Patent Office
Research and Development
Reports . . . . . No. 5

U.S. Patent Office. Office of Research and Development.

ORGANIZATION OF CHEMICAL DISCLOSURES FOR MECHANIZED RETRIEVAL

FOR OFFICIAL DISTRIBUTION

Prepared by
B. E. Lanham
Deputy Director

J. Leibowitz
H. R. Koller
H. Pfeffer
Patent Research Specialists
Staff Members

Office of Research and Development
U. S. Patent Office

June 14, 1957

Robert C. Watson
Commissioner of Patents

Sinclair Weeks
Secretary of Commerce

ORGANIZATION OF CHEMICAL DISCLOSURES FOR MECHANIZED RETRIEVAL[1]

This report will describe a system of organizing the total disclosure of a document to present a comprehensive, mechanically retrievable representation of the details and relationships of chemical subject matter. The method is being used in the preparation of data for patent searches which have been programmed for SEAC, the electronic digital computer of the National Bureau of Standards. Members of the Applications Engineering Section of the Data Processing Division of the National Bureau of Standards are working closely with the Patent Office Research and Development Group on the computer program and on more general aspects of the application of machines to the Patent Office search problem. At the time of writing, the program is undergoing the process known as "debugging." Details of the computer program itself will be described in a subsequent paper.

WHAT DOES "PATENT SEARCHING" CONTEMPLATE?

In "Information Retrieval," as the term is generally used, the subject matter desired is assumed to be present in the file being inspected. For example, if the file contains 1950 population data for all countries which are members of the United Nations and one wishes to find the population of France in 1950, the retrieval job resolves itself into locating and reproducing this particular information from the files.

However, in "Literature Searching" there is no presumption that the desired subject matter exists in the file, and in the example given one might search the file to find out whether or not the 1950 population of France is listed.

"Patent Searching" is a special type of literature searching which is performed by the Patent Office in determining the patentability of the claims of a patent application and it has peculiar characteristics, some of which are as follows. Confronted with the "claims" of a patent application, which set forth the limits of the area to which the inventor is attempting to acquire an exclusive right, the patent examiner, by means of a search of all available publications, must determine whether the subject matter set forth in the claims is novel. If it is novel, the examiner must determine whether it is sufficiently different from related or equivalent subject matter to be considered "inventive." The term "equivalent subject matter" is used in the sense of other embodiments of the same inventive concept represented by the claims, and these may be found by searching on the basis of the classes of which the concept claimed is a member. The standards for judging "invention" are complex and their bases are included in both statutory and judicial pronouncements; but since they are not relevant to the present discussion they will not be further explained here.

To determine both novelty and invention the patent searcher looks for disclosures of a particular concept rather than words. He is not primarily interested in the general subject matter of a document. Further, he is not interested in the entire document as an entity except as evidence that the subject matter of his search has been previously conceived. Of the totality of ideas in a collection of documents every one of them or any combination of them may be the basis on which someone may wish to make a search. This, together with the fact that it is impossible to predict just what questions will be asked of the system, makes it mandatory to include in the encoded file all of the disclosed technological details in each document, together with the relationships among them. This will permit the searcher to in effect synthesize whatever classification pigeonhole is pertinent to his needs.

Since generally, any concept can be described from multiple points of view, including a specific identification of it, the searcher must be able to find the disclosures in which he is interested by defining his field of search in such a way as to delimit his needs precisely and with whatever degree of specificity or breadth that he requires. He must be able to search for combinations or subcombinations of ideas and to vary his search on this basis as the system finds or fails to find answers to questions phrased more or less comprehensively. See Newman's article for a discussion of other phases of the subject of patent searching (1).

GENERAL CHARACTERISTICS OF THE MECHANIZATION SYSTEM

The computer search program which embodies the system is known as HAYSTAQ. (Have You Stored Answers to Questions?)

The documents first being incorporated into the system are chemical patents. These include within their scope not only compounds and mixtures of materials (referred to as compositions) but also processes for making and using compounds or compositions. Reasons exist for believing that the next large field of endeavor to which the system will be adaptable will be the electrical or electronic arts, since the structural forms of these disclosures are, to a reasonable extent, similar to those of chemical disclosures. Application to the mechanical arts is probably further in the future because of peculiar language problems that exist and the multidimensional interrelationships between disclosure elements. For example, in disclosures of machines, the elements are frequently described only functionally and the associations among mechanical elements are often described largely in

[1] Revision of paper presented before the Division of Chemical Literature, 131st Meeting of the American Chemical Society, Miami, Florida, April 8, 1957.

terms of complex interwoven sequences of events that occur when the machine is in operation.

(1) *The technological disclosures in the documents which will be incorporated into the system are structured.* Various levels of organization of subject matter exist, such as the entire document, the several discrete disclosed processes, the various independent mixtures of materials (these may be taught as being involved in some step of a process) and the individual compounds. If a single compound is disclosed as existing alone, organizationally it is treated as a composition. In addition, there are many relationships which may exist among the elementary ideas disclosed, such as combinations, positive and negative relationships, alternativeness, equivalence, exclusion (identified in the field of patent law by the term "consisting"), inclusion (identified by the term "comprising"), and sequence, either in time or space.

(2) *The encoded form of the disclosure closely follows the structure of the disclosure in the document.* Neither a narrative summary nor a transliteration of English words into machine code is aimed at, nor is the order of presentation of ideas used in the document preserved in the code. Analysis and coding of a disclosure involve much sifting and reordering of the subject matter so that the final coded format reflects a logical arrangement of non-intermeshed structural groupings of the teachings of the document.

(3) *An encoded question* is in the form of a model answer and it reflects the structure which such information would have if it were a disclosure. In addition to the model answer, the question contains data for use by the machine in screening disclosures for relevance to the search question as well as data used in selecting the appropriate parts of the search program. The latter is necessary because the program includes facilities for searching under a very large variety of conditions, and less than all of these conditions will obtain for any one search. At present, approximately 2^{40} variations of the search are provided for and this is entirely apart from variations due to the use of different code words.

The program may be described as treating each search question as though it were in the following form: is it true or false that the file contains an encoded disclosure of X? (X is the subject matter sought). If false, give some indication that there is no such disclosure, but if true, produce the evidence by identifying the document containing the disclosure of X, such as by printing out the patent number.

In stating the search question, X is specified in exactly the scope desired to be found. However, it is to be understood that for a question which is generic in scope, it is generally assumed that what is sought is either a disclosure of any member of the genus or a teaching of the genus. (Facilities for searching questions which require as answers disclosures of a teaching of a genus are contemplated for incorporation into the program.) The machine is not permitted to approximate the answer, nor is it permitted to vary the scope of X. Thus, if a combination of A+B+C is asked for, answers corresponding to a question for A+B are not acceptable, nor are answers corresponding to a question involving A+B+D acceptable. One may state whether only disclosures of exactly A+B+C are desired or whether disclosures which include something in addition to A+B+C are also acceptable. Should no answers for A+B+C be found, it would be desirable to have the machine ask itself additional questions for less than the entire combination, such as A+B. Also, it would be desirable to have the machine determine and find equivalents of A+B+C, such as a disclosure of A+B+D together with another disclosure which teaches that D is the equivalent of C. Searching according to either of these last two variations is contemplated for the future but has not yet been incorporated in the program described in this paper. In any event, such search facilities will be designed with control of the extent of variation from the original question determined by the searcher, rather than by the machine, for it can easily be demonstrated that unrestricted "free choice" in this regard can lead to finding logically equivalent answers which are actually not at all desired. For example, a request for A+B+C might result in finding D+E+F, where D is equivalent to A, E is equivalent to B and F is equivalent to C, but the total effect of the D+E+F combination may be too unlike A+B+C to be of utility to the searcher.

THE PROBLEM OF ENCODING THE TERMS USED IN A DOCUMENT

One using the system, either by encoding data or asking questions, should not have to bridge the gap between the words of a document and a universal basic concept language and then bridge the gap between that language and a machine code language. Therefore, there will be provided a dictionary of machine codes representing the various subject matter concepts provided for in the system. Entries will be listed under common language terms. In the dictionary the single concept denoted by a code should be unambiguous. The concept invoked by a word or expression in its context should be considered as important rather than the word per se. When several expressions have the same meaning they should have the same code.

For example, one wishing to encode the expression "boy" should find in the dictionary the same code as one looking up the expression "young human male." On the other hand, the expression "ring" would have several codes, and the correct one to be used would depend upon whether the context in which it was used indicated its meaning to be (1) the sound of a bell, (2) a chemical configuration, (3) an annular piece of jewelry worn on the finger, (4) an association of counterfeiters or (5) a prize fighting stage.

Provision must be made in the dictionary of concepts for which codes are provided to indicate generic and specific relationships as well as the relativity of these concepts. Thus, if A is generic with respect to species B, B may be at the same time generic to species C.

Provision must also be made for including all species under a genus, and miscellaneous categories are useful under each generic catagory to include all species not specifically provided for.

The HAYSTAQ program has been designed to be independent of vocabulary problems. The solutions to problems in this area should be embodied in the coding schedules.

THE ORGANIZATION OF A DISCLOSURE

The basic unit of searchable information in the system, which is called an "item," represents all the pertinent information about a single chemical compound. Included in an item are a specific unique identification tag called an "index number," the empirical formula (where known), descriptors of the chemical characteristics of the compound, physical characteristics, source characteristics (where the compound is present as part of an extract or other naturally occurring mixture) functional or use characteristics, process involvements of the compound and its structural formula. (See Fig. 1).

Item A

1. *Index number* for Item A
2. *Empirical formula* for Item A
3. *Chemical descriptors* for Item A
4. *Miscellaneous descriptors* for Item A (e.g., physical characteristics or source from a named botanical genus)
5. *Functional descriptors* (e.g., use as an antimalarial)
6. *Process descriptors* (e.g., starting material for reaction X in time 1)

Figure 1

Each of these is called a "descriptor" of the item. If a compound is named in a disclosure but is taught to be excluded from a mixture whose components are named, the negative character of the item representing such compound is indicated. When several compounds are disclosed as alternatives in a composition, this relationship among items is indicated. The further relationship of equivalence among alternative items is shown by the fact that they share a common function. Should a compound be disclosed generically but with certain negative limitations, the positive and negative descriptors are each identified. Generally, each descriptor in an item is encoded in a single unit of computer storage, called a "word." However, where a single word is insufficient, as may occur in the empirical formula, chemical descriptors or structural formula, provision is made for treating a group of words as a unit. The several words which together constitute an item and the several items which together represent a composition are grouped together by "heading" words, which contain information used by the machine in selecting the appropriate parts of the search program.

A process step is considered to be defined by a statement of its mixture of starting materials and its mixture of final products. Thus, two compositions are required to define a process step. Each of the compounds involved is set forth as an item, and as noted above, an item includes the codes describing process involvements of the compound. The several steps involved in a total process include each of the compositions involved in each step and such compositions form an organizational grouping known as a "process string." Several independent process strings may be grouped together in the largest grouping recognized, which is called a "document" or "patent."

Since HAYSTAQ has been devised for use with a computer which has a serial input and a serial type of internal operation, inspection of a large file might be an uneconomically lengthy operation unless devices were incorporated into the system to accelerate the search. Four main approaches to the problem of increasing the rate of scanning the file are provided.

1. The Rule of Progression

In making a search, one question item is compared successively with each item in a disclosure composition. Only if an answer to a question item is found is the next question item called up, and then it is successively compared with each item of the disclosure composition. If any one question item fails to find an answer, no further time is spent in inspecting the same disclosure composition. Instead, the next disclosure composition is immediately called up and the process of comparing the first question item with each item of the new disclosure composition is begun. The same principal is followed at all levels of search. Thus, within an item, before a question descriptor can be compared with a disclosure descriptor, the previous question descriptor must have been matched and if any one question descriptor fails to find a match in a disclosure item, the next disclosure item is called up and a match for the first question descriptor again is sought.

The general rule is that as soon as any unit of disclosure is discovered to be not an answer to any part of the question, that unit of disclosure is no longer inspected to determine whether it satisfies the rest of the question, and the machine immediately progresses to the next higher level unit of disclosure which, until proven otherwise, seems to be more promising.

2. Ordering

The several kinds of descriptors contained in an item each have characteristic identifying indicia. Thus, the first digit of a code word for an index number is 1, for an empirical formula 2, for a chemical descriptor 3, and so forth, for each type of descriptor used.

The descriptors are arranged in ascending order of their first digits. In searching, when a descriptor in a question item is being compared with a descriptor in a disclosure item, initially the first digits are compared. If the first digit of the question descriptor is larger, the next disclosure descriptor is called up and its first digit is compared with that of the question. If the first digits match, the descriptive or "substantive" portions of the two descriptors are then compared. Should the first digit of the disclosure descriptor be larger than that of the question, it is apparent that it is no longer possible to find a match in this disclosure item. Therefore the rest of the disclosure item is ignored and the next disclosure item is gone to at once. In a similar manner the several elements listed in an empirical formula are arranged in ascending order of atomic numbers. Descriptors of the chemical characteristics, as many as four of which appear in a single computer word, are similarly arranged. Other orderings are provided for similar purposes. For example, if several process strings are disclosed in the same document, the largest one is inspected first. If it is screened out because the question deals with a larger process, it is apparent that no other process string of this document could suffice and the remainder of the document is ignored.

3. Screens

Screening is a procedure which permits making a rapid determination of whether it is worthwhile asking a larger number of more specific questions of a disclosure unit. For example, if one were interested in finding a particular structural formula, it would be well to make a detailed structure search of only those file entries which have an empirical formula inclusive of that of the question compound. Other bases for screening include: various general chemical structure concepts, the presence or absence of a negative disclosure when one is being sought, the presence or absence of a disclosure of alternatives when a teaching of equivalence of two or more things is wanted, an indication of whether or not any process is disclosed when the search is for a particular process, and the number of compositions involved in a process.

In a sense, because of the rule of progression, every part of a question has some screening value. However, a true screen is a concise, redundant statement of information which is also presented in greater detail elsewhere in the disclosure and question codes. A screen may concern itself not only with such information as is likely to be the subject matter of a real search, but also with mechanical details, such as the form or size of the code for a given organizational unit of information. An example of this may be the number of negative descriptors in a disclosure item, which could be compared with the number of negative descriptors in the question item. A screen should have a large measure of ability to separate the wheat from the chaff, that is, it should easily and effectively eliminate the need for close scrutiny of many nonpertinent disclosures while accepting for detailed inspection at least all those which are pertinent. Those disclosures which "pass the screen" but are not really pertinent will be later eliminated in the more detailed search following the screening.

How are screens selected? According to one theory, a screen would be well chosen if it permitted acceptance of 50% of the disclosures for further searching and eliminated the other 50%. A plurality of screens would be provided and the whole array used in every search, each one being used on those disclosures which passed the preceding screen. The proportion remaining after several successive tests would be very small. It is easy to see that if 4 such screens were used, after the first one only 50% of the original number of disclosures would remain, after the second one only 25%, after the third 12 1/2% and after the fourth 6 1/4%. In fact, it has been proposed that for some uses a searching system would be adequate if a predetermined number of screens were employed without any more detailed searching. However, the larger the file, the longer the documents and the more complex the questions in the system, the less satisfactory would the system be which made all determinations of what documents are to be selected strictly on the basis of probability. In such a large collection as the Patent Office deals with, it is necessary to make a very precise selection of the pertinent documents, that is, the selected group must include all those which are pertinent, in that they contain the disclosure sought and the total number selected must be as nearly equal to the number of pertinent ones selected as possible. It is for this reason that the benefits of screens are employed in this system, namely to theoretically cut down the size of the file, and to facilitate making the ultimate selections on the basis of more exact criteria.

The probability or statistical approach in which desired subject matter is converged upon by searching solely in terms of multiple descriptors without showing relationships among them is considered to be useful as an auxiliary tool in a system directed to precise and specific searches of a very large file. Intensive detail specification, such as structural formulas, and specific identifications, such as index numbers, will be relied upon for precise pinpointing of subject matter.

A second theory of screening differs from the first in that it requires rejection of a larger proportion of the file and permits the use of a large variety of available screens, the ones employed in a particular search being selected in terms of the nature of the search question at hand. If the 4 screens used in the previous example could each eliminate 90% of the references, only 0.01% would remain instead of 6 1/2%, which is an improvement by a ratio of 625 to 1, resulting from a judicious selection of the screens applied. At the present time various screens are being tested and only a large amount of experience with the system in actual use will permit real evaluation of particular examples of them.

4. The Relative Placement in Storage of the Structure Codes as Compared With That of the Remainder of the Codes.

The space required to store one large composition disclosure is more than that available in the internal memory of the computer. Therefore each composition will be stored externally on two magnetic tapes, the codes for the general information about the composition being on one tape and the codes for the structural formulas being on the other.

First, the general information will be read into the internal memory of the computer and searched, and only afterwards, if the disclosure is tentatively selected, will information on the structure tape be read and searched. In practice, this means that if during the search on the general information tape a composition disclosure fails to answer the question for any reason, both tapes can be advanced to the next composition, the structure tape being neither read in nor inspected. This is important because time spent in reading in data without processing it is wasted time, as far as productive results are concerned.

An example of a disclosure and a schematic representation of its code follows. An X reaction between A and B forms C and D, then C and D by a Y reaction form E and F and, finally, E and F by a Z reaction form G.

On Tape No. 1

Patent No. 1
Composition I
Item 1
1. Index number for A
2. Empirical formula for A
3. Chemical descriptors for A
4. Miscellaneous descriptors for A (e.g. physical characteristics or source from a named plant genus)
5. Functional descriptors (e.g. use as an antimalarial)
6. Process descriptors (e.g. starting material for reaction X in time 1)

Item 2
1-5. (Similar to Item 1)
6. Starting material for reaction X in time 1

Composition II
Item 1
1-5. (Similar to Item 1 of Composition I)
6. Final Product of reaction X in time 1
6. Starting material for reaction Y in time 2

Item 2
1-5. (Similar to Item 1 of Composition I)
6. Final product of reaction X in time 1
6. Starting material for reaction Y in time 2

Composition III
Item 1
1-5. (Similar to Item 1 of Composition I)
6. Final product of reaction Y in time 2
6. Starting material for reaction Z in time 3

Item 2
1-5. (Similar to Item 1 of Composition I)
6. Final product of reaction Y in time 2
6. Starting material for reaction Z in time 3

Composition IV
Item 1
1-5. (Similar to Item 1 of Composition I)
6. Final product of reaction Z in time 3

On Tape No. 2

Patent No. 1
Composition I
Item 1
Structure codes for A
Item 2
No structure codes for B
Composition II
Item 1
Structure codes for C
Item 2
Structure codes for D
Composition III
No Structure codes for E or F
Composition IV
Item 1
Structure codes for G

In addition to the information represented in this illustration, negative descriptors and negatively disclosed items will be indicated, as will also those groups of items which are disclosed as being alternatives for each other. By comparison between Tape No. 1 and Tape No. 2 it will be seen that not every item (e.g., Item 2 of Composition I) nor every composition (e.g., Composition III) will have codes for structural formulas. Whether or not one is coded depends solely upon whether the document reveals such information. The format of the code for a question will be similar to that for a disclosure. Included in the "heading words," which are the words in the example which identify the number of the patent, composition and item, will be the information for use by the machine in selecting and using the appropriate portions of the total available search program.

SEARCHING FOR A PROCESS

The system provides facilities for searching for disclosures of a process, which may be defined in the search question in as much detail as desired in terms of the starting materials, intermediates, end products, reactions involved, reaction conditions, or any combination of these factors. The identification of a process step, as by name (e.g. diazotization, neutralization, etc.), is coded as a descriptor of each item involved in this step, and associated starting materials and final products are also identified. In a process involving several steps, intermediate compounds are identified as both the final product of an earlier step and a starting material (reactant) for a later one. The time sequence is denoted by assigning numbers in sequence to the descriptors of the process steps.

The example of the process disclosure shown above may be represented in summary as shown in Figure 2.

In this representation, A, B, C, D, E, F, and G represent all of the non-process descriptors for items A, B, C, etc., SM represents "starting material" and FP represents "final product."

The time sequence numbers are relative, not absolute. If one were to search, for example, for a Y reaction between C and D to make E and F, followed by a Z reaction between E and F to make G,

Composition I	Composition II	Composition III	Composition IV
Item 1	*Item 1*	*Item 1*	*Item 1*
A	C	E	G
SM - X - 1	FP - X - 1	FP - Y - 2	FP - Z - 3
	SM - Y - 2	SM - Z - 3	
Item 2	*Item 2*	*Item 2*	
B	D	F	
SM - X - 1	FP - X - 1	FP - Y - 2	
	SM - Y - 2	SM - Z - 3	

A = all descriptors for cpd. A except process and structure
B = all descriptors for cpd. B except process and structure, etc.

Figure 2

he would specify the time number for Y as 1 and that for Z as 2. The machine would compare the numbers associated with Y and Z for their relative values, namely that Z's number is greater than Y's (which would indicate its later occurrence in time), rather than looking for the same absolute values as are stated in the question.

Actually, because of the "interlocking effect," which will be described below, the question need not specify sequence (or relative time) numbers unless it is of the type referred to as a "generic" process question. Such a process question is one which specifies the desired sequential relationships among some of the steps in a process but permits a hiatus between other steps. For example, as in Figure 3, a question may be directed to an X reaction between compounds A and B to make compound C, followed by reaction Y, in which C forms compound E, which in turn is used to make compound G. In this question, it is not specified whether (1) E is the immediate precursor of G, or whether (2) several process steps intervene. This is called a "generic" process question because disclosures of both of these types, (1) and (2), will be acceptable. When a disclosure appears to meet all of the general requisites of the generic process question, sequence numbers for the parts of the disclosure which bracket the hiatus in the question will be compared. In the example, these would be the numbers associated with the process descriptors for the E and G items.

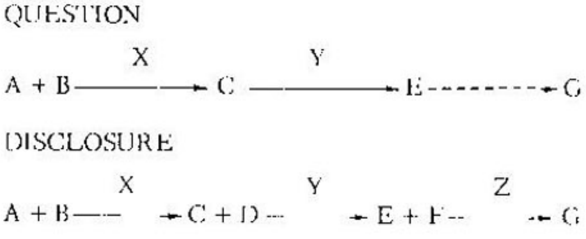
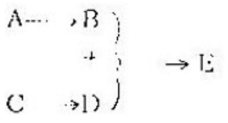


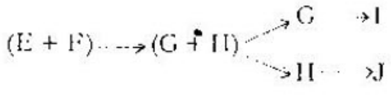
Figure 3

The "interlocking effect," is illustrated by compositions II and III of Figure 2. If a question identical to this disclosure were asked, the sequence numbers would be omitted. Since compositions II and III are involved in two steps of the process, Composition II would have the effect of positively linking together Compositions I and III, and Composition III would have the effect of linking together Compositions II and IV. This being the case, the question need not specify the time sequence numbers.

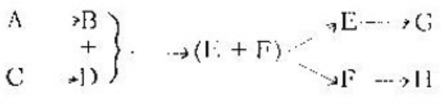
Straight linear processes are not the only ones which can be treated by the system. Thus, branching processes which are convergent, divergent or mixed, may be handled. An example of a ***convergent process*** would be one in which A makes B, independently C makes D and then B and D are reacted to make E.



A ***divergent process*** might involve a reaction of E and F to form G and H, separation of G from H, and separate reactions of G and H to yield I and J respectively.

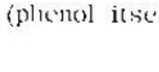
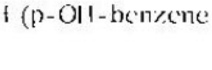


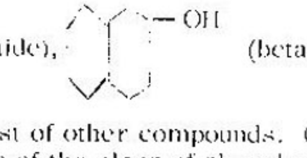
A ***mixed process*** might be one of the type:

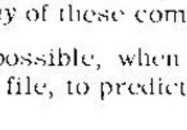


THE TOPOLOGICAL STRUCTURE SEARCH

Structure searching is often performed in terms of a fragment, large or small, which denotes a class of compounds. This is important in patent searching for two reasons. Many generic concepts are defined in the claims of a patent application in terms of chemical structure. In addition, structural fragment searching enables one to find equivalents of claimed specific compounds. For example, the class of compounds containing the group

 includes, among others, 

(phenol itself),  (p-OH-benzene

sulfonamide),  (beta-OH-naphthalene)

and a host of other compounds. One searching for members of the class of phenols would be satisfied by finding *any* of these compounds.

It is not possible, when encoding documents to make up the file, to predict which groups of atoms will be the ones later sought. For some searchers the common functional groups or radicals *may* not be of any more importance than other groups for which there is not a well known designation. It is therefor necessary to provide for making an atom by atom search to permit the searcher maximum flexibility in designating his particular interest.

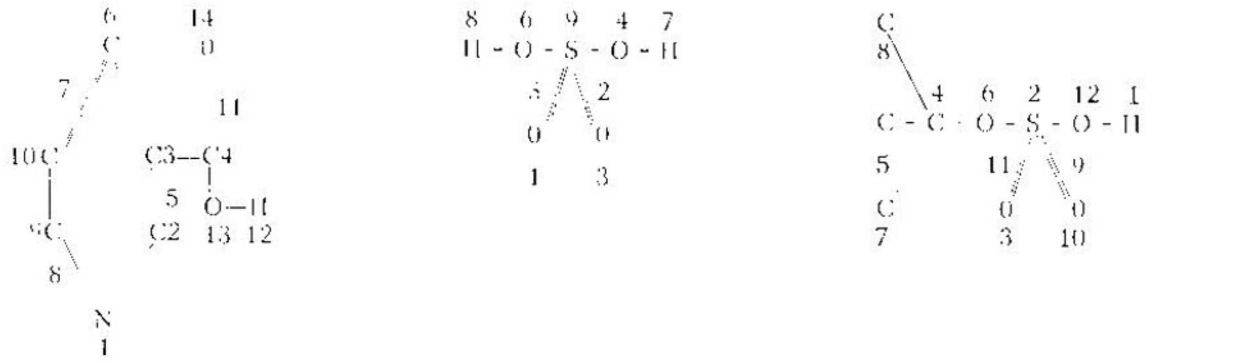
The method of coding structures employed in this system is equally applicable to organic and to inorganic compounds. Each atom (except non-significant hydrogens, which are the ones attached to the carbons of the hydrocarbon skeleton) and each significant bond (bonds other than single) in the structure is assigned an arbitrary serial number or "interfix."[2] Consecutive numbers beginning with 1 are assigned to the bonds and elements in a completely random manner. The complete notation for each atom or bond coded is contained in one computer word and includes a symbol identifying it (e.g., by atomic number), its serial number and the serial number assigned to each of the atoms or bonds to which it is connected. Both questions and disclosures are similarly coded. The serial numbers assigned to the bonds and elements in a question will seldom be the same as those assigned to a corresponding structural fragment of a disclosure compound containing the fragment sought. However, correspondence of numbers in this respect is not required by the computer program, as will be explained further.

Some exemplary structures and their codes are shown in Figure 4. A question asking for all compounds containing the sulfate radical might have numbers assigned as shown in Figure 5.

The search would result in the retrieval of two of the compounds shown in Figure 4, as well as all other compounds in the file which contain the sulfate radical. Although the codes asked and those found differ from each other, they are recognized as being topologically equivalent, that is, the several equivalent codes represent fragments containing like pieces with like connectivities.

During the search, the machine inspects the first atom listed in the question code and examines each of the atoms in the disclosure code being considered until it finds a match. The atoms connected to this question atom are then determined and their correspondence to those connected to the first acceptable disclosure atom are compared until further matches are found. This continues until either all of the bonds and atoms in the question, together with their proper connective relationships are matched in the disclosure compound, or until it is determined that no such total match is possible. Note that the carrying forward of these comparisons is based upon identification of elements and bonds and the identification of the adjacent elements and bonds. The arbitrary serial numbers act merely as guides to the machine in progressing from piece to piece in the two structures being compared.

[2] A previous report describes another coding method employing interfixes (2).



NICOTINIC ACID SULFURIC ACID MONO-TERTIARY BUTYL SULFATE

6 C 14 O 7 11 10 C C3—C4 5 O—H 9 C C2 13 12 8 N 1
8 6 9 4 7 H - O - S - O - H 5 2 O O 1 3
C 8 4 6 2 12 1 C - C - O - S - O - H 5 11 9 C O O 7 3 10

Serial No.	Connec-tions	Symbol	Serial No.	Connec-tions	Symbol	Serial No.	Connec-tions	Symbol
1	2-8	N	1	5	O	1	12	H
2	1-5	C	2	3-9	=	2	6-9-11-12	S
3	4-5-6	C	3	2	O	3	11	O
4	3-11-13	C	4	7-9	O	4	5-6-7-8	C
5	2-3	-	5	1-9		5	4	C
6	3-7	C	6	8-9	O	6	2-4	O
7	6-10	-	7	4	H	7	4	C
8	1-9	=	8	6	H	8	4	C
9	8-10	C	9	2-4-5-6	S	9	2-10	=
10	7-9	C				10	9	O
11	4-14					11	2-3	=
12	13	H				12	1-2	O
13	4-12	O						
14	11	O						

Figure 4

Serial No.	Connec-tions	Symbol
1	2 - 4	–
2	1	0
3	4	0
4	1 - 3 - 5 - 6	S
5	4	0
6	4 - 7	–
7	6	0

SULFATE RADICAL

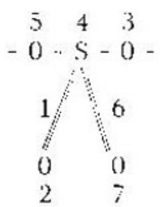


Figure 5

The use of these numbers in tracing through a path is carried on independently in each of the two structures, thus fulfilling the search despite the use of completely arbitrary and inconsistent numbers in each occurrence of the code for a particular structure.

In structure searches conducted on SEAC, which is considerably slower in operation than many of the newer and larger computers, a file of 250 complex steroid structural formulas was completely searched in 8 minutes. Further experiments to determine the best method of asking a question have shown that it is possible to lower this time to approximately 8 seconds.

AUXILIARY TASKS WHICH WILL BE DONE BY THE MACHINE

For the future, in addition to performing the search, it is planned to take advantage of the computer's ability to make logical decisions and to perform repetitive clerical tasks rapidly and without error in accomplishing certain other operations, as follows:

(1) The encoding of disclosures can be facilitated by the machine in several ways. As a disclosure item is encountered for the first time an "index card" will be prepared which will contain all syn-

onyms for the item, a picture of the structural formula (if known), the codes for index numbers, codes for the empirical formula, codes for those chemical or functional descriptors which are "fixed" (in the sense that they will always be coded with the item) and codes for the structure details. The empirical formula will be generated by the machine when the index cards are prepared and this will be computed from the topological structure code. Punched cards corresponding to the index cards will be made for use in the preparation of the tapes on which the data is finally stored.

In preparing an abstract, the abstractor will manually write out an indication of the organization of the various levels of disclosure, the index numbers of the items included in various compositions, indications of the compositions in the various processes, the "accidental" descriptors (accidental in that they are to be coded for particular items only because of the environment disclosed for that item in the particular document being treated), and indications of alternativeness, negativeness, conjunction and other relationships. The subject matter contained in the abstract will then be encoded and transcribed onto punched cards.

To compile the complete code for the disclosure, the machine will be programmed to record on magnetic tape all codes derived from the abstract and to intersperse in all proper places the codes from the punched cards representing the index cards which have been previously prepared. The tape so prepared will then be processed by the machine according to an "editing" program to generate and insert into the code, data for the mechanical screens. Following the editing routine, the encoded data will be checked for mechanical and certain logical errors by use of still another computer program.

(2) Questions to be searched will be compiled, edited and error checked by the machine in a manner similar to that used in preparation of the disclosure file.

(3) Use can be made of the machine in dividing the disclosure file into smaller files, each of which would take less time to search than the entire file. As will be explained, this would not be a separation of documents into collections based on their general subject matter, as is the effect obtained in conventional classifications, where one collection of documents might relate to domestic clothes washing machines and another collection might relate to ion exchange resins. Patent searchers are well aware of the situation in which some of the documents in the washing machine collection disclose synthetic resins used in the protective coating for the casing of the machine, and these resins may be chemically identical with a resin claimed in a patent application for use as an ion exchanger.

A scheme to permit the searching of files smaller than the total collection would begin operations with a single, all-inclusive file. Records would be kept of questions actually asked of the system. As the total file grows by the incorporation of additional documents into the system, the records of previously asked questions would be analyzed to determine frequently asked types of subject matter. It would then be possible to set up a separate collection of encoded documents dealing with, for example, ion exchange resins, by using this subject matter for a search question and having the machine copy off onto another tape all documents in the general file which disclose such subject matter, regardless of the general nature of the disclosure of each of the documents copied. Note that the copied documents still remain in the original file and continue to be available for search on the basis of other subject matter which they disclose. All future searches for particular ion exchange resins would be made only on the smaller file with the assurance that all disclosures in the total collection which are pertinent had been investigated.

(4) In the main searching operation, it would be desirable to avoid repeating searches already made. If search questions previously asked were filed together with identifications of documents found in response to such questions and the identification of the last document included in the file when the previous search was made, the machine could search the "previous question" file before inspecting the main disclosure file. If the present search question is found to be a repetition of one asked previously, the machine would print out the previously found results and then update the search by making the ususal search of those documents in the general file which had been added since the earlier search was performed. Where inspection of the "previous question" file reveals that the question has not already been asked, the entire general file is searched, and the question and the results of the search then added to the "previous question" file.

(5) The previously mentioned plans for permitting the machine (a) to vary the scope of the search question and (b) to seek a valid combination of references which together may represent the answer to a complex question, will be investigated. In many cases, operation of the machine in such a manner could avoid the necessity of making successive searches on each of a series of closely related questions.

(6) It is contemplated that future programs will make use of "look-up" tables. For example, it is expected that there will be in the file many documents which disclose the same generic concept together with a list of details or specific concepts, such as the term "halogen" followed by a recital of each of the elements which make up this class. It would be convenient if, in each of such disclosures, only the general term were coded together with an indication of where the table specifying what is implied could be found. The table would be stored but once and only where the details are important need the machine refer to them. Con-

siderable storage space could be saved by such a device.

(7) Two additional types of searches made in the Patent Office could be performed by the use of systems analogous to those described here for making searches of chemical subject matter. The first is the so called "interference search," which is made by the patent examiner immediately prior to "allowance" (notification to the inventor of the patentability) of an application. Its purpose is to determine whether any other pending application is directed to the subject matter claimed in the allowable case. If one is, a proceeding called an "interference" is instituted to determine which of the applicants is entitled to the patent. The interference search is limited to pending applications, but since well over 200,000 applications are now pending before the Patent Office it is no easy task to make a complete and conclusive interference search. Encoding the subject matter of each pending application and searching a file so derived with the subject matter of an allowable application as the question could be accomplished by a mechanized system on all fours with the system here described for making patent searches.

An additional system would be one for locating legal decisions pertinent to the prosecution and examination of a particular application. Such a system could be set up to permit finding such judicial statements in terms of the legal principles involved in a particular case as well as on the basis of the technological subject matter of the application or patent involved in the litigation which gave rise to the decision.

ACKNOWLEDGMENTS

The contributions of Mrs. E. C. Marden, Mr. L. C. Ray, and Mr. R. A. Kirsch, of the National Bureau of Standards, to many phases of the work described in this paper are gratefully acknowledged.

REFERENCES

(1) Newman, S. M.--"Problems in Mechanizing the Search in Examining Patent Applications"--Patent Office Research and Development Reports, U. S. Patent Office, Washington 25, D. C., 1956.

(2) Lanham, B. E., Leibowitz, J. and Koller, H. R.--"Advances in Mechanization of Patent Searching"--Presented before the Division of Chemical Literature, 129th Meeting of the American Chemical Society, Dallas, Texas, April 11, 1956. Reprinted in J. Patent Office Society, 38, 820-838, December 1956.

54048--U.S.Dept.of Comm--DC--1957