with Sorter and Collater. Some of the simpler parts of *Chemical Titles* can be made on the IBM 407. However, the 407 is primarily an accounting machine and it cannot make logic choices. For the type of logic choices required, some time on a Computer is necessary.

The heart of the operational planning for literature searching and data retrieval is the IBM 1401 Computer. The 1401 was chosen because it offers a number of advantages. The operators are freed from the necessity of fixed-field coding. Ever since 1946, it has been possible to write ciphers unique for the structures of organic compounds, but equipment was not available which was capable of free-field operation and able to write lower-case letters, subscripts and superscripts. The 1401 fills both needs; it reads the entire card, from column 80 back to column 1, it provides 120 characters including lower-case letters, subscripts, superscripts, plus and minus signs, underlined numbers (essential to the cipher), and a number of special characters. It is no longer necessary to waste record-space by reserving fields for data not immediately available. It is no longer necessary to arrange data in columns, nor count spaces. It is no longer necessary to devise cumbersome codes for two-letter symbols of the elements; they can be written conventionally, with an upper-case and a lower-case letter. It is not necessary to segregate letters from numbers; the IBM 1401 can intersperse them at will. In addition, the 1401 is relatively easy to operate. It has a big memory (8000 characters) and more memory can be added, in successive 8000 character units.

From the searchers' point of view, the most exciting thing about the 1401 is the speed of the tape system. The large (10.5 inches in diameter) reels of magnetic tape (2400 feet) have a capacity of 16,012,800 characters in memory; in most records, about 14,000,000 are usable. Arithmetically, a reel equals about 175,000 cards; actually this is a little bit high because record design, and the kind of blocking maintained, influenced the capacity.

Preliminary counts, made in several ways from existing indexes, led to the conclusion that the number of known compounds is grossly underestimated. Probably the system must have capacity to handle 2.0 to 2.5 million compounds, of which 1.8 million are organic. Assuming that this estimate is correct, it will require only 10 reels to store the structures and molecular formulas of all organic compounds. Depending somewhat on the design of the record and the programming, and on the high-speed accessories available on the given 1401, the computer is potentially able to scan an entire reel in 3.5 to 4.0 minutes. At worst, it could be no longer than 10 minutes. This means that it is possible to scan all known organic compounds in 25 to 30 minutes, or at worst 100 minutes. Simple searches should take no longer than 30 minutes if the records are properly block-sorted. Moreover, preliminary indications are that three similar searches can probably be done at one time.

IMPLICATIONS OF THIS WORK

The aid of high-speed machines can now be available to the chemist in the laboratory. The research scientist can be freed from the repetitive tasks of assembling and organizing background material related to a project. This will make it easy to do a type of correlation that was previously possible only by hours and hours of chemist-time. For example, the limits and uses of a reaction can be explored by searching decks 1 and 5. The relation of structure and physical properties can be explored by searching decks 1 and 2. The effect of structure(s) on physiological action can be explored by searching decks 1 and 6. The relation of physical properties and physiological action can be derived from decks 2 and 6. All the indicative signs from such searches can point the way to new compounds and new uses of old compounds. Moreover, because of the unique properties of the cipher, and the mathematical relations inherent to it, there is now available a way to search for a given group or series of groups in complex structures and fused-ring systems. Finally, the entire information on any such search can be converted, by machine, to conventional typed pages. If necessary, it can be converted to a new tape or a new deck of cards.

BIBLIOGRAPHY

(1) G.M. Dyson, *J. Chem. Doc.*, 1, 24 (1961).
(2) G.M. Dyson and Elizabeth F. Riley, *Chem. and Eng. News*, 29, 72 (1961).

Experiments with the IBM-9900 and a Discussion of an Improved Comac as Suggested by these Experiments[1,2]

By MORTIMER TAUBE
Received August 24, 1961

This paper will present the system and machine considerations which led, in the first instance, to the development of the design of a Continuous Multiple Access Comparator; a description of the IBM 9900, which is IBM's reduction to practice of the COMAC principle; certain improvements in design of the COMAC to make it a more efficient low-priced device for the storage and retrieval of information; and, finally, the concept of using a wholly new method of logical comparison in order to achieve maximum efficiency from the use of E.D.P. equipment in the storage and retrieval of information.

In the literature on information storage and retrieval there is a generally recognized dichotomy between scanning systems and look-up systems, which have also

(1) This work was done for the Directorate of Mathematical Sciences, Air Force Office of Scientific Research, under Contract No. AF 49(638)-91.
(2) Presented before Division of Chemical Literature, American Chemical Society, St. Louis, Missouri, March, 1961.

been called conventional systems and inverted systems. In conventional (scanning) systems, which the information storage and retrieval art took over from standard data processing systems, a physical record—a card or a segment of a tape—is used to record an individual item and its characteristics. In such a system, when any item is characterized by many characteristics, it becomes impossible to order the file by indexing terms except by duplicating the total item card. If only a single item record is made, a search for items defined by any characteristic or set of characteristics involves a search of the total file. In principle, such a search is inefficient and this fact is generally recognized in the literature. Attempts have been made to overcome the inefficiency of a total linear scan by increasing the speed with which the recording medium is scanned and by packing information onto the recording medium more densely. These measures are usually not sufficient, at least in the present state of the art, to justify scanning for answers to particular questions. Usually this method is made economical by batching, that is to say, retrieval questions are saved until enough of them are accumulated to make it economically feasible to search the total store. However, this method does involve increasing the arithmetical and logical apparatus of the device and is only possible with quite sophisticated E.D.P. equipment. Batching of questions is not possible with ordinary punched card equipment. Hence, the use of item records and scanning with punched card devices is severely limited and must be restricted to small systems.

There are other difficulties with punched card scanning systems, *e.g.*, the limited coding area, which tends to restrict the number of characteristics or terms which can be used in indexing any item, in addition to the problems of superimposed coding, which is designed to obviate the necessity of fixed field coding and searching. The limitation of card area has a particular relevance to the problem of indexing chemicals. The attempt to use punched cards for searching information on compounds and similar chemical information developed at a time when it was supposed that scanning was the only possible method of using punched cards. If the punched cards are to be scanned and selected in terms of matched characteristics, then the names of chemical compounds must be coded so that they can fit on the punched cards and be compared, one against the other, in fixed fields. This requirement gave rise to the various attempts to develop chemical codes and ciphers, such as the Dyson code, the Wiswesser code, the CBCC code, *etc.* It is not generally realized that in an inverted or look-up system, the requirement for coding chemical compounds essentially disappears, since in look-up systems, comparisons are made on the basis of common item numbers and a human being selects the names of the particular decks to be compared. This particular relevance of inverted systems to chemical searching has, so far as I know, never been noted in the literature. Even though inverted systems are now widely used in both manual and machine forms, this has not led to a re-examination of the need, if any, for chemical codes and ciphers.

There is no need, in this paper, to describe the way in which inverted or look-up systems operate. Such systems have been exemplified in standard Uniterm indexes, in Batten or matrix systems, in the collating system used by CBCC, in the COMAC system used by the du Pont Engineering Department and Documentation Incorporated, in the 709 system used by the Aircraft Gas Turbine Division of General Electric, *etc.* All such systems involve an increased input cost to provide for posting, that is, prefiling of the item numbers under appropriate terms in an array of terms, but they permit rapid retrieval by comparison of selected portions of the file.

The COMAC was first conceived as a refinement of ordinary collation as used for the storage and retrieval of information by such organizations as the Chemical-Biological Coordination Center. In collating systems, searches are performed by collating items in one deck against items in another deck in order to determine items in common, that is, when the search is for a logical product. If the search is for a logical sum, the search is for the totality of items in both decks, less duplicates, and if it is for a logical complement, it is for the items in one deck which are not represented in the other. Although a collating system thus seems to be performing logical operations, actually, standard collating equipment performs arithmetical operations, as a result of which it makes certain decisions which enable the user of the system to regard the results of these operations as members of certain logical classes. The significance of this distinction between arithmetical and logical operations will be referred to again and its significance will become apparent when we discuss more advanced types of COMAC systems.

Although the COMAC system already has been described in the literature, a brief description of it here will serve to introduce some of the difficulties encountered in using a Mark I and the reasons for advancing to a faster Mark II and a Mark III tape model which uses logical, as opposed to arithmetical, comparison.

The collator was developed by business machine companies to provide for interfiling of records and for matching identical records. As we have said, it performs both interfiling and matching by making arithmetical comparison of high, low, and equals. The standard collator compares two decks at a time and it uses the arithmetical decision as instructions to advance one deck or the other. This process results in an intermittent advance of cards from either deck and results in a mechanical limitation of collating speed. Since a collator is designed for interfiling, the cards in a collating deck can represent only single items. For example, if one deck contained Items 2 and 4 and another deck contained Items 1 and 3 and we wished to interfile the two decks in numerical order, this would prove to be physically impossible if Items 2 and 4 were on one card and Items 1 and 3 on another. In information storage and retrieval problems, there is no requirement for interfiling the several decks of the system. On the contrary, interfiling or selecting out of the decks disarranges the decks and they must be restored to their original order before they can be used again. If the requirement for interfiling is eliminated, it becomes possible to put as many items on a card as the coding space allows. In the original proposal for the COMAC, we suggested using binary coding, which would have permitted up to forty 6-digit numbers on a card. The IBM 9900 uses Hollerith coding and thus has a maximum limit of twelve 6-digit numbers on a card plus a term number and certain housekeeping

columns. But even with this limitation a COMAC deck is less than one-tenth the size of the standard collating decks which have been used for information storage and retrieval problems. When multiple items are encoded on a single card, the indication of an item or items selected from the file cannot be performed by the physical selection of cards, but must be done by providing a print-out. This requirement is, of course, an advantage of the COMAC system, since the original decks are not disarranged and there is no return to base problem. In summary, the COMAC system provided these advances over the traditional scanning and collating systems: As compared with scanning, it provided for indexing by an unlimited amount of terms; it eliminated the requirement for coding of terms; and it obviated the requirement for searching the total deck. As compared with collating systems, it reduced the size of the decks by over ninety per cent; it eliminated the problem of refiling selected cards; and it provided a print-out of the answer as an input for a computer to retrieve data on the items indicated by address only on the COMAC output.

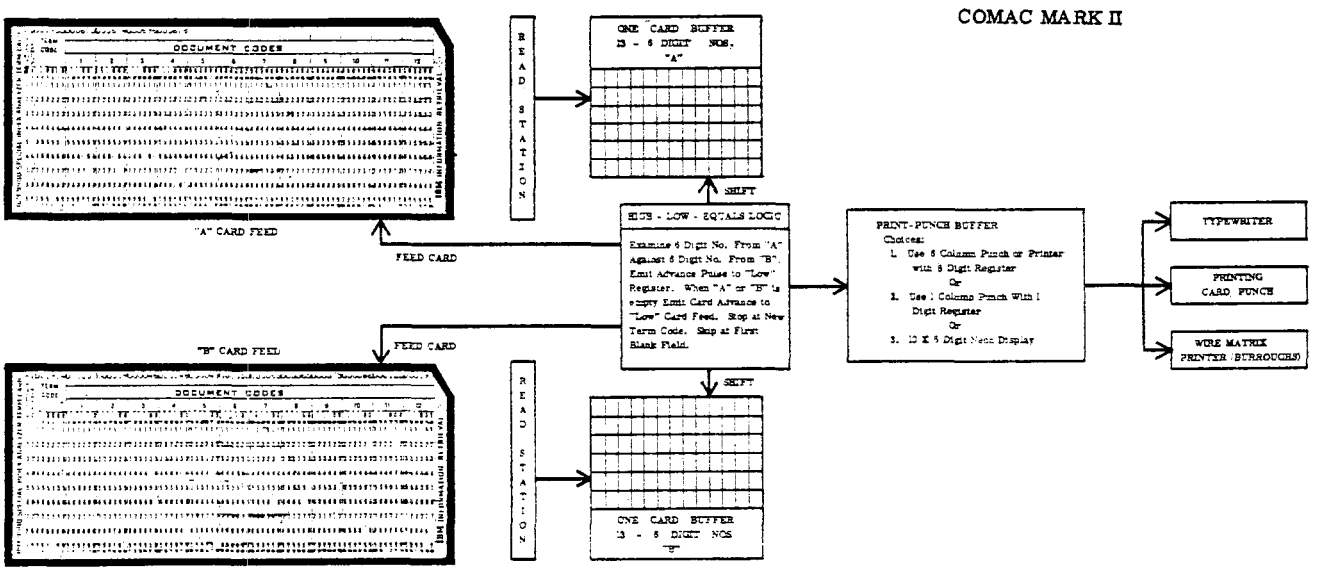
The major difficulties which remained in COMAC-type systems were the intermittent feed, which the COMAC shared with other collating systems, and the relative difficulty of making logical sums, which the COMAC shares with almost all other devices. In the Mark II COMAC, we plan to eliminate a percentage of the stop-start motion and in the Mark III, we propose to eliminate it almost entirely and also, by a change from arithmetical comparison to logical comparison, to provide an efficient method of making logical sums. The advantages of eliminating intermittent feed are obvious, but the problem of sums needs some discussion. If a COMAC device is used for making products and more than two terms are involved, the successive logical operations require less and less time because the number of the items in the product class is equal to or less than the items in the separate classes. If we make a product of A and B and then make a further product of A, B, and C, the second operation will take less time than the first operation unless the Class A and the Class B are identical. On the other hand, if we are searching for sums, the Class A or B is larger than the Class A or the Class B unless, again, the two classes are identical. Hence, if a question involves a search for sums, as well as products, the search time for each operation will tend to increase. The du Pont people, who have a great deal of operational experience with the COMAC, find that the problem of sums limits severely the number of questions they can answer per day. This is especially true with reference to the 9900, since this device uses a paper tape as an intermediate record between the card input and the card output, that is, the first class is written on a paper tape from the first deck of cards; the second class is then compared against the tape and a product tape or a sum tape is written, depending upon the type of question. The product tape gets shorter; the sum tape gets longer; and since the paper tape is relatively slow, if a sum of many terms is required, answers are delayed while the tape gets longer and longer.

In the Mark II COMAC which we are now designing, the paper tape is eliminated in favor of the two-card read stations which we had originally proposed, but we are also proposing two registers, each of which is large enough to accommodate the data on a total card. In these registers the numbers will be compared electronically and thus the intermittent motion is reduced from a movement per number to a movement per card. This should provide a considerable increase in speed, especially for product searches. A schematic drawing of the Mark II is presented in Fig. 1. We believe such a device can be produced at a very reasonable cost and by reasonable we mean comparable with present collators or the 101 and it will provide a high level of efficiency for small and medium size information systems.

When we advance from card to electronic or tape systems, the advantage of prefiling, which is so clear and obvious with card systems, becomes marginal, since the great possible speed of tape systems is lost by virtue of the requirement for intermittent stopping and starting of the tapes. This fact has led, in many cases, to decisions to use tape scanning systems and to compensate for the requirement of searching the total file by batching. Such a palliative, however, is only possible if the operating environment permits batching and there is no requirement on the information service for immediate answers to individual questions as they are received by the service. Hence, it seemed to us that neither scanning nor high, low, and equal comparison provided maximum efficiency of tape-type equipment for the storage and retrieval problem. This led us to re-examine the total theory of systems structure. The result of this re-examination has been the development of the notion of dedicated enviroment comparison, which will lead to a system and device which we will call COMAC Mark III.

The existence of systems which are prefiled, and thus resemble inverted systems structurally but use the technique of scanning, rather than collation, and thus resemble conventional systems dynamically, indicates that the dichotomy of inverted and conventional systems is not general enough to encompass all structural and dynamical possibilities of IR systems. An illustration will make this point clear. The Minicard system utilizes prefiling of duplicate copies of Minicards and to this extent resembles a collating system in which each deck of Minicards represents a term in the system and all items indexed by that term. However, the large coding area of the Minicard makes it possible to code each Minicard with all the terms used to index any individual item. Hence, in searching the Minicard system it is not necessary to collate one deck of cards against the other. Each deck, since each Minicard has all codes, can be scanned by a reading head for selection by any logical combination of codes on the individual Minicards. This method of scanning prefiled decks eliminates the bad dynamical features of stop-start motions and dual feed required in collation. On the other hand, it increases the size of the coding area necessary in the total system by a factor equivalent to the average number of descriptors or terms per item indexed. In other words, it has been thought that the only way to avoid the dual simultaneous feed of collation with its stop-start operation has been to increase the size of the code or store by an order of magnitude.

Suppose one asks, at this point, whether or not it is possible to develop a new method of comparing items for retrieval which does not involve either the dynamics of alternate stop-start movements which characterize standard collating systems or an increase in code storage



which characterizes multiple copy systems like the Minicard system or the necessity of examining the whole file which characterizes scanning systems. Peek-a-boo systems do seem to provide the required method, namely, they do not involve sequential comparison and stop-start motion, since the items on one card are compared in parallel with items on the other; they do not involve an increase in code storage, since each item is represented by a single hole in a matrix of holes; and, finally, being prefiled systems, they provide a retrieval mechanism which does not necessitate a search of the total apparatus. But matrix systems have an additional difficulty. Their advantages are achieved at the cost of providing, on each record, dedicated space for the totality of items in the collection. This is not a serious matter for special data files of limited vocabulary and high density posting, but for general IR problems, it has been our experience that approximately one per cent of the coding area of matrix systems will be used and the rest will be excess dedicated space.

Parenthetically, it may be observed at this time that in writing this paper, I am putting down as parallel considerations a number of ideas and problems which came to us over an extended period of time as we worked with IR problems and IR systems. When all the considerations which occurred through an extended period are put down at one time, the recommended solution also seems to emerge almost as a logical implication of the statement of the problem. It would only be honest to say that we understood these problems long before the solution which I am now about to present occurred to us.

What we are after is the following: We wish to provide a type of comparison as efficient as that provided by a Peek-a-boo system. On the other hand, we wish to avoid the dedication of space and the inefficient storage of Peek-a-boo systems. We wish to provide prefiled systems but we wish to avoid the bad dynamics of stop-start comparison. At the same time, we are not willing to return to a linear scan involving the search of a total file plus coding problems which are eliminated in prefiled systems. These desiderata, once fully understood, virtually provide the design of the COMAC Mark III.

In the COMAC Mark III the basic store will be in the form of an inverted file with item numbers under terms. Think of this basic store as a set of tapes or a segment of tapes. Let us suppose that we are required to find the logical product of the classes represented by two segments of tape. The usual method involving stop-start motion is to use arithmetical comparison, which finds the product class and moves the tapes by determining high, low, and equals. Suppose, instead of this method, we take a particular class and distribute it in another tape in dedicated positions in the second tape. The same procedure can be carried out and perhaps understood more easily if one thinks of the addresses on a disc memory in a RAMAC-type device. Using such a disc memory, the items on any segment of tape under any term can be distributed to addresses corresponding to the item numbers on the disc. This writing can be continuous and does not involve any stop-start motion. In effect, the distribution of the item numbers to addresses converts the basic dense store to a matrix-type store. The second class can also be distributed in the same set of addresses and, depending upon the requirements of the question, the sum, product, or complement can be immediately read out. The high, low, equal comparison has been replaced by occupancy of the same dedicated areas, as in a Peek-a-boo system. On the other hand, the matrices are not regular and permanent parts of the store but are set up as required in the searching process. We thus achieve the dense store of non-dedicated systems in the total system and the rapid and simple logical processing of matrix systems when the system is used to answer a question. Intermittent stop-start is eliminated, since the total class is written into the comparing medium before it is compared with any other class. Finally, although we have illustrated the comparison and interfiling of numbers only, using dedicated addresses, there is no limit in principle to the amount of information which could be stored at each dedicated address, *e.g.*, a name of a person, an item, a description of a person or item, *etc.* It is not possible to use dedicated positions in ordinary Batten systems for more than a number, since this would increase the enormous waste of storage space

in such systems; but, since dedicated space in the COMAC Mark III system is used only for the logical operations and not for permanent storage, any type of information now handled by collating and interfiling systems using the arithmetical operations of high, low, and equals can be handled in this system using dedicated environmental comparison, and logical, rather than arithmetical, instructions.

Although we have every reason to believe that the COMAC Mark III will be an efficient storage and retrieval system for many different types of IR problems and environments, one of the most important things about this development is not its practical utility, but the stimulus it gives to a renewed theoretical consideration of IR structures and systems. In essence, we have attempted to optimize an IR system based upon the selection of positive characteristics of existing systems and the elimination of their negative characteristics. This very attempt has opened up the real prospect of a new and more general theory of the structure and dynamics of IR systems in terms of which existing systems and devices can be understood as special cases. The important term above is "dynamics." Most of the existing theory to date has concerned itself with the structure of systems, namely, whether they are inverted, linear, multiple, *etc.* What we have now recognized is that the manner in which comparison takes place within a system and how the reading head is brought into juxtaposition with the store constitute the dynamics of the system, which is every bit as important as the structure. We can say now, for example, that the COMAC Mark I system has a good structure and bad dynamics, whereas a Peek-a-boo system has good dynamics and a bad structure. It is our belief that the COMAC Mark III has both a good structure and good dynamics, but this remains something to be substantiated by future study and experiment.

REFERENCES

Mortimer Taube, "An External Index to a Computer Store of Items and Transactions as Illustrated by Project ECHO," AFOSR TN 60-8, December, 1959.

R.W. Murphy, "The IBM 9900 Special Index Analyzer," International Business Machines Corporation, November 17, 1958.