

# Information Retrieval

By THOMAS WILDS\*

*Union Carbide Corporation*

AFTER World War II a few American scientists began to ask themselves if recent advances in electronics and automation could be applied to information—to the storage, handling, and retrieval of scientific knowledge. They were vitally interested in better ways to handle information because conventional systems could not keep up with the swelling flood of scientific literature and much scientific information was being lost. Hoping to invent machines to replace conventional systems, the scientists applied mathematics and logic to file systems and library science, and the resulting intellectual ferment became known as information retrieval.

Although originally intended for scientific information, information retrieval can be applied to any area of human activity. For information retrieval is nothing more than a scientific approach to recordkeeping and librarianship, and it will find uses wherever records and books are found. Its application to records may be widespread and revolutionary, and consequently information retrieval is of major interest to archivists and records managers.

## THE NEED FOR INFORMATION RETRIEVAL

The clarion call for new ways to retrieve information was sounded by Vannevar Bush in 1945, in his famous and truly seminal article "As We May Think."<sup>1</sup> During World War II, as Director of the Office of Scientific Research and Development, Dr. Bush was responsible for coordinating the war work of 6,000 leading American scientists. He had seen a growing mountain of technical literature overwhelm traditional methods of handling it, so that many scientists were left unaware of work done by other scientists on problems similar to their own. In his article he proposed the invention of a special machine to accommodate scientific records and books on microfilm. The machine would provide a scientist with all pertinent

\* The author is an analyst with the Records Administration Department, Management Services, Union Carbide Corp. He was formerly on the staff of the Maryland Hall of Records and holds degrees from the University of Michigan in oriental languages and history.

<sup>1</sup> *Atlantic Monthly*, 176:101 (July 1945); reprinted in Vannevar Bush, *Endless Horizons*, p. 16-38 (Washington, Public Affairs Press, 1946).

©Thomas Wilds 1961

information on a given problem at the flick of a button. He pointed out that building the machine was relatively easy and that the real problem lay in designing a workable indexing system.

Dr. Bush's point about the indexing problem turned out to be crucial. The new electronic computers were wonderful for handling masses of numerical data, but the most important scientific information is in graphic form. Much of it is prose with shifting semantic significances hard to pin down to the rigid logical requirements of machine indexing. But we are getting ahead of the story. First let us examine how traditional approaches to information handling led to the dilemma Dr. Bush encountered. These traditional approaches may be described in terms of three basic ways to handle information—by consecutive arrangement, indexing, and classified arrangement.<sup>2</sup>

*Consecutive arrangement* is the arrangement of information in a linear order such as alphabetical, numerical, or chronological. Examples are chronological correspondence files and the internal arrangement of telephone books. Consecutive arrangements do not endow their subdivisions with categorical or subject significance. They maintain their integrity at any volume, but are virtually useless as keys to information content other than by name, number, or date. Consequently they have little value for information retrieval unless they are combined with indexing or classified arrangements.

*Indexing* means the creation of new and separate records that refer to the consecutive or classified locations of the information itself. Examples are library subject card catalogs, file cross-index sheets, and *Chemical Abstracts*, the monumental index to chemical literature.

Conventional indexes have the fault of being indirect; that is, they provide, not information, but only clues to its location. Another disadvantage is that the searcher can consult only one index characteristic at a time. If he wants information that has a combination of indexed characteristics he must wade through superfluous references that refer to only one of the characteristics he needs. For example, a search for information on "automobiles" in "Texas" means consulting index entries under both terms and then eliminating all entries that do not concern both.

The major shortcoming of conventional indexes is the time and labor consumed to create and use them. As published literature and

<sup>2</sup> Leo E. Montagne, "Historical Background of Classification," in Maurice F. Tauber, ed., *The Subject Analysis of Library Materials*, p. 16-28 (New York, 1953); J. W. Perry and Allen Kent, *Documentation and Information Retrieval; an Introduction to Basic Principles and Cost Analysis*, p. 116-130 (New York, 1957).

private reports increase and as new indexable characteristics are conceived, the demands in time and labor increase geometrically, indexing falls behind, and literature searches become more laborious.

*Classified arrangement* is the arrangement of information in predetermined categories having significance as subjects. Examples are most library shelf arrangements and the classified telephone book. Classified arrangements also have their limitations. Whoever creates the categories must accurately assess the future use of their information content, something that proves difficult in practice. Books and records that straddle categories can be placed in only one. Most important, classification systems tend to collapse under great increases in volume.

An example of a classification system in trouble is provided by the U. S. Patent Office. The Patent Office categorizes all patents by class. When an application is received all patents in the appropriate class must be searched to make sure the application is truly novel. As time goes on searches become increasingly difficult because of the rising volume of prior patents in each class. Adding more searchers becomes less and less economical and unfeasible for other reasons. The problem is therefore cumulative and cannot be solved unless radical new means of searching are found. Despite many ingenious adjustments in the system, patent searching is falling years behind and becomes a real barrier to scientific progress. Here is a classic example of a classified arrangement breaking down under stress of increasing volume.<sup>3</sup>

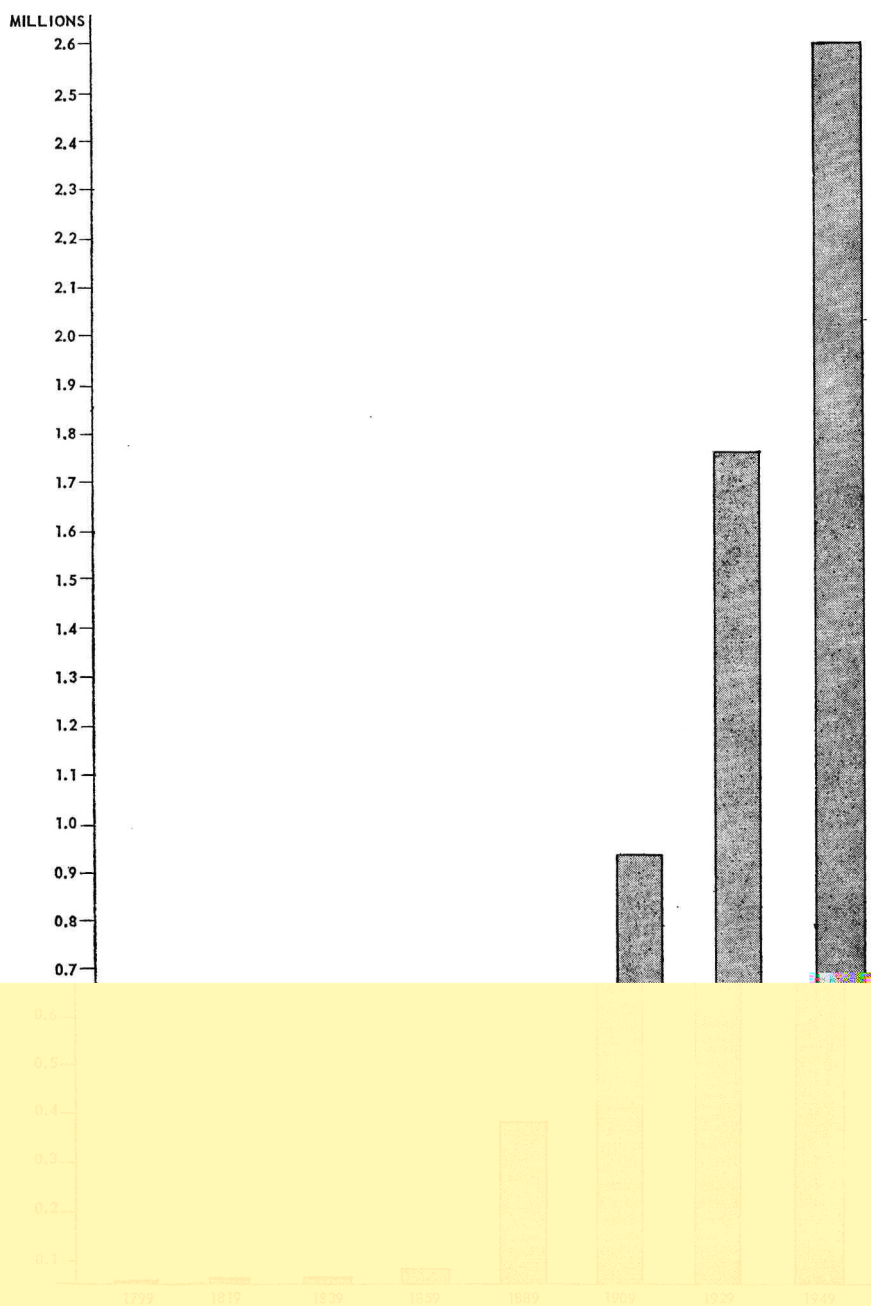
The defects of traditional approaches to consecutive arrangement, indexing, and classification were scarcely noticed in the past, when the volume of information to be stored and retrieved was low. But the quantities of information increasing in almost geometric proportion year upon year made these defects glaringly apparent. Traditional methods had to be changed; there would be new systems, new concepts, new machines, and, inevitably, new problems.

#### NEW SYSTEMS<sup>4</sup>

*Edge-notched cards* were an early means to improved information retrieval. They are still the simplest and easiest to use. Known also as edge-punched cards and by the trade names McBee, E-Z Sort, and Flexisort, these cards have many uses in accounting systems as well

<sup>3</sup> Advisory Committee on Application of Machines to Patent Office Operations, *Report to the Secretary of Commerce* (Department of Commerce, Washington, 1954).

<sup>4</sup> Much of the information on these new systems has been taken from Jesse H. Shera, Allen Kent, and James W. Perry, *Information Resources; a Challenge to American Science and Industry*, p. 160-161 (New York, 1958), and from Perry and Kent, *Documentation and Information Retrieval*, p. 123-128.



GROWTH OF PATENTS GRANTED BY U.S. PATENT OFFICE  
CUMULATIVE TOTALS 1799-1949

as in information retrieval. Edge-notched cards differ from older information handling systems in several important respects. All indexing characteristics of a single record are punched on one card. This card may contain the entire record if microfilm card stock or windows are used, but most edge-notched cards represent the record with an abstract and location reference. Edge-notched cards may be arranged in any order—classified, consecutive, or random.

Each card has holes prepunched around the edges, and each hole signifies one indexing characteristic. To code a card one punches out the appropriate holes so they become notches. The searcher sorts a deck by passing a long needle, like an icepick, through the holes and lifting the deck on the needle. All cards with notches where the needle passes through then fall from the deck and constitute the retrieved information. By using more than one needle the searcher can sort for combinations of indexing characteristics. To find information on automobiles in Texas the searcher puts one needle through the holes for automobiles and another needle through the holes for Texas. Then he lifts both needles and the cards concerning neither characteristic or only one remain in the deck, while cards concerning both characteristics fall out.

Ideal for very small applications, edge-notched cards soon proved useless for major information retrieval problems. Physical difficulties in needling the cards and certain indexing limitations confined the practical number of cards to about 10,000. Below this quantity their many advantages make them attractive. Unlike more complex systems, they carry all or much of the information itself on the indexing medium and are conducive to browsing.

*Aspect cards*, including Uniterm, Batten, and Peek-A-Boo cards, differ from edge-notched cards in that there is one for each index characteristic rather than one for each record. When an aspect system is used the original records are filed in consecutive order by number. Each aspect card has a standard grid with spaces for each record number. Searching for automobiles in Texas, the searcher takes the card for automobiles and the card for Texas and superimposes one on the other. Light passing through a hole common to both cards indicates a record that combines both index characteristics. The first aspect cards were hand-manipulated, but mechanical systems are under development. The National Bureau of Standards prototype machine to mechanize its Peek-A-Boo system was described at the Boston meeting of the Society of American Archivists.

*Cross-filing* systems start with a classified file. Unlike conventional classification systems, which accept a record in only one place, cross-filing accepts the record or its duplicate in as many places as

necessary. A record of automobiles in Texas would go under either "automobiles" or "Texas" in a conventional classification scheme. A cross-file system duplicates the record and places copies under both headings.

An outstanding example of the cross-file technique is the Titanium-Aircraft file at Battelle Memorial Institute.<sup>5</sup> This file brings together all available information about titanium uses in aircraft. It deliberately excludes information about titanium not related to aircraft and information about aircraft not related to titanium. To create the file, scanners read all available literature in appropriate fields and extract pertinent passages. The extracts are then typed on multilith masters, reproduced on 5x8 cards, and filed under all applicable categories. The beauty of the system is that a searcher can find all he needs to know about any subject category, including interrelationships with other categories, merely by reaching into the file and pulling out a handful of cards. In all other systems, handling cards or equipment is merely a first step. Cross-filing also is more conducive to browsing than any other system.

The drawbacks of cross-filing must be readily apparent. Bound books and thick reports cannot be cross-filed for physical reasons. A record to be cross-filed must be standardized, miniaturized, and duplicated—with all the time and expense this implies. A record might well go in 50 places, but a marked increase in duplicate copies could swamp the system. But if applied within definite limits and for a specific purpose, like the Battelle file, cross-filing can be very effective.

The Battelle system is similar to and may be modeled after the Yale University Human Relations Area File started in 1949.<sup>6</sup> The Yale file has been built up in the manner described above, but has a broader subject matter and faces difficulties from a growth of 14 percent annually. This growth is forcing the file's custodians to abandon the basic concept of a searcher's finding all his information in his hand and to turn to microfilm and computers for help.

### NEW CONCEPTS

Attacks on older ways of information handling have been advanced with much vigor by three men at the Western Reserve University School of Library Science—Allen Kent, J. W. Perry, and Jesse H. Shera. Typical of their viewpoint is a 1952 article by Mr.

<sup>5</sup> Author's notes on visit to Battelle Memorial Institute, Columbus, Ohio, Oct. 3, 1957.

<sup>6</sup> Frank W. Moore, "Social Science Documentation," in *Special Libraries*, 49:421-426 (Nov. 1958).

Shera,<sup>7</sup> in which he declared that conventional library methods had become obsolete. The hierarchical classification of knowledge was too rigid; knowledge itself was constantly growing and shifting and could not be relegated to a set of pigeonholes. A book that looked significant in one category today would be far more important in another category tomorrow; 50 index cards for 50 different subject characteristics could not foresee a 51st, nor could they provide for interrelationships among characteristics. Librarians, said Mr. Shera, must abandon their traditional emphasis on physical forms such as books. Instead of classifying books they should classify the ideas in them.

Shera proposed an entirely new approach based on symbolic logic. Citing Leibnitz, George Boole, Bertrand Russell, and Alfred North Whitehead, he averred that the exact methods of mathematics are applicable not only to the study of quantities but to any realm of knowledge, particularly to relationships between classes and propositions. This symbolic logic or mathematical language could be used to manipulate the relationships of ideas in the same way ordinary algebra could manipulate the relationships of numbers.

Basic to Shera's proposal was the fact that the new mathematics employed a binary algebra with only two values, 0 and 1. This "yes" and "no" language was, of course, the language of the electronic computer. Once a classification system was translated into the equations of symbolic logic it could be manipulated by existing hardware to search any desired logical combinations of references or characteristics. Here is one kind of logical problem a computer can solve:

It is known that salesmen always tell the truth and engineers always tell lies. G and E are salesmen. C states that D is an engineer. A declares that B affirms that C asserts that D says that E insists that F denies that G is a salesman. If A is an engineer, how many engineers are there?<sup>8</sup>

If computers can solve this, runs the thinking of Shera and others, why can they not solve a problem like retrieving information on automobiles in Texas, or even blue automobiles in Texas except for air-conditioned Cadillacs?

With these new concepts in mind, the leaders in information retrieval research turned to punched card machines, computers, and microfilm manipulating devices.

<sup>7</sup> "Classification; Current Functions and Applications to the Subject Analysis of Library Materials," in Tauber, *Subject Analysis*, p. 29-42.

<sup>8</sup> David O. Woodbury, *Let Erma Do It; the Full Story of Automation*, p. 238 (New York, 1956).



## NEW MACHINES

*Punched card machines*, manipulating Hollerith cards, are the familiar machines used in everyday accounting operations. The use of punched cards for information retrieval is essentially the same as that of edge-notched cards, except that searching is done by machine, the number of index entries is unlimited, and less graphic information can be recorded on the card itself. The volume of workable punched card files varies from 15,000 to 75,000. They can provide relationships among indexing characteristics more complex than edge-notched cards can, but they are inferior to computers and advanced microfilm systems in this respect.

The fact that Hollerith cards are manipulated by automatic accounting equipment designed for routine data processing imposes many limitations on converting them for information retrieval applications.<sup>9</sup> These limitations and the expense of the machines limit their practical applications to relatively narrow fields such as spectrometric analysis and engineering drawings. The latter application uses a punched card with a microfilm "window" insert containing the drawing. This permits machine searching of the cards and retrieval of full-size drawings by blowing up prints from the "windows."

*Electronic computers* offer many advantages for information retrieval. They accept an unlimited number of index entries per document and will search unlimited complexities of relationships among the entries using the concepts of symbolic logic. Unfortunately, the standard commercial computers designed for accounting purposes turned out to be hard to adapt for information retrieval.<sup>10</sup> Attempts to use them for information retrieval have often proved inefficient and prohibitively expensive, and the answer now seems to lie in developing special computers aimed solely at information retrieval applications. Work in this area is now in the research and development stage. One such project is the GE-250 Searching Selector, a tape-fed machine of computer-like design developed by Western Reserve University and General Electric.<sup>11</sup>

Punched cards and computers suffer from one important limitation, which may in the long run disqualify them as information retrieval tools. They cannot conveniently store large quantities of graphic information. They are primarily designed and are best

<sup>9</sup> Perry and Kent, *Documentation and Information Retrieval*, p. 127.

<sup>10</sup> *Ibid.*, p. 128.

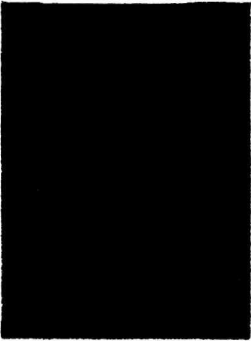
<sup>11</sup> National Science Foundation, *Current Research and Development in Scientific Documentation*, no. 7, p. 50, 83, 90 (Nov. 1960).



1W676	229	C1/C1	1W741	FL ASSY-ELECTRICAL CHASSIS	10
DRAWING NUMBER		FRAME NO.	ED ON	TITLE	Dist. No.

**ORIGINAL CAMERA FILM**



**THE RECORDAK PRECISION  
ENGINEERING DRAWING SYSTEM**

**SAMPLE**

Reduces Costs  
Improves Quality  
Provides Faster Service

**RECORDAK CORPORATION**  
(Subsidiary of Eastman Kodak Company)

DISTRIBUTED BY RECORDAK CORPORATION      FILMSORT (R) U.S. PAT. NOS. 2,811,859; 2,812,106; 2,807,022 OTHERS PENDING

151-403405

PUNCHED CARD WITH MICROFILM WINDOW

(The microfilmed engineering drawing is indexed by the punched holes.)

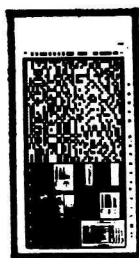
sited for handling digital information, raw data expressed in numbers. Yet the information retrieval problem centers around books, reports, and other media containing prose and graphs as well as figures.

*Microfilm* was an essential part of the system that Vannevar Bush envisioned in his prophetic article. He did not intend graphic information to be reduced to codes or stored in a file system separate from the index system; he wanted a single machine that would store, find, and produce or reproduce the graphic report. He chose microfilm because it preserved the graphic record in a miniaturized and standardized form convenient for storage and machine handling.

Soon after publication of his article attempts to design such a machine were made at Massachusetts Institute of Technology and elsewhere. At first experimenters used conventional microfilm rolls, but soon they decided to cut the rolls into manageable pieces called bits. They set aside part of each bit for one or several page images and left part for indexing codes. They made unlimited

---

INDEX CODES →



← INFORMATION

THE EASTMAN KODAK MINICARD

---

indexing possible by adding more bits entirely devoted to codes. Then they built equipment similar to punchcard machines to manipulate the bits and designed special electronic computers to search the codes. These computers had the same ability to handle logical relationships that ordinary computers have, but they were specially designed for information retrieval use. This special equipment uses optical sensing devices activated by passage or nonpassage of light through small squares on the index parts of the bit.

Microfilm retrieval systems will search unlimited complexities of relationships among index entries. Unlike punched cards or conventional computers, the result of a search is not just a list of document numbers but hard copies of the records themselves reproduced from microfilm. Though now so expensive that only the Federal Government can afford them, microfilm retrieval systems are being

developed rapidly by several companies and their price tags promise to drop from an original \$1,000,000 per complete system to a more reasonable figure. Eastman Kodak is a leader in developing microfilm systems and now has three in actual operation in Federal agencies. Other organizations working to develop microfilm retrieval systems are Itek, Inc., Bell and Howell, AVCO Corp., General Precision Laboratory, Inc., the National Bureau of Standards, FMA, Inc., Magnavox Co., and more.<sup>12</sup>

### NEW PROBLEMS

As government agencies, universities, and businesses learn more about information retrieval techniques they see two major problems emerging—the coder problem and the semantic problem.

All the new retrieval systems have one thing in common with conventional file and library systems: human beings have to do the classifying and indexing, and the systems are therefore no better than the people who feed them information. The drawback is especially significant in punchcard machine, computer, and microfilm systems that remove the information from the searcher and leave it accessible only through the indexing. The indexing is done by coders who are not always the intellectual equals of the searchers and who have difficulty in foreseeing the searchers' needs. Moreover, the coders' job is a dull one—just the sort of job that automation is supposed to eliminate.

One solution to the coder problem is an optical scanner such as the one being developed by H. P. Luhn of IBM. His scanner is intended to "eliminate the intellectual and manual tasks still existing in the preparation and processing of documents for mechanical retrieval."<sup>13</sup> Luhn's device can scan a printed book or record and select key sentences to be printed out as an "auto-abstract." The auto-abstract is stored in a computer memory, which the searcher consults by submitting a typewritten question to the scanner for computer input. The computer analyzes the question for relationships among words similar to those in its memory and retrieves all related auto-abstracts. Even if the auto-abstract idea should prove unfeasible, it would still be possible to use Luhn's scanner to index on the basis of single words or "keywords."<sup>14</sup>

The second problem—semantics—is the one most frequently

<sup>12</sup> Author's notes on visit to Eastman Kodak's Hawkeye Plant, Rochester, N. Y., Oct. 10, 1957, to inspect Minicard equipment; National Science Foundation, *Current Research*, p. 81-89.

<sup>13</sup> National Science Foundation, *Current Research*, p. 30.

<sup>14</sup> Charles L. Bernier, "Documentation in the Field of Science," in *Special Libraries*, 49: 419 (Nov. 1958); Woodbury, *Let Erma Do It*, p. 256-257.

cited by information retrieval researchers. If words had strictly defined meanings that everyone understood, all would be well. The problem of matching words to meanings is a notorious hindrance to communication among people, to say nothing about communication among people and machines. Again using our well-worn example, suppose we want information on automobiles in Texas and the information is indexed under motor vehicles and Lone Star State—what then?

One answer is the keyword system already mentioned, where a machine recognizes and uses only a restricted vocabulary of certain words. Another is Itek's development of L-Indexing, which translates conventional language into language expressible in symbolic logic. Lockheed Aircraft is searching for a "normalized" English in cooperation with Itek. MIT and Ramo-Wooldridge are working on a system that will search natural language (our confusing and illogical everyday English) without many special language changes. Planning Research Corporation is working on a "simplified language." The Patent Office has come up with a "meta-language" amenable to logic called "Ruly English." Remington Rand-Univac is one of the many organizations putting together a "machine thesaurus" to translate illogical ordinary language into logical machine language.<sup>15</sup> The erection of this latter-day Tower of Babel is evidence enough that the semantic problem is serious, that it is being attacked from many sides, and that its solution may well be the key to successful information retrieval on a large scale.

Strangely enough, it seems easier for machines to translate one language into another than to translate any language into logical terms. Programs are underway for machine translation into English of Russian, French, German, Italian, Japanese, Czech, Polish, Serbo-Croatian, Chinese, and Arabic. These programs are mostly restricted to technical literature, and some can already produce unorthodox but understandable translations.<sup>16</sup>

#### THE FUTURE OF INFORMATION RETRIEVAL

In spite of many advances in theory and technique, information retrieval is not yet a practical success. It certainly has not solved the problems of scientific information. A 1958 survey<sup>17</sup> noted that

<sup>15</sup> National Science Foundation, *Current Research*, p. 17; Bernier in *Special Libraries*, 49:419.

<sup>16</sup> National Science Foundation, *Current Research*, p. 57-78; Woodbury, *Let Erma Do It*, p. 253.

<sup>17</sup> Burton W. Adkinson, "United States Scientific and Technical Information Services," in *Special Libraries*, 49:407 (Nov. 1958). For a critical view of information retrieval see I. A. Warheit, "Machines and Systems for the Modern Library," in *Special Libraries*, 48:357-363 (Oct. 1957).

scientific information handling had deteriorated further in the 13 years since Vannevar Bush's article. Technical information had increased rapidly: technical journals were 24,000 in 1924, 50,000 in 1952, and would be 100,000 by 1979. Technical libraries were swamped and had virtually abandoned hope of covering their respective fields. Fourteen of the leading abstracting and indexing services were covering only half the literature in their fields. These libraries and services still had to depend on conventional systems because machine systems developed thus far were too costly or just not so good.

Though information retrieval has not yet accomplished what it set out to do, there are good reasons for believing it will in the future. First, it is badly needed. All information, technical and nontechnical alike, is constantly increasing in volume and complexity, and new ways must be found to handle it. Second, many American electronics and business machines companies have committed their scientists and their money to information retrieval research. Much of this research is encouraged or supported by the Federal Government. Third, the products of this research are gaining acceptance as they appear. A 1958 list of American organizations using machines for technical literature searching shows 10 government and defense organizations, 9 universities and private institutions, and 14 business organizations.<sup>18</sup> Finally, we should expect machines to assume more and more of the monotonous tasks of indexing and research just as they have assumed repetitious operations in our factories.

As new methods of information retrieval are applied to scientific and other areas of human activity, we can expect many innovations in records and recordkeeping. These innovations may be revolutionary. One could be centralized recording and machine-searching of land records in State capitals. Another could be the reduction of major genealogical source records to a single computer installation for accurate and instantaneous searches. Another could be the automation of business records such as those of credit-rating companies. Will not innovations of this magnitude call for changes in our own field of archives and records management?

How should we prepare for these innovations? I think we can prepare ourselves in three ways. First, we should keep informed of every advance that might improve our records. Second, we should keep our minds open to the changes that will inevitably re-

<sup>18</sup> Allen Kent and James W. Perry, *Centralized Information Services; Opportunities and Problems*, p. 152-154 (New York, 1958).

sult when these advances are put into practice. Third, we ourselves should play active roles in introducing these changes and in adapting information retrieval advances to records—the heart of our profession.

### ***Plain and Comprehensive***

Book-Keeping is a system of recording the transactions of business in any occupation in which a person may be engaged, so as to show in a plain and comprehensive manner, its condition and progress, and thereby enable that person to know his pecuniary situation, possess ability to substantiate his claims and protect his property, and at death leave behind him, evidence that will enable his friends to understand his business relations and engagements, and settle his affairs in a satisfactory manner.

—*Gray's Manual of Reference for Business Men*, p. 157 (Springfield, Ohio, 1879).

### ***Lost If Not Recorded***

Communication may leave no trace. A conversation, whether direct or over the telephone, will be lost if it is not recorded. But in our time a dozen new means of recording have been invented to supplement or replace handwriting and print. We now have typing with carbon copies, mimeographing, microfilming and tape-recording of radio talks and TV appearances. Probably we are only at the beginning of this new series of recording and communicating devices; and already we are keeping a record of a far larger portion than ever before of a total of human utterances which is greatly increasing in volume.

It would take an atomic world war to reduce this accumulating mass of stored information to the dimensions of those fragmentary records that have come down to us from the less recent past. Meanwhile, the increase in the output of documentation is fantastic. In Britain, for example, the number of cubic feet of official documents producing during the Second World War and not destroyed by “conventional” (i.e. pre-atomic) air bombardment is said to have been greater than the number of cubic feet of surviving documents from the whole previous history of the kingdoms of England and Scotland.

—Arnold Toynbee, “The Gulf of Ignorance: Can Computers Help Bridge It?” in *The Information Explosion*, sec. 11 (advertisement), *New York Times*, Apr. 30, 1961. Quoted by permission of International Business Machines Corp.; ©1961 by International Business Machines Corp.