching techniques will supplement, and will be supplemented by the library card indexes.

4. The fine working relationship with Computing Engineering has been a major advantage to the program. The cooperation and interest in our common problems have resulted in a system we feel is unique—one that is both user, Library *and* machine oriented.

At this point, what is operational is proceeding well and is being received with satisfaction. We feel that we have the flexibility to expand and grow. We are looking forward to the development of the third and most important phase of our program, the ability to handle literature searches mechanically.

REFERENCES

(1) "The Uniterm System of Indexing, Operating Manual," Documentation Incorporated, Washington, D. C., 1955.

(2) The words: "subjects," "uniterms," "terms," "descriptors" and "unit concepts" are used interchangeably in this report.

(3) Bunnow, L. R., "Study of and a Proposal for a Mechanized Information Retrieval System," Report No. SM-37418, Douglas Aircraft Co., Inc., Santa Monica, Calif., May, 1960.

(4) Koriagin, G. W. and Bunnow, L. R., "Mechanized Information Retrieval System for Douglas Aircraft Company, Inc., Status Report," No. SM-39167, Douglas Aircraft Co., Inc., Santa Monica, Calif., Jan., 1962.

(5) Unless otherwise stated, all information refers to the Missile and Space System Engineering Library.

(6) Luhn, H. P., "Selective Dissemination of New Scientific Information with the Aid of Electronic Processing Equipment," November 30, 1959, International Business Machines, Advanced Systems Development Division, Yorktown Heights, New York.

PACIR: Practical Approach to Chemical Information Retrieval*

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

Received May 18, 1962

I. INTRODUCTION

This paper describes a "Practical Approach to Chemical Information Retrieval" (PACIR). The object of the PACIR system is to provide a flexible universal approach to specific compound retrieval that is not limited to the unique charateristics of a particular area of chemistry, that is adaptable to various types of hardware, that employs a maximum of machine assistance in the analysis, that provides a file that may be organized according to the need of the user, and that is practical from the economic viewpoint.

It will be apparent that emphasis is placed on problems of efficiency and accuracy in analysis. To a great extent, it is in this area that the economic feasibility of a given approach is determined. Complications in analysis, imposed by complexities of translation of the analyzed subject matter to useful machine recognizable symbology, can easily add confusion and inaccuracy to the analysis step. This can well render a system that is theoretically sound an empirical failure. The PACIR system attempts to keep these rigidities and dangers to a minimum.

The keywords at all steps in creating the system have been flexibility, clarity, and simplicity.

II. EVOLUTION OF THE SYSTEM

Over the past five years, the Office of Research and Development of the U. S. Patent Office has experimented with and put into operation a number of chemical retrieval systems[1-8] employing a variety of machines and approaches. A description of the contributions of these projects is necessary for full understanding of the reasons behind many of the features of the PACIR approach.

Early efforts were made to solve the problems of searching the rapidly growing and important area of steroid chemistry. This led to the use of a composite one-punch-card-per-document approach which has been revised and updated and is now in general use, both in

*Presented before the Division of Chemical Literature, ACS, National Meeting, Washington, D. C., March 23, 1962.

the Patent Office and in industry.[1-3] Its retrieval logic is based on the accepted numbering system for the complex steroid nucleus, so the system is limited to only that narrow area of chemistry. The increase in false drops by the use of compositing further limits the size of the file. This limit is not a serious problem in the steroid system, but it probably would be prohibitive in a more general area of chemistry.

Apparent success and general acceptance of this effort led to further study in the area of organo phosphorus chemistry that culminated in a punch card mechanization system (Project CAMP) for these compounds.[6,7,9] Here two punch cards are used to record the subject matter of each document and the logic for retrieval of relationships is based on the distance in chemical fragment units of a given fragment from the central phosphorus nucleus. These relationships are coded in a matrix format with the various types of fragments as coordinates. Although effective in retrieving organophosphorus compounds, and probably applicable to any art that involves a unique atom in every compound, this system is similar to that for steroids in that it lacks broad applicability. Both the steroid and phosphorus systems indicate that compositing can be used to keep analysis costs within reasonable economic bounds. This procedure also avoids the errors inherent in systems employing additional or more complicated steps.

Since these approaches were art-limited and employed a limited punch card memory, simultaneous efforts were undertaken to devise systems and principles more universal in scope. Previous projects in the Office of Research and Development had demonstrated the possibility of using an expanded memory for each document[10,11] and the use of the interfix to show relationships.[11,4] These concepts were incorporated in the first resin system[4] along with the use of roles to store relationships. This project was severely hampered by a poorly organized analysis operation and by a complicated wiring procedure for the search hardware, the ILAS.[12] Eventual loss of user confidence in the accuracy and reliability of the file combined with the complicated search procedure resulted in the system being withdrawn from operation. Although the first resin system had little practical use as a search tool, it indicated the effectiveness of the logic employed. The interfix and role technique were found to be powerful retrieval aids in a large, deep index. (This system should not be confused with current efforts of the Office of Research and Development in the resin art.)

Efforts to devise a useful random access system were brought to fruition in the organophosphorus project, RAMP.[6] This project avoids false drops by coding each compound as a separate unit. The computer input consisted of a linear formula for each fragment of the individual structures. A comparison of this system with its punch-card-memory associate, CAMP, is presented in another paper.[13] RAMP also used an underlining analysis approach that had been developed earlier[14] to produce lists of the specifically named compounds in the document,. This project is the direct forerunner of the PACIR approach.

All of the projects summarized above pointed out the need for flexibility in the initial analysis stage. It is here that unexpected problems can be solved and efficient and useful revisions made without deleterious effect. As a new difficulty arises the crystallizing system should incorporate the best solution. Lack of flexibility in the rules at this point would hamper its refinement and subsequent utility as an effective search tool.

A dominant lesson from these experiences is the need for careful organization and strict control of the analysis operation. Duplicate coding of each document by two different analysts, or some similar method of checking the analysis work, was found to be necessary in all cases to provide a reliable file. Motivation, supervision, and guidance of the staff of analysts were seen to be critical factors in formalizing a system since the entire group must think with "one mind" in applying and developing the written rules. All systems indicated the desirability of uncomplicated, lucid analysis procedures to ease the tedium and complexity of analysis. The analyst has a difficult enough task understanding the documents, which range over several decades of chemistry, without being burdened by requirements to interpret them in the light of complex system rules. System rules that cooperate with an efficient analysis organization are seen as a necessity. These points were considered carefully during the creation and initiation of the PACIR analysis and retrieval procedures.

Of course, a necessary concomitant to the success of any system is the availability and use of competent analysts. This type of person is valuable and is every bit as important as any of the mechanization techniques or hardware used. The successful system in theory and the successful system in application are worlds apart and the analyst is an indispensable connecting link. The experiences noted above have shown that this problem may be under-emphasized to the detriment of the system at hand.

PACIR, to a great extent, is an amalgam of the advantages of these prior experiences. It is basically an interfix approach. It employs roles for identification of ancillary subject matter. A limited compositing is employed by grouping similar compounds into a single composited generic structure (the "Markush" compound of Patent Office terminology) whenever possible to prevent unnecessary repetitious coding. It provides for either random access or serial searching. It uses a matrix of fragment relationships. It uses semi-automatic procedures for encoding. Finally, it incorporates major consideration of the practical problems of analysis, as well as problems of retrieval, in its system rules.

III. CHARACTERISTICS OF THE SYSTEM

The PACIR approach has the following characteristics: (1) non-art limited; (2) capable of storing and retrieving compounds, processes, compositions, biological data, etc.; (3) both specific and generic search potential; (4) compatibility with several machines, including both random access and serial computers; (5) segmentable in accordance with need; (6) unlimited dictionary (open ended); (7) machine assisted encoding; (8) economic practicality.

The PACIR approach incorporates two different types of analysis that allow for selective depth of coding for a given type of subject matter in the chemical art being analyzed. One of these approaches is an underlining and role-assigning technique that produces catalogs of the ancillary subject matter of the art. The other procedure provides for specific coding in depth of any selected type of compound that can be represented by the conventional chemical structural formula. Both of these procedures are universal in scope as they are not limited by the need for a unique nucleus (as in the steroid and CAMP phosphorus projects mentioned earlier). The versatility of the latter procedure is limited only by the need for definable chemical structures. The underlining and role-assigning technique has wide applicability to structures, utility, properties, and any desired topic that is described by particular words within the document. In the application of PACIR to any particular chemical art, the analysis procedures are allocated to the various types of chemical subject matter within the art depending on considerations of cost and search need.

IV. PACIR and the PESTICIDE ART

The PACIR approach is being applied to the organic pesticide art, a group of about 5000 patents from class 167, subclasses 22, 30, and 33.[15] Patents are classified to the pesticide area without regard to structure type, so the area contains compounds that are representative of most of the areas of organic chemistry. This group of patents should provide a comprehensive test for the effectiveness of the PACIR approach in the analysis and retrieval of the entire scope of organic chemical structures.

Pesticide patents generally teach the use of mixtures of chemicals to kill selected organisms. One or more of the components of the mixture is an "active ingredient" which imparts toxicity to the mixture. These active ingredients are the most important parts of the documents and their specific structures are extracted and coded.

The remaining components of the mixture play various roles. Some provide the medium and are termed "solvents" if liquid or "carriers" if solid. Other chemicals serve as wetting agents, emulsifiers, etc. These non-active ingredient components of the pesticidal mixture, along with certain teachings of utility, are extracted and classified by the underlining and role assigning technique.

This allocation of the PACIR analysis procedures will afford highly specific retrieval of the toxic agents of the pesticidal mixtures as well as retrieval of the remaining components. The following description of the PACIR procedures will employ examples from the application of the system to the pesticide art.

V. UNDERLINING AND ROLE ASSIGNMENT

This procedure consists of the selection of specific teachings in the document that conform to predetermined roles, underlining them, and assigning the applicable role to the underlined term. The underlined terms with their roles are key punched onto the standard IBM punch card. These cards then are used to prepare catalogs and dictionaries of terms. The procedure is relatively inexpensive and is an effective technique for subordinate subject matter that is not of sufficient value to the examiner in this art to warrant the more costly structure analysis described in section VI.

Since the pesticide search is normally directed primarily toward the structure of the toxic component, the underlining and role assignment method is applied to the nontoxic components of the pesticidal mixture. Four roles are used: (S) SOLVENT.—This is the pesticidal mixture component that provides the medium or acts as a diluent. Solid "carriers" as well as liquid solvents are assigned the role "S." (A) ADJUVANTS.—All compounds in a pesticidal mixture that are not active ingredients or solvents (carriers) are considred adjuvants and given the role-letter "A." (F) FORM.—Terms that denote the form of the active mixture, *e.g.*, paste, spray, dust, are assigned the letter "F." (U) USE (EFFECT, LOCUS).—Information which is descriptive of the utility of the pesticidal mixture is given the letter "U." Specifically, this includes the *effect* of the pesticidal mixture, its *locus of application*, and the *type of pest* affected. Latin species names are excluded.

The following example shows the application of this procedure to a typical pesticide teaching:

> The compound of my invention, N-bromo-succinimide, may be admixed with petroleum oil and a minor amount of N-butyl-oleate and sprayed on rosebushes to prevent blue mold disease.

This procedure allows the machine preparation of: (a) an alphabetical list of the ingredients in each patent; (b) an inverted file list in which the ingredients are classified by role; (c) an alphabetical listing of the ingredients which contains, listed under each material, the document numbers in which the term appears. These lists may be used either manually or in machine retrieval.

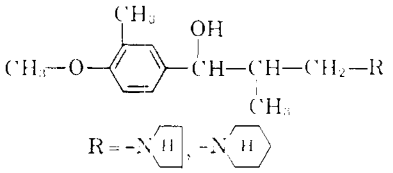
Once the analyzed subject matter has been transformed to punch card form any number of mechanization techniques may be applied to form an organized retrieval file. In the U. S. Patent Office pesticide application of PACIR, a random access computer is used as hardware. The extracted terms with their roles are stored intact in the computer. Generic search is provided by a classification of each of the terms in the dictionary according to the descriptors. These descriptors are independent of the analysis and may be replaced in accordance with the user's requirements. Listings of the underlined subject matter descriptors and the structural formula fragmentation descriptors used in the Patent Office application of PACIR to the pesticide art are available from the authors upon request.

All pertinent descriptors are applied. Maximum machine use is incorporated here for efficiency and accuracy. The original dictionary is assigned its descriptors manually. As the file is expanded the machine is directed to look up the input term in the existing file and assign the applicable codes if the term is already recorded. If the term is not recorded, the machine asks for the codes which are then assigned manually. In this way, each term is coded but once, the computer being used to incorporate subsequententries of the same term by automatic encoding.

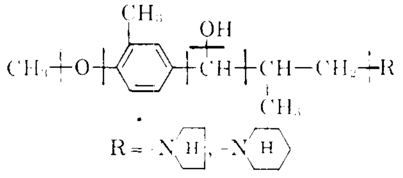
VI. STRUCTURAL ANALYSIS

The following procedure is the heart of the PACIR approach. It is an attempt to provide a practical, effective method of analyzing, storing, and retrieving specific or generic chemical structures with a minimum of false drops. It will be apparent that a key factor in this method is the organization and analysis of the subject matter of the patent. The procedure attempts to omit all unnecessary analysis and to provide reasonable rules and guidelines for the analyst. The method has three main steps: drawing the formulas, converting them into linear formulas, and translating the linear formulas into numerical codes with maximum machine assistance.

1. Drawing the Formulas.—The object of this step is to draw a structure, or a series of structures, that contain all of the relevant compounds taught by the document. In the pesticide analysis, this includes all of the "active ingredients" of the pesticidal mixtures. After studying the document to gain full comprehension of its disclosure, the analyst attempts to group similar structures together into one generic compound ("Markush" formulas). In grouping structures in generic form, the analyst has a wide degree of freedom. A few limitations necessary to the linear encoding procedure are imposed on the analyst, but these are rarely the source of trouble. A typical example of a structure extracted from a patent is



The structure is then divided into its component fragments.

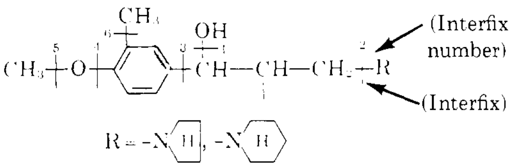


The fragments produced are those that would generally be considered a unit by organic chemists. Groups falling within the following definitions are treated as fragments: (a) rings, (b) hydrocarbon chains, (c) any combination of nonring-members as N, S, O, carbonyl carbon, and the $-N=\overset{|}{C}-N<$ group, (d) phosphorus-X groups, where X is O or S. A list of the fragments found in the first 500 patents of the pesticide analysis is generated by machine from the analyzed patents.

The fragmentation rules attempt to record the largest possible units that make up the structure. This increases the specific retrieval power of the system and thereby reduces the number of false drops incorporated in the analysis. In order to provide a clear record of the source of any given structure, a compound number is assigned each structure and recorded near the teaching in the patent that generated the structure. This aids in checking the initial analysis and in ascertaining the reasons for questionable search answers.

This procedure leaves the analyst free to extract the pertinent structures without unnecessary formal limitations. The only requirement is that all the pertinent information be extracted. The analyst may choose the degree of genericity to be used in grouping the structures depending primarily on the nature of the subject matter and its interpretation. Thus, each patent is treated according to the depth and scope of its teaching, and unique situations may be treated with versatility and efficiency. The analyst is left free to devote his thinking to the interpretation and recording of the chemical structures of the document with little regard to a formal set of rules necessary to the machine procedures. The most difficult problems appear at this stage. It is here where the accuracy and reliability of the file is determined. The relatively uncluttered analysis requirements of PACIR in this regard are designed to cooperate with efficient, accurate, and flexible analysis.

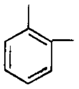
2. Linear Encoding.—This portion will describe the encoding procedure for the fragmented structural formulas. Once the proper subject matter has been extracted and depicted, there remains the task of translating this information into machine compatible linear code. The patents are received by the encoder in analyzed form with the pertinent structural formulas drawn on separate sheets of paper. The structures are interfixed, but no interfix numbers have been assigned. These numbers are entered by the encoder in any pattern, so long as a number is not used more than once.

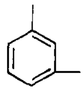


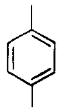
The fragments of each interfixed formula are written linearly, preceded by their interfix numbers, and followed by any of the applicable symbolic codes in the order listed below. A dash is interposed between the last interfix number and the start of the fragment, *e.g.*, in the structure above, the ring would be coded "3,4,6-BENZENE-CH-IND-H-W/."

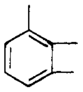
1	-CYCN	this denotes "cyclic-nitrogen" and is used when the nitrogen of a fragment, *e.g.*, carbamate, is also part of a heterocyclic group
2	-Q	nitrogen with a valence of 5
3	-R	ring connected
4	-Ring Position Number	point of attachment of a fragment to a ring according to the preferred numering systems in the Ring Index
5	-CH	chain connected
6	-S	saturated hydrocarbon
7	-U	unsaturated; carbon-carbon double bond
8	-UT	unsaturated; carbon-carbon triple bond
9	-ST	straight hydrocarbon chain
10	-BR	branched hydrocarbon chain
11	-T	terminal hydrocarbon chain
12	-NT	non-terminal hydrocarbon chain
13	-IND	independent ring
14	-SP	spiro ring system
15	-FUS	fused ring system

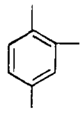
16 Benzene Codes (letters denoting configuration of substituted benzenes)

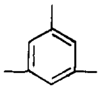
-D 

-E 

-F 

-G 

-H 

-J 

17 -O ortho
18 -M meta
19 -P para
20 Frequency Letter designates number of occurrences of the fragment in the compound:
-W one occurrence
-X two occurrences
-Y three occurrences

After the frequency letter is assigned, a slant is drawn and the next fragment is coded, and so on until all of the fragments for a given compound are coded.

The final code(s) for all compounds is S5-a/., where a is equal to the number of fragments in the structure. Where there is a variable fragment total, as may be the case in some Markush formulas, each numerical value for a is written separately, *e.g.*, the coding for a structure with alternatively 8, 9, or 10 fragments would be concluded with

S5-8/
S5-9/
S5-10/.

The period is placed only after the last code and denotes the end of the compound in the machine program.

Following this procedure, the codes for the compound above would be:

1-OH-CH-W/
1;2,3-C4-R-1-CH-S-BR-NT-M-P-W/
2-PYRROLIDINE-CH-IND-W/
2-PIPERIDINE-CH-IND-W/
3,4,6-BENZENE-CH-IND-H-W/
4,5-O-R-CH-O-P-W/
6-C-R-S-T-O-M-X/
5-C-CH-S-T-X/
S5-7/.

The above encoding is entered on punch cards using as many cards as are needed. The computer recognizes the end of a compound by the appearance of a period.

When a Markush formula, *e.g.*, the structure above, is linear encoded, the fragment to which the Markush group (R, above) is attached must be encoded prior to the coding of any of the Markush fragments. Thus, in the example above, the C4 group is coded before either of the rings represented by R. This sequence will then allow proper machine interpretation of the interfix relationships.

For consistency in the linear depiction of nonhydrocarbon double and triple bonds the following rules have been established.

1. One valence dash is used for each valence unit of the multiple bond. Thus carboxy is written linearly as C--OOH, cyano as C---N.
2. Note that valence dashes are not used to depict single bonds at any time.
3. Valence dashes must never precede the fragment, thus the imide group must be coded NH--, not --NH. Similarly a solitary keto oxygen fragment (=O) is coded O--, not --O. In practice, these groups are actually followed by three dashes, the third dash being the usual separation mark between the fragment and its codes. Thus a ring connected keto with an interfix number of 1 would appear as 1-O---R-W/

The following examples represent the encoding of typical specific and Markush structures:

EXAMPLES

1. Specific

$$Cl \overset{1}{+} CH_2—CH_2 \overset{2}{+} NO_2$$

1-Cl-CH-W/
1,2-C2-CH-S-ST-NT-W/
2-NO_2-CH-W/
S5-3/.

2. Markush (middle fragment encoded first)

$$R \overset{1}{+} CH_2—CH_2 \overset{2}{+} NO_2$$

1,2-C2-CH-S-ST-NT-W/
1-Cl-CH-W/
1-F-CH-W/
1-Br-CH-W/
1-I-CH-W/
2-NO2-CH-W/
S5-3/.

This section, in combination with section VI-1, shows the PACIR approach to the achievement of a standard analysis. At this point any number of manipulations of the linear formula are possible to produce files that are in accordance with the user's need.

3. **Machine Encoding.**—This section describes one method of translation of the linear formulas of the preceding section into a more useful machine language. The generated codes utilize the entire depth of analysis and may be used to set up either a random access or a serial file.[16] The procedure has two main parts; one stores information about the nature of the chemical fragment, the other stores information about the relationship between fragments.

In the first part the linear codes for the compounds to be included in the file are machine processed to produce a list, with duplicates eliminated, of all the different fragments involved. The codes from the user's dictionary are manually applied to each of these fragments and this is stored in the computer memory. The entire linear encoded deck is then fed into the computer and by suitable program the proper codes are assigned to each linear fragment. The codes are then used to create the actual file.

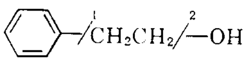
The relationships between fragments are stored by matching the interfix numbers that appear in each linear fragment. All fragments fall within one of the generic headings in the code dictionary. Next to each heading is a letter that is used to store the generic relationships within a compound. Thus in a halogen-alkyl relationship both fragments would have the same interfix numbers assigned to them. The machine recognizes this in the matching operation and stores the proper letter combination, EA, associated with the pertinent compound and patent numbers.

It should be remembered that the format here is arbitrary, to be selected by the user according to his need, and that it is independent of the PACIR analysis technique.

One of the significant advantages of the automatic encoding system, besides the obvious advantage of obtaining the codes desired quickly and accurately, is the ability to change codes at will by machine methods without going back to the original document. For example, in the linear encoding the term "piperidine" may be used. In one system all that may be desired are the codes for "nitrogen heterocyclics." At a later date, or in some other system, it may be desirable to store more about the nature of piperidine, *e.g.*, 1-Nitrogen-containing 6-membered ring, saturated, *etc.* The codes stored in the machine may then be changed to the codes desired. The original linear encoded cards are passed through the computer and a new deck is generated containing the new codes for piperidine. In this manner reversion to the original document to get the new codes is unnecessary. It is possible to change codes completely or partially by machine methods, so long as the original starting units are acceptable. An attempt has been made to make our original building blocks as fundamental as possible within practical limits and to conform with those generally accepted by organic chemists.

If a term is not in the dictionary, the computer will automatically print out "term not in dictionary." The term and its codes are then manually inserted in the dictionary and the linear encoded card is sent through the machine. Future entries of the term will be automatically encoded.

It should be noted that if the same term is linearly encoded by different names the actual machine search codes will be the same. For example, if the ring in this compound:



had been linearly encoded by two analysts as:

(1) 1-PHENYL-CH-IND-W/ and
(2) 1-BENZENE-CH-IND-W1

the machine would automatically encode both with the same codes. In one case the identifying term would be benzene and certain codes would be located in the memory for this term. In the other case phenyl would be the identifying term and its codes would also be in the machine memory. These two sets of codes would be identical, and graphically illustrated as:

Benzene	7	14	22	11
Phenyl	7	14	22	11

Machine memory, (codes 7, 14, 22 and 11 may have any desired meaning)

Thus it can be seen that some of the need for consistency is alleviated and even misspellings are less of a problem. The codes for the same term will be identical regardless of its being named differently by different documents or analysts.

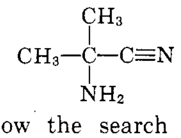
One of the major advantages of this type of automatic encoding is the fact that it makes compatible many information coding and retrieval systems. Many coding systems have appeared recently but, in each case, the analysis has been useful only for that particular code system. Different analyses are required for different systems. The PACIR procedure gives an analysis product that may be used to generate different types of files that may be searched on different types of machines.

In summary, the linear encoded fragments representing the compound are used as the basic input file. Each user supplies his own coding dictionary to translate the linear formulas into a searchable specific or generic (or both) file. The applicable codes from the user's dictionary are then manually assigned to each different linear code. This dictionary of assigned codes is stored in the computer. The entire linear encoded punch card deck is then fed into the computer and by means of an appropriate program the assigned codes are applied to each linear fragment. The computer generates a punch card for each linear fragment, in machine language. These cards may be used as a basis for input in many types of systems with various types of computers depending on the user's choice. They may be converted to magnetic tapes if so desired. Thus a standard analysis from any source is used to generate a system peculiar to the user's need.

In the Patent Office application of PACIR to the pesticide art, two search files are being developed, one employing a RAMAC 305 computer, the other using an IBM 101 statistical sorter.[16] Note that one system is random access and the other is a serial card approach. It should be apparent that the analysis product is so designed that a user may adapt the PACIR analysis to his particular needs almost solely by machine methods.

VII. SEARCH STRATEGY:

For the purpose of this paper, the search strategy will be described using a random access system and the pesticide patents. However, the principles are generally applicable and are not limited to a particular art or machine. A search for alpha-aminoisobutyronitrile.



will be used to show the search strategy. A graphic representation of the machine's memory for the relevant fragments might be:

20660	20403	20259
C---N-CH-W	NH2-CH-W	C3-CH-S-BR-NT-W
12051	12051	12037
12363	12363	12051
13297	12765	12363
14031	14031	13287
14225	14225	14225
15001	15001	15001

When instructed to find a compound (1) "containing one NH2 group, one CN group, and one 3 membered chain connected, saturated, branched, non-terminal alkylene chain" and (2) "containing C3-CN and C3-NH2 relationships," the computer goes to the address 20403 (NH2-CH-W) and to the address 20660 (C---N-CH-W) and then compares the five-digit numbers and stores the numbers that are the same:

12051 12363 14031 14225 15001

It then goes to address 20259 (C3-CH-S-BR-NT-W) and compares the stored numbers with those under this address. This results in

12051 12363 14225 15001

which is then transformed into the patent number by a dictionary look-up. The last digit in each number is the compound number. The first four digits comprise an accession number that corresponds to the patent number.

One of the additional features of the new system is the use of interfixes in the random access file. From the above, the computer tells us in a fail-safe manner the document number and the compound number in that document. The program of the computer has been so written that the computer seeks out the compound by number in the specific document and then tests for the relationships between the fragments (interfixes). These relationships have been previously stored by the automatic encoding procedure. In answering part 2 of the question, if the compound previously designated by the computer does not contain these interfixes, it is rejected and does not answer the question. However, if the compound does have the relationships it is selected as an answer. This feature enables the user to ask not only for a compound containing certain fragments, but in addition ask for the presence of relationships between the fragments.

In this example, it may also be desired to find out whether the above compound (alpha-aminoisobutyronitrile), employed in a composition containing water as a solvent, is active against nemas which infest soil containing tomato plants. The underlining and role assigning procedure mentioned in section V has provided an analysis product that allows retrieval of this secondary subject matter. The following is a graphic representation of the computer's memory for this information:

Compound A	H_2O-S	Solution F
12051	12056	12057
12363	12365	12368
14225	27984	27895
15001	15067	

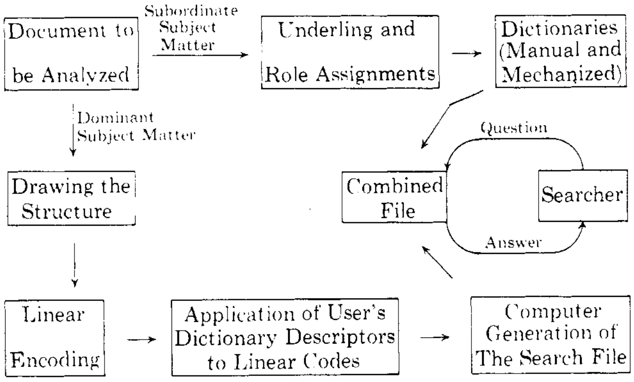
Nemas-U	Soil-U	Tomato Plants-U
12059	12052	10047
12364	12367	12059
40791	27462	12366
		28984

In the above search, documents 1205 and 1236 would answer the question since, upon comparison, the 1st four digits (accession number) are the same.

Questions can be as specific or generic as desired. For example, a user may only want to know whether nitriles have been used to destroy nemas. In actual practice, the user need not employ all the pertinent descriptors but only those which he feels will satisfy his specific query.

VIII. CONCLUSION

The following is a schematic representation of the PACIR procedures described herein:


Document to be Analyzed
Subordinate Subject Matter
Underling and Role Assignments
Dictionaries (Manual and Mechanized)
Dominant Subject Matter
Question
Drawing the Structure
Combined File
Searcher
Answer
Linear Encoding
Application of User's Dictionary Descriptors to Linear Codes
Computer Generation of The Search File

This system attempts to provide a standard analysis and linear and semi-automatic encoding technique that allows flexible and varied use of the analyzed subject matter. The approach has eliminated the dependence of the analysis on a particular type of file organization, coding, or machine. Wide acceptance of this technique would allow various users to interchange their analysis efforts without loss of the particular codes and systems that are most useful to each user.

The authors wish to express appreciation for the aid and cooperation of Diane B. Russell, Ellen D. Lewis, Lorraine T. Kendell, John Liptock, John Price and Maude Bender of the U. S. Patent Office, who have given unstintingly of their time and technical knowledge in the development and application of the techniques involved in this project.

REFERENCES

(1) J. Frome and J. Leibowitz, "A Punched Card System for Searching Steroid Compounds," Patent Office Research and Development Report No. 7, Department of Commerce, Washington, D. C., 1957.

(2) J. Frome and J. Leibowitz, "A Manual for Coding Steroids," Patent Office Research and Development Report No. 11, Department of Commerce, Washington, D. C., 1958.

(3) J. Frome, "Revised Steroid Search System Coding Manual," Patent Office Research and Development Report No. 19, Department of Commerce, Washington, D. C., 1961.

(4) J. Frome, J. Leibowitz and D. D. Andrews, "A System of Retrieval Compounds, Compositions, Processes and Polymers, Patent Office Research and Development Report No. 13, Department of Commerce, Washington, D. C., 1958.

(5) J. Leibowitz, J. Frome and D. D. Andrews, "Variable Scope Patent Searching by an Inverted File Technique,"

Patent Office Research and Development Report No. 14, Department of Commerce, Washington, D. C., 1958.

(6) J. Frome, "Mechanized Searching of Phosphorus Compounds," Patent Office Research and Development Report No. 18, Department of Commerce, Washington, D. C., 1961.

(7) J. Frome, M. A. Gannon, P. T. O'Day and F. M. Sikora, "Manual for a Punched Card Retrieval System for Organic Phosphorus Compounds," Patent Office Research and Development Report No. 22, Department of Commerce. Washington. D. C., 1962.

(8) J. Frome. H. R. Koller. J. Leibowitz and D. D. Andrews. "Recent Advances in Patent Office Searching: Steroid Compounds and ILAS," Patent Office Research and Development Report No. 8. Department of Commerce. Washington. D. C., 1957.

(9) J. Frome. "Mechanized Searching of Phosphorus Compounds." *J. Chem. Doc.*, **1**, 76–87 (Jan., 1961).

(10) M. F. Bailey, B. E. Lanham and J. Leibowitz, *J. Patent Office Society*, **35**, No. 7, 566–587 (1953).

(11) J. Leibowitz, J. Frome and D. D. Andrews, "Variable Scope Search System: VS3," preprints of papers for the International Conference on Scientific Information, National Academy of Sciences-National Research Council, Washington, D. C., 1958, pp. 291–316.

(12) D. D. Andrews, "Interrelated Logic Accumulating Scanner (ILAS)," Patent Office Research and Development Report No. 6, Department of Commerce, Washington, D. C., 1957.

(13) J. Frome, "Mechanized Searching of Phosphorus Compounds," *J. Chem. Doc.*, **1**, No. 1, 88–90 (1961).

(14) J. Frome, "Semi-automatic Indexing and Encoding," Patent Office Research and Development Report No. 17, Department of Commerce, Washington, D. C., 1960.

(15) Classification Bulletin of the U. S. Patent Office, Class 260-Chemistry, Carbon Compounds, Bulletin No. 200, Revision I.

(16) J. Frome, "Generic Mechanized Search System," *J. Chem. Doc.*, **2**, 15–18 (1962).

Photocopying by Libraries of Copyrighted Documents. A Proposal for Revision of the Present Copyright Law*

By GEORGE E. McCARTHY and EDWARD H. VALANCE
Geigy Chemical Corporation, Ardsley, N. Y.

Received May 21, 1962

In July of 1961, the Register of Copyrights reported to the House Judiciary Committee on general revision of the United States laws governing copyright material. This was the culmination of several years of study and some thirty-four interim reports on the varied aspects of author–composer–architect and artist protection. Many, if not all, of these reviews touch upon the work of the library, librarian and researcher. Specifically, this paper is concerned with the area dealing with the photoduplication of copyright materials by libraries. The recommendations include the following:

> "The Statute should permit a library, whose collections are available to the public without charge, to supply a single photocopy of copyrighted material in its collections to any applicant under the following conditions."
>
> "(a) A single photocopy of one article in any issue of a periodical, or of a reasonable part of any other publication, may be supplied when the applicant states in writing that he needs and will use such material solely for his own research."
>
> "(b) A single photocopy of an entire publication may be supplied when the applicant also states in writing, and the library is not otherwise informed, that a copy is not available from the publisher."
>
> "(c) Where the work bears a copyright notice, the library should be required to affix to the photocopy a warning that the material appears to be copyrighted."[1]

The inclusion of the clause "are available to the public without charge" has, of course, met with a number of objections. It limits photocopying of copyrighted materials to the public libraries. The university, the business-connected research group, the government library could not reproduce such papers without the time-consuming permissive grant from the copyright owner for each specific request. On the face of it, such a policy would seem over-restrictive and almost unenforceable.

Perhaps a glance at the present practices of libraries and research centers is in order.[1] The joint libraries Committee on Fair Use in Photocopying investigated the policies and photocopy work of government, university, and public libraries.[2] The making of copies for library clients is not only common practice but is a simple extension of the traditional library service, available because of the development of equipment capable of making such extension possible. Librarians and others have been offering such services under the "fair use" policy. They protect themselves, at times, with a signed statement from the patron that it is for his own use. On only rare occasions does a copyright notice appear on the copy. Does this lack of notice relieve the patron of responsibility if he makes as many additional copies as he finds convenient?

Corporate libraries, of course, also do a fair share of photocopy work. Few, if any, do any policing of the copyright material. The library user may request a copy or simply make use of the machine himself. And who is to say how many such copies are made? And, of a given paper, how often? It is much easier to make a copy to be read at leisure than to consume time taking notes. And, at today's cost of a chemist's time, isn't it better? Library notes are replaced by copies whose information is later transferred to laboratory punch or code cards. The danger of missing an important point in a paper is eliminated. Note-taking is becoming an old-fashioned and out-of-date art.

* Presented at 141st meeting, American Chemical Society, Division of Chemical Literature, Washington, D. C., March 24, 1962.