REPORT NO. RSIC-510

METHODS FOR SATISFYING THE NEEDS OF THE SCIENTIST AND THE ENGINEER FOR SCIENTIFIC AND TECHNICAL INFORMATION

by

Hubert Murray, Jr.

January 1966

Distribution of this document is unlimited.

REDSTONE SCIENTIFIC INFORMATION CENTER
REDSTONE ARSENAL, ALABAMA

JOINTLY SUPPORTED BY

U.S. ARMY MISSILE COMMAND

GEORGE C. MARSHALL SPACE FLIGHT CENTER

11 January 1966

Report No. RSIC-510

METHODS FOR SATISFYING THE NEEDS OF THE SCIENTIST AND THE ENGINEER FOR SCIENTIFIC AND TECHNICAL INFORMATION

by

Hubert Murray, Jr.

Distribution of this document is unlimited.

Research Branch
Redstone Scientific Information Center
Research and Development Directorate
U. S. Army Missile Command
Redstone Arsenal, Alabama 35809

ABSTRACT

This report covers, in brief form, the methods for satisfying the needs of the scientist and the engineer for scientific and technical information. A model of method for scientific communication is discussed. The equipment and techniques which are being developed in an effort to handle information more effectively are discussed in the light of present and future input, storage, and output devices.

FOREWORD

This report is a reprint of a presentation by the author to the Working Group on Analytical Chemistry of the Chemical Propulsion Information Agency in Cleveland, Ohio, on 3 November 1965.
Valuable assistance given by Mr. Gus J. Caras of the Redstone Scientific Information Center is acknowledged.

ILLUSTRATIONS

The importance of scientific communication cannot be over emphasized. The rate at which scientific developments can proceed frequently depends on the facility, speed, and accuracy with which technical information can be procured and disseminated. The time needed to search for information is a function of the amount of information to be searched. If nothing is done to handle information more efficiently, the user will have to wait longer and longer for his answer as the store of information grows larger, and a point may soon be reached where it would be cheaper and easier for scientists to rediscover something that had already been discovered than to search for it in the literature.

Like raw materials, capital, and people, information is a great national resource -- indeed an international resource. Much of what we do or fail to do is based on information, or unfortunately, is based on misinformation. Since the Federal Government is the primary supporter of research and development, it is concerned with the efficiency of scientific communication. A report entitled, "Science, Government, and Information: The Responsibilities of the Technical Community and the Government in the Transfer of Information" was prepared by the President's Scientific Advisory Committee Panel on Science Information under the chairmanship of A. M. Weinberg. On the legislative side, Vice-President Humphrey (then a senator) and the Senate Committee on Government Operations studied the problem extensively and have issued several reports dealing with it.

In order to emphasize the problem, it is worthwhile to cite some figures. The rate at which technical documents are produced at the present time is estimated to be well over 500,000 per year. In every 24-hour period approximately 20,000,000 words of technical information are being recorded. A reader capable of reading 1,000 words per minute would require $1^1/_2$ months, reading 8 hours every day, to get through 1 day's technical output, and at the end of that period, he would have fallen $5^1/_2$ years behind in his reading! Even in attempting to read the portion of the literature in a single subject field such as chemistry, he would find himself falling behind an estimated 850,000 pages per year. This production rate of scientific information will undoubtedly increase as countries such as China and India begin to produce technical work commensurate with their size.

Prior to a discussion of equipment and techniques which are being developed in an effort to handle information more effectively, a model of scientific communication is shown by Figure 1. This system is a simplification of the process based on the concept that scientific information is produced by Researcher and transmitted to Researcher.

Researcher is also a producer of information, and the information he produces can be transmitted to Researcher, among others, thus closing the loop. The system becomes more complex when all researchers are substituted for Researcher and Researcher, and when the information must pass through such barriers as time, distance, discipline, and language.

Books, reports, journals, and so forth are the more common methods used in transmitting information. Oral communication is important, but has not received as much attention in recent years as printed communication. Scientists attend conferences, talk with associates, and in general interact with other members of their group.

Since no one knows when and who will need the information produced by a given researcher, the material must be stored where it will always be available to those who need it. This storage place is normally a library or an information center. Other less familiar places where information can be recorded and stored are magnetic tapes, punched cards, and so forth. Although a reel of tape and a shelf of books in a library have no physical similarities, in many respects they serve the same function since both represent storage media. Storing of information in some manner is extremely important because of the relatively small amount of communication of one scientist to another.

The information system is the connecting link from one scientist to another in the exchange of information. Figure 2 represents the major functions of a system where material is collected, processed, and made available to the user through various classification and indexing schemes. In practice, however, the flow of information is not as simple as depicted in this figure because of barriers through which information must be transmitted. One of these which is easy to understand and represent is the linguistic barrier. Figure 3 shows how information might be transmitted between scientists who speak different languages.

Our recorded knowledge has been compared to dinosaurs whose extinction can be partially attributed to the fact that they grew so big that the brain could not control the unwieldy body. Our collection of information can also become so huge that its parts cannot be controlled efficiently.

Because of the increasing amount of materials and the need for effective control, education and training have become urgent matters. In the past, librarians were capable of handling information quite well, but now that volume and complexity have increased so much, it is impossible for them to cope with the problem themselves. The fact that technical advances now make it possible to process information many times faster than before means that the boundaries of information handling have widened enormously. As a result, such disciplines as mathematics, chemistry, electrical engineering, computer technology, and psychology are contributing their technologies for the solution of the information problem. During the last few years a new science has emerged known as information science, and several schools throughout the country have graduate degree programs in this new area of academic study. Information science may be defined as the science that investigates the properties and behavior of information, the forces governing the flow of information, and the means of processing it for optimum accessibility and usability. The training of information scientists is designed to bridge the gap between the librarian and the scientist.

In addition to education new techniques and equipment are being developed that promise to speed or simplify the flow of information. Frequently these techniques are among the most spectacular developments of modern science and engineering. The computer is one of the outstanding developments in equipment. Its significant role in the mechanization of information handling has opened a wealth of possibilities because it provides capabilities enormously more powerful than those of any other mechanical device. Equipment developments are discussed in relation to digital computers, assuming that computers will become the central mechanical elements in the mechanization of information processing. The general types of devices necessary for automatic handling of information are input, storage, and output devices.

In order to introduce data into an electronic processor, the data must first be converted into a language suitable for machine processing. The conversion of information into machine readable form presents a major problem because this operation is largely manual, and as such, it is time-consuming and expensive. The type of code particularly important for machines is the binary code which involves two basic symbols, represented as 0 and 1, called "bits." Familiar forms of data suitable for machine processing are punched cards, punched paper tape, and magnetic tape. Currently, the type of information that can be coded conveniently is limited to letters, digits, and a few symbols. Chemical structures, pictures, drawings, and the like are also a part of our natural language, but at present these are much more difficult to represent in machine language.

Much research is currently in operation in an effort to develop techniques whereby original data from the printed page can be converted automatically to machine language without human intervention. Character-recognition devices have been developed which are capable of reading alphanumeric characters in a prescribed format. The two most popular character recognition techniques in use involve either optical scanning or magnetic ink reading.

The optical scanning technique involves an image of the character by reflected light to a lens system. The optical system sharpens the image and transmits it on through two intersecting slits to a photomultiplier. The white parts of the image transmit light, and the black portions of the image do not. The photomultiplier converts light patterns into electrical signals. In theory, if the combinations of spots of light for each image are different, then the converted electrical signals will also be different. These differences, in effect, give the technique its power to discriminate between numeral characters. As the photomultiplier produces electrical signals, they are sent to a special purpose computer for comparison with all of the possible combinations of electrical signals, each of which identifies a given numerical character. After a number has been identified on the basis of the comparison, computer circuits are activated which trip the appropriate character on a printer, or display it on an electrical print-out device.

In magnetic ink reading, characters are imprinted by using ink containing magnetic oxide, and the characters are then passed through a magnetic field where they become energized or charged. A magnetic read-head senses the signals which are generated by the thin layer of magnetic ink. Each character produces a different signal which is amplified and transmitted to correlation networks where it is used to punch cards, write on magnetic tape, and print. Only certain formats can be recognized at the present. Neither optical scanning nor magnetic ink reading, in their form, can be used as a general purpose character recognition system for examining a typical page from a book. Work on devices to which books can be fed and ones with mechanical arms to turn pages are now in progress, and the developments are not as far-fetched as might be believed. Successful experiments have been conducted on reading even ordinary handwriting. Research on devices which can recognize speech sounds is also being done to determine if a machine can discriminate phonetic sounds automatically and produce a satisfactory digital code for machine input.

Storage techniques can be broadly divided into several major types depending on the medium used such as perforated, magnetic, photographic, and so forth. Perforated media such as punched cards and paper tape have a relatively low storage density, which is the number of bits of information that can be stored in a unit area of medium, and they are not likely to be used widely as a basic medium for recording and storing information. These have certain advantages, however, and they will continue to be used for specific applications. Cards, for example, can be easily sorted, interfiled, or otherwise rearranged, while paper tape is used by magazines and newspapers for the transmission of information from editorial offices to various publishing points distributed throughout the country.

Magnetic devices include tape, drums, discs, cores, and thin films. In general these devices provide very high storage densities. A reel of 2400 feet of magnetic tape may contain over 100 million bits of information. Methods to retrieve information from these storage media have also advanced accordingly. There are tape-handling devices which transport the tape past the reading and recording heads at extremely high speeds; the transfer rate is very high, in some cases over 100,000 characters per second.

Photographic storage media include microfilm, microfiche, film strips, and many others. Microfilm is used extensively for space reduction of infrequently used material such as newspapers and periodicals. These photographic media require a viewing device, and some of these viewing devices combine the ability to print so that the user can make a paper enlargement on the spot.

Photochromics are light-sensitive organic dyes which can be placed as a molecular dispersion in suitable coatings that can be applied to almost any surface. The thing that gives photochromic microimages a powerful advantage in film technology is the ability of exposing the molecular coating with ultraviolet radiation and erasing it with white light. Storage densities are extremely high (Figure 4) on the order of 2,000 lines per millimeter, or a potential storage capacity of over 3 billion bits per square inch. This is nearly 30 times as much as on a whole reel of magnetic tape.

More recently, a thermoplastic device was announced by the General Electric Company. Thermoplastic films can record at high rates and produce optically visible images of very high resolution. The film requires no chemical development and can be erased and re-used. Photoplastic recording (Figure 5) is a photographic process in which images are recorded on a photoconductive film in the form of surface deformations. The image is developed by heat, and by melting the film, the image can be erased and used again. The essential component of the recording medium is the photoplastic film which is sensitive to light and becomes deformable on heating, and the image can be read optically or electrically. If the film is transparent, schlieren optics permit direct viewing. The developed image can be erased by heating the film above the development temperature.

The output devices include a great variety of high-speed mechanical printers, cathode-ray tubes for displaying information in a TV screen fashion, and various photocomposing devices such as the Photon. High-speed mechanical printers usually print 100 to 1,000 lines per minute of 100 to 150 characters per line. Output printers based on the use of cathode-ray tubes for the generation of individual characters are capable of extremely high speeds, up to 5,000 lines per minute.

The Photon is typical of photocomposing machines which operate on photo-optical principles. As currently used, the Photon accepts instruction from punched tape or from a keyboard and prepares a photographic negative for subsequent printing. It also provides some composing flexibility since mixing of style and sizes in the same line is possible. There are at present a number of less sophisticated photocomposing machines on the market.

Another type of hardware important in the area of information processing is the communication devices by which information can be transmitted from one point to another. The present transmission systems fall into three major classes:

1) Electrical transmission. Its most common form is the voice over the telephone which is slow (about 200 words per minute), but offers some advantages such as two-way conversation.

2) Digital class which refers to the transmission of information in digital code.

3) Picture transmission of graphical material. The principle consists of breaking up the picture into a succession of parallel lines. An electrical wave is generated for each line, varying in magnitude with the intensity of light from the line as it is scanned from one end to the other. Ordinarily, the electrical wave analogs are transmitted in succession, and when translated at the receiving end into variable light intensity, the picture can be reconstituted.

The techniques currently used in processing information are varied and are of great importance. Retrieval of information cannot ever be any better than indexing. Indexing has received much attention, and, therefore, many experiments are being conducted to evaluate and improve the various indexing systems. The coordinate type using uniterms or keywords has been used in many mechanized information systems. In most systems these uniterms do not show relationships, a condition which can result in the recall of irrelevant documents. For example, no distinction can be made between "venetian blind" and "blind venetian." There are, of course, certain techniques by which relationships among terms can be indicated.

Another type of indexing, which is normally used for current awareness rather than retrospective searching, is the so-called "keyword-in-context." This technique lists the significant words of a title in alphabetical order with the added refinement that each indexing word includes the words by which it is surrounded. One disadvantage of this index is that, unfortunately, titles are not always descriptive of the material contained in the document. The chief advantage is that the index can be prepared by a computer quite inexpensively which makes the information available to the user promptly. Many other types are in use today, such as author indexes, formula indexes for chemical compounds, and so forth.

A technique designed to provide current awareness is the Selective Dissemination of Information or "SDI," which is based on interest profiles of the users of the information system. For many years librarians have kept track of their clients' interests, and when they received new material that they felt was pertinent, they sent it to them. SDI is based on the same principle except the operation now may be mechanized. The users' interests are expressed by keywords or descriptors. The descriptors of incoming documents are compared with those of the interest profiles, and notification cards with abstracts of the documents are sent to those clients who might be interested. SDI, in other words, is like an electronic traffic director of information, analyzing it and routing it to those who have a need to know.

A large portion of the world's scientific literature is published in foreign languages which enlarges the linguistic barrier. Since manual translation is very slow, much work is going on in the area of mechanical translation. Translation of Russian into English has received the most attention in recent years, although at present several other languages are under study. Perhaps the most notable advance is the increased size of the dictionaries. In the past dictionaries were limited to a few thousand words, but now there are a few mechanically translated dictionaries of 100,000 words or more. One of the major problems in mechanical translation is that of words with multiple meanings, and very little progress in this area appears to have been made. If the computer does not know a word, it prints it out in Russian, and if it does not know which is the right translation of three possible translations of a word, for example, it will give all three and leave a choice to be made. Most mechanical translations leave much to be desired as far as grammar is concerned. CIA through Georgetown University has turned out several mechanically translated books.

Some attention has also been devoted to the development of a common international language which would be logically constructed. It could be learned in a fraction of the time required for most national languages, and it would be a basic language for the different symbolic computer languages. Such a language would be useful as an intermediate language for mechanical translation, although eventually it would be desirable for scientific literature to be published and read in this language without translation.

Automatic indexing and abstracting experiments using computers have been going on for some time. These terms as currently used do not imply the substitution of machineable procedures for human intellectual efforts normally required to identify, classify, index, and abstract documents. Automatic indexing and abstracting are based on statistical methods and the assumption that certain words and sentences in an article contain more information than others. These words and sentences are identified and written verbatim as they appear in the original text. "Abstracting" is, therefore, a somewhat misleading term, perhaps "extracting" would be more appropriate.

Mechanization techniques are being applied in many other areas, but none of these applications at the present time assumes any intelligence on the part of the machine. In a sense, these are not more sophisticated than some routing applications with which computers have been assisting such as routing of periodicals, circulating, ordering, journal renewing, and so forth. If one wished to know which is the most widely used plastic and requested all works on plastics or uses of plastics from an automated system, it would then be necessary to go through thousands of references to obtain the answer. It is hoped that future systems can produce direct answers to questions submitted, for example, "Polyethylene" might be the answer to the question, "Which is the most widely used plastic?" Even this type of information retrieval is fairly elementary compared to a truly creative retrieval method which is envisioned. Suppose that the information system has the two sentences "Socrates is a man." and "Man is mortal." and not another word about Socrates. The obvious question "Is Socrates mortal?" might follow, and since the system fails to find a stored answer, it must deduce the correct response. In other words, it is hoped that information systems in the future will be able to use deductive reasoning. It is assumed that progress will continue, and that considerable improvements may be seen in the methods of handling information.

Automation most likely will be applied at the source. Now in most cases the conversion of information to machine language requires human intervention, but most likely, in the future, reports will be prepared in a format suitable for automatic reading by character-recognition devices, or as the report is written, a machine-language by-product will be produced at the same time. In this way the report will be immediately transmitted in digital code to the information center, and the center will have the information even before the report is published.

Another refinement of the information center of tomorrow is remote interrogation of the system which will enable the user to request information by telephone rather than visit the center. He can receive the answer in a typewriter-like message or a visual display on a cathode ray screen, and if the information is in a foreign language, it will be automatically translated. The center may also be capable of receiving information in oral form. In this manner the research material can be orally communicated to the center as soon as the work is performed. An elaborate system of selective dissemination will render this information immediately available to all other researchers who must know the results.

Since machine methodology and information processing techniques are feasible and attainable within today's engineering environment, this kind of an information center may not be very far in the future.

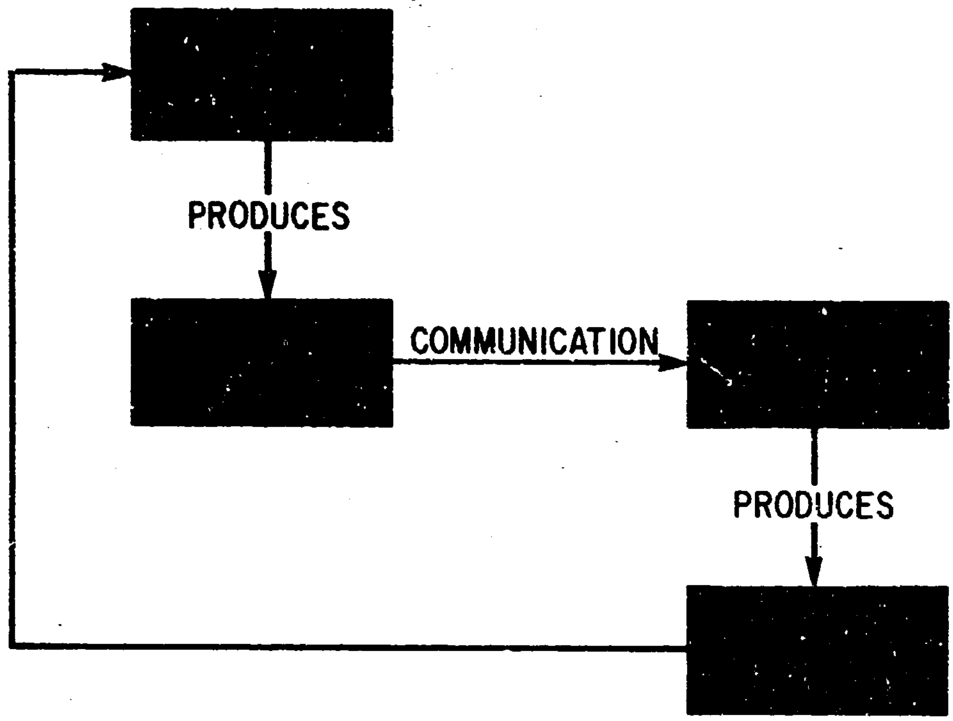

Figure 1. Scientific Communication

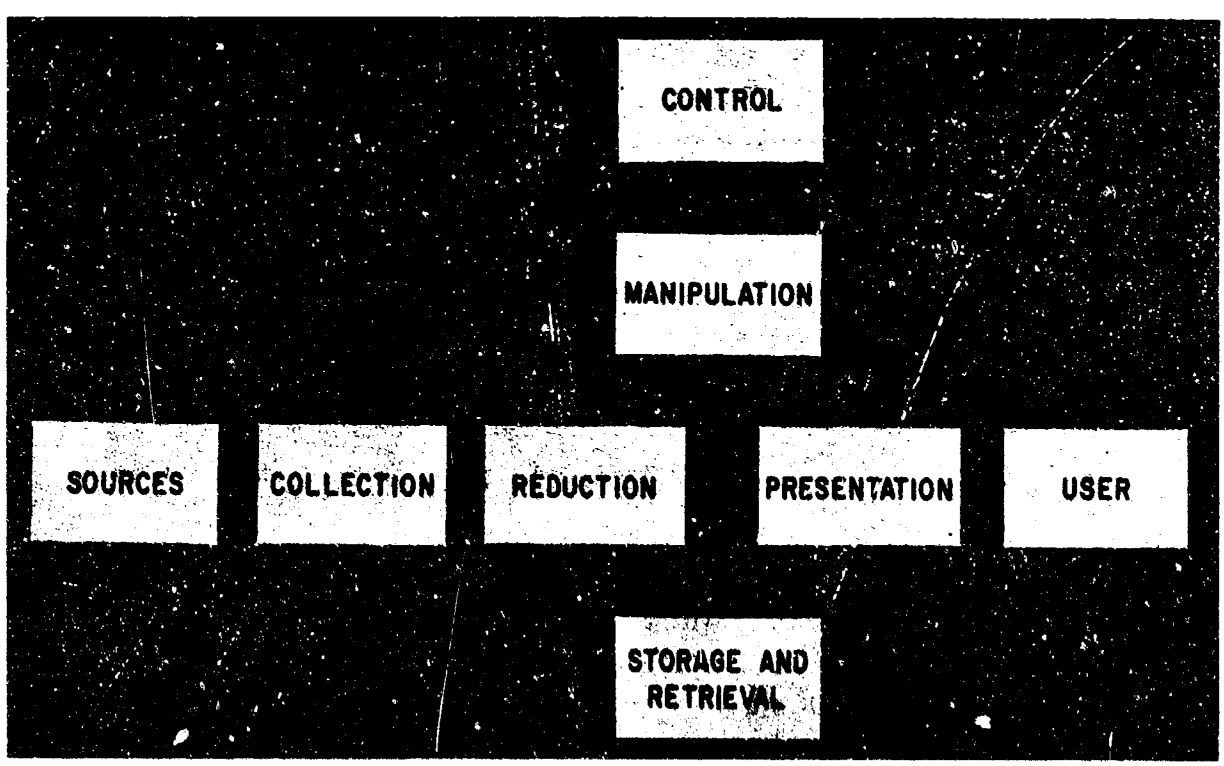

Figure 2. Functional Diagram of an Information System

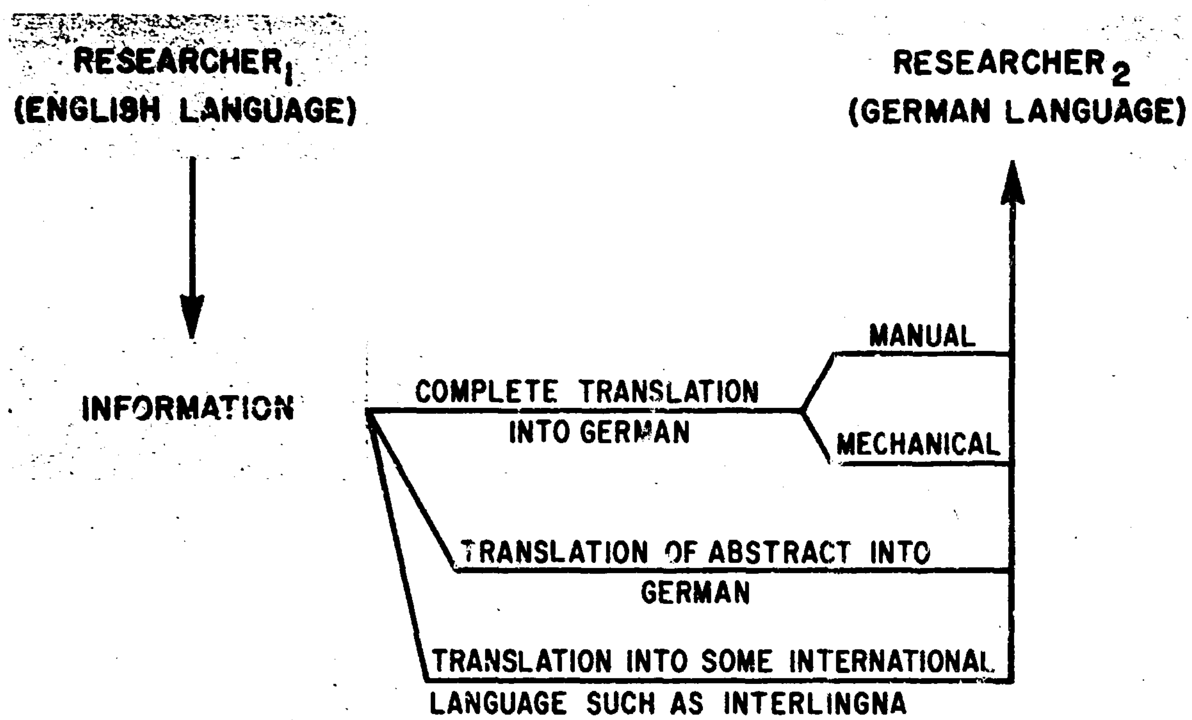

Figure 3. The Linguistic Barrier

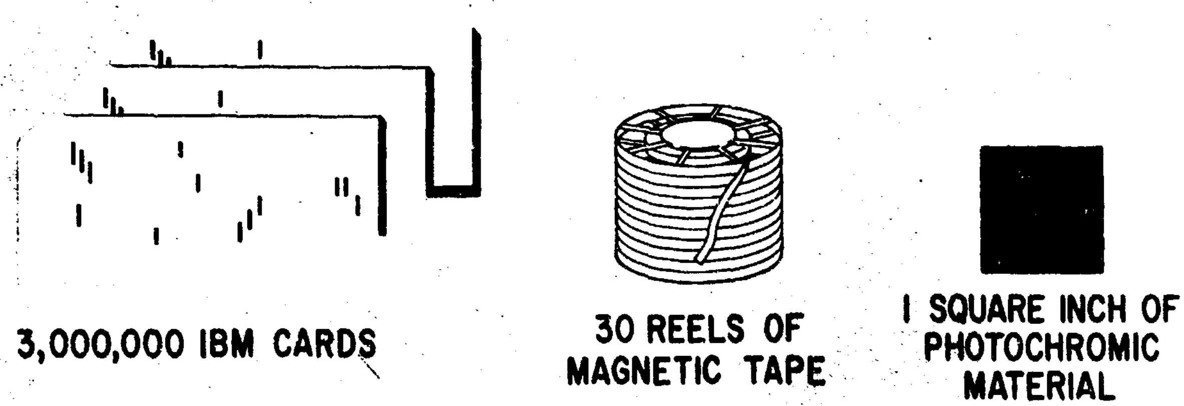

Figure 4. Storage Capacities of Various Media

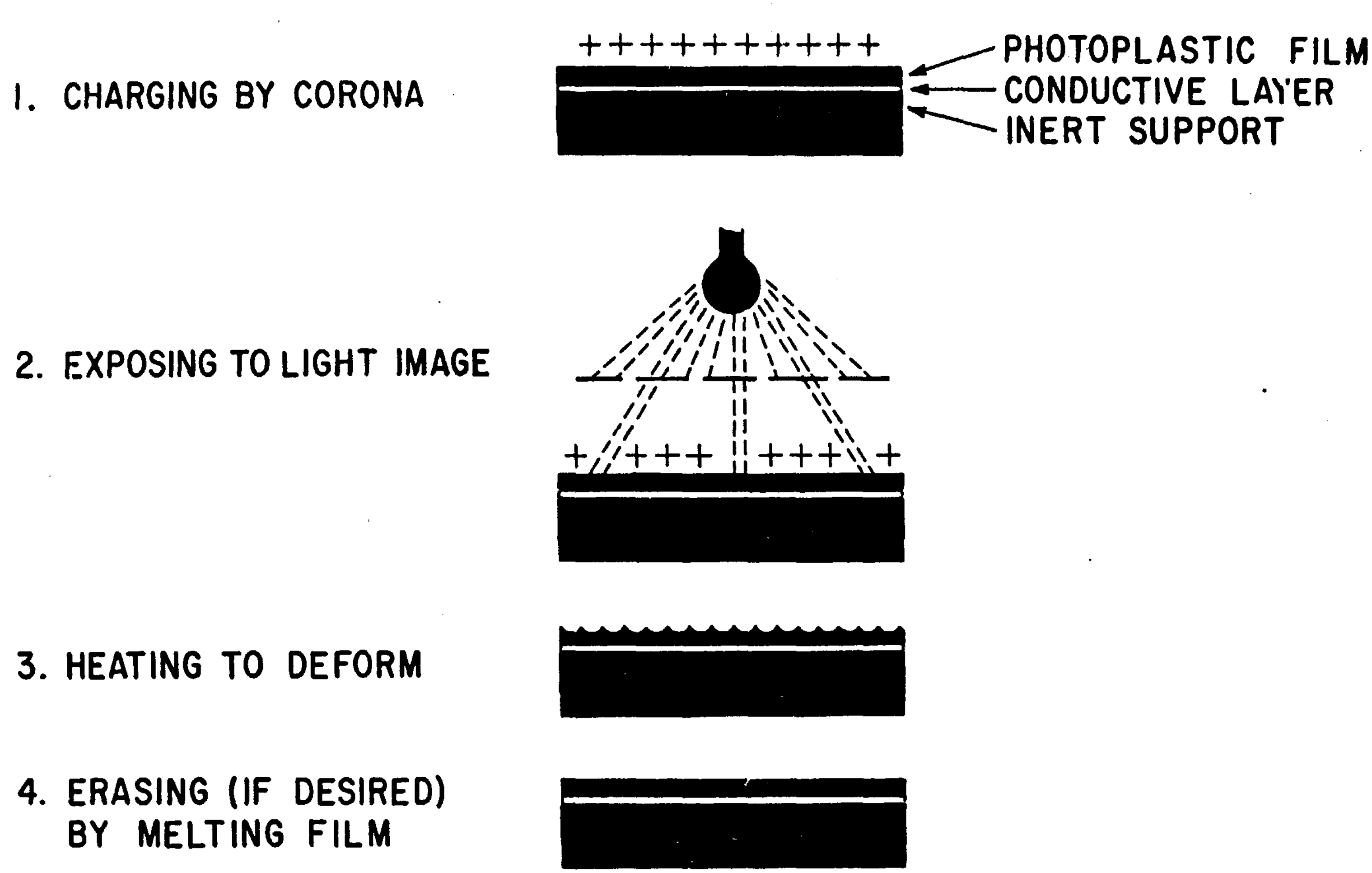

Figure 5. Photoplastic Recording Process

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Redstone Scientific Information Center Research and Development Directorate U. S. Army Missile Command Redstone Arsenal, Alabama 35809	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b GROUP N/A

3. REPORT TITLE

METHODS FOR SATISFYING THE NEEDS OF THE SCIENTIST AND THE ENGINEER FOR SCIENTIFIC AND TECHNICAL INFORMATION

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(Last name, first name, initial)*

Murray, Hubert,Jr.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
11 January 1966	15	None

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S) RSIC-510
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	AD________

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Same as No. 1

13. ABSTRACT

This report covers, in brief form, the methods for satisfying the needs of the scientist and the engineer for scientific and technical information. A model of method for scientific communication is discussed. The equipment and techniques which are being developed in an effort to handle information more effectively are discussed in the light of present and future input, storage, and output devices.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Scientific communication Information system Magnetic devices Photographic storage Selective dissemination of information						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (*corporate author*) issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers (*either by the originator or by the sponsor*), also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through ________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through ________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through ________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring (*paying for*) the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (*TS*), (*S*), (*C*), or (*U*).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.