Machine Retrieval of Information

MORTIMER TAUBE

THE ROYAL SOCIETY Scientific Information Conference in 1948 included a working party on mechanical indexing. This working party considered the various systems of mechanical selection which were brought to its attention: edge-notched cards; the Batten system; Hollerith and Power-Samas punched card systems; the Samian punched card system; the Rapid Selector; the Univac; Zatocoding; and the combination of punched cards and microphotography.[1]

On the basis of the deliberations of the working party and its recommendations, the Conference concluded "that in the field of subject indexing and selection, designers of apparatus are well ahead of users in the facilities they offer, or plan to offer in the near future. The present need, therefore, is for experiments on a realistic scale using available appliances. . . ."[2]

There are two points which should be especially noted with reference to the considerations and conclusions of the Royal Society Conference on this matter of mechanical selection. In the first place, for almost ten years the list of available systems and appliances has remained constant. Anyone who today undertakes a survey to determine what is available in the field of mechanical selection will find almost the same possibilities considered by the Royal Society Conference in 1948. This seems a curious phenomenon in an age which prides itself on its rapid technological advances. But apparently this stagnation is not something which characterizes the machine designers. The second noteworthy point in the conclusion of the Conference is that the designers of appliances had advanced way beyond the willingness of the librarians, information officers, and documentalists who use manual systems to experiment with and utilize the appliances that are available. The Conference did not conclude that better machines were necessary but that those charged with maintenance of manual systems were lagging in their readiness to utilize

Mr. Taube is President of Documentation Incorporated.

what was already available to them. To be sure, the Conference recognized that such utilization would be experimental but it thought that such experimentation on a wide scale was justified and necessary because of the prima facie superiority of available mechanical systems to existing manual systems.

Although many who fancy themselves as pioneers and innovators have been impatient of those who seem quite content with the status quo, the author does not recall in the whole literature of documentation a stronger condemnation of a profession than that implied in this conclusion. With machines available, manual methods are still used; and ten years after the Royal Society Conference's condemnation, the same machines are still available and no one or no institution has yet carried out the required experiments on a realistic scale.

If this writer may for a moment take on the unfamiliar role of a defender of the status quo, he would like to point out that the conclusion of the Royal Society Conference was very foggy and not worthy of its own high standards and traditions. It is not enough to call for an experiment; one must also so design an experiment that it will prove what it is supposed to prove. If there is overwhelming theoretical evidence against certain conclusions, experiments are not necessary to reject them. If someone insists on an experiment in the face of such theoretical evidence, he must indicate why he thinks the theoretical considerations are not substantial or conclusive, and how they might be modified by the results of the experiment. Now, R. R. Shaw has calculated that in order to handle the daily reference load of the Library of Congress, 8,333 Univacs would be needed at an investment of close to a billion dollars.[3] What type of experiment is required to prove Shaw wrong? There have been dozens of "experiments" with edge-notched coding systems and with punched-card systems reported in the periodical literature and in the Casey-Perry volume on punched cards.[4] But although the author has searched the literature assiduously and has asked assistance of several expert bibliographers in the field of documentation, he has not succeeded in finding one reference to a controlled experiment which demonstrates the superiority of commercially available mechanical equipment as compared to traditional manual systems. One doesn't need an experiment to determine that cards can be selected from an edge-notched deck or a file of punched cards; one doesn't need an experiment to determine the speed with which a Univac will search a magnetic tape, or a Rapid Selector will search a roll of film.

If Zatocoding, the Batten system, and the Rapid Selector are ex-

cluded from the devices and appliances considered by the Royal Society, the other types of equipment obviously were not developed for the purposes of literature searching or mechanical indexing, but for accounting and computing. Their existence was not an indication that the designers of equipment for literature searching had advanced faster than the willingness of users to test such equipment. These machines having been produced for accounting problems, it remained for the equipment people to demonstrate that their machines could be used for purposes quite alien to those for which they were designed. The burden of proof here is on the designer of equipment not on the information or documentation center.

The other chapters in this book indicate that librarians have not been backward in seeking mechanical solutions for their various problems. Therefore, their failure to use machines for information retrieval is not attributable to their unwillingness to experiment and try new techniques but to the prima facie evidence that the available machines are not satisfactory and cannot perform information retrieval functions as adequately as a card file or a printed catalog.

Librarians have always been aware of the cost of filing cards in an alphabetical or classified array, and of the time required to read through a group of references to select the reference or group of references which will answer any particular problem. The simple fact that edge-notched cards and interior-punched cards can be selected from a random file and at speeds faster than a librarian can read and select cards has confused a great many people.

Actually, an ordered file permits a librarian to make his selection from a small segment of the file; whereas, with random files the total file must be scanned for each selection. Effective sequential information searching devices will be available when it is possible to find any name by a random search of a telephone book as fast as it is possible to find a name by manual search. Further, since there are multiple phone books so that many people can look up names at the same time, any busy library would have to have many machines, each one of which could search the whole catalog as quickly as a reference assistant could find a particular card under a particular heading in an ordered array. This requirement underlies Shaw's conclusion that the Library of Congress would require 8,333 Univacs to match its present service. Mechanized information searching is held back not by the unwillingness to experiment but by the absence of devices which are even remotely suitable to the reference needs of a large and busy library.

It is necessary to emphasize that the inefficiency of sequential searching is a matter of principle because there are those who hope to overcome this inefficiency by spending more and more dollars for data processing equipment with faster and faster rates of search. That is to say, the obvious inadequacy of punched-card equipment for large-scale storage and retrieval systems has turned some minds in the direction of the general purpose digital computer as the solution.

The efficiency of any machine system or device is always measurable in terms of cost. Our railroads changed from steam to diesel because by so doing they raised the efficiency, i.e., lowered the cost of transportation. Similarly, when hand tools are changed to power tools the purpose is always to do more work at less cost. To be sure, the first model of a power tool to replace hand labor may involve a capital expenditure which is not immediately recoverable as lower cost. But even first models must be able to operate more efficiently and at lower cost than the hand labor they are built to replace. It is inconceivable that anyone would ever use a machine which was less efficient than the hand operation for which it was substituted. In fact, this very notion has been responsible for a classic of American cartooning—the Rube Goldberg machine.

All of this is indeed so obvious as to be trite. Yet neglect of just these obvious facts, coupled with unguarded enthusiasm concerning the potentialities of digital computers, has led to a failure to consider the theoretical and practical efficiency of computers as information storage and retrieval devices. This does not mean that those who imaginatively extend the utility of computers from the mathematical problems for which they were devised to the fields of mechanical storage and retrieval systems, do not recognize the need for bigger and better memories with rapid access time and low input and output costs; but there is an absence of a more general concern with the basic question of the suitability of computer design and techniques for the purposes to be served by storage and retrieval systems.

The mechanization of any art of activity, whether it be computation, intelligence analysis, or the storage and retrieval of information, involves the union of two lines of research: the determination of the essential logical pattern of the activity, and the discovery or invention of a mechanism which will be the physical analog of the logical pattern. In the field of computers, there is a special case—the logical pattern of computation, i.e., mathematics, is a highly developed science. Computers were built and used efficiently as soon as the necessary gears, holes, dots, electronic elements, etc., could be

assembled. But the fact that computers found a logical structure ready and waiting has led to serious confusion in other fields. The erroneous assumption was made that computers could be universally applied even to fields whose logic was undeveloped or which did not involve numerical data or the processing of digits. Thus, futile attempts have been made over and over again to force non-numerical and non-digital problems of identification, search, and information recovery, with special logics into the narrow possibilities of digital computers.

Studies of the logic of information storage and retrieval have demonstrated that it is not data processing or rapid addition, subtraction, and multiplication that is required for storage and retrieval systems, but random access, instantaneous recognition, and direct display of any item permanently stored in a static memory. The crucial fact is that not binary digits are input to be computed but terms and logical relations which are appropriate to the storage and retrieval problem.

Suppose in an information center or library that the main concern is the rapid recovery of any item in the library or any fact recorded in any item. Suppose further that costs were not a consideration. The most effective storage and retrieval device for a collection of 1,000,000 reports, documents or other items might then have the following components:

1. One million people each holding and studying a single item.
2. A system of communication connecting the million "storage" points with a central reference bureau or input microphone.
3. A system of facsimile transmission connecting the "storage" points with a central output.

A question sent over the central microphone would start all "storage mechanisms" studying their items and the one holding the desired document could send a copy by facsimile to the central output station located adjacent to the input microphone.

This fanciful supposition has a serious purpose. It depicts exactly what has been done when a million abstract cards are substituted in a catalog for the live storage elements and the searcher then is asked to walk through the catalog system to find the card which suits his purposes; or when cards are brought to the searcher by means of sorting machines, tapes or drums. Certainly the shift from live storage elements to cards or magnetic dots is only significant as a saving in input costs without a proportionate increase in output costs.

Steam shovels and automobiles may have many parts in common—motors, gears, batteries, wheels, etc. But no one would suppose that the way to get better automobiles is to concentrate on designing and building bigger and better steam shovels or vice versa. And the fact that the development of more efficient storage batteries for steam shovels may be, at the same time, the development of more efficient batteries for automobiles is still no argument for the identity of steam shovels and automobiles. Even though steam shovels may be self-propelled and could be used, if efficiency were no object, to carry people from place to place, they are not primarily vehicles for transportation. If all this seems as fanciful as the storage device with one million live storage elements, it is, nevertheless, as reasonable to use a steam shovel for locomotion rather than digging as it is to use a computer as a storage and retrieval device rather than for computation. It can be done but someone driving from Washington to New York in a steam shovel would certainly receive many stares, and similarly the ingenuity of using a good computer as a poor storage and retrieval device should be met with little enthusiasm.

The one factor most responsible for the confusion between the entirely distinct functions of computers and storage and retrieval systems is the common interest in devising ever larger storage devices which will store more information at a lower bit cost.

A computer system processes data which may be fed into the system directly by an operator or indirectly from a memory in order to arrive at certain mathematical quantities. As is manifest from its name, a computer performs an arithmetical computation or a series of them in order to arrive at an arithmetical value. A storage and retrieval system does not perform any arithmetical operations or even logical operations. It searches a memory or storage device to select or identify data in accordance with specific questions put to the system. The form of the question may involve a logical operation, e.g., one can ask for the logical product, PQ. That is, a storage and retrieval machine is not asked to calculate, compute, or derive logical products, but only to find anything stored in the system which can be identified in terms of a logical product. In making identifications for selection, a storage and retrieval machine may use the type of components used by a computer to make computations, that is, reading heads, switches, relays, diodes, holes, magnetic dots, etc. But this similarity of components again should not obscure the essential difference between the two types of systems. Advances in the computer art may certainly

be significant for information systems and vice versa. Thus, the development of the tapedrum "memory" with its large storage capacity is significant for both computers and information systems alike. But the tapedrum is not a computer. In describing the uses of the tapedrum, its developers state: "The tapedrum can be used as an auxiliary memory or storage device for large scale computers."

In short, the similarity of components and subassemblies, the similarity of problems of storage costs, random access times, etc., should not obscure the basically different purposes and functions of computer systems and storage and retrieval systems. The consideration which determines that this is an essential difference with superficial resemblances, rather than essential similarity and superficial differences, is that a highly efficient computer may be a very poor storage and retrieval machine, and a highly efficient storage and retrieval machine may be completely useless as a computer.

Much of the research in the past four years has indicated the possibility that the efficiency of computer systems and storage and retrieval systems varies inversely and now there is reasonable expectation that this relation can some day be quantified and expressed in an equation. It is known, for example, that the efficiency of a collation operation is determined by the ratio of the balance file and the transaction file; that is, if two decks of cards to be collated are of equal size and one deck is evenly distributed throughout the other, then machine collation reaches maximum efficiency. But if one deck contains 1,000,000 cards and the other contains 100, collation becomes highly inefficient. This situation can be generalized into the homely expression, "The less *looking* (searching) and the more operating, the more efficient is a collator or a computer," or, "The smaller the required memory and the larger the number of operations, the more efficient is a computer operation." On the other hand, in a given increment of time a storage and retrieval system should perform the maximum amount of searching and the minimum amount of operating.

There is nothing in what has been said above which in any way denigrates the possibility of efficient storage and retrieval devices. In fact, the building of such machines is certainly a practical possibility. But this practical possibility will never be realized if we continue to emulate the March Hare. Butter isn't any good for a watch even if it's the *best* butter, and digital computers are not storage and retrieval devices even if they can compute in milliseconds.

References

1. For a description of the mechanics of edge-notched cards and interior-punched card systems, see Casey, R. S., and Perry, J. W., eds.: *Punched Cards, Their Application to Science and Industry,* New York, Reinhold Publishing Corporation, 1951, pp. 5-75. The only difference between such devices and the advanced machines using magnetic tapes, drums, microfilm, etc., is greater compression of data and faster scanning speed. For a description of the mechanics and operations of the Batten System, see Wildhack, W. A., *et al.:* Documentation and Instrumentation. *American Documentation,* 5:223-237, Oct. 1954. For a description of the general problem of machine collating and searching, see Taube, Mortimer: *Studies in Coordinate Indexing,* Washington, Documentation Incorporated, 1954. Vol. 2, pp. 11-18.

2. Royal Society of London. Scientific Information Conference, 1948. *Report and Papers Submitted.* London, The Society, 1948, p. 204.

3. Shaw, R. R.: Management, Machines, and the Bibliographic Problems of the Twentieth Century. *In:* Shera, J. H. and Egan, Margaret E., eds.: *Bibliographic Organization.* (University of Chicago Studies in Library Science) Chicago, University of Chicago Press, 1951, pp. 200-225.

4. Casey, R. S., and Perry, J. W., eds.: *Punched Cards, Their Applications to Science and Industry.* New York, Reinhold Publishing Corporation, 1951.