Information Retrieval

Through support and research we should develop a science upon which to build a better technology.

Ralph R. Shaw

A few weeks ago one of our local physicians received a frantic telephone call. A man who had been working in his garage had accidentally drunk weed killer instead of the beer he thought he had picked up. While a member of the family drove the patient to the hospital, Dr. P. had the wife identify the container from which the poison had come. He then telephoned the Poison Control Center in New York City, giving them the trade name, and by the time the patient reached the hospital the antidote was ready.

Less spectacularly, when you need a new lamp for your stoplight or some antifreeze for your automobile, the filling station attendant turns to a manual that tells him what kind of lamp or how much antifreeze is required for the make and model of your car. Your broker has a number of manuals to help him supply information about the history and current operations of the corporation you inquire about, and your banker turns to Dun and Bradstreet for financial information on any of tens of thousands of individuals and firms. Your library has handbooks and other tools that provide the list price of any periodical or give the phase of the moon on 7 October 1902. Telephone directories around the country provide the phone number of almost any individual you select or of organizations to perform almost any service you require, and the telephone company has additional reference tools which translate street addresses into names and telephone numbers. Thus, in tens of thousands or possibly even millions of ways, each day, for purposes varying from idle curiosity to the saving of lives, much information is promptly made available to users ranging from semi-skilled laborers to the most advanced creative research workers.

There are at least two major areas in which we do reasonably well in providing published information. The first of these includes all the fields in which the frequency of demand for information has justified the production of printed reference tools and their distribution on a mass basis—tools such as road maps, engineering handbooks, technical data sheets, indexes, abstracting journals, and thousands of similar publications. We use such tools constantly as retrieval mechanisms, but we do not think of all of them as "reference books."

The other area in which we have done comparatively well encompasses fields in which society has recognized the urgency of the need for prompt information service and has made provision to satisfy it, regardless of the relative efficiency of the tools used to satisfy the need. The Poison Control Center is an example.

In addition to tools for retrieving specified factual data, numerous devices (including periodicals such as *Science*) have been developed to provide "current awareness" service. The objective is to let people know what is going on in their areas of specialization and in associated areas where the pertinence of the information to their own field is less demonstrable.

Information is transmitted formally through such devices as conferences, and informally through the contacts that occur every time one individual meets another, writes a letter, telephones, or communicates in any other way. This type of communication represents a combination of "current awareness" and "retrospective search," and to think of it in terms of a 1:1 ratio of communication is to underestimate its value. Each of us carries a store of information gathered from many other people, and whatever we communicate is a synthesis of the writings and statements of others, modified through our own personalities and understanding. This has been recognized and exploited to a certain extent in the structure of our information services. Any special librarian worth his salt has a file of "walking encyclopedias" to whom he turns when the printed indexes or reference books are inadequate. Government and industry alike are setting up positions for people whose primary job is to keep in touch with what is going on, or has gone on, in a given area of specialization and to transmit pertinent information to the group they serve. Publications such as *Who Knows What* further attest the value of these "walking and talking and writing encyclopedias."

In view of the fact that vast amounts of information are now available and readily retrievable through various devices, the outcries we have been hearing about the alleged crisis in information retrieval seem to be somewhat extreme. No one would contend that our information services do not need improvement, or that improvement is not the normal order of life in any viable organization or system. But we have gone astray in this field by equating improvement with change, and emphasizing the means rather than the end.

Requirement and Search

The only purpose of information service, whether of the retrospective-search or the current-awareness type, is to satisfy the user's need for information under the conditions under which he is working. The cycle in each case starts with an "information requirement." Often the user does not know exactly what information to look for, and he must begin by browsing. When he does know what to look for, he may not know that it is possible to get what he needs from sources available to him. Sometimes he comes to the information service with a specific request, knowing that he can ask for information and get it. The general system for providing the information is the same in these several situations, regardless of the mechanisms or tools that are used. Its object is to provide the information the user requires when he needs it, where he needs it, and in a form in which he can use it.

It is manifestly difficult, if not impossible, to index and retrieve informa-

The author is professor in the Graduate School of Library Service, Rutgers University, New Brunswick, N.J.

tion from a document that nobody has ever seen. Hence, the first indispensable step in bibliographical control is the assembling of source material. This material must then be described in some way for identification. Sometimes this involves no more than stamping a serial number on the document or on each page or on each paragraph or each book or each film or tape. Sometimes it involves a detailed analysis, to determine the author's best edition or to describe a particular copy. The item must then be analyzed and classified according to subject or other pertinent content or attribute; whether this is done by Uniterm or Thesauri, by the Universal Decimal Classification, by alphabetical subject headings, or by some other system is a matter of detail. Not until these three steps have been taken can a search for information begin.

The path then followed depends upon which tools or types of tools seem to be the most likely to provide the information required by the user. This is a matter of search strategy—a little-explored field that is currently entirely dependent upon the thinking of the individual who interrogates the system, whether he makes his search by turning pages or by pushing buttons. Whether we need detailed indexing, in order to go directly to the information required, or more general indexing to lead us to a number of volumes likely to contain the needed information depends on the amount and type of material involved and the uses to which it is to be put.

Regardless of the system (or systems) selected to achieve our purpose, the measure of its efficiency is the effectiveness with which it provides the information required and its ratio of input to output in providing the information. Unfortunately, this approach discloses, in the clearest possible fashion, the present lack of scientific basis for such selection. To determine the most efficient method through a management-engineering analysis based on times and costs and frequency of use of the alternative methods should be relatively simple, but we have little hard information about the relative efficiencies of these tools, or about the various input and output approaches and devices, upon which to base such an analysis. The field, therefore, remains one in which claim is matched by counterclaim, neither being supported by objective evidence.

Efficiencies of Some Existing Systems

In the past few years we have seen the first accumulation of objective data obtained through application of the scientific method to the research tools and systems of information retrieval.

Let us take, for example, the field of classification or indexing, or coding. The Cranfield study (*1*) indicates that there is no significant difference in the efficiency of retrieval of the four indexing schemes studied—Alphabetical Indexing, Universal Decimal Classification, Uniterm, and Facet Analysis—and that these gave on the order of 60 to 80 percent retrieval, with about 25 percent relevance under test conditions. This conclusion is borne out by Ann Painter's findings (*2*). She had the indexers in three government agencies using so-called modern systems re-index publications that they had previously indexed, and had the staff of a fourth, using conventional indexing, do the same. The efficiency of retrieval was found to be on the order of 60 percent for all four systems. Similarly, Judith MacMillan and Isaac D. Welt (*3*) showed, in a study of duplicate indexing of articles, that only 20 percent of the articles were indexed with the same subject headings and the same number of headings in the two tests, even though the indexers were specialists in the subject field. In view of the mounting evidence that, even under optimum conditions, neither the coding schemes used in the past nor those currently proposed have a high degree of reliability, one is inclined to wonder about the usefulness of the deluge of literature in this field—a literature which rarely provides objective evidence of the usefulness of the schemes discussed. As for those who think that machines, as such, will alter the picture, I refer again to the Cranfield study. In the study, the results of a machine retrieval test, made with a General Electric 225 computer by a group from Western Reserve University (WRU), were compared with the results of retrieval with a manual index. According to the report, "It was frankly a surprise to find that the WRU system was not particularly effective either from the point of view of recall or relevance . . . In fact, neither with recall nor relevance did it equal the performance of the Cranfield index . . ."

It appears reasonable to ask, now that we know the level of consistency that can currently be expected, whether there is any point in going to more and more esoteric (exotic if you will) coding schemes so long as we do not consistently operate above the level of 60 or 70 percent efficiency. Here, as in physics or mathematics, it appears that there is little to be gained through carrying calculations to more decimal places than are provided by the original data.

Unsupported Claims

One of the characteristics of the literature in this field is the presentation of claims, for whatever system is being advocated, on the basis of facts that are not given and by comparison with the worst possible alternative. By definition, any method of doing a job should be more efficient than the worst possible alternative. But literature in this vein continues to pour forth. Three examples of this approach, selected not because they are the worst that can be found but because they are recent and fairly typical, are given here.

Carolyn Kruse (*4*), in an article on the use of electronic computers for information retrieval, agrees that the earlier use of the 701 computer was inefficient and goes on to report on the current use of the IBM 7090. Maintaining that the present method is more efficient than manual processes, she states, "I have no comparative figures on the cost of a manual search and the subsequent typing of a list from the catalog cards. It would, perhaps, appear offhand that a clerk could do this as cheaply as a computer. However, at NOTS (Naval Ordnance Test Station) there is an extreme labor shortage so that comparable costs would still favor the machine search." Since there are no data for comparable manual costs, it is a little difficult to see any basis for the assertion that the machine search is cheaper. Also, no consideration is given to alternative methods of offsetting the shortage of typists. Since there are cameras available that can copy 3- by 5-inch or 4- by 6-inch cards at less than 1 cent per card, it is a little difficult to see how one can argue that a shortage of typists necessitates use of an IBM 7090 as a substitute. This is simply justification of an uneconomical method by comparing it with the most uneconomical method rather than with the best available method. But perhaps the most revealing sentence in the whole article is

Carolyn Kruse's statement, "Meanwhile, there was considerable pressure on the library to adopt some method of mechanized retrieval."

Ellis King (*5*), in a comparison of electronic facsimile transmission for interlibrary loans, limits the use of facsimile transmission to periodical articles and compares the cost with the cost of lending whole bound volumes. He cites the disadvantages of lending whole bound volumes as justification for further work on facsimile transmission but fails to mention the alternative of providing microfilm or photoprint copies of articles. It is a much more common practice to provide such copies than to lend whole bound volumes when an article is wanted; microfilm or photoprint copies provide many of the "intangible values" that King claims are offered by electronic facsimile transmission of articles. When his arithmetic does not favor electronic facsimile transmission, even by comparison with the lending of bound volumes, King says, "Yet, as was observed earlier, the system offers certain intangible values."

In 1962 the *Times Literary Supplement* ran a series titled "Freeing the Mind," which it has reprinted as a separate publication (*6*). In summing up, with the argument that "eventually the automatic preparation of bibliographies seems to have more useful possibilities" (p. 66 of the reprint), the author of these papers says, "Where the scholar wastes time is not in handling a comprehensive index but in looking for one possibly quite simple item (a man's name, say) in a large number of indices which have to be located and brought to him, whether or not they contain it." Since the most rudimentary reference service in the smallest public library normally provides information service of this kind, the argument that we need computers to do this work simply ignores alternative methods of obtaining the information. Also, the process of addressing and printout does not use the logical capacities of the computer, thus use of the computer for this work is inefficient. Those who advocate it show naiveté about computers and about conventional library services as well.

The summary goes on to say, "It is filling out slips and accumulating a pile of books that takes both the scholar's and the library's time." This again is an instance of ignorance or ignoring of the facts. For the scholar to do the clerical work and the messenger work himself is the worst possible method. Any book can be brought from a collection of a million volumes in a median time of less than 5 minutes by conventional messenger service. The volumes would have to be brought by messenger in any case, after the computer had indicated which books were needed. The entries can be copied by clerks or cameras. If the scholar must waste time doing these things, that is because the administration of the library service is poor or because support is inadequate; such short-comings must be corrected before any system can be expected to function effectively.

Returning to Carolyn Kruse's article, we must agree that there certainly is continuous pressure on libraries and information centers to adopt one or another mechanized system. Granting agencies are under this same pressure, and they are generally (and probably correctly) considered to favor requests that are based on machine application. This is a result of the emphasis placed on mechanization by scientists and administrators, and by congressional committees which, without having any facts at their disposal, have been deluged with claims to the effect that mechanization is the answer to all problems in the field of information retrieval.

While making a study for a great industrial organization, I found a pitiful example of the inefficiency that can result from pressure of this type. The organization had a documentation group using an IBM 601. Study of the work of this group showed that in some 6 years they had, at a cost of about $100,000 per year, built up a total file of approximately 16,000 indexed articles. This was a very small and unevenly distributed fragment of the literature required or of the large amount of literature available in their library (operation of the library cost less per year than operation of the documentation center). During the period studied —the year of highest use—only 60-odd questions were put to the machine. For a number of questions no answer was found. Interviews with the scientists for whom questions were answered revealed that in no case had information been provided that could not have been obtained in other ways, and that in every case the machine search had to be supplemented by conventional search. The head of the major unit under which this operation fell, agreeing that it could not be justified since other means of accomplishing all that it could accomplish, and more, were available, pleaded that it should be continued anyway, since its discontinuance would go against the general trend in the industry and the organization would be considered backward if it did not have a mechanized information retrieval system.

As one more example, let us turn to the hearings on the Science Technology Act of 1958 (*7*). On page 183 of that report we find the cheering statement, "Although documentation research institutes such as the Center for Documentation and Communication Research are currently able to service adequately the research requirements, it is foreseeable that . . . coordination of such activities on a national scale would be advantageous."

Some 4 years later (November 1962), at the Conference on Bibliography at Pennsylvania State University, Jessica Melton, assistant director of the Center for Documentation and Communication Research, announced that the American Society for Metals literature service had passed the experimental stage and that searches for a single subject would be made for ASM by the Center for $500 per search, the items to be searched numbering something less than 100,000.

Discussing this same service, which has been the subject of a deluge of articles over the last 5 years, Marjorie R. Hyslop (*8*), speaking for the American Society for Metals, says, "At the same time, experience to date has convinced us more fully than ever before that by the intelligent use of machine capabilities, the American Society for Metals is able to offer an information searching service that is infinitely better, faster, more up to date, more complete, more thorough, and above all cheaper, than any similar service ever offered that is based on traditional library methods."

In view of the fact that the file that is searched consists of something under 100,000 items—a little less than 1 year's listings in the *Bibliography of Agriculture* and substantially less than 1 year's listings in either the *Index Medicus* or *Chemical Abstracts*—and in view of the fact that any librarian with experience in the field should be able to make a single-subject search in 1 year's listings of any of the foregoing bibliographies in less than 15 minutes, the claim that machine search is "above all cheaper" could make sense only if we were paying librarians and other bibliographers about $2000 an hour.

Machine and Man

Lest it be claimed that the machine provides something that man cannot provide in this particular instance, let me quote another passage from the Hyslop article (p. 51): "All abstracts, before being mailed, are carefully scrutinized by the Documentation Center staff. . . . This requires, of course, a lot of time and attention on the part of the metallurgically trained people who examine the abstracts retrieved in a search."

Sooner or later we must recognize the elementary fact that answering a question by running punched cards or by running a computer is not a miracle. It is exceedingly doubtful that there is anything that can be done manually that cannot be done by machine if it can be described in sufficient detail. Similarly, it is exceedingly doubtful that there is anything in the field of information retrieval that can be done on a computer or with punched cards that cannot be done by hand. The method is not the end. The method is simply a management-engineering choice, based on the available methods, the times, and the costs, including times and costs for all the factors involved for the complete cycle of operations. This is a matter of methodical management-engineering flow-process charting, which must include all of the steps in the complete cycle of the process. Given that (something we practically never get in the literature of information retrieval), it would be easy to decide which tools and procedures are the most efficient for providing the information required at the level of quality and speed required.

This is not a call for complacency, nor is it an argument against the use of new tools, methods, or approaches when they can be shown to offer improvement in service. We do not always provide the best information that is available. We do not perform information services as well or as promptly or as efficiently as we should or could. It is important that each of the necessary steps be taken better and faster and more cheaply, with whatever resources are at our disposal, in order that more and better information service may be provided. Those responsible for information services must, therefore, continue to study all alternative tools and systems.

This is a far cry, however, from change for the sake of change. It has yet to be demonstrated that those who have been crying havoc and calling for vast expenditures on new technologies have anything to offer that will currently increase the effectiveness of our information-retrieval services. On three occasions over the last 2 years—one of them the Gordon Research Conference on Information Retrieval in 1962—I have asked groups of experts on computers and punched cards, information systems, information retrieval, and the related arts, to name a single application of the computer to information retrieval for which it can be shown that the computer is currently accomplishing anything of significance that cannot be done faster and more cheaply by hand. No one at these meetings or since has named one. This is not an academic question. We are faced with the opportunity to present such a demonstration to a tremendous audience, and if any of the readers of this article can cite such an application I shall greatly appreciate it if he will do so.

If computers can retrieve information more efficiently than it can be retrieved by other methods, we ought to be using them more widely. If they cannot, then in the interest of service to scholarship—whether in science or in other fields—we should stop making irresponsible claims for these systems, regardless of whether the claims are made by librarians, information officers, documentalists, government officials, administrators, engineers, scientists, or others.

Three Programs

In the current state of the art of information retrieval it appears reasonable to approach the improvement of information services through simultaneous operation of three major programs.

As indicated at the beginning of this article, we are currently providing large amounts of information by conventional means. Our support of systems that have proved capable of providing these services has not kept pace with the demands made upon the systems. So, first of all, very great improvement in information services could be effected immediately, were libraries and other existing information services (including primary publication, abstracting, reviewing, and information centers) supported to a degree consistent with the greatly increased volume of research, and then held accountable for providing services proportional to this increased support by means of the most efficient techniques available. Until we support the basic library and dissemination functions more adequately, no other devices can possibly be effective, since they are based upon and dependent upon these.

Second, it is doubtful that we can radically improve the technology in this field so long as it must rest as heavily as it does on purely empirical foundations. This means that we should be putting massive effort into the development of a science upon which a better information technology may be built.

Finally, technological proposals and essays in this field should be subjected to rigorous objective investigation to determine whether they do provide something demonstrably useful which cannot be provided by other known means; or, lacking that, whether they do offer improvements over known methods for achieving the same objective.

References

1. C. W. Cleverdon, *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems* (Her Majesty's Stationery Office, London, England, 1962).
2. A. F. Painter, thesis, Rutgers University (1963).
3. J. T. MacMillan and I. D. Welt, *Am. Doc.* **12**, 27 (1961).
4. C. J. Kruse, *Spec. Libraries* **54**, 90 (1963).
5. E. F. King, *Am. Doc.* **11**, 32 (1960).
6. "Freeing the Mind," London *Times Lit. Suppl.* (1962).
7. U.S. Senate, Committee on Government Operations Science and Technology Act of 1958, 85th Congress, 2nd Session, *Document No. 90* (Government Printing Office, Washington, D.C., 1958); (memorandum, Center for Documentation and Communication Research, Western Reserve University, pp. 178–188).
8. M. R. Hyslop, *Am. Doc.* **12**, 49 (1961).