PB 161586

NBS

Technical Note 85

A SURVEY OF COMPUTER PROGRAMS FOR CHEMICAL INFORMATION SEARCHING

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

THE NATIONAL BUREAU OF STANDARDS

Functions and Activities

The functions of the National Bureau of Standards are set forth in the Act of Congress, March 3, 1901, as amended by Congress in Public Law 619, 1950. These include the development and maintenance of the national standards of measurement and the provision of means and methods for making measurements consistent with these standards; the determination of physical constants and properties of materials; the development of methods and instruments for testing materials, devices, and structures; advisory services to government agencies on scientific and technical problems; invention and development of devices to serve special needs of the Government; and the development of standard practices, codes, and specifications. The work includes basic and applied research, development, engineering, instrumentation, testing, evaluation, calibration services, and various consultation and information services. Research projects are also performed for other government agencies when the work relates to and supplements the basic program of the Bureau or when the Bureau's unique competence is required. The scope of activities is suggested by the listing of divisions and sections on the inside of the back cover.

Publications

The results of the Bureau's work take the form of either actual equipment and devices or published papers. These papers appear either in the Bureau's own series of publications or in the journals of professional and scientific societies. The Bureau itself publishes three periodicals available from the Government Printing Office: The Journal of Research, published in four separate sections, presents complete scientific and technical papers; the Technical News Bulletin presents summary and preliminary reports on work in progress; and Basic Radio Propagation Predictions provides data for determining the best frequencies to use for radio communications throughout the world. There are also five series of nonperiodical publications: Monographs, Applied Mathematics Series, Handbooks, Miscellaneous Publications, and Technical Notes.

Information on the Bureau's publications can be found in NBS Circular 460, Publications of the National Bureau of Standards ($1.25) and its Supplement ($1.50), available from the Superintendent of Documents, Government Printing Office, Washington 25, D.C.

NATIONAL BUREAU OF STANDARDS

Technical Note

85

FEBRUARY 1961

A SURVEY OF COMPUTER PROGRAMS FOR CHEMICAL INFORMATION SEARCHING

Ethel C. Marden and Herbert R. Koller

DISTRIBUTED BY

UNITED STATES DEPARTMENT OF COMMERCE

OFFICE OF TECHNICAL SERVICES

WASHINGTON 25, D. C.

Price $2.25

TABLE OF CONTENTS

ABSTRACT

The authors describe twelve computer systems for searching chemical literature. Preceding discussion of the computer systems, a brief description is given of different chemical notation systems, indexing and abstracting procedures, punched card systems (which were the forerunners of the computer systems), and special purpose literature searching machines. A short discussion of the difficulties (linguistic and other) attendant to literature searching terminates the paper.

A SURVEY OF COMPUTER PROGRAMS FOR CHEMICAL INFORMATION SEARCHING

by

Ethel Marden and Herbert R. Koller

INTRODUCTION

Chemists, more than anyone else, have been concerned with the need for mechanical aids to literature searching. They have evidenced their interest by the publicity given over the years in their literature to the efforts made in the notation, indexing, and abstracting areas and, in later years, to the various projects for literature searching by means of punched card machines, special purpose machines of several categories, and computers. Workers in the chemical area have been made aware of the need for mechanical aids because the chemical literature has been increasing at a more rapid pace than most of the other technological literature, in an era which has seen an explosion of technical publications in all fields. Chemists particularly have been aware of some of the problems which are encountered in setting up systems for mechanical literature searching, which range from those of a linguistic nature, through orientation of the planned-for system to the specific objective desired, to considerations of requisite equipment. They have sought solutions, both by encouraging the attempts of others and, in some cases, by giving up their own professional careers in the field of chemistry to join the workers in the area of documentation.

Recognition was recently taken by a Senate Committee of the importance of scientific information retrieval, with its allied problems in related areas. Among the interested organizations which supplied testimony to the Committee on Government Operations of the U.S. Senate were several chemical companies, who described their work with chemical searching operations. In addition, the testimony of others reflected some applications in the field of chemistry. For

a timely account of the status of scientific information retrieval work in all areas, the reader is referred to the report of the Committee's hearings. [167]

The terminology among information retrieval workers has not been standardized, which leads to some difficulty in appraising the content of the literature which has been published on this subject. (In this report, for the most part, the language of the authors of papers described has been preserved; this procedure naturally results in a certain lack of uniformity in the present work.) Several projects have attempted to do something about the lack of standardization, among them the American Standards Association, the Western Reserve University Conference Committee, and, to a lesser extent, the UNESCO and IFIPS groups interested in information retrieval.

The scope of the present survey is limited to the searching of chemical literature and stored chemical information by general purpose electronic computers.* There are quite a few computer projects actively under way at this time. However, before the present stage of development of literature searching in this area was reached, a great deal of pioneering effort had to be expended in several fields tangential to machine searching, and it is desired to give recognition to some of this work.

Notation for Chemical Structures

Ten years and more ago, several independent groups proposed notation and coding systems for the representation of chemical structures. [17], [18] As early as 1946 Dyson discussed a chemical cipher notation, [40] and published a revised version of the original system in 1947. [39] Also in 1947, M. Gordon, C. E. Kendall, and W. H. T. Davison proposed use of the chemical ciphering notation system, which came to be known as the "GKD Notation". [65] Most such notation schemes were developed as a means of designating structures in an unambiguous way without resorting to structural formulas, and as

---

* No attempt has been made to include in the survey the classified literature with respect to chemical literature searching.

a means of developing an alphabetical index for the structures. Although the principal purpose was not to assist in mechanical aids to searching, at least three of the early schemes were said to be applicable to such mechanical searching, and appeared to be directed primarily to punched card systems, though mention was sometimes made of the possibility of use of computers for such applications. In fact, a special purpose computer, to be used in conjunction with a card sorter, was actually proposed by the Gordon-Kendall-Davison group, [37] in connection with their system of chemical ciphering. Specifications for this computer, to be called the Electronic Structural Correlator (ESC), were described, but a prototype model was never built. [36] Another pioneering effort was the numerical designation code developed by Frear, et al. [51] The basic principles of the Frear code were adopted as the nucleus of the notation scheme used by the Chemical-Biological Coordination Center (CBCC) of the National Research Council, Washington, D. C. [25] These early activities were followed shortly by notation and coding systems proposed by W. Gruber, [71] J. G. Cockburn, with the Newcastle System, [34] J. A. Silk, [157] and W. J. Wiswesser. [192] In the early 1950's, the National Research Council conducted extensive tests of the notation systems proposed by Dyson, Gruber, Wiswesser, and Silk. Based on the conclusions reached by NRC as a result of these tests, E. M. Crane and M. M. Berry developed a composite notation system which, in part, combined the best features of the systems tested by NRC. They proposed use of a "classification code" as well as a basic structural cipher. [18] In 1957, Wheeler, Andrews, Fallon, Krueger, Palopoli and Schumann, from the Wm. S. Merrell Co., Cincinnati, Ohio, published a simpler code designed for an 80-column punched card machine. [187] They assigned a unique serial number to each compound, and provided space to accommodate as many as 999,999 compounds. A modified version of the Wiswesser notation has been used by Elbert G. Smith, at the University of Hawaii in Honolulu,* for searching a catalogue of 50,000 compounds in a punched card operation. [159]

---

* Now at Mills College, Oakland, California.

Several more notation schemes for chemical structures have been proposed in recent years, among them Isbell's system for carbohydrate structures, [83] Luhn's system for identifying geometric patterns by description of their envelopes, [106] Chodosch's ciphering system, [29] and Feldman's stereo numbers. [46]

In connection with plans for mechanized searching, a group at VINITI (the All-Union Institute for Scientific and Technical Information, Moscow) is working on a notational coding scheme for structural configurations. Since there is considerable interest in correlating physical or chemical characteristics with structural features, the coding is directed to only that part of the total configuration which gives rise to the property of interest. The system is said to be "simple and rational," and, while longer than, for example, the Dyson notation, it is claimed that it is more logical than the Dyson system. It is also claimed that it will be applicable to subject matter in mathematics, linguistics and physics, as well as chemistry. [11], [74]

Another system of notation, developed by a group at the Moscow State University under Professor A. P. Terentiev, called the TKZP system (Terentiev, Kost, Zuckermann and Potapov), is designed to record both the spatial configuration and the connectivities of the various atoms or groups of atoms in the structure. In a paper describing this system, the authors also propose a scheme for causing a machine to convert a formula or compound name to an atom-by-atom recording, formula, or code, without manual aid. [197] Kent et al [93] also report (see footnote 7 of the Kent et al paper) that this group of workers is developing codes similar to those of Opler and Norton [129], [125] for use with computers.

Vleduts, in a long monograph on notation, indexing, and nomenclature, discusses the Dyson system and compares it extensively with the systems proposed by the VINITI group and another proposed Russian system. [171] In another paper, he discusses a linear language for organic chemistry suitable for machine searching. [172]

A proposed international standard notation system, based upon the Dyson notation, has been recently published by the International Union of Pure and Applied Chemistry. [82]

By a fortuitous accident, or by inspired foresight, some of the work on notation systems anticipated some of the major considerations in devising systems for computer searching. One of the principal and continuing investigations facing would-be machine searchers is that of linguistics. Certainly the early notation schemes resulted in symbols which offer a working intermediate stage between English language text and machine language in the representation of chemical compounds. It is not intended to imply in any way that the linguistics problems have been solved, even in the area of chemistry, but only to state that a basis has been formed for the conducting of simple machine searches, on limited size libraries or catalogues of compounds. The topological nature of the structure was also recognized by the early devisers of structure codes, and groups were represented in some codes with bonds indicated, in others not. Even in cases where no direct use was made of any one notation system, the thinking which led to devising the systems, on some of which several years were spent, has formed a springboard for further exploration. The work which has gone into them has undoubtedly influenced the thinking of every person who is engaged in setting up systems for chemical literature searching.

Indexes and Abstracts

The abstracting and indexing services have formed two of the most important aspects of information retrieval, both before and after the rise of computers in this field. Abstracts make available to the scientist much more information than he would be able to scan otherwise; indexes give him his best route to the recovery of information. Because of the recognized importance of the work in these two areas, there has been widespread activity in both abstracting and indexing in this country and abroad. It is impossible to give credit here to the fine work accomplished by many devoted scientific personnel working in these two areas.

Eugene Garfield, who has been diligent in the indexing area of information retrieval, made the statement that "The abstracts section of Chemical Abstracts is a device for dissemination while the index is for retrieval." [58] Welt has combined the two, which results in a "lengthy hybridized index entry," and has used such an index for the handling of 13,000 (in May 1958) papers, in diverse languages, in the Cardiovascular Literature Project of the National Research Council. [183]

Although Chemical Abstracts does a monumental job in summarizing the chemical literature, many large problems still reside in supplying service of this type and scope. The importance of continuing research in these areas was recognized by the recent International Conference on Scientific Information, which devoted one of its seven areas to indexing and abstracting problems. [79] It is important to note the accomplishments in these areas, both for the purpose of evaluating the applicability of the methods used to machine procedures, and (more important) for the purpose of making firm decisions to deviate from them for mechanized practices, where necessary to do so. If recovery of all information is required, abstracting becomes a very much more difficult procedure. It is at once more selective and more inclusive. Information not particularly pertinent to the main body of the article being abstracted may increase in value tremendously when evaluated in a different context, or it may increase in significance because of new developments. Computer systems tend to make multi-branched and greatly expanded usages of stored file data, which are usually not anticipated to the fullest extent when such files are prepared. For such purposes, abstracts which represent summaries may not be all-inclusive enough for mechanized systems. A pursuit of this reasoning reveals that only a multi-level indexing which lists every subject will be adequate.

Various indexing proposals have been made for chemical literature. Bernier proposes that, using techniques of alphabetical correlative trope indexing, arbitrary symbols or names be used to index organic groups, [16] to form a "correlative chemical-group index." Among the proposals that have been put into operation is the one which

the American Society for Metals and the Special Libraries Association, working together, have developed. This is an indexing and classification system known as the ASM-SLA classification for edge-punch cards which covers not only the field of metallurgy, but various aspects of metallurgy found in the areas of chemistry, physics, and engineering. 3/, 50/, 90/ A computer analysis of an indexing system for 40,000 documents has been made by Remington-Rand, with particular interest given to the number of descriptors used to describe documents, frequency of use of descriptors, combining power of descriptors, etc. 152/

Citation indexes have been proposed by Eugene Garfield both in the area of chemical patents 59/ and for other scientific literature 58/, 60/, where the index would be accumulated and oriented to the user's frame-of-reference. On the other hand, for quickly supplying users of information with knowledge of current developments in their field, as an alternative to abstracting services, Garfield supplies photocopies of the contents pages of all current periodicals in their particular area of interest. These are sometimes made available by the journal publishers prior to actual publication of the journal. 61/, 5/

The task of analyzing a document and extracting from it the significant portions to be included in an abstract or a file is a very demanding one, and normally requires the skill, knowledge, and experience of technical personnel. Experiments in delegating this task to computers have been tried by H. P. Luhn of IBM, who employs statistical procedures on the IBM 704 to extract sets of significant words, based on the frequency of their occurrence within a document. 107/, 146/ Planning Research Corporation's experimental studies in automatic indexing and abstracting have employed the "relative-frequency approach" to the measurement of the significance of words, word groups, and sentences; they use the frequency of both "uniterms" and "multiterms" as a function for the production of "auto-indexes", and multiterms for the production of "auto-abstracts". As in Luhn's work, the auto-abstracts take the form of "extracts" from the author's own language. 43/ Similar auto-abstracting work is being

carried on by V. A. Agrayev and V. V. Borodin (Gorkey). [1] General purpose computers or "information and translation machines" can be used for the work. Only significant words are taken into account; tests of several methods are based on (a) frequency of occurrence of words and (b) probability of reciprocal combinations of words within a sentence. As a result of analysis of inflections during the summarizing procedure, some of the more significant sentences are isolated from the text. Indexing, largely as related to notation and nomenclature schemes proposed by Dyson and certain Russian workers, is discussed at some length in a monograph by Vléduts. [171]

Ohlman (of Systems Development Corporation) and Luhn have also accumulated bibliographies by "permutation indexing" (Ohlman) or "key words-in-context" (Luhn) experiments, using direct computer output for the bibliographies. [85], [108], [31]* In this connection, Chemical Abstracts is initiating the publication of a bimonthly serial publication of a permuted title index from, in the beginning, some of the 8000 journals received in their office each year, which includes, in addition, the list of titles, authors, and references. [41], [26] A bibliography of papers on the subject of permutation indexing going back as far as 1856 has been compiled by Ohlman. [127]

As an aid to indexing, Eugene Wall, of the Engineering Service Division of du Pont, has been concerned for several years with the arrangement, grouping, and indexing of material making up a store. He believes that more effective indexing will materially assist in recovering information which has relatively few "handles" for retrieval, and he proposes that the indexer be provided with reminder lists of words generally associated in context with those used in the text. He calls such a list a "Word-association Matrix"; [35], [177] this matrix consists of a list of associated terms, listed in order of frequency of association under the main term, and could be considered in this light a coordinate indexing arrangement.

---

* See also discussion under DIFFICULTIES, LINGUISTIC AND OTHER.

There will be available next year an A. I. Ch. E. Thesaurus, based on one which the du Pont Company has for internal use, but modified and enlarged from that version. In addition, the Institute publications will incorporate new style abstracts and indexing terms for each article in each issue, for easier use of the Thesaurus. This work was done under the auspices of the Standards Committee. [44]

Punched Card Systems

Although the principal content of this report is not concerned with punched card systems per se, it is interesting to note the effect of some such systems on the development of electronic computer systems. In some cases, punched card systems were the forerunners of similar computer systems, and some projects, with added sophistications in some cases, were transferred from punched card equipment to computer installations, without a complete transformation in the systems. In these cases, the punched card systems were simply the forerunners of the computer programs, but they determined or defined in some cases the directions taken by their successors, which were handled by computers.

Many systems lend themselves successfully to handling by punched card equipment, and for that reason are still carried on by such machines. In some cases, an electronic computer could speed up the search, but otherwise could add little to the advancement of the caliber of the program now in existence. Indeed, in some instances where random access and variable branching are not involved, punched card systems are superior because of the capability of executing parallel searching on punched cards. Because of the decision-making ability of punched card equipment at the time of input of the card, this kind of system can be faster than computer operation of a similar nature, in cases where this kind of operation is particularly suitable. Where punched card equipment is suitable for the particular application, punched card operation is generally considerably cheaper, both in terms of equipment needed and in kinds and numbers of personnel required.

The U.S. Patent Office makes prior art searches by means of a punched card system in connection with patent applications directed to steroid compounds and their synthesis. [52] The cyclopentanohydrophenanthrene nucleus is assumed to be present in each structure and the coded description of each compound is recorded on one card. Twenty-two substitution positions are provided for; at each of these points on the basic structure may be found any one of nineteen common types of substituents, double bonds, alpha or allo configurations, or miscellaneous groups. Each such substituent and position has a specific punching position in one of the columns 1 to 48 on an IBM card. In columns 60 to 69 are recorded general descriptors not identifiable with any particular position on the nucleus. Searches are made serially on a multi-column sorter, although a single column sorting machine may be used.

For the purpose of testing the relative efficiencies of such a serial file and an inverted file, a group of workers at Documentation, Inc., converted the codes for the same steroid documents into an inverted file on Matrex cards, and made some comparative studies of the relative searching efficiencies. [116]

Much of the information retrieval research being carried on at universities makes use of punched card equipment; this is true of some of the Groth Institute work being carried on at The Pennsylvania State University in connection with chemical and physical data on crystals; [70], [134], [170] and the work on properties of hydrocarbons and related compounds at Carnegie Institute of Technology. [2] Using the Wiswesser notation, a punched card process for recording and recovering chemical information was described at Remington-Rand as long ago as 1953. [14] Chemical companies, such as Dow, [126] Ethyl Corporation, [182] American Cyanamid, [49] and du Pont, [42] of course, made early use of punched card equipment for searching for chemical compounds, as did pharmaceutical companies such as The Sharpe and Dohme Company, [151] (now the Merck, Sharpe and Dohme Research Laboratories) and Smith, Kline, and French. [144] Linde Company links chemical terms to "role indicators", and calls such

combinations "structerms;" there is a dictionary or glossary of such terms, together with the role the term plays in the context. [186] Linkage of the documents with the structerms is by means of the cards themselves, which are known as "docuterm cards". The punched cards, dealing with internal technical reports, principally chemical laboratory reports, are collated for the structerms, which include role indicators. The Chemical Department, Experiment Station, of the E.I. du Pont de Nemours and Company, has a punched card system for the indexing of chemical compounds and polymers, where the card for each compound and polymer component carries additional information, such as types of reaction, properties, etc. [42]

An important early contribution to information retrieval was made by the Chemical-Biological Coordination Center, who devised a system for searching chemical and biological literature in such a way as to correlate the chemical and biological characteristics of such material in the files. [193] There were several outgrowths as a result of the early pioneering work of the CBCC (some of which are now developing into computer programs). Among them was the work of Maloney, at Fort Detrick, who developed a medical and chemical system patterned after the CBCC work. [113] Welt developed one of the CBCC ramifications, which has subsequently become the Cardiovascular Literature Project. [132], [184] A somewhat related project dealing with carcinogenic effects of chemical compounds is being carried on at the Cancer Chemotherapy National Service Center at the National Institutes of Health in Bethesda. [56]

Smith, Kline and French Laboratories have also developed a system in which biological effects are correlated with characteristics of the materials that cause them. [133] They have made syntactical and semantic studies, and from these analyses have developed a simplified grammar with rigid rules of word order for expression of information for machine indexing. The area of use is in the fields of biology, chemistry, and clinical medicine. Although the work is being conducted on an IBM 101, the logic of the system has been developed with application to electronic computers in mind; and Smith, Kline and

French apparently expect to employ computers for the work sometime in the future, according to their statements.

The Netherlands Patent Office (Octrooiraad) is engaged in experimental work on a retrieval system for documents which deal with carburetors for internal combustion engines. 95/ The fluid employed in the carburetor is an integral part of the procedure, and for this reason the work has interest for chemists as well as engineers. Approximately 16,000 documents are in the collection under study. Coding of characteristics is done on horizontal rows of conventional 80-column punched cards; relationships between characteristics are shown by interfix punches in columns 48 to 59. Each characteristic is represented by a series of characters, each containing four binary bits. One character specifies the main aspect (e.g., cooling device, mixing device, throttle); up to four characters specify details of the characteristic; and one character denotes the type of fluid which is flowing through the carburetor element being described. The system is suitable for use on a machine such as the ILAS, of the U. S. Patent Office.

Dr. J. M. Ganeff, of the Institut International des Brevets (The Hague), is experimenting with an internally slotted card system for searching documents dealing with electroluminescent materials. 57/ A special device has been developed to agitate the deck of cards to permit those with slots in needled positions to drop and to be separated from the rest of the deck. Each card represents one document and each slot represents one characteristic or category on a classified list of topics of interest in this field. The subject matter is characterized from many points of view.

Most punched card systems for searching in the area of chemical applications seem to serve the specialized needs of those who instituted the systems. There have been few, if any, large-scale generalized punched-card systems developed which would be all-encompassing enough to serve the common need of all users of mechanized equipment for searching of chemical literature. Among the few systems which sought to embrace a larger territory was that developed by the

U. S. Patent Office, [10] and it was motivated in this by the necessity of making many kinds of searches of large and heterogeneous files. For a number of years, the Australian Patent Office has been searching applications by means of a punched card system. It is not in the province of this paper to report on all such punched card systems. However, there have been two editions published of an excellent work by Casey, et al, which do present a much greater coverage in this area. [23], [24]

Special Equipment for Literature Searching

Commercially available machines specifically designed as literature searching machines have not yet come into widespread use. This is not surprising, since there is no general agreement on the form of data for storage or on searching systems design, nor have there been developed any really satisfactory comprehensive general theories which can explain, correlate or predict the characteristics of such systems. [80]

For the most part, mechanized searching schemes that are in actual use have employed general purpose machines or machines designed primarily to perform other types of tasks. [121] Some writers have discussed characteristics which would be desirable in information retrieval machines and have shown why general purpose machines are probably not "the answer." [97] It should be stated, though, that strong arguments have been advanced in support of the use of general purpose computers, at least until such time as there is developed better information about user requirements. [128] However, a number of proposals for information retrieval machines have been made and these will be described briefly below. It is not known how many of these machines have been used for chemical literature searching, but it is possible that any one of them has been, and all certainly have the potential. It was therefore decided to include a fairly comprehensive list of them for their possible interest to the reader.

Probably the most widely known proposal for a mechanized information retrieval system is the MEMEX of Dr. Vannevar Bush. 21/ Dr. Bush's early contributions in this field are well known, both his work on the Rapid Selector 9/, 141/, 169/ and as Chairman of the Advisory Committee on Application of Machines to Patent Office Operations; 168/ his thinking has been an inspiration to many later workers in the field.

Edge notched cards have been used by a large number of groups in retrieval applications; specialized equipment for use with this type of card has been developed by several organizations. A good review of available equipment is given by Casey et al. 23/, 24/

Equipment for implementing the type of searching system which is known in this country as the "Peek-a-Boo" system has been designed by and is in use at the National Bureau of Standards. 189/, 190/ Commercial equipment intended for use with peek-a-boo cards is sold under the names TERMATREX or MATREX. 87/

Punched card sorting machines (IBM type cards) have also been built especially for information retrieval purposes. The ILAS machine, developed at the U. S. Patent Office, provides facilities for searching hexadecimal codes recorded on the horizontal rows of an 80-column punched card and for relating codes to each other by grouping signals, interfixes and modulants. 4/ A storage and retrieval scheme for use with this machine has been developed by the Patent Office. 54/ The system has been applied to documents dealing with ethylenic unsaturated homopolymers and copolymers. The information content of each document is recorded on punched cards, compounds being identified by unique numbers and classes of compounds being similarly identified. Functions of compounds are recorded as "modulants" or modifiers of the subject matter (compound) codes. Relationships between the several codes for a compound, between the several compounds associated together in a process, and between all of the codes in a document, are denoted by grouping signals which follow the last code of the group involved. Interfixes are used to show relationships among the various groups. By assigning "weights" to the parts of a question and

specifying a minimum total weight for acceptable answers, it is possible to retrieve documents which fully meet the search criteria as well as those which meet them only partially. It is possible to search for compounds, functions, and processes.

In collaboration with the Textile Fibers Department of du Pont, this system has also been experimentally programmed for the Bendix G-15D computer. 54/* The information is recorded in two types of words, one of which includes modulants and subject matter codes, the other interfixes and grouping signals. One more character in words of the first-mentioned type is added for questions. This character indicates to the computer whether or not modulants are being searched, and whether the subject matter codes are to be matched on an equality or an "included in" basis. The output is a printed table which states whether or not the full "weight" or the minimum "weight" of the question has been achieved, whether all subjects and interfixes required have been found, and whether all specified relationships were found.

The International Business Machines Corporation offers the 9900 Special Index Analyzer, which is a form of collator for performing coordinate indexing searches. 120/ For the 9900, IBM cards are punched, one for each term in a glossary, with codes for the documents to which that term applies being punched on the card. Decks of cards are operated upon to find sets of documents representing, for example, the logical union, intersection or intersection with complement, and the results are automatically typed. Searching systems designed for use with this machine have been developed by M. Taube. 163/

The Row-by-Row Scanning attachment is a device which permits the IBM ESM-101 to analyze a pattern represented within a single row of a card, independently from patterns in other rows, and of storing the incidence of a match until all 12 rows of a card have been analyzed in a similar fashion. 109/ The 101 with the scanning attachment may

---

* This work is discussed in greater detail under COMPUTER SYSTEMS.

also be programmed to detect given patterns in predetermined rows, or to determine that certain patterns precede others (in the order they are found amongst the various rows). These variations from the normal behavior of the 101 are produced by changing the functions of some of the recode selectors. A large variety of coding schemes are possible for row coding, as contrasted with the more conventional Hollerith column codes, fixed field, or superposed random codes. It is of interest to note that a 101 was specially modified to operate in this manner for the Patent Office in 1950. [10]

The IBM 9310 Universal Card Scanner is a later development of the type of machine represented by the earlier X-794 and "Luhn Scanner" machines, which used as one example Dyson's System of Notation of Organic Compounds in a demonstration of the machine. For these machines a code in a file card is punched on a vertical column, with each permissible code being represented by a pattern of exactly five punches. Questions are asked by means of cards in which the 7-hole complements of the desired codes are punched. Matching is done by an optical "blackout" technique; coincidence between a file code and its complementary question code indicates a match. Thus, holes in file codes must all match "no holes" in the question codes, and vice versa. [110]

The WRU Searching Selector,[92] which uses relays internally and has a punched paper tape input, was developed at Western Reserve University. Questions are entered through a special typewriter, and the file, which is stored on the tape, is read into the machine for comparison with the stored question. A number of searches can be made simultaneously. The subject matter logic employed permits use of a number of levels of grouping of information, such as words which are contained in sentences, sentences which are contained in paragraphs, etc. A number of other relationships, such as "negation," "AND," and "OR," may be indicated, as well as ranges of numerical information. Output is by typewriter. This equipment is being used at Western Reserve University in connection with the metals information service conducted there, as well as with other projects at

WRU. [143] Several existing general purpose computers have been programmed to perform searches of the types being made on the WRU Searching Selector; [89] these include the IBM 650 and 709 machines. Information to be searched by the WRU system has also been prepared in a format which will make it acceptable to the Minicard system. General Electric initiated a transistorized version of the WRU machine, the General Electric Information Searching Selector (GE-250), [72] with a magnetic tape input. However, by mutual agreement between GE and WRU, a general purpose computer, the GE-225, which embodies the searching characteristics of the GE-250 in addition to the potential of a computer, was substituted to replace the GE-250, which was not put into production. [63], [88]

For the purpose of maintaining U.S. Air Force personnel files for more than 1.5 million individuals, an auxiliary special purpose machine, the GE-260 High Density Information Selector, has been proposed for use in conjunction with a computer. The principal distinguishing characteristic of the GE-260 is a thermoplastic tape memory for storage of the 1.5 million personnel records, where it is expected that each personnel record will contain approximately 1000 characters. Four simultaneous searches can be performed, with a search time of under 20 minutes. [64] There are two modes of operation: (1) under program control of the computer and (2) by means of inquiries put directly to the "inquiry buffer", which has a capacity of 288 characters. Results may be obtained by means of an electric typewriter.

Computer Control Company has designed and built the Index Searcher for the Rome Air Development Center. [94] This machine stores the file to be searched on magnetic tape. Questions may specify up to ten criteria for search and the elements of the question may be grouped into as many as 15 phrases by plugboard wiring. The machine can be operated in any of six different modes, identified as document search, regenerate tape, question insert, search, print, and (tape) edit. 5200 words, each containing 42 bits, can be scanned per second. The average document is encoded in 20 words, about 68,500 documents being storable on a 2400-foot reel of tape; about 260 documents can be searched in one second.

The Magnacard System, of the Magnavox Company, [188] while not specifically designed for information retrieval purposes, is a sufficiently interesting data handling device to require mention here. Characteristics of the system suggest its use for searching by employing the sorting and file-manipulation facilities that it offers, as well as its potential use as a computer input device. Magnacards are 1"x3" magnetic oxide coated mylar cards, each capable of storing over 5,000 bits. The cards are made to adhere to transport drums by a pneumatic system, which has the effect of converting the transport drums into magnetic drums. Magnetic reading heads are provided for sensing the coded data.

Several machines have been devised for retrieval purposes which make use of photographic storage elements. In each of them, frames of film are presented to a reading station at which binary coded information is sensed; each frame contains both photographically recorded pages of documents and binary codes describing their contents. Upon recognition of the desired code factors, an image of the pictorial part of that frame is projected for viewing or copying. The Minicard system of the Eastman Kodak Company employs discrete pieces of film which are manipulated in a manner similar to the handling of punched cards. [99] The relative amounts of space devoted to pictorial and coded information are completely flexible. The Filmorex System is generally similar; however, one difference is that the pictorial and coded areas are fixed. [32], [145] Both FLIP (Film Library Instantaneous Presentation), a product of the Benson-Lehner Corporation, [15], [194] and the Rapid Selector employ continuous reels of film. The Rapid Selector was first constructed by Vannevar Bush, but ideas concerning it were previously discussed by many people (among others, Draeger and Davis), both during and prior to its inception. Ralph Shaw was subsequently interested in its development; it was later improved by a group at Yale and is at present being further developed at the National Bureau of Standards for the Navy's Bureau of Ships. [9], [141], [169], [118], [156] The Minicard, Filmorex, FLIP, and Rapid Selector systems each sense the coded patterns optically. A similar machine described by Peter James of IBM uses as a file storage medium a photographic film

containing images of document pages adjacent to an alterable magnetic strip on which are recorded the binary coded data descriptive of the document. 86/

An apparatus for performing coordinate indexing searches mechanically is described by Plankeel. 137/ Punched tapes are used rather than the cards usually employed in this type of system. Each document in the collection is assigned a number which occupies a unique position on any tape which refers to an indexing term related to that document; each tape represents a single indexing term. Every punching position, starting at some fixed point on the tape, corresponds to a single document number, and each channel on the tape represents a different series of document numbers. A punch in any one position signifies that the correspondingly numbered document contains the concept of the indexing term to which that tape refers. Only those tapes are searched whose indexing terms are required. The equipment compares the pertinent tapes simultaneously for coincidences of punches in corresponding positions, and automatically records the numbers of the documents which "possess all the terms represented by the tapes." A mechanism is also proposed for correlating such a searching technique with a document copying device. In this form, the tape search is correlated with a film strip which contains micro-images of document pages. Recognition of an answer to the search question results in the printing of enlarged photocopies of the documents found by the system.

Each of the searching tools discussed so far requires some form of physical movement of the file or motion by the machine used to effectuate the search. The time required for making the search is dependent upon the size of the file, or upon the length and complexity of the question, or upon both of these factors. This may be attributable to the fact that, in most cases, searching systems have been designed for use with equipment designed for other purposes, such as punched card machines and computers. Searching systems designed for use with these more or less conventional types of machines have employed various techniques to alleviate the time problems involved. Generally,

these have involved screening techniques or various attacks on the file organization problem, both intended to preclude the necessity for detailed inspection of all information in the file. Some systems which resort to inspection of only some of the documents in the file are those employing classified decks of punched cards, or classified tapes in the case of conventional serial machines, or inverted files of either the peek-a-boo type or the Patent Office's DRAM system. [101] In the HAYSTAQ system [136] (which is a serial search on a stored program digital computer), after an externally stored encoded document is read into the internal storage it is "screened" by inspecting information which gives general information about the document contents. If it is determined that there is no possibility of an answer, details of the document are not examined. Another proposal is that a serially inspected file be so organized that the "most likely" documents are looked at first. None of these approaches completely eliminates the defects inherent in systems having time span characteristics resulting from the serial nature of the searching operation.

A logical development, therefore, is the series of proposals for systems and machines which employ a static file and interrogate all portions of the file simultaneously. In effect a search question is broadcast to all documents in the file at one time and those which satisfy that question all respond simultaneously. A group at the Benson-Lehner Corporation has proposed such a system and has suggested hardware for its implementation. [62], [181], [195] The Sandia Corporation has also come forward with such a suggestion. [114] Documentation, Inc. has designed the EDIAC, which is a small model of a simultaneous searching machine. [164] Some requirements and advantages of this kind of development in automata have also been discussed by groups at the Patent Office and the National Bureau of Standards. [97]

At least two devices are known which will locate microfilm copies of documents in a file and project an image or reproduce it photographically. Both of these require that the exact physical

location of the desired document be specified, and they should be considered as a special form of retrieval equipment, but not as searching machines. One of these is the automatic micro-image file proposed by M. L. Kuder of the National Bureau of Standards. In this system, up to 10,000 miniature photo-images are recorded on a 10-inch-square sheet of microfilm. When coordinates of desired images are specified, the machine, operating continuously, photographically prints out one frame every two seconds. 6/, 98/ A further development of this type of apparatus, called the VERAC 903, has been made by the Crosley Division of AVCO Corporation. 8/, 20/

The second machine, AMFIS (Automatic Microfilm Information System), stores images on 20" microfilm strips, a multiplicity of which are contained in "scrolls". Access to these strips is obtained in a fashion comparable to sliding a hair comb out of its carrying case. A number of such scrolls are mounted on the face of a drum. The address of the desired document is dialed from a keyboard and access is achieved by three simultaneous movements: movement of the drum along its axis, rotation of the drum to locate the appropriate scroll, and the unwinding of the scroll to position the desired area of film between a light source and an optical system for projection on a viewing screen. 7/

One additional special-purpose machine which should be mentioned is the FLEXINDEX of the I. H. and R. Co. 66/ The system is centered about a coding device which takes common words entered from a keyboard and automatically encodes them in a greatly compressed form. Searchable files made by use of this device are stored on magnetic tapes and questions asked by specifying the pertinent terms via the coding device.

While information concerning the development of information storage and retrieval machines in the USSR has been somewhat incomplete and sometimes conflicting, it is known that at least two such machines have been built or are under development.

In 1954 the Institute of Scientific Information of the Academy of Sciences (now known as VINITI, the All-Union Institute for Scientific and Technical Information, headed by Prof. A. I. Mikhailov) produced

the Experimental Information-retrieving Machine, designated as the EIM. This device operates upon punched cards which are divided into four fields, each of which employs a distinctly different type of coding system. The cards are the conventional 12-row, 80-column variety. The first 25 columns are devoted to alphabetic and numeric information, recorded in a manner similar to the well known Hollerith code. Columns 26 to 33, which include 96 punching positions, are directly coded, each position representing a distinct characteristic of the document being coded. Columns 34 to 40 contain random superposed codes. Columns 41 to 80 make use of the horizontal rows to record data, each row containing eight characters made up of 5 bits each for numeric information. By using columns 39 to 80 in this manner, 6 alphabetic characters containing 7 bits each may be recorded on a row. The horizontal code patterns in this part of the card are recognized by the machine regardless of the row in which they appear. Questions are set up by means of plug-board wiring and console switches. Cards are scanned at the rate of 420 per minute. The comparison apparatus is said to employ "cheap and reliable" elements, rather than electronic valves or relays, while the apparatus for registration of the results "does comprise several electronic recording and relay elements." A special offline printer is provided for recording the bibliographical information on cards selected by the search. 139/, 140/

A brief mention is made by Kent and Iberall 93/ of a special purpose information machine being developed for the faculty of organic chemistry at Moscow State University, scheduled for delivery in 1959. No details of this machine are given by Kent et al and it is not certain whether this equipment is different from either the EIM or the machine described in the next paragraph.

An "information machine" which is apparently intended for retrieval, translation, and other forms of non-conventional information handling processes is under development at the Laboratory for Electric Modeling in Moscow. This laboratory is a part of the Institute for Precise Mechanics and Computational Techniques, which is also attached to the Academy of Sciences. The machine development work

is under the direction of Dr. L. I. Gutenmakher. The file of information is recorded on sheets of paper about 7 1/2 x 14 inches in size, on which are printed metallic spots which act as capacitors. A pattern is recorded by punching out some of these spots. The sheets are compressed into blocks or "books", and the pattern circuits on such a collection of sheets are interconnected by bus bars. Interrogation of the entire block is made simultaneously; no physical movement of the file occurs during the searching operation. It is claimed that very large volumes of information can be searched in a very short time, one hour being required to search the equivalent of 4 million pages of ordinary text. The storage responds to descriptors, rather than to addresses, and the individual pages or books constitute, in effect, a non-addressable storage. 93/, 73/, 74/ Recent visitors to this Laboratory have reported that a 100,000-bit capacitor storage is now in operation and that this type of store is ready for production by industry. 179/

Vléduts et al have taken the position that "While general-purpose computers are quite suitable for experimental work in this field, specialized information machines are necessary to obtain serious practical results." He describes the machine memory recently developed under the leadership of Prof. L. I. Gutenmakher, and states that it is particularly suitable for chemical information searching, since it makes it "possible to consider the creation of a large informational-logical machine for information on chemical compounds and chemical reactions." He states further that in this connection, it is possible to simulate certain aspects of "chemical thinking" to aid in methods of synthesizing compounds not yet produced. 173/

A brief report in Electronics 28/ mentions a special machine, under development for the Council for Cybernetics of the Academy of Sciences, which will perform chemical information retrieval tasks. It is stated that the machine will use an algorithmic technique, "translating chemical characteristics into mathematical expressions by matrix inversion and other linear methods."

The authors have attempted to describe the major computer chemical searching systems which have been developed, either on an experimental basis or to be used as productive tools in research. No distinction has been made between those which have been developed and later abandoned, those which are quiescent, or those which are in active operation; it is sometimes dangerous to make the distinction, even when it is supplied by the originators of the systems, because the status of such projects is subject to change, either in whole or in part. The dormant stage of such work, as well as the more active stages, sometimes contains valuable lessons for others. Information concerning work which has not been described herein is invited.

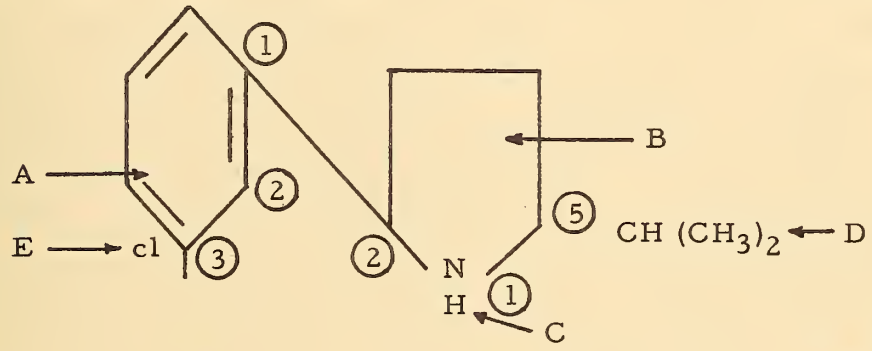
The Dow Chemical Company

Ascher Opler*, together with Ted Norton, when at the Dow Chemical Company, developed a topological system for coding and searching chemical structures. 125/, 129/ The earliest programs for this system were written around 1952. Several commercially available general purpose computers have been used in the course of the years in this work, among them the Datatron, IBM 650, 701 and the 704, on which the work has been performed exclusively since 1956. The file included approximately 15,000 structures, stored on magnetic tape. The chemical groups which are coded are intermediate in size between single atoms and very large general groups. Thus, the authors reject the idea of coding the benzene ring as six single carbon atoms and also the idea of coding benzoic acid as a single group. Something over 300 different structural groups are provided for in the code schedules, each identified by a 3-digit number. In coding a structure, each such group is assigned seven digits. If group A is being coded, the first digit identifies the position in a previously coded group (B) to which A is attached. The second digit specifies the position of A which is attached to B. The third, fourth, and fifth digits are the identification

* Now with Computer Usage Co., Inc., New York, New York.

of the A group. The sixth is a number arbitrarily assigned to group B in this compound and the seventh is a similar arbitrary number assigned to group A in this compound. Rules are provided for the order of enumerating the code groups.

An example for the code of one such structure is as follows:



A	B	C	D	E
0110601	1210512	1121123	5200324	3106115

106 = BENZENE
105 = CYCLOPENTADIENE
211 = HETERO -NH-
003 = PROPYL
061 = CHLORINE

In making a search, five criteria must be met, and failure to meet any one terminates the search. The first test is for the empirical formula,* and that of the compound being examined has to be at least equal to the one requested. The second test is for the presence of the groups specified by the question. These are called "keys". The third test ascertains whether stated keys, taken in pairs, are indirectly connected through a third group. The fourth test is for relative

---

* The authors have attempted to employ the same terminology used by the originators of the research being described; the use of the term "empirical formula" in some instances and "molecular formula" in others reflects this practice.

positions of pairs of keys sought in the third test. The final test determines whether stated pairs of keys are directly attached.

It is possible to ask for a class of groups, such as "alkyl", rather than for a specific one, and such a search is satisfied by finding a structure containing any one of the 30 or so specific alkyl groups provided for in the schedule. The technique used here is a look-up operation, in which the computer, after noting that "alkyl" is a class term, refers to a stored list of specific groups which are of this type, and, in effect, substitutes these specific terms for the generic term used in the question. It is possible to perform up to five independent searches simultaneously by a "multiplexing" technique. The output of a multiplexed search is a list of answers for all of the questions, each answer being marked in terms of which question it satisfies. A subsequent routine sorts these answers into separate lists and translates the arbitrary numbers given in the original output lists into conventional systematic names for the compounds.

In collaboration with Norma Baird, Opler has also devised a routine for enabling the 704 to display structural formulas on a cathode ray tube; when this is desired, some other form of description of the compound is used as input. 27/, 130/ A large number of small structural groups are stored in the machine in a coded form. When an appropriate description of the compound desired for display is given, the depiction program calls up the applicable structure codes, converts each one into a pattern of dots which corresponds to the structural configuration of that group, assembles the dot patterns in their correct relative positions, and displays the entire structural formula on the face of the cathode ray tube. A camera may be coupled with the CRT to record the displays, or an output magnetic tape may be written containing the dot pattern words to be used in a later display. Output rates are about two structures per second for photographic copying, or five per second for recording on magnetic tape. Opler also found it desirable to modify the output routine of the topological structure searching routine (described above) so that instead of producing lists of structure names, the pictorial depiction codes could be brought in,

converted to dot patterns, and the structure displayed on the CRT.

In addition to the work described above, Dow Chemical Company is using a Burroughs 220 electronic computer in its Midland, Michigan, plant, in a program to develop new rocket fuels, textile fibres, plastics, and other chemical products.

Midwest Research Institute (MRI) [22], [47]

The Midwest Research Institute is an independent non-profit research institution, financed largely by gifts from businessmen in the geographical area of Kansas City and St. Louis. A study is under way at Midwest (supported initially by a National Science Foundation grant, the Basic Research Associates program of MRI, and International Business Machines Corporation) to determine the "structural and physical attributes which correlate with the uses of chemical compounds."

Of approximately 750,000 different chemical compounds known at this time, only about 3,000 have recognized commercial and medical uses. MRI is attempting to determine applications of the remainder by the use of high-speed computers.

MRI wrote to manufacturers in the United States and Canada and through them acquired information concerning commercial products for their research. They have collected data on more than 2,000 commercial compounds with known uses and they want to determine the factors which are keys to those uses. Then, on a statistical basis, they seek to discover whether other compounds with those factors are likely to have the same uses. Their routine seeks to determine, then, not only new uses for compounds for which some uses are known, but uses in addition for new compounds which have been synthesized but whose utility has not yet been discovered.

Data for 2,000 of the compounds have been coded and are acceptable to the IBM 704 computer in the form of punched cards. MRI is applying the system to several major categories of chemicals.

Midwest has a list of "usages" for the compounds; in some instances there are as many as fifty for an individual compound, although some have as few as three. Certain physical properties were known for most of the compounds under study, and they were made a part of the input data. They are: Molecular weight, melting point, boiling point, refractive index, density, and empirical formula. The use of two-digit numbers to represent physical property values was made possible by the following procedure. From Lange's Handbook 2,000 compounds were chosen randomly as a representative base group; each of the physical property values for this group of 2,000 compounds was ordered and the range was separated into equal groups. A two-digit code number was then assigned for each physical property value, corresponding to the percentile of the particular constant in the group. This procedure did not sacrifice accuracy to any required degree.

An example of this code for dimethyl phthalate is shown below:

Property	Actual Value	Code Percentile
Molecular Weight	194	49
Melting Point	5.5°C	17
Boiling Point	285°C	78
Refractive Index	1.516	53
Density	1.189	43
Empirical Formula*	$C_{10}H_{10}O_4$	CC10HH10004

*Where only one letter is required for the element, it is repeated in order to standardize spacing.

Four different data tapes were used for the IBM 704 computer routine. One of these (1) contained the coded physical properties data; one tape (2) contained the coded "use" data, consisting of six-digit numbers, made up of a three-digit number for the basic function of the compound and a three-digit number for the type of material on which the compound functions; another tape (3) carried the name of the compound, together with an arbitrary identification number; and the last tape (4) contained the chemical structure and the identification number. The name tape is useful for post-search identification, and the other three are working tapes.

The structural code is that used by Norton and Opler in the Dow Chemical Company research, with some modifications. [125], [129] Seven-digit code numbers are assigned to commonly occurring structural groups and linkages in organic compounds, of which there are something over 300, but provision is made for expanding the number if need arises. Of the seven digits making up each number, three refer to identification of the particular group and the remainder relate positions of mutual attachment between the subject and the preceding group.

The Midwest researchers have continued Opler's practice of "multiplexing"; that is, they handle simultaneously ten sets of search criteria, each of which consists of a series of six-character words describing physical constants or uses. A separate IBM card is punched for each such set of test criteria, with the remaining six-character words on the card supplying data for the structure search.

Each compound is classified first into broad classes, and subsequently into successive sub-classes, up to a limit of three for the sub-classes. One such example is:

Solvent	(Broad class)
Rubber solvent	(Sub-class)
Organic rubber solvent	"
High-temperature organic rubber solvent	"

If there are no usages specified for the compound being searched, that search is skipped; the same procedure applies to the absence of any physical constants. The broadest classification that can be used as a test criterion is the basic function of the compound, with the finer classifications refining the search to more restrictive classes of material. Five types of search are possible:

(1) Compounds having usage in a specified class or sub-class, where four such levels are possible.

(2) Compounds with physical constant value
   (a) Specified, or
   (b) Within limits

(3) Compounds having a specified minimum or maximum number of specified structural groups.

(4) Compounds having a specified minimum number of direct connections between specified structural groups.

(5) Compounds having a specified minimum number of indirect connections between specified structural groups.

Compounds are grouped according to use, and once it is deemed that a sufficient number has been accumulated under a use, statistical analysis is made of the group. Use categories are incorporated in a tape together with the encoded properties and characteristics of new compounds, and the compounds are tried against the use categories.

It is possible to specify as an addition a special output which is a frequency distribution for any specified physical constant or chemical element. Both a frequency distribution table and a cumulative frequency table are obtained as a result, and the tables are printed out, after which a general or a weighted distance function analysis is made of the results.

Not only physical properties are distributed, but also the frequency of occurrence of functional groups, inter-relations of functional groups, types of atoms in the compounds, etc., until a great many correlations have been made. When certain compounds have physical properties which diverge greatly from the majority in that use category, the anomalies are investigated in an effort to "unlock new areas for developmental research."

The MRI researchers say that "the first function of this program is to describe the statistically significant requirements for each use in terms of physical properties and structural characteristics. The second is to screen old and new compounds for possible uses." They expect the screening to indicate the most promising areas for the laboratory work necessary to confirm the ideas so developed. Not only do they expect to decrease development costs and discover better chemicals for many uses, but they expect that a significant result of their research will discover concealed correlations and "substitute a methodology for the current confused and varied approach to use development."

Monsanto Chemical Company

Monsanto's stated aim was the design of a total mechanized documentation system, with the completion to be realized in small increments. An integral part of the system design is the manner of encoding, storage, and retrieval of chemical structures, including indexing the structures and their retrieval by classes, when desired.

Monsanto's research at this time is concerned with nearly 30,000 compounds. Each compound has a unique number assigned to it. In the early work, [175] four magnetic tape files were maintained, where each of the classes of information listed below was contained on a separate magnetic tape, in variable length records:

1. Compound's structure and number
2. Compound's name and number
3. Compound's test results and number
4. Compound's molecular formula and number

Later work [174] makes use of the first three listed; the last is computed and stored with the structure data.

Initially, the coded information describing the structure is punched on 80-column IBM cards, together with the unique identification number. A maximum of ten cards may be used for each record. The program for performing different types of searches is contained on a deck of cards, and included with it are the necessary control cards to perform the specific search required. The search was previously carried out on an IBM 702, but the details of the work are reproducible with either the 702 or the IBM 705 computer; and the IBM 704 is at present being used for searching.

There is little redundancy in the system, which Monsanto feels is a limitation with respect to certain types of error-detecting procedures. Proof-reading is employed in an effort to catch human errors. Some such errors are found occasionally by chance by trained chemist users of the output data. Monsanto does employ certain known characteristics of the data as safeguards in several

editing and error checking procedures, e.g., a test as to the oddness or evenness of numbers of characters in a row, or of number of rows of data for a structure, and tests which make use of chemical valence rules.

For the "application-research"* carried on at Monsanto, a new method of encoding structures of chemical compounds was developed for automatic storage and retrieval. The encoding scheme used by Monsanto is a relatively simple one, and the encoding can be performed by trained clerks instead of professional chemists. Structures are set down on cross-hatched paper with numerals used for chemical bonds, single character letters for the elements, and the letter "M" used for all metals. The elements and bonds are written in the squares of the paper in the same arrangement as the original structure written by the chemist. One example of a structure so encoded is shown as follows:

Encoded data
C1C2C10
2 1
C1C2C

Under the topological principles employed in the philosophy of the system, the chemical elements can be thought of as hard cores, which are joined by the different types of chemical bonds, thought of as flexible links, and the configuration may be twisted into a large number of shapes. Provision is made for recognition of the compound, regardless of divergence among chemists in forming and encoding the structural diagram of the request.

The computer calculates the molecular formula, and the complete tape record contains: compound number, molecular formula, controls and structure data. Control data include the molecular formula requirements and the substructure. A chemist may search the files for a specific substructure or chemical moiety, instead of the complete

---

* As defined by Monsanto Chemical Company: "that type of work in which an attempt is made to find a chemical to do a specific job."

structure, if he desires. The search strategy may use the molecular formula as a screen, before continuing further detailed searching.

Monsanto also uses the computer for report writing. A "screening report" is prepared on each compound tested, and the running of tests for specific applications requires the distribution to interested personnel of about 20,000 different screening reports each year. Results of the tests are stored in the files in the form of one of three statements:

1. No further interest
2. Some activity, a lead for further study
3. Active, needs secondary screening or field testing

For the report writing, a standard preprinted form is not used; rather, for purposes of flexibility and revision, the output format is pre-programmed and printed out each time by the machine. The computer searches name and structure for the input compound number, then prints out on the desired format the information pertinent to the test being reported. Six copies are made, and are ready to be mailed at output. The output from the computer is in some cases a two-dimensional printed structure which can be recognized by trained chemists. Monsanto chemists say the computer-printed structures are easy to read and interpret. The one-page laboratory report output contains the serial number of the compound being studied, its name, structure, the raw data recorded, and an evaluation.

For report writing, Monsanto recopies data in various combinations for progress reports, special reports, process reports, and others. Requests to regenerate stored data in repetitive situations are called "bonus questions" by Monsanto, as the cost of each such successive request is materially lower than the preceding requests for the same information. In addition, they have found it economical to accumulate lists of various combinations and cross-combinations of tests made under varying conditions, and they use the material so generated to form meaningful sets of tables.

In addition to the screening report, monthly summaries of data are made up and distributed to interested personnel of Monsanto, consisting of lists of results according to each of the three test findings. Special reports may be obtained, such as review of results for new synthesis leads, further testing, or collection of data for any one period.

Monsanto has made several hundred generic structure searches. They say that the data are ready to be searched for almost any two-dimensional structural fragment needed. The machine accepts the encoded structure, analyses it for searching purposes and updates the file. Monsanto considers that its work is now definitely out of the research stage, since it is now an integral part of their reporting system and information retrieval program. [176]

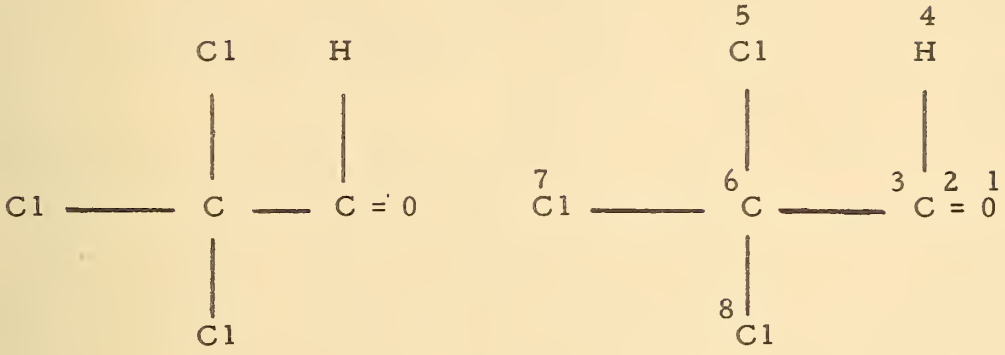
Ray and Kirsch Work at NBS

The U. S. Patent Office and the National Bureau of Standards are engaged in a joint long-range research program to develop and apply automatic techniques of information storage and retrieval to the problems of patent searching.

An experimental program was written for SEAC, the electronic digital computer at NBS, to make atom-by-atom structure searches in the field of chemistry. [142] Codes were written to describe over 200 steroid compounds; this compilation was used as the search file.

The method of representing chemical structures in diagrammatic form lends itself to a consistent notational scheme. For the purposes of this experiment, the compounds were encoded by arbitrarily numbering serially each atom in a structural diagram. One computer word was designated to contain the number of the atom, plus the numbers of other atoms to which it is attached. The number of such attached atoms was limited to four. In addition, the element symbol was also included in the SEAC word, which thereupon contained six fields: the serial number of the atom, the element symbol, and four connection fields.

The coding for chloral, for example, is as follows:



The atoms and all bonds (other than single bonds) are numbered in any arbitrary order. The list of connections to each component of the structure is as shown below:

Component No.	Connections	Element Symbol
1	2	0
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

The SEAC search program seeks to make an atom-by-atom match between the specified question structure and each entry in the search file. It is possible to request a match of a fragment, and thereby to find a particular configuration contained within a larger structure, or to match all components of a complete structure. In this sense, the generic concept is introduced into the search program.

Each match made in the search program is considered tentative until the search of the first file structure is completed. Whenever a failure to match is discovered by the program, it tries to return to the previous match to make a new match. If all possible matches lead to irreconcilable mismatches, the program rejects this file structure and progresses to the next.

A data checking routine is a part of the program. The original file, prepared by the Patent Office, was in the form of about 2500 punched cards which described about 250 chemical structures. The data were checked for internal consistency (including the redundancy feature inherent in expressing the attachments of one piece to another) and for their adherence to the coding rules that had been established.

Ray and Kirsch suggest that, to speed up the searching procedure, the empirical formula be used as a screen, and that only those structures be examined which have at least the number of occurrences of atoms required.

HAYSTAQ

One result of the collaboration between the U.S. Patent Office and the National Bureau of Standards, carried out in accordance with a congressional mandate and recommendations by the Bush Committee,[168] is the electronic computer system known as HAYSTAQ. This program attempts to simulate the search which a patent examiner now makes manually; the scope of the search is confined to the area of chemical literature.[97], [136]

Patent Office searching problems present some peculiar difficulties, and the need for mechanization of some of their search procedures is growing more pressing because of the rapidity with which technology is advancing and technological publications are increasing. A statutory requirement is that patentability be predicated on novelty, utility, and inventiveness, and these factors must be taken into account in the design of any mechanized searching system.

The first computer routine devised for the HAYSTAQ system was a general searching routine which was written for the NBS computer SEAC. Patents or other documents are coded and are contained serially in the file. Each such document is divided into compositions (physical admixtures of materials), subdivided into items (ingredients) and further subdivided into descriptor words (see Fig. 1). Heading information for identification, screening, and housekeeping precedes the file information contained at each unit level.

The question is stored and remains in SEAC's internal storage during the duration of the search. Compositions are read in serially and by a matching procedure on the individual descriptor word level, an attempt is made to "find" a respondent among the disclosures* making up the file. Screening techniques are employed at each level of the search for the purpose of terminating the search as soon as it is discovered that a match cannot be made. The descriptor words within the items for both question and disclosure are ordered for the same purpose, in ascending order of initial digits according to descriptor types (see Fig. 1). The initial digits refer to the differing subject categories.

Examples of Descriptor Types

10000004852	-	Index number
2043C05E901	-	Empirical Formula
3----------	-	Chemical
4----------	-	Botanical
5----------	-	Zoological
6----------	-	Anatomical
7----------	-	Processes
8----------	-	Miscellaneous

Fig. 1

When a complete match is attained for a composition (on either an exact match or a "contained-in" basis) a record is made in a so-called

* A disclosure is any one of the many documents making up the encoded file which is being searched.

"hit word" and the search progresses to the next composition in the document. The total score making up the set of hit words is then analyzed to determine whether the combinatorial relationships among the several items are the ones sought and whether a sufficient number of real answers has been found. This is necessary because one question item may be answered by several disclosure items, or, more seriously, one disclosure item may be the only answer for several question items, where several disclosure items are required. Also, several disclosure items may be wanted only if they are in combination, and they may instead be in an alternative relationship. Provision is made in the search for AND, OR, NOT, and ABSENT relationships.

Provision is made in the search for a combination of five possible question item types with three possible disclosure types, for a total of fifteen different combination search paths (see Fig. 2). Determination is made in advance of the particular question and disclosure type combination, and the suitable search path applicable to this combination is chosen. The characteristics of the data are allowed to determine the choice of the proper subroutine, to the exclusion of all others; this technique permits direct access to the pertinent orders without lengthy tests or questions.

Q (question) Item Type	D (disclosure) Item Type
All Q descriptors positive	All D descriptors positive
Some Q descriptors positive, others negative	Some D descriptors positive, others negative
All Q descriptors negative	All D descriptors negative
Some Q descriptors positive, others absent	
All Q descriptors absent	

Fig. 2

One of the subroutines included with the general routine is an examination of the empirical formula to seek a match. Another subroutine involves a matching of chemical descriptors. As test cases were tried, the importance of the chemical descriptor loomed much

larger in the overall system, and it was decided to expand this part of the routine into an independent routine which could be used separately when desired. The latest HAYSTAQ routines are concerned with this expanded subroutine, which is much larger than the parent routine.

The generalized search routine is concerned with a small-scale, or bird's eye view of documents; the chemical descriptor subroutine brings up a magnifying glass for an intensive large-scale look at a smaller section. There are several problems which cannot be dealt with satisfactorily in the general system for several reasons. For one reason, it is an exploratory program from which it is hoped to gain experience as to future directions for exploration; for another, limitations of available computer capacity made it necessary to limit the scope of the search. Solutions for two disturbing problems which have not been adequately provided for in the first program were included in the chemical descriptor search: one is a method for handling structural formula "Markush" groups;* the other is provision for asking a question generically and accepting as a satisfactory answer any specific embodiment of the genus. In addition, provision is made for searching for an exact match, or for a fragment which may be contained within a larger configuration.

A study was made of two other systems which employ recognition of topological relationships among functional groups: that of Norton and Opler at Dow Chemical Company [125], [129] and the Ray and Kirsch system at NBS. [142] The Norton-Opler system with relatively large functional groups was considered to be too rigid for HAYSTAQ, and the Ray system was in too fine a detail, embodying as it does an atom-by-atom match, and was therefore too slow. Basic functional groups which make up the particular compound were chosen, and the topological relationships among them shown, without an indication of positions of attachment. Terms are also provided which represent combinations of two or more units of the basic functional groups; these are groups which have constant definitions and are recognizable by chemists.

---

* A "Markush" group is an art term used in the Patent Office for designating a synthetic genus by a listing of its members.

An additional category of terms which are generic in nature (halogen, ester, etc.) is also provided. In the stored and coded form of data the combination terms and the generic terms precede the functional group terms and they may be used either as screens for the functional group codes or they may be used as the subject matter of search questions themselves.

The data are ordered, both in the disclosure and in the question, in an ascending series, according to the numerical value of the codes representing the substantive information. Since the program is essentially a matching procedure, this permits the computer to determine (in searching for a particular term) at the earliest possible moment when there is no available answer.

Rings and alkyl groups, because of their frequent occurrence, are recognized and given special consideration in order to promote greater efficiency in the program.

The search proceeds in serial fashion through the two summary categories (combination and generic terms) and the topological-functional group terms. If it is determined that the required combination terms are present, the search progresses to ascertain whether the required generic terms are present, and, if so, whether they are properly defined in the more detailed data. If the searching criteria in these two sections are satisfied, the routine progresses to the topological search in the functional group data.

An "equivalence table" is kept of identifications of matching items on both the question and disclosure sides. If there are present more than one of any identity in the topological groupings, there is a possibility that the wrong such identity will be selected. If that occurs, when the mismatch is detected the program allows for a backup in another attempt to find a match and the data are examined to attempt to find another embodiment in the disclosure data which will satisfy the question.

In the case of a structural formula of the Markush type (a structure having a fixed nucleus with variable substituent groups), if any member of the group on the question side is encountered in the

disclosure data, it is considered to be a match. This concept allows a concentration of the description of many compounds in a single data entry; in fact, in one of the test cases which were tried, over 64,000,000,000 distinct embodiments were contained in the one data listing. This is not astonishing when it is considered that any one data entry may contain many such Markush groups, and the final number of compounds which may be represented by the one structural diagram will represent the product of all possible combinations.

Any failure to match after a backup, when there are no further entries of the kind sought, will automatically call up the next document against which to try the question. Any one question is tried serially against all the data in the file. The program makes extensive use of screening techniques at all levels of the search. In other words, information of a general character is examined before proceeding to more detailed inspection.

Auxiliary data preparation routines are being run against large amounts of raw data in an effort to accumulate a file which will be used for testing the efficiency of the system.

Use of the RAMAC-305

Because the IBM 305 (a general purpose computer with moderate speed) has a large random access internal storage, it is well suited for making searches of the coordinate indexing type. For such use, the documents to be searched are stored in the memory as an "inverted file", each record in the storage being based on one of the descriptors or terms of the vocabulary. At the address where each such record is stored are listed numbers identifying the documents to which the term pertains.

At the San Jose Laboratories of IBM, where the earliest work on this type of system was performed, Nolan and Firth [48], [124] experimented with 5,000 documents, most of which were company reports, articles, and papers ordered by the library at the request of laboratory personnel. Their dictionary included descriptive terms and

the address at which the applicable document numbers are stored. This address was automatically computed by treating the characters of the ten-letter descriptive words as numbers and:

1. Dividing the word into two five-character portions,
2. Adding them together,
3. Squaring the sum, and
4. Employing the three middle digits of this quantity as the address of the record.

Since each record (100 alphanumeric characters on the RAMAC discs) can store only a limited quantity of document numbers, provision is made for automatically gaining access to "overflow" records. Questions are asked by punching addresses of pertinent terms on standard 80-column punched cards and reading them into the RAMAC. The records stored at these addresses are then compared to find document numbers which are common to the several records being compared. It is possible to make searches where the terms are either in the logical "AND" or the logical "OR" relationships.

In the U.S. Patent Office, Leibowitz et al [101] have further developed a system of this type, which is called DRAM, and which has, at present, a file of documents on synthetic resins. In addition to the general features of the scheme used by the IBM workers, the DRAM system includes provision for additional relationships and for three levels of increasing specificity of the dictionary terms. Compounds are searchable as such, or in terms of their atomic groupings. Relationships between compounds involved in processes may be searched, as well as the roles, functions or uses of compounds used in any particular process. Provision is also made for searching in terms of numerical values of certain physical properties (e.g., melting points of reaction products) and reaction conditions (e.g., pH of reaction mixtures). Additional systems of similar nature are also being developed for searching the phosphates and other compounds. [53]

Grandine and others, of Kennett Computer Consultants, Inc. [67], [138], [147] (Kennett Square, Pennsylvania), have used the

IBM 305 in an experimental system based on the VS_3 system developed at the Patent Office. [102] The VS_3 work had been geared to ILAS, which is a special purpose punched card machine specially designed to handle interfix logic. Grandine's work included an extensive system for handling interfixes on the computer, somewhat more than 265 interfix fields being required by some documents. Positive and negative interfixes were employed, as well as "balancers" whose function was to remove some of the ambiguities resulting from the use of the positive and negative interfixes.

The Cancer Chemotherapy National Service Center and Documentation Incorporated are making use of the IBM 305 in connection with the former group's project dealing with reports of the screening of compounds for their possible use in treatment of cancer. [103]

Report Searching on the Bendix G-15D at du Pont

Since January 1959, the Patent Division of the du Pont Textile Fibers Department has been searching research reports by means of the Bendix G-15D. [68] Upwards of 13,000 reports (mainly chemical laboratory studies) with more than 125,000 index entries are in the file, and the rate of accretion to the file is estimated at 15 to 20 per cent per year. Names of compounds, uses, physical properties, and other data are recorded on magnetic tape. Each such indexing term is assigned a four-digit alphanumeric code; all of the codes for a document, together with identifying indicia, are recorded as a variable length block on magnetic tape. The searching program, which is stored on the drum of the computer, causes comparisons to be made between all question codes and all file codes. However, to speed up the operation, the codes for each document have been sorted numerically. The lowest number is stored, and subsequent differences between adjacent codes are stored, rather than the complete number in the succeeding instances. This permits a more rapid searching operation, since less than the theoretically required number of computer orders must be executed. The questions are read in from a punched paper tape. Each question code can be asked for positively or negatively; in the latter

mode, documents which contain negative codes are excluded. Relationships among the codes for a document are not recorded. Answers for up to 16 separate questions may be searched for simultaneously, provided that not more than 16 discrete codes are included. The workers report that "the chief cost and time-consuming bottleneck is the manual screening of answers to eliminate semantic noise from the results of the machine search."

Thermophysical Properties Research Center (Purdue University)

The Thermophysical Properties Research Center (TPRC) at Purdue University collects all available recorded information on the thermophysical properties of all substances, including the theoretical methods of determining the properties and the experimental techniques used in measurement. [166] One of the principal sources of the material is from the abstracting journals, which cover approximately 15,000 technical and scientific journals. Since the cost of searching the journals by highly trained scientific and technical personnel at TPRC is extremely high, an alternate search procedure has been developed there. Searchers locate and mark pertinent references; clerical assistants label them and assign code numbers to thirteen specific items of information selected to identify and describe the contents of the item. The thirteen coded items are contained in the first forty columns of standard IBM cards, and one or more cards are required for the code of each abstract; additional cards are used for more than one property or substance, or for more than one physical state for the material under study. The thirteen items contained on the card are as follows: Property, substance class, substance name, physical state, type of subject coverage, language of original article, serial number of the reference, journal name, journal volume, journal number, journal series, beginning page number, and journal year.

In addition to the codes listed above, a further breakdown is required for substance classification, as follows:

Series-000: Works not involving substance class

Series-100 and -200: Substances described by chemical formula

Series-100: Elements and compounds (except compounds containing both C and H)

Series-200: Compounds containing both C and H

Series-300: Ferrous metal alloys

Series-400: Non-ferrous metal alloys

Series-500 and -600: Substances which cannot be described correctly by a single chemical formula and are not metal alloys

Series-500: Chemical systems and combinations

Series-600: Commercial and natural products

TPRC uses a Datatron electronic computer, with two magnetic tape units and an Electrodata 500 Card Converter. The data cards containing the thirteen selected items of information concerning the document are sorted by property, by class within property, and by substance within class, and the information pertaining to each property is stored on a separate tape. Because more than one journal may refer to any given article or report, each new potential entry into the file is held until the file has been searched for duplicates; if the new entry is a duplicate, it is rejected.

TPRC states that the separate magnetic tape for each thermophysical property permits rapid retrieval for specific inquiries. Search queries must specify the name of the substance and the property desired for that substance. Once each year TPRC issues in book form an ordered reproduction of all information contained in its files: "Search and Retrieval Guide to Thermophysical Properties Information." The publication, which is issued to sponsors and others interested in having readily available the information which it contains, consists of three parts, and has an annual content of about 20,000 bibliographic items. Parts A and B, "Guide to Substance Classification and Code Designations" and "Classified Search Index," respectively, are reissued every year, with information added to that accumulated for previous years. Part C, "Master Bibliography and Author Index," is a permanent volume to which a new volume is added each year.

Fine Chemical Service (England)

A Pegasus computer, with magnetic tapes, has been used to search a collection of 25,000 organic chemicals, together with a file of relevant physical, chemical, and biological properties. [33] Certain structural features are stored in order to facilitate the selection of compounds with those features from the file.* The first 208 prime numbers were allocated to 208 of the structural features; then for any chemical compound, the primes corresponding to its structural features are selected and multiplied, and the number so obtained is used as the code for that chemical compound. To find a compound responding to a particular set of such structural features, the product (compound number) is divided by the required factors of the combination desired; a test for a zero remainder will indicate the presence or absence of any such factor. This procedure will select all compounds which have at least the given set of features. A similar system has been employed by IBM, [96] and by at least one chemical company in this country, one of whose representatives said informally that they had abandoned the scheme because they were "defeated by the division time."

Farbwerke Hoechst AG

The Farbwerke Hoechst AG (Frankfurt am Main/Hoechst) has developed a system for encoding chemical structures which enables searching for compounds and reactions on an IBM 705. The developers of this system (Fugmann, Braun, Schwalbach and Vaupel) state that they can handle structures containing variable components, searches containing excluding limitations, and searches which are generic in nature.

Reactions are described by coding the affected chemical groups as they exist both before and after the reaction has occurred. Auxiliary agents are coded by their elementary symbols according to the periodic system. Further, numerical terms are used to express the change in the degree of hetero-orientation of the carbon atoms concerned with the reaction.

---

* Examples of such structural features are ring systems, amines, chlorides, etc.

Coding terms of varying degrees of specificity are provided for more or less conventional structural groups as well as for such "quantitative characteristics" as chain length, position of ring substituents and relative positions of characteristic groups in chains. From the information available at the time of writing, it appears that the search involves a coordination of terms in the question and file, the mere presence of the requested terms in the code for the file compound constituting a successful search.

In a private communication, it was stated (Sept. 1959) that the tape file includes more than 20,000 compounds, that it will increase to nearly 100,000 within the year, [55] and that the system has been in use for over a year.

Badische Anilin und Soda Fabrik AG

Dr. Ernst Meyer of the Badische Anilin und Soda Fabrik AG (BASF, Ludwigshafen) has been developing a topological system of searching for compound structures employing a general purpose computer. [115] The "cells" chosen to represent the nodes of the structural network are the atoms. They are represented by arbitrary numbers, and arbitrary numbers are again chosen for the neighboring atoms. Identification of the element by a unique symbol (periodic table) is also recorded. One of the objectives of this project is to place as much as possible of the burden of preparing the encoded file on the machine, rather than on humans. To this end, advantage is taken of the fact that compounds with complicated structures usually contain large structural groupings which repeatedly occur in chemistry. Examples of such groups are anthraquinone and benzoylamino. Special symbols are recorded manually for these groups and the machine "computes" the atom by atom code. In one example, a structure was manually encoded in 8 lines, by this method, and the machine was used to work out the details, which resulted in 100 lines, The savings in time and reduction in errors are considerable over the same 100 lines done entirely by hand. The system provides for the handling of the so-called "Markush Structures," which include variable substituents on specified positions of a nucleus. Searching may be either for entire

specific compounds or for compounds which include a given structure as a fragment. The general pattern of searching through the question and file structure in a comparison for similarities is very much like the schemes applied in the HAYSTAQ structural search routine and that of Ray and Kirsch.

Gmelin Institute

The Gmelin Institute has as its general objectives the presentation and evaluation in a detailed and critical manner of all of the world's knowledge on inorganic and physical chemistry and related sciences. The ever increasing volume of the literature has compelled the Institute to study mechanical methods of classifying and filing abstracts. While as yet mechanized information storage and retrieval systems are not in use at the Institute, a research group headed by K. Schneider has produced several experimental systems. One searches documents on atomic energy on the IBM 650; [148] another searches similar subject matter on the Univac Fac-Tronic I; [149] and a third, also for use with the Fac-Tronic, has as its subject matter non-ferrous alloys. [150]

DIFFICULTIES, LINGUISTIC AND OTHER

Almost every literature searching project now in development is encumbered with some, at this time, unresolvable difficulties. Studies are under way by several groups engaged in special efforts on resolution of difficulties, which represent in some cases segments of the main problem. One primary difficulty embraces the entire area of linguistic considerations. Language in the form of discrete words is the usual means of communication in printed or written material. Language in this form is not understood yet by computers without an intermediary of some kind. A synthetic solution of the difficulty of communication between humans and computers is that of creating an intermediate language of symbols, where a unique symbol expresses a word or an idea, and the computer is programmed to substitute the symbol for the idea, or "thing", so expressed, and to recognize the concept or the object by means of the symbol expressing it. This sort of procedure, however, is practicable only in files of limited subject matter and size.

For several years, S. M. Newman, at the U.S. Patent Office, has been concerned with the creation of an unambiguous special language for the expression of ideas and objects, and their relation to each other; he calls this language "Ruly English". 123/

E. de Grolier has proposed the creation of a general machine language, which would also be a link to join different translating machines working with natural languages. 38/ He points out, however, that many scientific disciplines must be brought together for this work, including classification, linguistics, and logic. Among others who are working in this area are Ranganathan in India, Vickery and Foskett in England, Scheele in Germany, Durocq in France, and several Americans whose works are cited elsewhere in this report.

V. Yngve, at MIT, advises the use of an automatic programming system as an intermediary between the linguist and the computer, where the linguist employs in his research a notation or language called "COMIT"; that language in turn is converted by a conversion routine or

compiler. Through such a language, Yngve states that the linguist can direct the computer to analyze, synthesize, or translate sentences. He advocates the use of English text in literature searching because of the possibility of missing a reference when indexes or abstracts are used, in spite of the complexity which this procedure introduces into the searching. [196] A group at Lockheed is studying word correlation with respect to the syntax of English; "a normalized English language is formulated in terms of which all other words are syntactically defined", after which correlation factors are computed by counting primitive words occurring in syntactic reduction strings. [117] Another group, at Ramo-Wooldridge, has been conducting research on word correlation and automatic indexing for information retrieval procedures (under the sponsorship of the Council on Library Resources). [165]

Thyllis Williams, at Itek Corp., advocates applying to natural language a "normalized representation" for the storage of index data; she calls this "L-representation" and employs an "L-dictionary" of special rules for correlating elements of non-normalized expressions. An L-thesaurus is also employed for the purpose of relating the normalized representations to stored index data. She maintains that the informative content of a document is usually represented incompletely in indexing it, and that under the system of L-indexing, it can be developed to a greater depth and specificity for a searching system employing selective indexing. [84], [100], [191]

Both Harris, at the University of Pennsylvania, [75] and Chomsky, at MIT, [30] have been engaged in linguistic research which may be applicable both to machine translation and to literature searching. Harris's work has resulted in computer programs for the UNIVAC, and some of his followers have applied his ideas on other computers. However, there has not as yet been a significant "breakthrough" to the extent that there exists a tool which can be employed productively for making literature searches independent of an intermediate symbolism. Bar-Hillel has written a critical review of work done in this and related areas of mechanization of information searching problems. [12]

In the linguistic problem, there are two primary aspects: semantics and syntax. Some researchers have approached their study initially through semantics, some through syntax; some have used an empirical approach which incorporates both aspects. Much more work is needed on both. Any approach to the study of language solely through semantics is a dangerous one. Among the problems encountered are semantic changes which take place with the passage of time; there are additional complications in the changing conditions in science, such as, e.g., changes in the use of physical constants, changes in the accepted values of precision in measurement, or even in the standards of measurement employed.

A great deal of current research is being conducted on character recognition devices, which should assist eventually in the reading of English (or foreign) language text by computers. A certain amount of development in this field has made available some commercial machines for this kind of activity. 162/

Difficulty of developing a system for literature searching varies, of course, from project to project, and depends partially upon the objective of the search, the magnitude of the file or library which must be searched, and the fineness of discrimination which must be built into the search program. The scale of the search might run from a simple hunt for a matching name or title to an all-encompassing search similar to those required by the U.S. Patent Office.

While it is likely that one could develop a mathematical model of any existing retrieval system, no one has yet set forth a comprehensive model which might be generally applicable. The value of such a model would lie in the ability to predict the characteristics of systems under development; to correlate properties of existing systems so as to permit comparative studies to be made, and, hopefully, to serve as a guide to the establishment of really useful and economical comprehensive search systems. A number of research workers have concerned themselves with developing general theories in this field and the recent International Conference on Scientific Information devoted an entire session to this problem. 81/, 119/

An enormous difficulty to be surmounted in designing any system for literature searching is that of preparation of the file to be searched. Every stage of the journey from natural language and figures, illustrations and diagrams (the form in which it is assumed most literature would be found initially), through the analysis by professional personnel to extract pertinent material (or to abstract or to transfer material en toto), to the final form of the information on some magnetic medium, is studded with formidable obstacles. Thyllis Williams's concern regarding the preservation of the informative content of a document is a very real one. Although the analyst of a particular document may extract all information which is pertinent to requirements envisaged for any particular system, or for others in the foreseeable future, some information content is lost when a complete document is not retained, and the value of the missing material may depend upon inventions not yet in existence and whose development cannot reasonably be anticipated. [45] It is extremely important, but extremely difficult at this stage of development, to insure that there will be adequacy of analysis of documents, and proper coverage of all material contained in them. It is perhaps for these very reasons that Yngve insists so strongly on preservation of the natural language input as one prerequisite for proper subsequent retrieval.

The accuracy of any data which have been handled by humans at any point is suspect; on the other hand, machines which handle data cannot use "human judgment", and it is difficult, if not impossible, to formulate rules for processing every situation when diverse kinds of data are handled by machines. There simply does not exist any foolproof method at this time for transferring hard copy information directly to a magnetic medium with any assurance of obtaining an error-free file.

The basic approach to the design of a system for information retrieval varies widely. For some retrieval problems, a set of descriptive terms called "descriptors" is stored for each entry in the file. Some retrieval attempts require the specification of identical descriptors in order to recover the stored information. A broad or

general specification might deluge the questioner with answers if (a) the file is extremely large or (b) the question is not asked in sufficient detail to screen out all of the unwanted items having duplicate descriptors. One criticism of this kind of file is its unwieldiness because of size; another is the difficulty of phrasing questions in order to avoid large numbers of "false drops." One fashion of forming a file for such a system is by means of an "inverted file" structure, where, instead of listing under a document the descriptors applying to it, the file categories are made up of characteristics of material in the documents, and under each such descriptive heading is a list of the documents which contain material of such a category.

A more succinct file can be obtained by the use of precise definitions for wanted information from the file, but this procedure is not possible in very many applications. A chemical structure lends itself to representation in this fashion very readily, since it permits exact definition in terms of molecular (empirical) formula and topological arrangement. Where an exact definition can be specified, any question which is formed for the purpose of retrieval must be equally exact if only pertinent responses are to be elicited.

Sometimes arrangement of the files in terms of the most frequently desired items is preferable, particularly if a serial procedure of searching is employed. The greatest activity is then concentrated in the area where the information most often required is clustered, and a more efficient operation results. Most searches, regardless of the manner of arrangement of material in the files, are serial in nature, according to the plan for retrieval. An ideal system would permit an inquiry to be made of the complete file simultaneously, with a simultaneous response from every entry which is being called for; at this time neither equipment nor systems for retrieval have reached this stage of development. An approach to it, however, is in the concept of selective use of random-access storage in terms of addressability, where symbols for information coincide with the address where the information is stored, so that instead of serial searching, selective searching is employed. Sir Robert Watson-Watt has stated:

"When he asks for a random-access memory, the client usually wants to have whatever may be the converse of a serial-access memory. Physically the answer has reduced to a very rapid-access (and still, most frequently, a serial-access) device, but philosophically, in the end it need not, and practically it should not, do so." He suggests that the operation which is usually desired is a "Parallel partial-identifier search" and he further points out that what is wanted is "assured timely individual retrieval by selection, preferably by parallel interrogation of the system's 'memory'". [180] Sir Robert referred to a paper by Slade and McMahon [158] which holds promise of providing the kind of parallel rapid-access storage which is needed for literature searching: "(1) capacity for storage of a large number of fixed-length words, (2) parallel interrogation with simultaneous access to the whole memory and (3) many interrogate operations and relatively few erase and write operations."

Until the development of such storage has been completed, however, a simulated parallel storage can be partially realized by more sophisticated system designs which permit interrogation of only likely respondents by selective addressing or summary screening techniques. Semarne has applied symbolic logic to screening techniques. [131], [154]

Vléduts has specified as a necessity "large long-time high-speed machine memories.....capable of long-time storage or large volumes of information and of operating without mechanical motion, by using purely electric counting principles." Such a machine is said to have been developed in Russia by Prof. L. I. Gutenmakher, at the Electric Simulation Laboratory, U. S. S. R. Academy of Sciences. [173]

A large amount of effort has gone into automatic indexing and abstracting of English language text, because of the sheer magnitude of this task. P. B. Baxendale has said that the "abstracting and indexing of input documents represents at least 80 per cent of the effort of current literature-searching systems as against 20 per cent devoted to retrieval." [13] In an effort to force computers to assist in this one-sided effort, H. P. Luhn, together with a group associated with him

at IBM, has for several years been concerned with "auto-abstracting" of documents by computers, where the abstracted portion is based upon a statistical analysis of the frequency of occurrence of words, in context. 76/, 85/, 111/, 146/ The abstracts are made up of whole sentences from the text of the document, where the sentences are chosen not only with respect to the presence of certain words, but also with respect to the relationship of those words to each other in terms of their location in the sentence. Word lists are formed, excluding articles, prepositions, conjunctions, etc., and the most frequently used words are assumed to be of "high significance." From the number of these words in a sentence, and how closely they are clustered among the other words in the sentence, a "sentence significance factor" is computed. Those sentences with high scores based on these factors are extracted from the document to form an Auto-Abstract.

Some information retrieval systems have adroit indexing schemes for relating discrete portions of files when a unified response is desired. In this way, segments or fragments may be recovered in the context in which they are physically located, or they may be tied together in diverse ways for several levels of maneuvers. Complexity of such arrangements of course requires more intricate designs for the manipulation of data so arranged.

Vléduts discusses the application of modern mathematical statistics to chemical information (as well as to experimental physics) as a means to a more compact representation of the information contained in the tremendous archives which have accumulated in research institutions, but he states that it is effective only when statistical methods are used not only for the final processing, but in the planning stages as well (arrangement of material, choice of number of parallel measurements, etc.). 173/

Another form of disposition of files is that proposed by Newell, Simon and Shaw, 155/ where data making up an entity are not necessarily arranged together in contiguous storage locations; instead,

the data form a "trail" where each datum carries the association with what precedes it and what succeeds it.* This particular arrangement allows facility of amplification of data entries without massive rearrangements of data in storage. It also permits new entries or inserts, without dislocating former entries. In the same manner, it facilitates deletions of unwanted or no longer pertinent information. The unwanted locations are simply returned to storage for later use by successive entries. Unfortunately, with present equipment, this type of arrangement demands a certain amount of interpretive programming and requires a very large random access storage. Information processing language (IPL) VI was employed to describe the operations defined in the paper cited above; the authors are now engaged in developing IPL VII, which is expected to be able to handle more complex situations.

Most of the chemical literature search systems discussed in this paper do not take advantage of some of the advanced techniques with respect to data arrangement and format. The logical problems of setting up such a system for computers are formidable enough, without the application of several levels of sophistication in programming techniques and data arrangement. In addition, the incorporation of such intellectual advancements to programs now in existence is probably not worthwhile until a plateau of development in literature searching is reached, because of their high cost in both time and money. Some of the improvements which could be put into programs now in existence might well be abandoned before they have proved themselves because of advances in the state of the art.

One of the important features to build into future searching programs, once there has been a changeover from research to development, is the incorporation of some sort of learning technique in order to avoid multitudinous repetitive searches. This is a very important consideration which should be kept in mind, although its advent will be in a later phase, after the initiation of continuing

---

* See also 97/, p. 1163, 1164.

programs. A program making use of selective storage recall operations and employing certain elementary learning methods, called IQ, has been written by M. E. Stevens, of the National Bureau of Standards. It is of interest to note that this program has been used experimentally in conjunction with an output routine for an information retrieval system. 161/

Although rapid advances continue to be made in the speed with which arithmetic calculations can be performed, computer manufacturers have been slow to incorporate into computers special devices which would aid materially in making searches. At least one computer manufacturer now is incorporating "search units" into a general purpose computer. 77/ The inclusion of such equipment in computer design (perhaps as an optional feature) could well become a trend, and would not only cut down the search time by a very large amount, but would make unnecessary some of the refinements in programming techniques, which are costly, but will become increasingly necessary in the absence of special hardware characteristics.

Certainly one of the difficulties encountered in all literature searching programs is the lack of appropriate automata with which to conduct the searches. Computers were originally designed to compute; there has been a lag in satisfying the need for automation in other areas where the need is just as great as for computers. Extremely large random-access storage with rapid "symbol manipulators" are required. 69/, 112/, 122/, 135/, 160/, 178/ Since those who are designing systems for literature searching do not in general have available the kind of automata which specifically meets their requirements, they must simulate such equipment by the use of computers. It is presumed that even this difficulty will be surmounted in the not too distant future, particularly since so much interest has been awakened in information retrieval in the last few years.

CONCLUDING STATEMENT

There is a distinct possibility that the authors may have missed important work which should have been recognized in this survey. If that is the case, reports of such work, either complete or in progress, are hereby invited. There has been no intentional omission of any work in the area delineated at the beginning of this survey.

Two invaluable series of reports issued by the National Science Foundation should be mentioned. One of these, "Current Research and Development in Scientific Documentation," which is issued semi-annually, contains considerable information on active research projects in information retrieval and closely related fields. The other, "Non-Conventional Technical Information Systems in Current Use", reports from time to time on searching systems in actual use.

In addition to the many fine bibliographies and works otherwise cited in this report (see particularly Casey, et al 24/), a number of other bibliographies are pertinent to subjects discussed above. 19/, 78/, 91/, 104/, 105/, 153/, 185/

It is desired to give credit for assistance in the preparation of this survey to Robert T. Moore, Forest Miller,* and Frances Neeland, of the National Bureau of Standards.

---

* Now with Union Carbide Corp., Oak Ridge, Tenn.

BIBLIOGRAPHY

of

A SURVEY OF COMPUTER PROGRAMS FOR CHEMICAL LITERATURE SEARCHING

by

Ethel Marden and Herbert R. Koller*

National Bureau of Standards
Data Processing Systems Division
Washington, D. C.

1. Agrayev, V.A., and V.V. Borodin. "The problem of automatic abstracting and possible solutions." [Abstract] U.S. Joint Publications Research Service. Abstracts of the Conference on Mathematical Linguistics, Leningrad, 15-21 April 1959. (Soviet developments in information processing. JPRS:893-D) New York, 31 August 1959. p. 90.

2. American Petroleum Institute. Research Project 44 on Data on Hydrocarbons and Related Compounds. Annual report for year ending 30 June 1958. Pittsburgh, Chemical and Petroleum Research Laboratory, Carnegie Institute of Technology [1958]

3. American Society for Metals. ASM-SLA metallurgical literature classification, prepared by a joint committee of the American Society for Metals and the Special Libraries Association. Cleveland, American Society for Metals, 1950.

"ASM-SLA literature classification revised," Metals Review, 30:2 (February 1957) 6-9.

---------- Watertown, Mass., Ordnance Materials Research Office, 1957. 3 p. ASTIA doc. no. AD-134 620 (a).

4. Andrews, Don D. Interrelated Logic Accumulating Scanner (ILAS). (Patent Office Research and Development reports, no. 6). Washington, Department of Commerce, 25 June 1957.

---------- "ILAS". Frome, Julius [et al] Recent advances in Patent Office searching: Steroid compounds and ILAS. (Patent Office Research and Development reports, no. 8). Washington, Department of Commerce, 1957. p. 13-15.

* U.S. Patent Office

5. Andrews, Don D. "U.S. Patent Office. Office of Research and Development." National Science Foundation, Office of Science Information Service. Current research and development in scientific documentation, no. 4. [Washington] April 1959. p. 30.

6. "An automatic micro-image file," Journal of the Franklin Institute, 262:5 (November 1956) 396-398.

"An automatic microimage file searches film record and makes photographic print," Technical News Bulletin, National Bureau of Standards, 40:7 (July 1956) 89-90.

7. Avakian, Emik A., and Eugene Garfield, "AMFIS -- The Automatic Microfilm Information System," Special Libraries, 48:4 (April 1957) 145-148.

8. Avco Corporation. Crosley Division. VERAC 903, A system for data indexing, searching, retrieving and reproducing. [Brochure] 1329 Arlington St., Cincinnati 25, Ohio. [1960]

9. Bagg, Thomas C. Status of the Rapid Selector development at the National Bureau of Standards. Reported at the Annual Meeting of the American Documentation Institute, Lehigh University, 23 October 1959. Preprint, 7p.

10. Bailey, M.F., B.E. Lanham, and J. Leibowitz. "Mechanized searching in the U.S. Patent Office," Journal of the Patent Office Society, 35:7 (August 1953) 566-587.

11. Baker, D.B., and M. Hoseh. "Soviet science information services," Chemical and Engineering News, 38:2 (11 January 1960) 70-73, 75.

12. Bar-Hillel, Yehoshua. Some theoretical aspects of the mechanization of literature searching. Jerusalem, Hebrew University, April 1960. 74 p. Contract no. N62558-2214 (ONR). Technical report no. 3. ASTIA doc. no. AD-236 772.

13. Baxendale, P.B. "Machine-made index for technical literature -- an experiment," IBM Journal of Research and Development, 2:4 (October 1958) 354-361.

14. Benson, Frederic R. "Recording and recovering chemical information with standard tabulating equipment," Abstracts of papers, 124th Meeting, American Chemical Society, Chicago, 6-11 September 1953. p. 5G-6G.

15. Benson-Lehner Corp. "Introduction to the FLIP (Film Library Instantaneous Presentation)," American Documentation, 8:4 (October 1957) 330-331.

16. Bernier, Charles L. "Correlative indexes. IV: Correlative chemical-group indexes," American Documentation, 8:4 (October 1957) 306-313.

17. Berry, Madeline M., and James W. Perry. A review of notational systems for designating organic structural formulas. Cambridge, Mass., Massachusetts Institute of Technology, [1951]. 431 p.
Submitted to the Codification Commission of the International Union of Pure and Applied Chemistry.

--- "Notational systems for structural formulas," Chemical and Engineering News, 30:5 (4 February 1952) 407-410.

18. Berry, Madeline M., and Edward M. Crane. "Which notation?" Chemical and Engineering News, 33:27 (4 July 1955) 2838-2843.

19. Bourne, Charles B. Bibliography on the mechanization of information retrieval. Menlo Park, Calif., Stanford Research Institute, 1958. 22 p.

--- Supplement I. Menlo Park, Calif., Stanford Research Institute, 1 February 1959. 25 p.

20. Bowker, K. [et al.] Technical investigation of elements of a mechanized library system. (Final report no. EW-6680). Boston, Avco Corp., Crosley Division, Electronics Research Laboratories, 11 January 1960. 110 p.

21. Bush, Vannevar. "As we may think," Atlantic Monthly, 176:7 (July 1945) 101-108.
Also pub. in his Endless horizons. Washington, Public Affairs Press, 1946. p. 16-38.

22. Carpenter, R.A., C.C. Bolze, L.D. Findley. "A system for the correlation of physical properties and structural characteristics of chemical compounds with their commercial uses," American Documentation, 10:2 (April 1959) 138-143.

--- Reprint. Kansas City, Mo., Midwest Research Institute. (MR1300)

23. Casey, Robert S., and James W. Perry (eds.) Punched cards -- their applications to science and industry. New York, Reinhold, 1951. (See also 2d ed., item 24 below.)

24. Casey, Robert S., James W. Perry, Madeline M. Berry, and Allen Kent (eds.) Punched cards -- their applications to science and industry. 2d ed. New York, Reinhold, 1958. (See also earlier edition, item 23 above.)

25. Chemical-Biological Coordination Center. The Chemical-Biological Coordination Center of the National Academy of Sciences -- National Research Council. Washington, September 1954. 33 p.

--- A method for coding chemicals for correlation and classification. Washington, National Academy of Sciences - National Research Council, 1950. 98 p.

26. "Chemical literature gets a quicker index," Chemical and Engineering News, 38:14 (4 April 1950) 27-28.

Chemical Titles: current author and keyword indexes from selected chemical journals, no. 1, 15 April 1960. [Washington] American Chemical Society.

27. "Chemical structures by computers," Chemical and Engineering News, 36:7 (28 April 1958) 108-109.

28. "Chemistry information retrieval will be carried out by a new special computer under development in the USSR for the Council for Cybernetics of the Academy of Sciences. [Newsletter]," Electronics, 32:5 (18 December 1959) 11.

29. Chodosch, Robert. An artistic extension of conservative descriptive chemical nomenclature. New York, The Author, (1212 Grand Concourse) 1959.

30. Chomsky, Noam. "Logical structures in language," American Documentation, 8:4 (October 1957) 284-291.

--- "On certain formal properties of grammars," Information and Control, 2:2 (June 1959) 137-167.
Also pub. by [Research Laboratory of Electronics] Massachusetts Institute of Technology, Cambridge, Mass., in cooperation with the Institute for Advanced Study, Princeton, N.J., 28 October 1958. Contract DA 36-039-sc-78108. ASTIA doc. no. AD-221 311.

--- Syntactic structures. (Janua linguarum Nr. 4) The Hague, The Netherlands, Mouton and Co., 1957. 116 p.

31. Citron, Joan, Lewis Hart, and Herbert Ohlman. A permutation index to the "Preprints of the International Conference on Scientific Information." Santa Monica, Calif., System Development Corp., [1958] 140 p. (SP-44)

--- Rev. ed. Santa Monica, Calif., System Development Corp., 15 December 1959. 37 p.

32. Claridge, P. R. P. "Information handling in a large information system." International Conference on Scientific Information, Washington, D. C., 16-21 November 1958. Washington, National Academy of Sciences -- National Research Council. Preprints. 1958. Area 5, p. 377-394.

----- Proceedings. 1959. Vol. 2, p. 1203-1220.

33. Cockayne, A. H., and E. Hyde. "Prime number coding for information retrieval," Computer Journal, 3:1 (April 1960) 21-22.

34. Cockburn, J. G. The Newcastle system of representation of organic compounds. Private communication to the [Codification Commission] of the International Union of Pure and Applied Chemistry, [1951].

35. Costello, J. C., Jr., and Eugene Wall. "Recent improvements in techniques for storing and retrieving information." Taube, Mortimer (comp.), Emerging solutions for mechanizing the storage and retrieval of information. (Studies in coordinate indexing, v. 5) Washington, Documentation Inc., 1959. Ch. 8.

36. Davison, W. H. T. Private communication. 13 February 1959.

37. Davison, W. H. T., and M. Gordon. "Sorting for chemical groups using Gordon-Kendall-Davison ciphers," American Documentation, 8:3 (July 1957) 202-210.

38. de Grolier, Eric. "Problems in scientific communication," IBM Journal of Research and Development, 2:4 (October 1958) 276-281.

39. Dyson, George Malcolm. A new notation and enumeration system for organic compounds. London and New York, Longmans, Green, 1947. 63 p.

---------- 2d ed. London and New York, Longmans, Green, 1949. 138 p.

40. ---------- Private communication. 25 October 1960. Report on a lecture held under the auspices of The Chemical Society, The Society of Chemical Industry, The Royal Institute of Chemistry, London, 1946.

41. ---------- "Research expansion at the Chemical Abstracts Service," Chemical and Engineering News, 37:36 (7 September 1959) 128-131.

42. Edge, Eleanor B., Norman G. Fisher, and Lucy A. Bannister. "System for indexing research reports using a punched card machine," American Documentation, 8:4 (October 1957) 275-283.

43. Edmundson, H. P., V. A. Oswald, Jr., and R. E. Wyllys. Automatic indexing and abstracting of the contents of documents. Los Angeles, Planning Research Corp., 31 October 1959. 133 p. Contract AF 30(602)-1748. Rept. no. PRC-R-126. RADC-TR-59-208. ASTIA doc. no. AD-231 606.

44. "Eliminating the hide-and-seek of indexing," Chemical Engineering Progress, 56:7 (July 1960) 35.

45. Fairthorne, R.A. "Automatic retrieval of recorded information," Computer Journal, 1:1 (April 1958) 36-41.

46. Feldman, Alfred. "Stereo numbers: A short designation for stereoisomers," Journal of Organic Chemistry, 24:10 (October 1959) 1556-1560.

47. Findley, L.C., C.C. Bolze, and R. A. Carpenter. A card controlled routine for searching chemical compound data with an IBM 704. Kansas City, Mo., Midwest Research Institute, 17 November 1958. 12 p.

48. Firth, F. E. An experiment in literature searching with the IBM 305 RAMAC. San Jose, Calif., International Business Machines Corp., 17 November 1958. 10 p.

---------- "An experiment in mechanical searching of research literature with RAMAC." Proceedings of the Western Joint Computer Conference, Los Angeles, 6-8 May 1958. New York, American Institute of Electrical Engineers, March 1959. p. 168-170.

49. Fletcher, J.H., and D. S. Dubbs. "Quick access to research records," Chemical and Engineering News, 34:48 (26 November 1956) 5888-5891.

50. Foster, Laurence S. "Revision of the ASM-SLA metallurgical literature classification," American Documentation, 9:1 (January 1958) 13-19.

---------- Watertown, Mass., Ordnance Materials Research Office, 1 May 1957. 26 p. ASTIA doc. no. AD-134 620.

51. Frear, D.E.H. "Comprehensive coding schemes for chemical compounds." Casey, Robert S., and James W. Perry (eds.) Punched cards--their applications to science and industry. New York, Reinhold, 1951. Ch. 22.

52. Frome, Julius, and Jacob Leibowitz. A punched card system for searching steroid compounds. (Patent Office Research and Development reports, no. 7). Washington, Department of Commerce, 8 July 1957. 5 p.

---------- A manual for coding steroids. (Patent Office Research and Development reports, no. 11). Washington, Department of Commerce, 17 November 1958. 20 p.

53. Frome, Julius. Semi-automatic indexing and encoding. (Patent Office Research and Development reports, no. 17). Washington, Department of Commerce, 10 December 1959. 32 p.

54. Frome, Julius, [et al.] A system of retrieval compounds, compositions, processes and polymers. (Patent Office Research and Development reports, no. 13). Washington, Department of Commerce, 17 November 1958. 14 p.

55. Fugmann, R. Documentation using the electronic computer IBM 705. [Abstract] Frankfurt, Germany, Farbwerke Hoechst AG, 1959. 11 p.
Presented at the 17th International Congress of Pure and Applied Chemistry, Munich, 2 September 1959.

56. Gamble, Dean F'. "A coordinate index of organic compounds," Abstracts of papers, 127th Meeting, American Chemical Society, Cincinnati, 29 March - 7 April 1955. p. 7G.

57. Ganeff, J. M. Private communication. July 1959.

58. Garfield, Eugene. "Citation indexes--new paths to scientific knowledge," Chemical Bulletin, 43:4 (April 1956) 11-12.

59. ---------- "Breaking the subject index barrier--a citation index for chemical patents," Journal of the Patent Office Society, 39:8 (August 1957) 583-595.

60. ---------- "Citation indexes for science," Science, 122:3159 (15 July 1955) 108-111.

61. ---------- "Eugene Garfield Associates, Philadelphia." National Science Foundation, Office of Science Information Service. Current research and development in scientific documentation, no. 5. [Washington] October 1959. p. 18.

Garfield (Eugene) Associates. "New bibliographic service in pharmaceutical publications," [Press release] American Documentation, 9:1 (January 1958) 71-72.

62. Garrett, Peter. A classification system for any data banking (information storage and retrieval) process. Los Angeles, Benson Lehner Corp., 15 June 1959. 16 p. Contract Nonr-266600. Research report no. 59-6. ASTIA doc. no. AD-219 090.

63. General Electric Co. Computer Department. The new GE 225 information processing system. [Brochure] Phoenix, Arizona, 1960.

64. ---------- Proposed implementation plan. Project AUTO MATE. [Brochure] Phoenix, Arizona, 20 May 1960.

65. Gordon, M., C.E. Kendall, and W.H.T. Davison. "A new systematisation of chemical species," Proc. 11th International Congress of Pure and Applied Chemistry, London, 17-24 July 1947. Vol. 2, sec. 3, p. 115-132.

---------- Chemical ciphering; a universal code as an aid to chemical systematics. London, Royal Institute of Chemistry, 1948. 47 p.

66. Gottesman, Jerome, and Edward Gottesman. "Machines, documentation, and automatic coding," American Documentation, 8:2 (April 1957) 129-133.

67. Grandine, J.D., 2nd, S.T. Polyak, and J. R. Schaefgen. Encoding chemical patents for electronic searching and retrieval. Kennett Square, Pa., Kennett Computer Consultants, Inc., 25 September 1959. 7 p.

68. Grandine, Joseph D., 2nd, Eva M. Starr, and Richard E. Putscher. Report index searching on the Bendix G-15D Computer. Wilmington, Del., Patent Div., du Pont Textile Fibers Dept., 3 September 1959. 16 p.
Presented at the 136th Meeting of the American Chemical Society, Division of Chemical Literature, Atlantic City, N.J., 15 September 1959.

69. Green, Julien. "Symbol manipulation in XTRAN," Communications of the Association for Computing Machinery, 3:4 (April 1960) 213-214.

70. Groth Institute. College of Chemistry and Physics. Pennsylvania State University. Abstracts of three papers on the status of the Groth Institute program. University Park, Pa., July 1959. Contract no. AF 49(638)-416. Report no. 35.
Contents. -- Organization and program, by Ray Pepinsky and V. Vand. -- Data handling procedures, by V. Vand and Ray Pepinsky. -- IBM 704 programming, by V. Vand [et al.]

71. Gruber, Wolfgang. "The Geneva nomenclature in code and its extension to ring compounds (abstract)," Angewandte Chemie, 61 (1949) 429-431; Beiheft no. 58, 72 p.

---------- Die Genfer Nomenklatur in Chiffren und Verschlage für ihre Erweiterung auf Ringberbindungen. Weinheim, Verlag Chemie, 1950. 84 p.

72. Gull, C.D., and P.O. Dodge. The transistorized Information Searching Selector. Arlington, Va., and Phoenix, Arizona, General Electric Co., Computer Dept., 1959. 33 p.
Presented at International Conference for Standards on a Common Language for Machine Searching and Translation, Cleveland, 6-12 September 1959.

73. Gutenmakher, L.I. "New types of statistical and information machines in the USSR," translated from Vestnik Akademii Nauk SSSR, Moscow, No. 10 (October 1956) 12-21.

---------- "The problem of an information machine." U.S. Joint Publications Research Service. Soviet developments in information processing and machine translation. (JPRS: 662-D.) New York and Washington, 24 April 1959. p. 7-10.

74. Gutenmakher, L.I., and G.É. Vléduts. The prospects for the utilization of informational-logical machines in chemistry (USSR). (Reports to the Symposium at the VIII Mendeleyev Congress on General and Applied Chemistry). New York, U.S. Joint Publications Research Service, 19 February 1960. (JPRS: R-331-D)
Translated from Problemy Vysshego Khimicheskogo i Tekhnologicheskogo, Obrazovaniya. (Problems of Higher Chemical and Technological Education). Moscow, State Publishing House, 1959.

75. Harris, Zellig S. "Co-occurrence and transformations in linguistic structure," Language, 33:3, Part 1 (July - September 1957) 283-340.

---------- "Linguistic transformations for information retrieval." International Conference on Scientific Information, Washington, D.C., 16-21 November 1958. Washington, National Academy of Sciences -- National Research Council. Preprints. 1958. Area 5, 123-136.

---------- Proceedings. 1959. Vol. 2, p. 937-950.

76. International Business Machines Corp. An experiment in auto-abstracting; auto abstracts of Area 5 conference papers, International Conference on Scientific Information, Washington, D.C., 16-21 November 1958. Yorktown Heights, N.Y., Information Retrieval Research Dept., IBM Research Center, 1958. 18 p.

77. International Business Machines Corp. INFORMER; pulse magnetic retrieval and data processing system. [Brochure] Washington, IBM, Federal Systems Division, [1960].

78. ---------- Lists of actual applications of IBM data processing machines to information indexing, cataloging and retrieving, by Technical Information Services, Dept. 673. Endicott, N.Y., 13 February 1958.

79. International Conference on Scientific Information, Washington, D.C., 16-21 November 1958. Washington, National Academy of Sciences--National Research Council. Preprints. 1958. (Area 2: The function and effectiveness of abstracting and indexing services. 207 p.)

---------- Proceedings. 1959. (Area 2, Vol. 1, p. 313-535.)

80. ---------- Washington, National Academy of Sciences--National Research Council. Preprints. 1958. (Area 6, p. 4: Proposed scope of Area 6.)

---------- Proceedings. 1959. (Area 6, Vol. 2, p. 1273-1274.)

81. ---------- Washington, National Academy of Sciences--National Research Council. Preprints. 1958. (Area 6: Organization of information storage and retrospective search. 123 p.)

---------- Proceedings. 1959. (Area 6, Vol. 2, p. 1269-1409)

82. International Union of Pure and Applied Chemistry. A proposed international chemical notation, tentative version. Prepared by the Commission on Codification, Ciphering and Punched Card Techniques of the IUPAC. London and New York, Longmans, Green, January 1958.

83. Isbell, Horace S. "System for classification of structurally related carbohydrates," Journal of Research of the National Bureau of Standards, 57:3 (September 1956) 171-178.

84. Itek Corp. A program of research and development on information searching systems. Summary. Waltham, Mass., Itek Corp., Information Services Dept., October 1958. 7 p.

---------- Waltham, Mass. Itek Corp., Information Services Dept., January 1959. 7 p.

85. James, Peter (ed.) Literature on information retrieval and machine translation--bibliography and index. New York, Service Bureau Corp., subsidiary of IBM, September 1958. 42 p.

---------- Literature on information retrieval and machine translation--bibliography and auto-index. 2d ed. New York, Service Bureau Corp., subsidiary of IBM, June 1959. 38 p.

86. James, Peter. "A photo-magnetic system for document and information retrieval," American Documentation, 10:4 (October 1959) 286-295.
Also pub. by IBM Research Center, International Business Machines Corp. Yorktown Heights, N. Y., 1 March 1958. 17 p. (RC-66)

87. Jonker Business Machines, Inc. Basic principle of Termatrex information retrieval systems. Gaithersburg, Md., April 1960. (A 11)

----------Main characteristics of Termatrex systems. Gaithersburg, Md., April 1960. (E 11)

----------Available Termatrex equipment. Gaithersburg, Md., April 1960. (B 11)

88. Katz, Charles. "General Electric announces 225," Datamation, 6:4 (July/August 1960) 44-45.

89. Kent, Allen. "Exploitation of recorded information. I: Development of an operational machine searching service for the literature of metallurgy and allied subjects," American Documentation, 11:2 (April 1960) 173-188.

90. Kent, Allen, Robert E. Booth, and James W. Perry. "Machine searching of metallurgical literature," Metal Progress, 71:2 (February 1957) 71-75.

91. Kent, Allen. Nonconventional retrieval systems in documentation. Preliminary comparative analysis. Cleveland, Western Reserve Univ., Center for Documentation and Communication Research, 24 June 1958. 25 p. Contract AF 49 (638) 357. Technical note no. 3. AD-158 396.

92. Kent, Allen and James W. Perry. "Searching metallurgical literature." Casey, Robert S., James W. Perry, Madeline M. Berry, and Allen Kent (eds.) Punched cards -- their applications to science and industry. 2d ed. New York, Reinhold, 1958. Ch. 11.

93. Kent, Allen, and A. S. Iberall. "Soviet documentation -- a trip report," American Documentation. 10:1 (January 1959) 1-19.

94. Kessel, B., and A. DeLucia. "A specialized library index search computer." Proceedings of the Western Joint Computer Conference, San Francisco, 3-5 March 1959. New York, Institute of Radio Engineers for the Joint Computer Committee, 1959. p. 57-59.

95. Koelewijn, G. J. "The possibilities of far-reaching mechanization of novelty search of the patent literature." International Conference on Scientific Information, Washington, D. C., 16-21 November 1958. Washington, National Academy of Sciences--National Research Council. Preprints. 1958. Area 5, p. 245-270.
------ Proceedings. 1959. Vol. 2, p. 1071-1096.

---------- "Mechanische Auswahl nach Begriffen und deren Beziehungen zueinander," [Abstract] Automatic Documentation in Action (ADIA), Internationale Arbeitstagung, Frankfurt, Main, 9 - 12 June 1959. Vorberichte. Preprints. Frankfurt, Main, 1959. p. 10-11.

96. Kokie, James E. An experimental numerical information retrieval system. Presented at 15th National Conference of the Association for Computing Machinery, Milwaukee, Wis., 23-26 August 1960.

97. Koller, Herbert R., Ethel Marden, and Harold Pfeffer. "The HAYSTAQ System: past, present, and future." International Conference on Scientific Information, Washington, D. C., 16-21 November 1958. Washington, National Academy of Sciences -- National Research Council. Preprints. 1958. Area 5, p. 317-353.
------ Proceedings. 1959. Vol. 2, p. 1143-1179.

98. Kuder, Milton L. Automatic information sorting system. U. S. Patent no. 2,907,011, issued 29 September 1959.

99. Kuipers, J. W., A. W. Tyler, and W. L. Myers. "A Minicard system for documentary information," American Documentation, 8:4 (October 1957) 246-268.
Also pub. in Shera, J. H., Allen Kent, and James W. Perry (eds.) Information systems in documentation. New York, Interscience, 1957. Ch. 27.

100. Kuipers, John W. A research program on information searching systems. Waltham, Mass., Itek Corp., 5 August 1959. 27 p.

101. Leibowitz, Jacob, Julius Frome, and F. D. Hamilton. "Chemical language coding for machine searching," Abstracts of papers, 135th Meeting, American Chemical Society, Boston, 5-10 April 1959. p. 3G.

Leibowitz, Jacob, Julius Frome, and Don D. Andrews. Variable scope patent searching by an inverted file technique. (Patent Office Research and Development reports, no. 14). Washington, Department of Commerce, 17 November 1958. 11 p.

102. Leibowitz, Jacob, Julius Frome, and Don D. Andrews. "Variable scope search system: VS_3." International Conference on Scientific Information, Washington, D.C., 16-21 November 1958. Washington, National Academy of Sciences--National Research Council. Preprints. 1958. Area 5, p. 291-316.
------ Proceedings, 1959. Vol. 2, p. 1117-1142.

103. Leiter, Joseph, Marvin Schneiderman, and Eugene Miller. Data processing program of the Cancer Chemotherapy National Service Center, utilizing the IBM 305 (RAMAC). Presented at the 136th Meeting of the American Chemical Society, Division of Chemical Literature, Atlantic City, N.J., 13-18 September 1959.

104. Levin, P. Tools and methods for searching the chemical literature: a selective bibliography. Thesis (M.S.L.S.) -- Drexel Institute of Technology, 1955. 41 p.

105. Loftus, Helen E., and Allen Kent. "Automation in the library - an annotated bibliography," American Documentation, 7:2 (April 1956) 110-126.

-----"An annotated bibliography." Perry, James W., Allen Kent, and Madeline M. Berry, Machine literature searching. New York, Interscience, 1956. p. 135-147.

106. Luhn, H.P. Identification of geometric patterns by topological description of their envelopes. (IBM technical report). Poughkeepsie, N.Y., International Business Machines Corp., 23 April 1956. 6 p.

-----A serial notation for describing the topology of multidimensional branched structures (nodal index for branched structures). Yorktown Heights, N.Y., IBM Research Center, International Business Machines Corp., 12 December 1955. 17 p. (Research report RC-27)

107. -----Auto-encoding of documents for information retrieval systems. Yorktown Heights, N.Y., IBM Research Center, 1958. 7p.

-----Boaz, Martha (ed.) Modern trends in documentation. New York, Pergamon, 1959. p. 45-58.

108. -----Potentialities of auto-encoding of scientific literature. Yorktown Heights, N.Y., IBM Research Center, International Business Machines Corporation, 15 May 1959. 22 p. (Research report, RC-101).
Also pub., without figures, in: Automatic Documentation in Action (ADIA), Internationale Arbeitstagung, Frankfurt, Main, 9-12 June 1959. Vorberichte. Preprints. Frankfurt, Main, 1959. p. 17-27.

109. Luhn, H. P. Information retrieval through row-by-row scanning on the IBM 101 electronic statistical machine (row-by-row scanning attachment). Yorktown Heights, N. Y., International Business Machines Corp., 17 November 1958. 11 p.

----Row-by-row scanning systems for IBM punched cards as applied to information retrieval problems. Yorktown Heights, N. Y., IBM Research Center, 8 May 1959. (Research report RC-100)

110. ----The IBM electronic information searching system. Yorktown Heights, N. Y., IBM Research Center, International Business Machines Corp., 29 May 1952.
Presented in part at the Symposium on Machine Techniques for Information Selection, Massachusetts Institute of Technology, Industrial Liaison Program, Cambridge, Mass., 10-11 June 1952.

----The IBM universal card scanner for punched card information searching systems. Yorktown Heights, N. Y., International Business Machines Corp., November 1958. 24 p.
Also pub. in Taube, Mortimer (comp.) Emerging solutions for mechanizing the storage and retrieval of information. (Studies in coordinate indexing, vol. 5). Washington, Documentation Inc., 1959. Ch. 7.

111. ----"A statistical approach to mechanized coding and searching of literary information," IBM Journal of Research and Development, 1:4 (October 1947) 309-317.
A revision of his paper, "A statistical approach to mechanized literature searching." Poughkeepsie, N. Y., IBM Research Center, International Business Machines Corp., 30 January 1957. 21 p. Research paper RC-3.

----"The automatic creation of literature abstracts (auto-abstracts)," IBM Journal of Research and Development, 2:2 (April 1958) 159-165.
Also pub. in IRE National Convention Record, 1958. New York, Institute of Radio Engineers, 1958. Vol. 6, part 10, p. 20-24.

112. McCarthy, John. "Recursive functions of symbolic expressions and their computation by machine. Part I," Communications of the Association for Computing Machinery, 3:4 (April 1960) 184-195.

McIlroy, M. Douglas. "Macroinstruction extensions of compiler languages," Communications of the Association for Computing Machinery, 3:4 (April 1960) 214-220.

113. Maloney, Clifford J. "A Remington-Rand punched card data retrieval system," American Documentation, 9:1 (January 1958) 1-12.

114. Marcus, Michael Barry. A computer system for simultaneous recall. Albuquerque, New Mexico, Sandia Corp., 1 June 1957. 33 p. AEC Contract AT-(29-1)-789. Tech. memo SCTM-173-57 (14).

115. Meyer, Ernst. Ein System zur topologischen Verschüsselung chemischer Strukturformeln für die mechanisierte Dokumentation. [A system for the topological coding of chemical structures for mechanized documentation.] Ludwigshafen A. Rhein, Badische Anilin- & Soda-Fabrik AG, Hauptlaboratorium (BASF), 10 November 1959. 11 p.

116. Miller, Eugene, Delbert Ballard, John Kingston, and Mortimer Taube. "Conventional and inverted grouping of codes for chemical data." International Conference on Scientific Information, Washington, D.C., 16-21 November 1958. Washington, National Academy of Sciences--National Research Council. Preprints. 1958. Area 4, p. 9-24.
------ Proceedings. 1959. Vol. 1, p. 671-685.

117. Mitchell, R.P., and E. Greer. Word correlation study I. Sunnyvale, Calif., Lockheed Aircraft Corp., Missile Systems Division, December 1958. 12 p. Contract no. AF 30(602) 1889. ASTIA doc no. AD-208 288.

Mitchell, R.P. Word correlation study. Sunnyvale, Calif. Lockheed Aircraft Corp., Missiles and Space Division. Contract AF 30(602)-1889.
---------- II. 1958. ASTIA no. AD-212 425.
---------- II. -Supp. 1958. ASTIA no. AD-212 425.
--------- III. 1959. ASTIA no. AD-210 727.
---------- Final report. 1959. ASTIA no. AD-230 408.

118. Mooers, Calvin N. "The next twenty years in information retrieval; some goals and predictions," American Documentation, 11:3 (July 1960) 229-236.
Also pub. by Zator Company, under Contract AF 49 (638) 376, Rept. no. ZTB-121. Cambridge, Mass., March 1959. 18 p. ASTIA doc. no. AD-212 225.

119. Mooers, Calvin N., and Ray J. Solomonoff. "Zator Company, Cambridge, Mass." National Science Foundation, Office of Science Information Service. Current research and development in scientific documentation, no. 5. [Washington] October 1959. p. 45-46.

120. Murphy, R.W. The IBM 9900 Special Index Analyzer. Poughkeepsie, N.Y., International Business Machines Corp., 17 November 1958. 24 p.
Also pub. in Taube, Mortimer (comp.) Emerging solutions for mechanizing the storage and retrieval of information. (Studies in coordinate indexing, vol. 5). Washington, Documentation Inc., 1959. Ch. 6.

121. National Science Foundation. Office of Science Information Service. Non-conventional technical information systems in current use. No. 2. Washington, September 1959. (NSF-59-49).

122. Newell, Allen, and F. Tonge. "An introduction to Information Processing Language V," Communications of the Association for Computing Machinery, 3:4 (April 1960) 205-211.

123. Newman, Simon M. Storage and retrieval of contents of technical literature -- nonchemical information. First supplementary report. (Patent Office Research and Development reports, no. 4). Washington, Department of Commerce, June 1957. 16 p.

---------- Second supplementary report. (Patent Office Research and Development reports, no. 12). Washington, Department of Commerce, 1958. 16 p.

----Linguistic problems in mechanization of patent searching. (Patent Office Research and Development reports, no. 9). Washington, Department of Commerce, 1957. 10 p.

----Analysis of prepositionals for interrelational concepts. Preliminary study. (Patent Office Research and Development reports, no. 16). Washington, Department of Commerce, 15 July 1959. 39 p.

124. Nolan, J.J. Principles of information storage and retrieval using a large scale random access memory. San Jose, Calif., International Business Machines Corp., 17 November 1958. 14 p.

---------- "Information storage and retrieval using a large scale random access memory," American Documentation, 10:1 (January 1959) 27-35.

125. Norton, T.R., and A. Opler. A manual for coding organic compounds for use with a mechanized searching system. Revised. Pittsburg, Calif., Research Dept., Western Div., Dow Chemical Co., 15 March 1956. 55 p.

126. Nutting, H.S., and S. P. Klesney. A diversified approach to the structure searching problem. The Author (Dow Chemical Co., Midland, Mich.).
Presented at the Division of Chemical Literature Symposium, American Chemical Society, Pittsburgh, 19-21 January 1958.

127. Ohlman, Herbert. Chronological bibliography of permutation indexing. The Author (System Development Corp., Santa Monica, Calif.) 1960.

128. Opler, Ascher. "Utilization of computers for information retrieval," Proceedings of the Fifth Annual Computer Applications Symposium. Chicago, Armour Research Foundation, Illinois Institute of Technology, 29-30 October 1958. p. 22-29.

129. ---------- "Dow refines structural searching," Chemical and Engineering News, 35:33 (19 August 1957) 92-96.

---------- "A topological application of computing machines." Proceedings of the Western Joint Computer Conference, San Francisco, 7-9 February 1956. New York, American Institute of Electrical Engineers for the Joint Computer Committee, 1956. p. 86-88.

Opler, Ascher, and T. R. Norton. A manual for programming computers for use with a mechanized system for searching organic compounds. Pittsburg, Calif., Research Dept., Western Div., Dow Chemical Co., 25 April 1956. 23 p.

---------- "New speed to structural searches," Chemical and Engineering News, 34:23 (4 June 1956) 2812-2814, 2816.

130. Opler, Ascher, and Norma Baird. "Display of chemical structural formulas as digital computer output," American Documentation, 10:1 (January 1959) 59-63.

Opler, Ascher. "On the automatic manipulation of representations of chemical structures," American Documentation, 10:2 (April 1959) 130-134.

131. "Organization of information for storage and retrospective search." International Conference on Scientific Information, Washington, D.C., 16-21 November 1958. Washington, National Academy of Sciences--National Research Council. Preprints, 1958. Area 6, 123 p.
------ Proceedings. 1959. Vol. 2, p. 1269-1409.

132. Patterson, Helen A. "A punch-card code for cardiovascular pathology," American Documentation, 9:2 (April 1958) 77-83.

133. Patterson, Helen A., Paul R. Ackley, and Nancy B. Rehmeyer. A system for context storage and retrieval of information from the published literature, applicable to both the IBM 101 and electronic computers. Presented at the Conference on Information Retrieval, International Business Machines Corp., Poughkeepsie, N.Y., 21-23 Sept. 1959. 25 p., preprint.

134. Pepinsky, Ray. Summary report on the Groth Institute. University Park, Pa., Groth Institute, College of Chemistry and Physics, 21 October 1959. 5 p. Contract no. AF 49(638) -416. Report no. 36. ASTIA doc. no. AD-228 902.

135. Perlis, Alan J., and Charles Thornton. "Symbol manipulation by threaded lists," Communications of the Association for Computing Machinery, 3:4 (April 1960) 195-204.

136. Pfeffer, Harold, Herbert R. Koller, and Ethel C. Marden. "A first approach to patent searching procedures on Standard's Electronic Automatic Computer (SEAC)", American Documentation, 10:1 (January 1959) 20-26.

---------- Another issue. (Patent Office Research and Development reports, no. 10). Washington, Department of Commerce, 28 January 1958. 15 p.

137. Plankeel, F.H. "Automation in documentation, a mechanized coordinate index system," American Documentation, 11:2 (April 1960) 128-134.

138. Polyak, S.T. A program for chemical patent searching on the IBM 305 RAMAC. Kennett Square, Pa., Kennett Computer Consultants, Inc., 15 September 1959. 25, 11 p. Contract CC-4488, U.S. Patent Office.

139. Rakov, B.M., and V. P. Cherenin. Experimental information machine of the Institute of Scientific Information, Academy of Sciences USSR (Eksperimental naya informatsionnaya masina Instituta Nauchboy Informatsii AN SSR). Moscow, 1955.
Review. Perry, James W., and Allen Kent. "The Russians have a machine for it, a review," American Documentation, 7:3 (July 1956) 233-234.

140. Rakov, B.M., and V.P. Cherenin. "Machines for retrieving information in the USSR," UNESCO Bulletin for Libraries, 11:8-9 (August-September 1957) 192-197.

---------- Translated into German.
"Maschinelle Dokumentation in der U. d. S.S.R.," Nachrichten für Dokumentation, 9:1 (March 1958) 30-34.

141. "The Rapid Selector, an automatic document retrieval device," Technical News Bulletin, National Bureau of Standards, 43:10 (October 1959) 178-179.

142. Ray, Louis C., and Russell A. Kirsch. "Finding chemical records by digital computers," Science, 126:3278 (25 October 1957) 814-819.

143. Rees, Janet, and Allen Kent. "Mechanized searching experiments using the WRU Searching Selector. Preliminary report," American Documentation, 9:4 (October 1958) 277-303.
Also published by Center for Documentation and Communication Research, Western Reserve Univ., under Contract AF 49(638)357. Technical note no. 1. Cleveland, 15 May 1958. ASTIA doc. no. AD-158 250.

144. Rockwell, Harriet E., Robert L. Hayne, and Eugene Garfield. "A unique system for rapid access to large volumes of pharmacological data; application to published literature on chlorpromazine," Federation Proceedings, 16:3 (September 1957) 726-731.

145. Samain, Jacques. The organization of documentation by the Filmorex technique. Paris, Filmorex, 1956. 7 p.

---------- "Documentation by the Filmorex technique." Shera, J.H., Allen Kent, and James W. Perry (eds.) Information systems in documentation. New York, Interscience, 1957. Ch. 26.

---------- "Filmorex--a new method of documentation," Nachrichten für Dokumentation, 9:1 (March 1958) 35-40.

---------- "The Filmorex system," [abstract] Automatic Documentation in Action (ADIA), Internationale Arbeitstagung, Frankfurt, Main, 9-12 June 1959. Vorberichte. Preprints. Frankfurt, Main, 1959. p. 32-34.

146. Savage, T.R. The preparation of auto-abstracts on the IBM 704 Data Processing System. Yorktown Heights, N.Y., IBM Research Center, 17 November 1958. 11 p.

147. Schaefgen, J.R. Encoding patents for electronic searching. Kennett Square, Pa., Kennett Computer Consultants, Inc.; 1 August 1959. 9 p. Contract CC-4488, U.S. Patent Office.

148. Schneider, K. "Der Einsatz der IBM 650 in der Dokumentation demonstriert an Beispielen der Dokumentation der Atomkernenergie," Automatic Documentation in Action (ADIA), Internationale Arbeitstagung, Frankfurt, Main, 9-12 June 1959. Vorberichte. Preprints. Frankfurt, Main, 1959.

149. Schneider, K. "Der Einsatz der Remington Rand Univac FAC-TRONIC I für die automatische Sortierung von Sachregistern, demonstriert an Beispielen für die Schriftenreihen des Referates Atomkernenergie-Dokumentation beim Gmelin-Institut," [Abstract] Automatic Documentation in Action (ADIA), Internationale Arbeitstagung, Frankfurt, Main, 9-12 June 1959. Vorberichte. Preprints. Frankfurt, Main, 1959. p. 39.

150. ---------- "Der Einsatz der Remington Rand Univac FAC-TRONIC I in der Dokumentation, demonstriert an Literatur über Buntmetall-Legierungen," [Abstract]. Automatic Documentation in Action (ADIA), Internationale Arbeitstagung, Frankfurt, Main, 9-12 June 1959. Vorberichte. Preprints. Frankfurt, Main, 1959. p. 35-38.

151. Schultz, Claire K. "An application of random codes for literature searching." Casey, Robert S., James W. Perry, Madeline M. Berry, and Allen Kent (eds.). Punched cards -- their applications to science and industry. 2d ed. New York, Reinhold, 1958. Ch. 10.

152. Schultz, Claire K., and Clayton A. Shepherd. A computer analysis of the Merck Sharp and Dohme indexing system. Philadelphia, Applied Mathematics Department, Remington Rand UNIVAC, Division of Sperry Rand Corp., 1959, 1 v. Contract Nonr-229700. ASTIA doc. no. AD-229 380.

153. Schutze, Gertrude. "Review of documentation literature and activities, 1957," Special Libraries, 49:2 (February 1958) 70-84.

154. Semarne, H. M. "Symbolic language in language engineering." Proceedings of the Western Joint Computer Conference, San Francisco, 3-5 May 1960. New York, Western Joint Computer Conference, 1960. p. 61-71.

155. Shaw, J. C., A. Newell, H. A. Simon, and T. O. Ellis. "A command structure for complex information processing." Proceedings of the Western Joint Computer Conference, Los Angeles, 6-8 May 1958. New York, American Institute of Electrical Engineers, March 1959. p. 119-128.
Also pub. by RAND Corp., Santa Monica, Calif. (RAND P-1277).

156. Shaw, Ralph R. "Machines and the bibliographical problems of the twentieth century." Bibliography in an age of science. Urbana, University of Illinois Press, 1951. p. 37-71.

157. Silk, J.A. A new system of organic notation. Private communication, 1951.

158. Slade, A.E., and H. O. McMahon. "A Cryotron catalog memory system." Proceedings of the Eastern Joint Computer Conference, New York City, 10-12 December 1956. New York, American Institute of Electrical Engineers, 1957. p. 115-119.

159. Smith, Elbert G. "Machine searching for chemical structures," Science, 131:3394 (15 January 1960) 142-146.

160. Smith, Joseph. "Syntactic and semantic augments to ALGOL," Communications of the Association for Computing Machinery, 3:4 (April 1960) 211-213.

161. Stevens, M.E. A machine model of recall. Presented at International Conference on Information Processing, Paris, 13-15 June 1959. Preprint no. UNESCO/NS/ICIP/J. 5. 4. 14 p.

162. ---------- A survey of automatic reading techniques. Washington, National Bureau of Standards. (in preparation)

163. Taube, Mortimer. "The COMAC: an efficient punched card collating system for the storage and retrieval of information." International Conference on Scientific Information, Washington, D.C., 16-21 November 1958. Washington, National Academy of Sciences--National Research Council. Preprints. 1958. Area 5, p. 419-428.
------ Proceedings. 1959. Vol. 2, p. 1245-1254. Ch. 5.

---------- Emerging solutions for mechanizing the storage and retrieval of information, compiled by Mortimer Taube. (Studies in coordinate indexing, v. 5). Washington, Documentation Inc., 1959.

164. Taube (Mortimer) and Associates. "The logic and mechanics of storage and retrieval systems," in their Studies in coordinate indexing. Washington, Documentation Inc., 1956. Vol. 3, Ch. 4.

165. Thompson Ramo Wooldridge Inc. Ramo-Wooldridge Division. Word correlation and automatic indexing. Submitted to the Council on Library Resources. Progress report no. 1. Los Angeles, 21 September 1959. (C82-9U9)
---------- Progress report no. 2. Canoga Park, Calif., 21 December 1959. (C82-0U1)
---------- Phase I final report. An experiment in automatic text searching. Canoga Park, Calif., 30 April 1960. (C82-0U4)

166. Touloukian, Y.S., [et al.] "Systems and procedures developed for the search, coding, and mechanized processing of bibliographic information on thermophysical properties." American Society of Mechanical Engineers. Thermodynamic and transport properties of gases, liquids, and solids. New York, McGraw-Hill, 1959. p. 78-91.

167. U.S. Congress. Senate. Committee on Government Operations. Documentation, indexing, and retrieval of scientific information. A study of federal and non-federal science information processing and retrieval programs. (Senate document no. 113. 86th Congress, 2d session. 1960) Washington, U. S. Government Printing Office, 1960. 283 p.

168. U.S. Department of Commerce. Advisory Committee on Application of Machines to Patent Office Operations. Report to the Secretary of Commerce. V. Bush, Chairman. Washington, Government Printing Office, December 1954. 76 p.

169. U. S. National Bureau of Standards. The Rapid Selector -- an automatic document retrieval device. (Summary technical report STR-2388) Washington, September 1959. 4 p.

170. Vand, V., and Ray Pepinsky. Specifications for the Groth Institute IBM 704 programs. University Park, Pa., Groth Institute, College of Chemistry and Physics, 23 March 1959. 8 p. Contract no. AF 49(638)-416. Report no. 23. ASTIA doc. no. AD-228 916.

171. Vléduts, G. É. Methods of improving chemical indexes. (Soviet developments in information processing and machine translation. New York, U.S. Joint Publications Research Service, 5 August 1960. JPRS: 3613. 61 p.) Translation of Russian monograph, Nekotoryye Voprosy Nauchnoy Informatsii v Oblasti Khimii (Some problems of scientific information in chemistry) Moscow, 1958, 78 p.

172. Vléduts, G. É., and V. K. Finn. "Concerning certain problems of chemical linguistics connected with machine search for information." [Abstract]
In Abstracts of the Conference on Mathematical Linguistics, Leningrad, 15-21 April 1959. New York, U. S. Joint Publications Research Service, 31 August 1959. JPRS:893-D. p. 84-86.

173. Vléduts, G. É., V. V. Nalimov, and N. I. Styazhkin. "Scientific and technical information as one of the problems of cybernetics", Soviet Physics Uspekhi, 2(69):5 (September-October 1959) 637-781.
Originally pub. in Uspekhi Fizicheskikh Nauk, 69:1 (September 1959) 13-56.
Partial contents: Part 4. Documentation in the field of chemistry (p. 647-652). --Part 5. Representation of results of experimental research in a form suitable for coding (p. 652-653).

174. Waldo, W. H., and M. DeBacker. "Printing chemical structures electronically: encoded compounds searched generically with IBM-702." International Conference on Scientific Information, Washington, D. C., 16-21 November 1958. Washington, National Academy of Sciences -- National Research Council. Preprints. 1958. Area 4, p. 49-68.
------ Proceedings. 1959. Vol. 1, p. 711-730.

175. Waldo, W. H., R. S. Gordon, and J. D. Horcher. "Routine report writing by computer," American Documentation, 9:1 (January 1958) 28-31.

176. Waldo, W. H. Private communication. 12 October 1960.

177. Wall, Eugene. "The grouping and arrangement of terms, items, and their codes." Taube, Mortimer, and Harold Wooster (eds.), Information storage and retrieval theory, systems, and devices. New York, Columbia University Press, 1958. Ch. 12.

---------- Terms, items, and codes in documentation, an introduction to a panel discussion. E. I. du Pont de Nemours, Wilmington, Del., The Author, 13, 5 p.
Presented at Air Force Office of Scientific Research, Professional Symposium on Information Storage and Retrieval Theory, Systems and Devices, Washington, 18 March 1958.

---------- "Information systems," Chemical Engineering Progress, 55:1 (January 1959) 55-59.

178. Wang, Hao. "Proving theorems by pattern recognition, I," Communications of the Association for Computing Machinery, 3:4 (April 1960) 220-234.

179. Ware, Willis H. (ed.) "Soviet computer technology--1959," by S. N. Alexander [et al.] Communications of the Association for Computing Machinery, 3:3 (March 1960) 131-166.

180. Watson-Watt, Sir Robert. "Are computers important?" Proceedings of the Eastern Joint Computer Conference, New York City, 10-12 December 1956. New York, American Institute of Electrical Engineers, 1957. p. 67-68.

181. Weber, Eberhard. A suggestion for an improved data banking system. Los Angeles, Benson Lehner Corp., 15 July 1959. 29 p. Contract Nonr-266600. Research report no. 59-8. ASTIA doc. no. AD-225 097.

182. Weil, Ben H., and Barbara Hildenbrand. "Information on fuel and lubricant additives from machine-sorted punched cards." Shera, J.H., Allen Kent, and James W. Perry (eds.). Information systems in documentation. New York, Interscience, 1957. Ch. 17.

183. Welt, Isaac D. "A combined indexing-abstracting system." International Conference on Scientific Information, Washington, D.C., 16-21 November 1958. Washington, National Academy of Sciences--National Research Council. Preprints. 1958. Area 2, p. 135-146.
------ Proceedings. 1959. Vol. 1, p. 449-459.

184. Welt, Isaac D. "The detailed indexing of biological effects of chemical substances," Bulletin of the Medical Library Association, 46:3 (July 1958) 367-380.

185. Western Reserve University. Center for Documentation and Communication Research. Bibliography of the publications produced by the staff. Cleveland, 1959. 10 p.

186. Whaley, Fred R. "A deep index for internal technical reports." Shera, J.H., Allen Kent, and J. W. Perry (eds.) Information systems in documentation. New York, Interscience, 1957. Ch. 21.

---- In Conference on Multiple Aspect Searching for Information Retrieval, Washington 12-13 February 1957. [Proceedings] Washington, Armed Services Technical Information Agency, 1957. p. 66-82.

------- Operational experience with Linde's indexing and retrieval system. Tonawanda, N.Y., Linde Co., Division of Union Carbide Corp. Preprint, 16 p. Presented at Information Retrieval Seminar, IBM Corp., Poughkeepsie, N.Y., 21-23 September 1959.

------ "Retrieval questions from the use of Linde's indexing and retrieval system." International Conference on Scientific Information, Washington, D.C., 16-21 November 1958. Washington, National Academy of Sciences--National Research Council. Preprints. 1958. Area 4, p. 101-108.
------ Proceedings. 1959. Vol. 1, p. 763-769.

187. Wheeler, K. W. [et al.] "A structure code for organic compounds," American Documentation, 9:3 (July 1958) 198-207.

188. Wiener, J. Magnacard -- a new high-speed filing system. Los Angeles, Magnavox Research Labs., 1957. 8 p.

189. Wildhack, W.A., J. Stern, and W. Smith. "Documentation in instrumentation," American Documentation, 5:4 (October 1954) 223-237.
Also pub. in Taube (Mortimer) and Associates. Studies in coordinate indexing, v. 3. Washington, Documentation Inc., 1956. Ch. 7.

190. Wildhack, W.A., and J. Stern. "The peek-a-boo system -- optical coincidence subject cards in information searching." Casey, Robert S., James W. Perry, Madeline M. Berry, and Allen Kent (eds.) Punched cards -- their applications to science and industry. 2d ed. New York, Reinhold, 1958. Ch. 6.
Also pub. in Conference on Multiple Aspect Searching for Information Retrieval, Washington, 12-13 February 1957. [Proceedings] Armed Services Technical Information Agency [1957] p. 101-134. ASTIA doc. no. AD-147 491.

191. [Williams, Thyllis] Translating from ordinary discourse into formal logic -- a preliminary systems study. Scientific Report. Alexandria, Va., ACF Industries, Inc., September-November 1956. 110 p. Contract AF 19(604) 1582. Tech. note no. AFCRC-TN-56-770. ASTIA doc. no. AD-98 813.

192. Wiswesser, W.J. Simplified chemical coding for automatic sorting and printing machinery. Reading, Pa., Willson Products Inc., 1951.

---------- A line-formula chemical notation. New York, Thomas Y. Crowell Co., 1954.

193. Wood, G. Congdon. "Biological subject-indexing and information retrieval by means of punched cards," Special Libraries, 47:1 (January 1956) 26-31.

---------- "A detailed biology code for storing, retrieving, and correlating chemical-biological data," American Documentation, 8:3 (July 1957) 168-180.

194. Worsley, Peter K. "Data retrieval with especial application to use of film library instantaneous presentation (FLIP) in literature searching." Boaz, Martha (ed.) Modern trends in documentation. New York, Pergamon, 1959. p. 70-73.

195. Worsley, Peter K. [et al.] A study of the fundamentals of information storage and retrieval. Los Angeles, Benson Lehner Corp., 20 August 1959. 97 p. Contract Nonr-266600. Final report no. 417. ASTIA doc. no. AD-229 709.

196. Yngve, Victor H. "A programming language for mechanical translation," Mechanical Translation, 5:1 (July 1958) 25-41.

---------- In defense of English. Presented at An International Conference for Standards on a Common Language for Machine Searching and Translation, 6-12 September 1959, Cleveland. 8 p., preprint.

---------- The COMIT System for mechanical translation. Presented at International Conference on Information Processing, Paris, 13-15 June 1959. Preprint no. UNESCO/NS/ICIP F. 4. 3. 10 p.

197. Zuckermann, A. M., and A. P. Terentiev. Chemical nomenclature translation. Presented at An International Conference for Standards on A Common Language for Machine Searching and Translation, 6-12 September 1959, Cleveland. Preprint, 8 p.

USCOMM--DC

U.S. DEPARTMENT OF COMMERCE
Frederick H. Mueller, *Secretary*

NATIONAL BUREAU OF STANDARDS
A. V. Astin, *Director*

THE NATIONAL BUREAU OF STANDARDS

The scope of activities of the National Bureau of Standards at its major laboratories in Washington, D.C., and Boulder, Colo., is suggested in the following listing of the divisions and sections engaged in technical work. In general, each section carries out specialized research, development, and engineering in the field indicated by its title. A brief description of the activities, and of the resultant publications, appears on the inside of the front cover.

WASHINGTON, D.C.

ELECTRICITY. Resistance and Reactance. Electrochemistry. Electrical Instruments. Magnetic Measurements. Dielectrics.

METROLOGY. Photometry and Colorimetry. Refractometry. Photographic Research. Length. Engineering Metrology. Mass and Scale. Volumetry and Densimetry.

HEAT. Temperature Physics. Heat Measurements. Cryogenic Physics. Rheology. Molecular Kinetics. Free Radicals Research. Equation of State. Statistical Physics. Molecular Spectroscopy.

RADIATION PHYSICS. X-Ray. Radioactivity. Radiation Theory. High Energy Radiation. Radiological Equipment. Nucleonic Instrumentation. Neutron Physics.

CHEMISTRY. Surface Chemistry. Organic Chemistry. Analytical Chemistry. Inorganic Chemistry. Electrodeposition. Molecular Structure and Properties of Gases. Physical Chemistry. Thermochemistry. Spectrochemistry. Pure Substances.

MECHANICS. Sound. Pressure and Vacuum. Fluid Mechanics. Engineering Mechanics. Combustion Controls.

ORGANIC AND FIBROUS MATERIALS. Rubber. Textiles. Paper. Leather. Testing and Specifications. Polymer Structure. Plastics. Dental Research.

METALLURGY. Thermal Metallurgy. Chemical Metallurgy. Mechanical Metallurgy. Corrosion. Metal Physics.

MINERAL PRODUCTS. Engineering Ceramics. Glass. Refractories. Enameled Metals. Constitution and Microstructure.

BUILDING RESEARCH. Structural Engineering. Fire Research. Mechanical Systems. Organic Building Materials. Codes and Safety Standards. Heat Transfer. Inorganic Building Materials.

APPLIED MATHEMATICS. Numerical Analysis. Computation. Statistical Engineering. Mathematical Physics.

DATA PROCESSING SYSTEMS. Components and Techniques. Digital Circuitry. Digital Systems. Analog Systems. Applications Engineering.

ATOMIC PHYSICS. Spectroscopy. Radiometry. Mass Spectrometry. Solid State Physics. Electron Physics. Atomic Physics.

INSTRUMENTATION. Engineering Electronics. Electron Devices. Electronic Instrumentation. Mechanical Instruments. Basic Instrumentation.

Office of Weights and Measures.

BOULDER, COLO.

CRYOGENIC ENGINEERING. Cryogenic Equipment. Cryogenic Processes. Properties of Materials. Gas Liquefaction.

IONOSPHERE RESEARCH AND PROPAGATION. Low Frequency and Very Low Frequency Research. Ionosphere Research. Prediction Services. Sun-Earth Relationships. Field Engineering. Radio Warning Services.

RADIO PROPAGATION ENGINEERING. Data Reduction Instrumentation. Radio Noise. Tropospheric Measurements. Tropospheric Analysis. Propagation-Terrain Effects. Radio-Meteorology. Lower Atmosphere Physics.

RADIO STANDARDS. High frequency Electrical Standards. Radio Broadcast Service. Radio and Microwave Materials. Atomic Frequency and Time Standards. Electronic Calibration Center. Millimeter-Wave Research. Microwave Circuit Standards.

RADIO SYSTEMS. High Frequency and Very High Frequency Research. Modulation Research. Antenna Research. Navigation Systems. Space Telecommunications.

UPPER ATMOSPHERE AND SPACE PHYSICS. Upper Atmosphere and Plasma Physics. Ionosphere and Exosphere Scatter. Airglow and Aurora. Ionospheric Radio Astronomy.


NBS