Reprinted from SCIENCE, October 25, 1957, Vol. 126, No. 3278, pages 814–819.

Finding Chemical Records by Digital Computers

Louis C. Ray and Russell A. Kirsch

Reprinted from SCIENCE, October 25, 1957, Vol. 126, No. 3278, pages 814–819.

Finding Chemical Records by Digital Computers

Louis C. Ray and Russell A. Kirsch

The National Bureau of Standards and the United States Patent Office are actively collaborating in a long-range program to develop and apply automatic techniques of information storage and retrieval to problems of patent search. An important preliminary phase of this program has been the carrying out of experiments with methods for locating information in large files of technological and scientific information.

In the granting of United States patents, it is necessary for patent examiners to refer to collections that may, in principle, contain from 10^6 to 10^7 documents. When an examiner conducts a literature search to determine whether a patent application represents a novel idea, which then must be tested against established criteria for patentability, he must search insofar as possible through all literature in the public domain that might possibly contain any information pertinent to the given application. It has been estimated that 60 percent of the time spent by an examiner in processing a patent application is devoted to searching the technical literature. In an attempt to reduce this expenditure of time, the National Bureau of Standards–Patent Office group has considered, among other techniques, the use of automatic data-processing systems.

By an automatic data-processing system (ADPS) is meant a collection of machines, usually but not necessarily electronic in nature, which have the ability to process information in accordance with internally stored programs and which can perform a whole data-processing task involving the use of data-storage facilities of diverse natures without the necessity for manual intervention. The system also includes devices for the preparation of input data and the reproduction of output data. SEAC, the NB*S* *E*lectronic *A*utomatic *C*omputer, is an automatic data-processing system; it has been used in successful preliminary experiments wherein a collection of over 200 descriptions of steroid compounds is exhaustively searched to answer typical questions that may occur in evaluating patent applications for new chemical compounds. This article (*1*) describes some theoretical ideas on the use of automatic data-processing systems for literature searching; these ideas have resulted from experiments in searching through chemical information.

In considering any attempt to automatize the searching of technical literature in the U.S. Patent Office, it must be remembered that the historical nonautomatic or manual method of searching which is presently in effect at the Patent Office utilizes the best intellectual efforts

The authors are on the staff of the Data Processing Systems Division, National Bureau of Standards, Washington, D.C.

of some 8000 examiners highly trained in diverse technologies and in the legal aspects of isolating significant information from large technical files. Consequently, there is no a priori reason to believe that there is any single over-all solution to the literature-search problem in the Patent Office which will function as effectively as this trained corps of examiners. With this consideration in mind, one area of the inventive arts was selected for initial experimental investigation in the hope that empirical solutions in that area could be put into productive operation. A bonus from the solutions has been the development of theoretical and experimental techniques which should prove applicable to the searching of technical literature in other areas with which the Patent Office is concerned.

The area selected for initial experimental investigation was that of "Composition of Matter"—that is, patents generally concerned with what may loosely be classified as chemistry. This area was recommended for initial investigation by the Advisory Committee on the Application of Machines to Patent Office Operations (*2*). Chemists have for a long time been concerned with information retrieval, and it was hoped that use could be made of some of the techniques that the chemists have developed. As it turned out, the experimental results obtained took advantage of a technique of chemistry that, historically, was probably not developed for the purpose of information retrieval—namely, the use of chemical structural diagrams for describing the chemical nature of matter.

It is of paramount importance to realize that in using automatic techniques for the retrieval of technical information, no more information can be obtained from a file than that information which is represented in the file according to a well-defined notational scheme. Because the method of representing chemical structures in diagrammatic form has just such properties, it was decided to experiment with the use of SEAC for searching through files of chemical structure diagrams in response to search requests fed into the machine.

Fig. 1. Fragment structure.

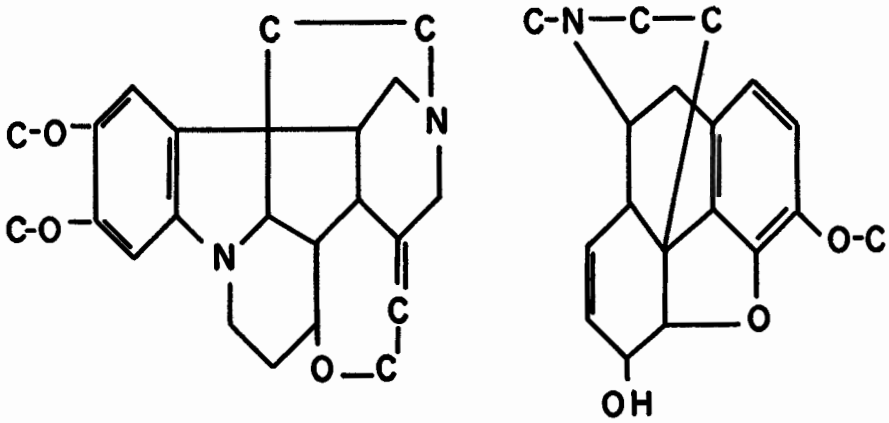

Fig. 2. (Left) Structure of brucine. (Right) Structure of codeine.

Searching Chemical Structures

Structure search problem. In the Patent Office, the examiners in the chemical arts have frequent need for performing generic searches through structure diagrams. As an example, it can be seen that the fragment structure which is given in Fig. 1 is contained in the two compounds brucine and codeine shown in Fig. 2. Unmarked vertices in the structures are understood to represent carbon atoms (C). Notice that the six-member ring with the nitrogen (N) occurs in codeine even though the diagram of codeine indicates the ring in a distorted manner. We can say that the two compounds shown in Fig. 2 share the generic property of containing this fragment. Some of the experiments performed on SEAC were concerned with developing a method for performing generic searches of this type through a file of structures taken from the art of steroid chemistry.

Structure search routine. The Patent Office search requires an unambiguous coding system in which any combination of atoms and bonds can be represented for purposes of retrieval. The traditional coding system (*3*) is not suitable for mechanized search because a given compound can be represented in conceptually different ways and because the system is so complex that it can be used only by a trained chemist. Opler (*4*) of the Dow Chemical Company has developed a code for use in machine searching. This code is flexible, but it is not suitable for Patent Office searches because it does not represent the most fundamental units of the chemical structure, the atoms and their bonds which are directly required in many typical patent searches. An example of a code suitable for machine searching was described by Mooers in the "Zatopleg" (*5*) system of ciphering structural formulas. Mooers' method of representing compounds provided the basis for representing the input data in the SEAC structure search routine described below. Methods for actually searching such data had to be developed.

In the system used on SEAC, each atom in a structural diagram is numbered serially in arbitrary order. One unit of computer storage, called a word, is given to each atom to represent its position in the structure. In each word are listed the numbers of the other atoms, up to four, that are attached to the atom represented by the word. The element symbol and the serial number of the atom are also placed in the word. Thus each atom word has six fields; the serial number of the atom, four connections fields, and an element symbol.

As an illustration, consider the compound, chloral, shown in Fig. 3. The coding would proceed as follows:

First, the atoms and all bonds other than single bonds are numbered in any arbitrary order as shown in Fig. 4. Then, a list of the connections to each component of the structure is made, as shown in Table 1. Finally, the element symbol of the atom the word represents is put in each word, as shown in Table 2.

The list (Table 2) represents the complete code for the structure. The struc-

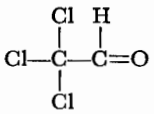

Fig. 3. Structure of chloral.

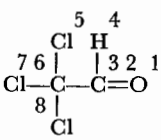

Fig. 4. Structure with atoms and double bond numbered.

ture may be easily drawn by using the code. The code for any structure is not unique since, by numbering the atoms in some other arbitrary order, a different code would be obtained. It may easily be seen, however, that all the possible codes are equivalent.

It is desired to search a file of coded structures for all structures which are identical with some compound in question or which have some generic property in the sense previously defined. To do this, the SEAC search program tried to make an atom-to-atom match between the atoms of the structure in question and the atoms of the first structure recorded in the file. Each match that is made is considered as tentative by the program until the search through the first file structure is completed. Whenever failure to match is discovered by the program, the program tries to go back to the previous match to make a new match. If the program finds that all possible first matches lead to irreconcilable mismatches, it will reject the first file structure and proceed on to the next. When a one-to-one correspondence exists between each of the atoms of the question and the atoms of part of the file structure that is being examined, the routine accepts the structure by printing on the computer output an indication of which structure was found. The search routine continues this process until the whole file has been searched.

The details of the search routine are given in the flow chart in Fig. 5. The symbols used in the chart are defined as follows: I, II, III, and IV denote the four connection fields in each atom word; Q_i is a question atom word; F_b is a file atom word; N_i is a temporary storage location for question atom words matched with corresponding file atom words (R); R_i R_j is a temporary storage location for file atom words matched with corresponding question atom words (N); α denotes fields of R (it can equal I, II, III, or IV), and β denotes fields of N (it can equal I, II, III, or IV). Standard flow-chart terminology in computer program is used.

Use of Screens. As fast as a high-speed electronic computer is, the fact remains that performing a detailed search of the type described would be altogether too time-consuming unless some short-cuts could be devised which would in no way compromise the exhaustiveness or accuracy of the search, while speeding up the process greatly. A technique is needed that will enable the automatic data-processing system to perform what is for the machine a cursory inspection of small pieces of data in such a manner that most structures that will not satisfy the search requirement will be rejected immediately. Such a technique is called a "screen" or a "screening device." It is essential that a screening device should never cause a structure to be rejected which does, in fact, meet the search requirement. It is acceptable, however, if the screen allows some structures to be considered further by the structure search routine even though they are subsequently rejected as failing to meet the search requirement.

By now, it will have become obvious to any chemist that one such useful screen is inherent in the empirical formula of a chemical structure. In other words, by storing in the file, along with a description of the chemical structure, a list of the number of occurrences of each atom in the structure, it is possible to find out whether there are enough atoms of the right type present to satisfy the search requirement. This screen was incorporated into the SEAC search program and on most searches it enabled the computer to reject quickly the vast majority of structures that would otherwise have been rejected only after a long computational procedure.

Other screens that were considered but not experimented with would make use of topological properties of the chemical structures. One such simple screen specifies the number of rings within the structure. Obviously, if one were searching for some generic structure having, say, three benzene rings, a structure having even 500 atoms in it could not contain the one being searched for unless there were at least three rings somewhere within the 500-atom structure.

Other topological properties that could serve as screens would be the longest non-self-intersecting path that could be traced through the structure, and a count of the number of atoms with each of the possible valences. However, these topological properties were not investigated experimentally.

The important properties to be sought for in devising screens for any type of searching, chemical or otherwise, are that the screens must be substantially independent of one another and that they must have universal applicability to all documents in a collection. These prop-

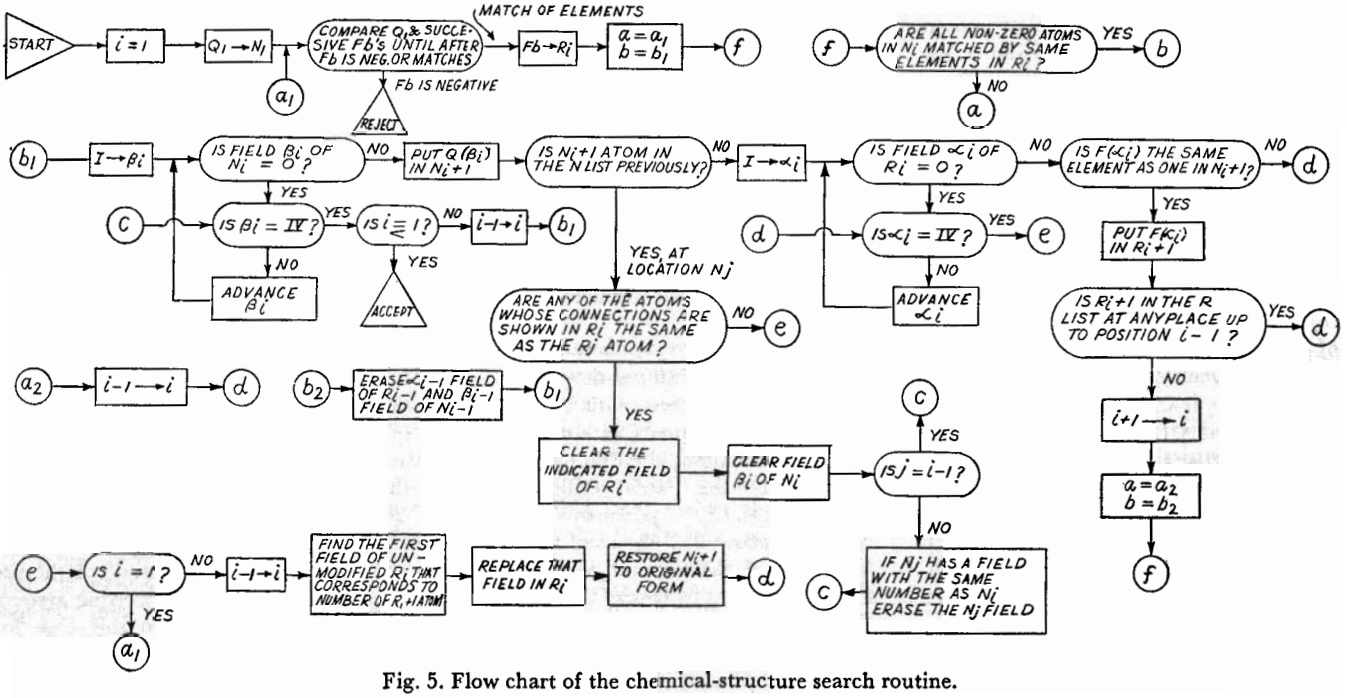

Fig. 5. Flow chart of the chemical-structure search routine.

erties are never completely achieved in practice, but they can be approached when the search is limited to narrowly defined subject matter.

Grouping of Information. In deciding upon a method for representation of structure diagrams for search by the automatic data-processing system, it was necessary to decide what data about a structure diagram would be represented in the code as it was stored in the machine file. Certainly an atom-by-atom description would give a completely general and flexible method for representing chemical information. But a difficulty arose in certain of the chemical arts where the distinction between one chemical and the next is very slight in the atom-by-atom description. On the other hand, very useful discrimination among the members of such a class could be made by including in the description certain characteristics that were peculiar to that particular class. The problem here is one that is undoubtedly a characteristic of systems for coding information wherever the information assumes diverse forms—namely, how much generality shall be sacrificed in order to provide a high degree of discrimination in certain isolated areas of the collection where most of the documents contain similar information.

A solution to the problem that is practical if ADPS searching is to be used, but which is probably not practical for use with simpler systems (for example, punched cards) is to have special coding systems applicable for certain documents in a collection. These coding systems serve as adjuncts to one (or more) coding system which describes the entire collection. If, in searching through a collection, the screening devices fail to reject a document for which a special coding system is applicable, then special instructions are automatically made available to the automatic data-processing system to enable it to search the codes for a special document in terms of the special coding system being used. It is quite practicable, when searching with an automatic data-processing system, to have documents scattered through a collection that have been described with special coding systems not generally used for the whole collection. This technique becomes costly of machine searching time only when the number of such special documents is so large that it becomes necessary for the machine system frequently to call up its auxiliary instructions to handle the special coding situation.

With these considerations in mind, some special coding methods suggested by the Patent Office were used for coding structures in the steroid chemical art. These coding rules permitted fine distinctions to be made between various compounds in the steroid art, even

Table 1. List of connections to each component of the structure shown in Fig. 4.

Component No.	Connections
1	2
2	1, 3
3	2, 4, 6
4	3
5	6
6	3, 7, 8, 5
7	6
8	6

Table 2. Complete code for the structure shown in Fig. 4.

Component No.	Connections	Element symbol
1	2	O
2	1, 3	=
3	2, 4, 6	C
4	3	H
5	6	Cl
6	3, 7, 8, 5	C
7	6	Cl
8	6	Cl

though the coding rules depended upon certain chemical structure properties which were quite peculiar to the steroid art and, therefore, inapplicable for searching other types of chemical structures.

Use as an Information-Retrieval System for Other Purposes

The experiments and theoretical considerations thus far described have been concerned with the use of the automatic data-processing system as a tool for actually performing a search through a file of information (in this case, chemical in nature). However, the use of automatic data-processing systems in retrieval of information extends considerably beyond the activity of actual searching. This section describes the use of such machine systems for processing data as part of an over-all information-retrieval system.

Checking data. It is evident that the file of information through which a search will be performed must be prepared without error, for any error in a recorded piece of information represents the loss of some information from the file. In the experiments with retrieval of chemical structures, the original file was prepared by the Patent Office in the form of about 2500 punched cards that described about 250 chemical structures. Since one of the features of the coding scheme used was its lack of reliance upon chemical knowledge for encoding structures, a group of punched-card typists were given the set of 250 pictures describing the structures to be encoded. The operators read the pictures and punched the descriptions on cards without the intervention of any supervision from a chemist. As was to be expected, there were some cards out of the 2500 that contained errors, and the problem was to find the errors and correct them.

Since the data were ultimately to be used as input to SEAC, they were first transcribed from punched cards onto magnetic wire (which was the principal SEAC input-output medium at the time of the tests described). Then a program was written for SEAC to take these data from the wire and check them. It is important to note that this data-checking program was entirely unrelated to the subsequent SEAC search program. The data were checked for internal consistency and for their adherence to the coding rules that had been established. The result was that about 50 punched-card errors were caught and that a copy of those parts of the file containing no errors was produced. This expurgated file could then have been used by another machine. If the coding system had been of suitable nature, other simpler mechanisms could have done the searching through data that had been checked by the automatic data-processing system.

Here, then, is an example of the use of an automatic data-processing system as part of an information-retrieval system although the machine system need not serve as the searching device. In preparing a large file for occasional use, where the cost of a large machine system would not justify its use as a searching tool, the machine system might still profitably serve the function of checking the initial data to be entered into the system.

Transliteration. Another function that an automatic data-processing system can serve as part of an information retrieval system is that of transliterating data to be entered into the system from the forms that are most convenient for manual preparation of the data into forms more suitable for searching by machine. In the structure search experiment, the data to be used by SEAC had to be arranged in a certain format for most efficient utilization of the SEAC memory space. This format was sufficiently complicated that the punched-card operators could not be expected to prepare the data in that format with any reasonable speed or accuracy. Consequently, a program was written for SEAC which accepted data in a format convenient for the punched card operators and transliterated the data into the form desirable for SEAC searching.

Again we have an example of the use of an automatic data-processing system, not for searching, but in this case for transliterating data from one form into another. It should be noted that both the

input data for this transliteration program and the output produced by the machine followed completely rigorous rules of organization and arrangement. The automatic data-processing system was converting data from one well-defined coded form into another. The fact that this can be done readily with an automatic data-processing system should not lead one to the conclusion that data expressed in some natural language (for example, English) can be *translated* by machine into a coded form suitable for machine search. The problem of *machine translation* is a formidable one to which much effort is being devoted (*6*) both in the United States and elsewhere. What is claimed here is that the comparatively simple problem of *transliterating* from one code into another can be conveniently handled by an automatic data-processing system.

Another example of transliteration by SEAC which is the subject of some current experiments is its use for generating chemical structure descriptors for use by a simpler searching machine. It has been suggested by Mooers (*7*) that, for purposes of retrieval, complex structures such as chemical diagrams can be represented in terms of a list of, say, all the triples of atoms and bonds occurring within the structure. Thus, chloral (Fig. 3) would be described as consisting of combinations of the triples in the following list: Cl—C; C—C; —C—; —C=; C—H; C=O.

A simple search mechanism performing simple comparisons between similar types of triples in the search request and the file could retrieve complex structures without the necessity of doing the complex processing of data of the type performed by the structure search routine described earlier. It is not within the scope of this article to discuss the merits of such a search system. However, this example is mentioned to demonstrate that an automatic data-processing system can generate the *N*-tuple descriptors from a complete representation of the structure. If a large file of chemical structures is to be searched by a very simple mechanism using Mooers' *N*-tuple descriptors, the original file may be prepared in, for example, the form required for the SEAC structure search routine. It is then a fairly straightforward job to program SEAC to generate the *N*-tuple descriptors and consequently to produce as output a transliterated file all prepared for searching by more simple mechanisms.

Some present SEAC experiments are devoted to generation of the *N*-tuple descriptors from the file of steroid compounds previously described. It is intended, then, to run comparative tests on a file of chemical structures coded according to the two coding schemes. Any results obtained from such a comparison will be less significant, however, than the fact that such an experiment demonstrates the feasibility of simulating on a large automatic data-processing system the searching procedure of a simpler mechanism for purposes of comparative evaluation.

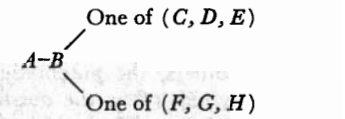

Fig. 6. Structure with alternatives.

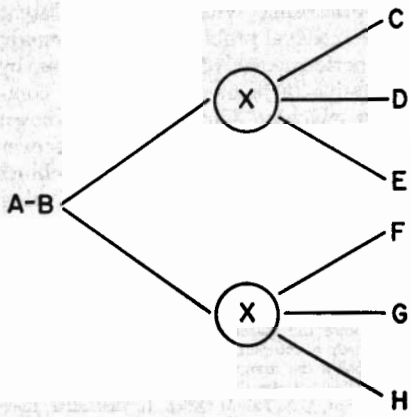

Fig. 7. Intermediate dummy element.

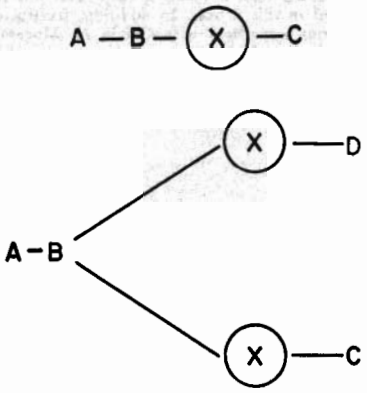

Fig. 8. (Top) Structure included in Fig. 7. (Bottom) Structure not included in Fig. 7.

Exploration of complex logical situations. In retrieval of information from Patent Office files, it often occurs that complex logical conditions must be imposed upon either the question or the file. A simple example of the way in which an automatic data-processing system can handle complex interrelationships occurs in the search for alternative patent disclosures.

It is quite common for an inventor to describe his invention in the following form: *A* plus *B* plus any one of the following three, *C, D,* or *E,* plus any one of the following three, *F, G,* or *H.* The elements *C, D,* and *E* are alternatives for each other, as are the elements *F, G,* and *H,* and any one of the first group may possibly be combined with any one of the second group for purposes of anticipating a subsequent patent claim. However, if someone were to claim the combination $A+B+C+D$, the patent described would not be relevant since *C* and *D* are disclosed only as alternatives for each other. It is desirable to be able to use a search machine that will not accept a patent like the first one described when searching for the latter.

One solution to this problem that has been proposed but does not appear to be practical is to code the alternative type of disclosure separately as each of the several types being described. Thus $A+B+$ one of (C, D, E) + one of (F, G, H) would be represented by $3\times3=9$ separate entries. In many real situations a number in the thousands would describe the number of possible combinations claimed in a chemical patent.

Another method for handling this type of situation on SEAC was suggested by the Patent Office; this method made use of the structure search routine. By the introduction of certain dummy elements into the chemical structure, several alternative structures could be coded in one large pseudo-structure. If the alternative structure of Fig. 6 is to be coded, it can be represented as shown in Fig. 7, where X (inside the circle) represents a dummy element. It can be seen that Fig. 8*a* is contained in the pseudo-structure of Fig. 7, but that Fig. 8*b* is not.

The handling of this complex type of alternative search is possible on an automatic data-processing system but difficult on a simpler mechanism. In a large information-retrieval system it may be possible to use simpler mechanisms than an automatic data-processing system for searching until a complex logical situation such as the alternative search arises, in which case the file may be made available to the automatic data-processing system for more complete searching.

There are other complex logical situations that arise in Patent Office searching for which it is not yet possible to announce experimental solutions. One particularly difficult one occurs when there is a reference that is complete except for a minor substitution of some component *A* for the desired component *B* and when, in an entirely separate patent, there is a statement attesting to the equivalence of *A* and *B* for the function concerned. It is often desirable to retrieve such a partly incomplete reference in conjunction with the reference stating equivalence. To date, however, no general solution to this problem is known.

Conclusions

The problems in information retrieval mentioned here have certainly been known to serious workers in the field for some time. Only recently, however, have

automatic data processing systems become sufficiently available to be considered as possible tools in an information-retrieval system. Thus the SEAC experiments indicate the practicality of using automatic data-processing systems for scanning a file of information at high rates. However, many mechanisms considerably simpler than an automatic data-processing system can also do such scanning, and the question remains open about the comparative advantages of an automatic data-processing system and simpler mechanisms for the actual process of looking at a properly organized file. In some retrieval situations, most notably in the Patent Office, the problem is of sufficient magnitude and complexity that the power of an automatic data-processing system to do more than just scan a file appears at first inspection to be a requirement. Where the machine system seems to offer a unique contribution is in the off-line jobs. For such functions as preparing a search prescription, editing a file, eliminating errors, transliterating from one code to another, exploring complex logical conditions imposed on the question and file, and probably many others, the automatic data-processing system offers the outstanding virtues of high speed and great versatility. Thus it is possible to use SEAC not only to test the utility of an automatic data-processing system for the Patent Office retrieval problem but also to study the performance of other devices by simulating them on SEAC. In the computing machine field it is a well-known phenomenon that machine users discover many new applications of these machines while they are in the process of using them. It is hoped that the experiments on SEAC will serve a similar purpose.

References and Notes

1. Because the search of Patent Office literature is such a complex task, we found particularly valuable the many informative discussions with members of the Office of Research and Development, U.S. Patent Office. In particular, their guidance helped to identify specific problem areas whose characteristics might be suited to machine processing and provided a background which led to a number of ideas that are contained in this article. In addition, fruitful discussions were held with Calvin N. Mooers of the Zator Co., whose helpful comments were derived from experience with related concepts in the field of information retrieval. The incentive to start on the computer search of chemical structures arose out of discussions with Ascher Opler of Dow Chemical Co. regarding his pioneering efforts with such techniques. Members of the Data Processing Systems Division of the National Bureau of Standards who contributed to the development of the results described here are Catherine E. Lester and Ethel C. Marden, who wrote the data-editing programs for SEAC, and Mary E. Stevens and Edwin K. Woods.

2. "Report to Secretary of Commerce by the Advisory Committee on Application of Machines to Patent Office Operations," 22. Dec. (U.S. Dept. of Commerce, Washington, D.C., 22 Dec. 1954), p. 37.
3. G. Malcolm Dyson, *A New Notation and Enumeration System for Organic Compounds* (Longmans, Green, London, ed. 2, 1949); Anonymous, "Which notation," *Chem. Eng. News* 33, 2838 (1955).
4. A. Opler, "A topological application of computing machines," Proceedings of the Western Joint Computer Conference, 7–9 Feb. 1956, pp. 86–88.
5. C. N. Mooers, *Ciphering Structural Formulas —the Zatopleg System* (Zator Co., Cambridge, Mass., 1951).
6. W. N. Locke and D. Booth, *Machine Translation of Languages* (Technology Press, Cambridge, Mass., 1955).
7. C. N. Mooers, "Information retrieval on structural content," in *Information Theory* (Academic Press, New York, 1956), pp. 121–134.