Patent Office
Research and Development
Reports . . . . . No. 14

Leibowitz, Jacob.

VARIABLE SCOPE PATENT SEARCHING BY AN INVERTED FILE TECHNIQUE

Prepared by

Jacob Leibowitz, Julius Frome, and Don D. Andrews

Office of Research and Development
Patent Office

November 17, 1958

Lewis L. Strauss
Secretary of Commerce

Robert C. Watson
Commissioner of Patents

For sale by the U. S. Department of Commerce, Washington 25, D. C. Price 10 cents.

VARIABLE SCOPE PATENT SEARCHING BY AN INVERTED FILE TECHNIQUE

INTRODUCTION

The United States Patent Office is conducting an experiment in mechanized searching of patent literature which employs a technique similar to one previously found not suitable for its literature search requirements. This technique is that of the coordinated or "inverted" file system for the coded information abstracted from the patent.

Previously,[1] this technique was found wholly inadequate for Patent Office operations and was abandoned in favor of the sequential or "normal" file arrangement. The inverted file at that time appeared to be unsatisfactory because of its apparent inability to retrieve information with precisely interrelated concepts as is required and yet allow a searcher to request a selection on the basis of either the generic or the specific scope of each term or concept being sought. Since that time, search systems have been devised by the Patent Office using the "normal" file arrangement which yield the required degree of precision for depicting interrelationship and yet allow the searching to be done at selected degrees of breadth or specificity.

The recent introduction of a small scale electronic computer having large random access memory has made possible the development of procedures for incorporating many of the precision and variable scope features of the "normal" file systems into an "inverted" file system.

The mechanized search system now being developed involves the following features:

(1) Parallel access searching in which only those portions of the file having subject matter pertinent to each set of search terms are isolated for mechanical processing as contrasted to serial searching in which a sequential processing of all portions of the file is required.

(2) Correlations are made amongst concepts or terms which individually are not restricted to the precise meaning of the terms as they appear in the dictionary but may be altered by instructing the stored program of the computer to generate within itself those files having the desired conceptual meaning.

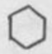
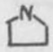
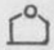
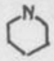
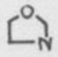
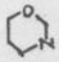
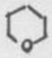
(3) Dictionary terms are generated from the language of those documents comprising the file without using a prearranged hierarchical system of terms.

The group of patents selected for this experiment comprises the chemical polymer art which involves organic and inorganic compounds as well as properties, functions and processes associated therewith. This is an extension of the types of information handled in the Variable Scope Search System (VS$_3$),[2] but the principles of recording precise interrelationships of subject matter and the recognition of genus-species relationships have been adhered to in a substantial way.

SHOWING RELATIONSHIPS IN SERIAL AND INVERTED FILES

It is believed that the present experiment may be best described by concrete illustrations of the manner in which relationships, both of the interrelational and the genus-species types are handled in the Patent Office.

Let us assume that our dictionary of descriptors consists of but ten terms; for simplicity in description. Furthermore, let these descriptors be limited to those applicable to certain chemical ring structures only, for the same reason. Such a dictionary might appear as follows:

Code	Name	Structure
1	phenyl	
2	pyrryl	
3	furyl	
4	pyridyl	
5	oxazolyl	
6	oxazinyl	
7	pyranyl	
8	a six-membered ring	
9	a nitrogen-containing ring	
10	an oxygen-containing ring	

The descriptors identified by 8, 9 and 10 are more generic in character than those identified by the numerals 1 to 7 since they may be properly applicable to one or more of the other descriptors.

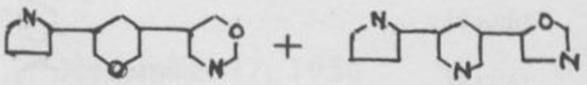
The descriptors 1 to 7 on the other hand, are actually "building blocks" or "fragments" of which one or more may be associated together to identify a chemical compound.

The disclosure of a chemical patent or other document is usually more than a mere listing of chemical compounds, for each of the chemical compounds is associated with certain of the other compounds to form a definite process or chemical reaction chain in which each compound may have a role such as, starting material, final product, solvent, catalyst and the like. A patent or document may also depict a number of different processes each having its own set of mutually related notions or descriptors.

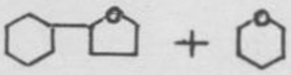
Neither a serial nor an inverted file would be satisfactory in the Patent Office if only a single level of association of the applicable descriptors were to be had. This results from the fact that a single level of descriptor association, i.e., the level of the entire document, would allow retrieval of a host of documents which are nonpertinent to a normal search request because there is no ability to associate those descriptors which together identify a particular "fragment" of one compound as distinguished from those descriptors in any other "fragment," or in any other compound or in any other process of that document.

This can be illustrated by recourse to a series of hypothetical processes involving hypothetical compounds selected from our dictionary.

Suppose hypothetical patent A were to disclose a first reaction process having two compounds as follows:



as well as a second different reaction process having two other compounds as follows:



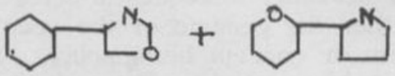
A serial file would consist of a heading identifying the patent number and followed or preceeded by all ten descriptors of our dictionary as follows:

$$\frac{A}{1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10}$$

An inverted file having only a single level of association would consist of a series of ten headings identifiable with each descriptor of our dictionary with each heading followed by the number of the patent as follows:

$$\frac{1}{A},\ \frac{2}{A}\ \ldots\ldots\ \frac{8}{A},\ \frac{9}{A}\ \ \frac{10}{A}$$

It is obvious that this patent would respond to every search request possible with our limited dictionary including such dissimilar requests as:



The serial file approach was the first to most readily adapt itself to a multi-level association of descriptors so that all these precise relationships of the building blocks of a disclosure, and no more, were recorded. A technique of the mathematician, i.e., the bracket, was employed to establish the proper relationships. Patent number A would thus be recorded as follows:

$$\left[\left\{\left[(2,9)\ (7,8,9)\ (6,8,9,10)\right]\quad\left[(2,9)\ (4,8,9)\ (5,9,10)\right]\right\}\quad\left\{\left[(1,8)\ (3,10)\right]\left[(7,8,10)\right]\right\}A\right]$$

In which [ ] = limits of a patent

{ } = limits of a process

[ ] = limits of a compound

( ) = limits of a fragment of a compound

Since the serial file was scanned one symbol at a time in sequence, it is relatively simple to cause a search machine[3] to recognize the codes for the descriptors as being grouped within or without any limit specified in a question. Thus a compound $\left[(3)\ (8)\right]$ could be recognized as present in patent A but not compound $\left[(1)\ (9)\right]$. The serial search system has proved itself to be satisfactory and is now operational in the Patent Office for a small portion of the polymer file. The chief difficulty is the apparent inefficiencies of scanning all the information contained in any file for each search request; which is time consuming for present day search machines, but future machine developments may possibly remove this handicap without adding excessive cost factors.

The availability of a large random access memory computer appears to remove many of the constraints which it was felt were present in the well known Batten card or coordinate index forms of implementing an inverted file. These constraints fundamentally resulted from the limited number of entries possible on a single card or sheet, the difficulty in posting new information and the manual manipulation of the cards or sheets.

In the inverted file system now being prepared interrelationships between building blocks or structural fragments are confined to the proper limits by adding to the end of the document number one or more arbitrarily assigned digits which reflect an association in a common process or compound[4, 5]. Genus-species relationships may be sought for as extensively as desired by machine generating a generic file from a series of specific files through programmed manipulations executed on the specific files.

Thus the inverted file for patent A of this simple example would be arranged as follows:

1	2	3	4	5	6	7	8	9	10
A-2-1	A-1-1	A-2-1	A-1-2	A-1-2	A-1-1	A-1-1	1	2	3
	A-1-2					A-2-2	4	4	5
							6	5	6
							7	6	7

Actually the relationships expressed under 8, 9 and 10 are wholly independent of the disclosure of any patent but may be regarded as a part of the chemical "grammar" common to all chemical documents. These may be considered as "high level" terms. As many different levels of meanings as may be convenient can be employed. Actually in the polymer file three levels were used.

With an inverted file system such as this a search for any compound would involve selecting only those portions of the total file which relate to the fragments involved in the sought for compound and correlating the complete document numbers thereunder to thereby identify those document numbers common to all the fragments. If the search is for two or more compounds associated in the same process, each compound is separately processed and the resulting lists of document numbers again compared, but this time ignoring the portion of the document numbers denoting the compound number. Similarly, if two different processes are sought in the same patent, correlations are made for each process as above and then followed by recorrelation of the document numbers ignoring both the compound and process numbers.

If one or more of the fragments of a compound are identified by a generic descriptor the search proceeds as before except each such generic descriptor produces lists of other descriptors which in turn produce a consolidated list of all compounds in the system meeting the description of the generic term. The searcher does not need to know which species are members of the selected genus.

The inverted file system now undergoing development utilizes three levels of descriptors in which the third or lowest level are descriptors of specific compounds identified in the documents while the second or intermediate level contains descriptors of specific structural fragments of these compounds and the first or high level terms describe various mutual attributes of those fragments. Since the third level terms are on a specific compound basis, it is not necessary to add to the document numbers more than the small arbitrary number indicative of the process in which the compounds are found. Furthermore, the patent numbers have been replaced by four digit accession numbers. This results in a file of five digit numbers, four for the accession number and one for the process number.

Thus the search routine is merely one of making successive correlations of lists of five digit numbers.

A computer program has been developed to recognize and make these correlations at the time of the search. This is done, in essence, by two routines called (1) merge and (2) match.

(1) Merge

The merge represents an "or" relationship. That is, things which are either 6 membered rings or nitrogen rings are discovered by merging the listings under 8 and 9 into one listing as follows:

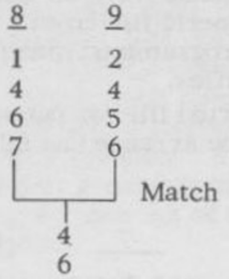
8	9
1	2
4	4
6	5
7	6

Merge

1
2
4
5
6
7

(2) Match

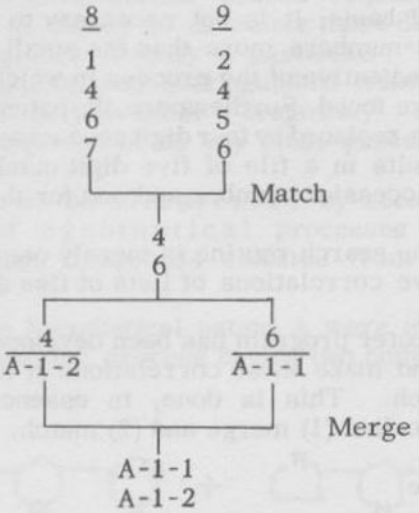
The match provides an "and" relationship. Rings which are both 6 membered and nitrogen containing are obtained by matching identities of fragments listed in 8 and 9.



Combinations of merge and match

These routines are ordinarily employed in combination. For example:

(a) To find–compounds having a 6 membered nitrogen ring.



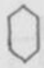
An illustration will now be given of the search operator using the hypothetical disclosures and dictionary set forth in Appendix A.

For convenience, one-digit codes are used to represent first level terms (generic), two-digit codes to represent second level terms (fragments) and three-digit codes to represent third level terms (compounds). Under each first level term is filed a series of two digit numbers identifying those second level terms correctly included as fragments within the genus of the first level term. Each second level term in turn heads a file of three digit numbers identifying all those third level terms corresponding to compounds containing that fragment. Each third level term finally heads a file of five digit numbers in which the first four digits represent the number assigned to a document and the fifth digit is an arbitrary identification of the particular chemical process in which the compound term on the third level is associated. Because of machine techniques the numbers representing these terms on all three levels are called "addresses" since they are used to locate the files in the memory of the computer.

For this example assume the search question as follows:

Find all documents in which compounds A and B are in the same process as well as compounds C and D are likewise in a common process and in which A, B, C and D are each specified as follows:

A--compound including both a 6-membered oxygen ring and

a  fragment.

B--compound including both a nitrogen containing ring and a

halogen

C--compound comprises C═C—C═C—Cl

D--compound includes both a 5 membered ring and a halogen

Symbolically this question would be represented as:

$$\left[\left\{\overset{A}{\left[(1,3)\ (10)\right]}\ \overset{B}{\left[(4)\ (8)\right]}\right\}\left\{\overset{C}{\left[(108)\right]}\ \overset{D}{\left[(2)\ (8)\right]}\right\}\right]$$

The machine to be employed is the RAMAC 305 for which a program is being developed which will accept a series of punched cards bearing the addresses of those portions of the file which are to be investigated as well as information showing the logical grouping required by the question.

The computer will then seek out the sets of data in its file corresponding to these addresses and perform a succession of merging, matching and reseeking operations until it arrives at the numbers of the documents satisfying the search requirement.

The specific steps performed by the computer are diagramed in Appendix B for this particular search question.

While not reported here, the actual search system being constructed for the polymer patents recognizes the role or function each compound plays in the total disclosure and is subject to retrieval on that basis as well as that of the compound identification.

CONCLUSIONS

The system described appears to offer a promising approach to the machine searching problem. Many problems remain to be solved however. For example, where it is required to find a process containing A+B and another process containing C+D, it is not yet possible to avoid retrieval of the invalid answer A+B+C+D, all in the same process. Similarly, a fragment answering two separate sets of descriptors will respond as an answer to both.

Also, while 3 search levels only have been described, it is believed that more levels of search can be provided in order to encompass a more extensive or elaborate hierarchy.

In addition, the system should be applicable, in principle, to subject matter outside the chemical field.

The dictionary of the 1st and 2nd levels is generated as needed from the actual terms used in the patents.

Each term in the dictionary is, therefore, in use and each term of the disclosure is therefore codable--in contrast to the situation involved in a pre-established dictionary.

Standardization of synonymous terms in the disclosure takes place through the 2nd level terms.

ACKNOWLEDGMENT

Acknowledgment is made of the assistance and cooperation of Floyd D. Hamilton, Rowena W. Swanson of the Patent Office and John Nolan and Barbara Wood of IBM.

REFERENCES

(1) Bailey, M. F., B. E. Lanham and S. W. Cochran, "An Experiment in Mechanizing Searching for Compositions of Matter." A paper presented before the 113th meeting of the American Chemical Society, Chicago, April 1948.

(2) Leibowitz, J., J. Frome and D. D. Andrews, "Variable Scope Search System: VS_3." Preprints of Papers for the International Conference on Scientific Information 1958, Washington, D. C., National Academy of Sciences--National Research Council. Area 5 pp 291-316.

(3) Andrews, D. D., *Interrelated Logic Accumulating Scanner (ILAS).* Patent Office Research and Development Reports. No. 6, 1957, Washington 25, D. C., U. S. Patent Office.

(4) Wadington, J. P., "Unit Concept Coordinate Indexing," *American Documentation*, April 1958.

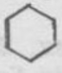
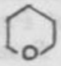
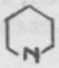
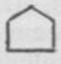
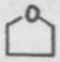
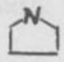
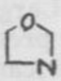
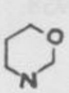
(5) Nolan, J. J., "Information Storage and Retrieval Using a Large Scale Random Access Memory." A paper presented before the 133rd meeting of the American Chemical Society, San Francisco, April 1958.

APPENDIX A

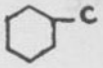
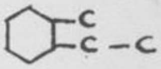
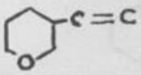
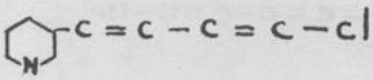
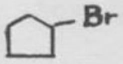
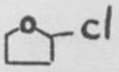
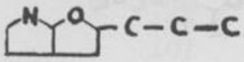
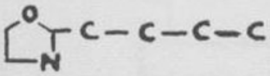
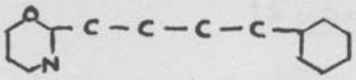
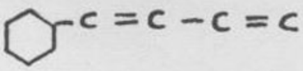
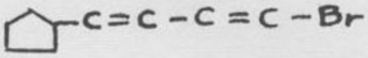
1st Level (Generic) Terms

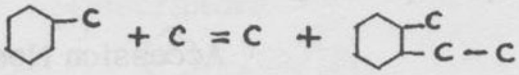
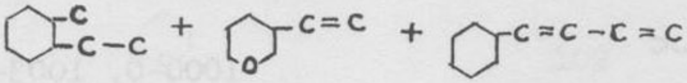
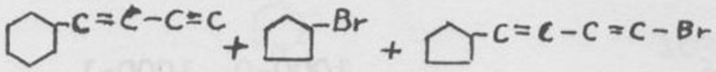
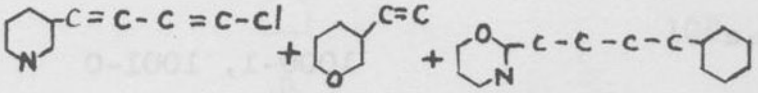
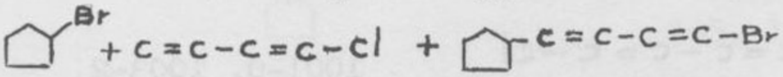
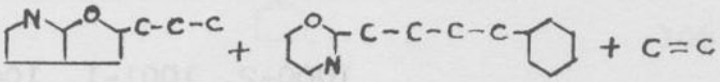
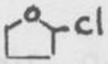
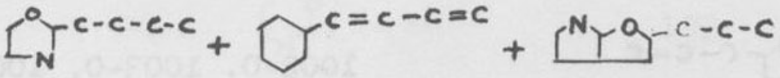
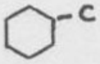
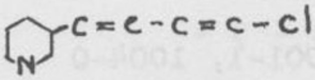
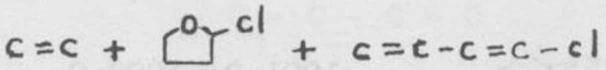
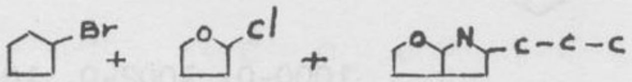
Address	Descriptors	Fragments
1	6 membered ring	10, 11, 12, 17
2	5 membered ring	13, 14, 15, 16
3	O containing ring	11, 14, 16, 17
4	N containing ring	12, 15, 16, 17
5	Alkyl	18, 19, 20, 21
6	Ethylenic unsaturated	22, 23, 24
7	Conjugated	23, 24
8	Halogen	25, 26, 27

2nd Level (Fragment) Terms

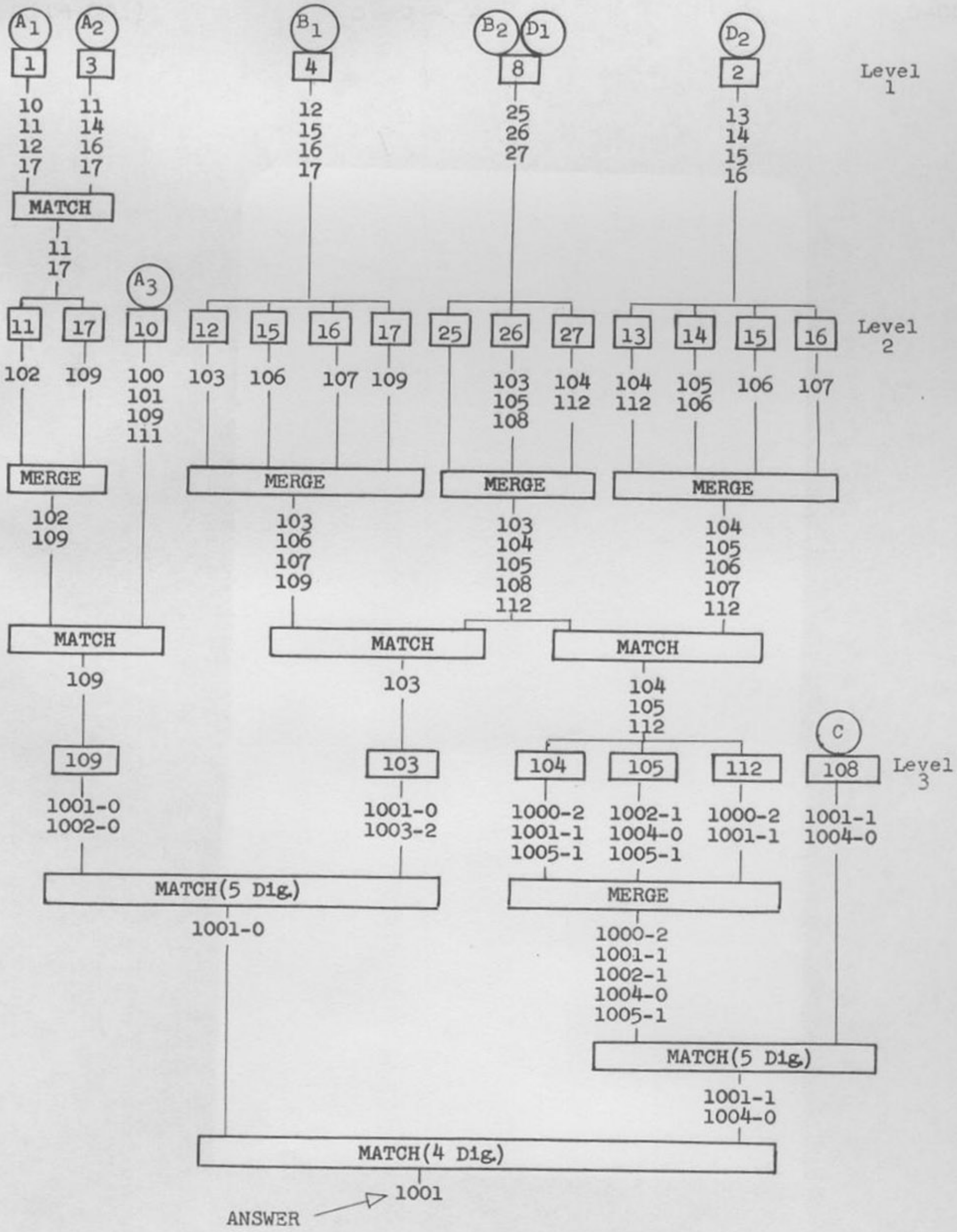
Address	Descriptors	Compounds
10		100, 101, 109, 111
11		102
12		103
13		104, 112
14		105, 106
15		106
16		107
17		109
18	- C -	100, 101
19	- C - C -	101
20	- C - C - C	106
21	- C - C - C - C	107, 109
22	- C = C -	102, 110
23	- C = C - C =	
24	- C = C - C = C -	103, 108, 111, 112
25	F	
26	Cl	103, 105, 108
27	Br	104, 112

3rd Level Terms
(Compounds)

Address	Descriptors	Accession Nos. of Documents
100		1000-0, 1003-1
101		1000-0, 1000-1
102		1000-1, 1001-0
103		1001-0, 1003-2
104		1000-2, 1001-1, 1005-1
105		1002-1, 1004-0, 1005-1
106		1002-0, 1003-0, 1005-1
107		1003-0
108	c = c - c = c - cl	1001-1, 1004-0
109		1001-0, 1002-0
110	c = c	1000-0, 1002-0, 1004-0
111		1000-1, 1000-2, 1003-0
112		1000-2, 1001-1

Accession Nos.	Processes in Documents	Compound Codes
1000-0		(100 + 110 + 101)
1000-1		(101 + 102 + 111)
1000-2		(111 + 104 + 112)
1001-0		(103 + 102 + 109)
1001-1		(104 + 108 + 112)
1002-0		(106 + 109 + 110)
1002-1		(105)
1003-0		(107 + 111 + 106)
1003-1		(100)
1003-2		(103)
1004-0		(110 + 105 + 108)
1005-1		(104 + 105 + 106)

APPENDIX B



43444--U.S.Dept.of Comm--DC--1958