THE DEVELOPMENT OF RANDOM ACCESS INFORMATION RETRIEVAL IN THE GE/MSD LIBRARY; AND USER INTERACTIONS

By

Lawrence I. Chasen
Manager - Missile and Space Division Library
Valley Forge Space Technology Center

SPACE SCIENCES LABORATORY
GENERAL ELECTRIC
MISSILE AND SPACE DIVISION

THE DEVELOPMENT OF RANDOM ACCESS INFORMATION RETRIEVAL IN THE GE/MSD LIBRARY; AND USER INTERACTIONS

Lawrence I. Chasen
General Electric Company, Valley Forge, Pennsylvania

ABSTRACT

In 1959, a study was undertaken on the feasibility of automating the Missile and Space Division Library. Plans for automation included mechanizing circulation and recall, descriptor retrieval, GE-235 combination, desk catalogs in lieu of card catalog and mechanized listing of subscriptions to journals. Implementation was accomplished by phases over a seven year period. This paper describes the organization of mechanized systems, random access, information retrieval and the GE DATANET 30 SYSTEM. User participation on the consoles is analyzed and Library-User constructive teamwork is shown.

For presentation at the Fourth Annual National Colloquium on Information Retrieval, May 3-4, 1967, Philadelphia, Penna.

PHASE I - HISTORY AND FIRST STEPS TOWARD MECHANIZATION

The General Electric Company's Missile and Space Division (originally known as the Guided Missile Department) was organized in Schenectady, New York, in 1945. The first major task of its young library was to translate for the U.S. Army Ordnance Corps the captured German war documents dealing with the V-2 guided missile, which Hitler had hoped would win the war for Germany.

The General Electric Company was one of the first large industrial concerns in the U.S. to pioneer guided missile research and development. During the early '50's, General Electric successfully launched such "birds" as "Bumper" and "Hermes". As guided missile technology grew, so did the size of the Document Library. By 1956, the Library had a collection of over 73,000 documents. In that year the Department moved to Philadelphia to begin working on the crash project -- the Atlas and Thor re-entry vehicles for the then Western Development Division, ARDC. Since 1956, the size, services and volume of the MSD Library expanded manifold as did many similar libraries and information centers in the U.S.

The problems of handling an ever-increasing work load challenged our staff, who felt like the legendary Sisyphus. By 1959, the size of our collection reached 109,000 documents. We commenced microfilming (3" X 5") large masses of our unclassified collection, to solve the space and storage problem. Later that year, the MSD Business Systems Operation and the Library staff began an intensive study to determine:

(a) Is the present conventional library methodology the most efficient?

(b) Can the present manual system be made more efficient (and less costly) by the use of data processing equipment?

(c) Can data processing equipment be used not only for information retrieval purposes but also as a means to handle a massive circulation program averaging well over 27,000 per year?

Our joint study disclosed that:

(a) Document circulation and recall was practiced in the conventional Library -- requiring many repetitious clerical operations, and thus was not only drudgery for the Library clerks, but documents were held up for days at a time, so that they could not be charged out to other borrowers.

(b) Mechanization was the only answer to a problem which could not be solved by practicing laissez-faire, and

(c) Data processing systems would lead to eventual total subject heading or "descriptor" mechanization, thereby reducing search time radically.

In September 1961, the MSD Library acquired an IBM-026 keypunch machine, an IBM Series 50 collator and IBM Series 50 sorter. Three files necessary to automatic circulation were created. A document accession number file, an employee number file (every MSD employee has a pay number) and a reserve card file (to handle requests for documents in circulation).

Last, but not least, a trained data processing equipment operator was hired by the Library. Bearing in mind the three types of punched card files above-mentioned, when a request for a library loan document is received, the document accession number file is examined to determine if the document is available for loan. If it is, the document is immediately loaned to the borrower (save for classified documents, where a "need-to-know" must be established). If not as will be seen later the request is placed on reserve. Since all requests from library borrowers must be submitted on the standard "Library Service Request Form" designed by the Library with necessary information such as the borrower's name, pay number, work location and the accession number there-on, this form can go directly to the operator for processing. In essence, the borrower need no longer wait to have posting and charge-out take place before he can have the document. For phone requests for document library loan services, a "loading" sheet is used to obtain the information, and then the operator includes this in her daily transactions.

For each document, the accession number, due data and employee number are punched on an IBM card. If the document is classified, the security classification and security control number are also punched. The cards are sorted on the last date due column. If a date due has been punched, the document is now in circulation. If the date due has not been punched, the document was not available and a reserve card has been created. The reserve cards are mechanically sorted by accession number and field. The loan cards are duplicated on the keypunch machine; the ORIGINAL cards are

sorted by employee number and merged into the employee file and the duplicates are sorted by accession number and merged into the accession number file. When borrowed documents are returned to the Library, the reserve file is checked. If another employee is waiting for the document, the reserve card is pulled and the date due is punched on it. The card is then duplicated - one filed by accession number and the other by employee number. The cards for the previous loan are destroyed. The employee number file is essential in processing terminating employees. Twice a month the employee number file is used to prepare overdue notices MECHANICALLY. At one time, it took the Library over four week of work to issue 10,000 overdue notices. This process is now done in 30-45 minutes! Bi-annually, the document number file is used to mechanically inventory our large holdings of classified documents in the collection.

ADVANTAGES OF MECHANIZED CIRCULATION

(a) Better Utilization of Library Personnel

Prior to mechanization three record clerks spent 120 hours per week manually creating and manually maintaining the Library's circulation files. The automation system eliminates 95 of these 120 hours. The record clerks are now assigned to service Library customers, where their knowledge is put to good advantage.

(b) Faster Dissemination of Newly Arrived Documents

Formerly 76 hours a week were spent on preparing circulation cards and pockets for new documents. This meant that there

was a corresponding delay in getting the document into the hands of the scientist or engineer. There was also a delay in generating the Library's Technical Abstract Bulletin of New Acquisitions. Because of the automation program, conventional circulation cards and book pockets have been eliminated.

(c) Improved Control of Classified Documents

Previously each classified document received two numbers - a library accession number and a security control number. The accession number is used for circulation and permanent identification. The security control number is used for accountability in the document control program. Relating the two different numbers (for the same document) at any given point in time was a time consuming procedure involving the classified document log book, the circulation card file, the originating agency file and if necessary, the subject heading file. Under our new system, the security control number and accession number are on the same punched card. For purposes of document inventory, it is now possible to obtain a listing of 5-6 thousand classified documents out in circulation in less than 30 minutes. At one time inventory took hundreds of hours.

MECHANIZED CIRCULATION AND RECALL - PHASE II

The benefits derived by circulation and recall mechanization include not only man-hours saved; but also the elimination of many library practices, which may have stood the test of time in the ante-space age, but cannot

compete in today's literal surge of technical information. Phase I enabled us to steer directly into Phase II "DESCRIPTOR RETRIEVAL."

PHASE II - INFORMATION RETRIEVAL

Utilizing our existing data processing equipment, a series of experiments were undertaken - to determine feasibility of keypunching, subject headings or descriptors, transferring this information to a magnetic tape and attempting to use the GE-225 Computer to retrieve technical information. Briefly, the retrieval system is based on the mathematical Theory of Sets.

Since the computer is mathematically oriented, it follows that a solution to retrieval problems should have a foundation in mathematics. By using Set Theory, we took advantage of its basic flexibility, which provides alternatives to the searcher while remaining sufficiently restrictive to minimize the "noise factor." It has been our experience, in the past four years that even with a computer system, the value of theinformation retrieved is directly dependent on the indexing of the information. At the present time, close to 80,000 descriptors are on magnetic cores, and the thesaurus created by theindexes has grown to 9000 descriptors. A search will yield a listing of accession numbers in the Library's file which may answer the requestors need. Examples of descriptors follow:

(a) All document owned and cataloged by the MSD Library on the pressure in the wake of a vehicle moving at speeds ranging from low subsonic to hypersonic.

(b) Drag effects on a hypersonic test vehicle in a turbulent boundary layer where heat transfer and vibration are guiding factors.

(c) Radar echo areas from a vehicle where the dominant factors of interest are velocity, radiation patterns, distance and effectiveness of test methods.

(d) Wind tunnel test results on the Apollo where Mach No. 18 and above is essential.

A configuration given to the computer for searching is composed of descriptors and set theoretical relationships. We consider all those accession numbers that meet the condition specified by a search request to be a set. We term this set of accession numbers as the "Solution Set." The elements of a solution set are always determined by the relationship between the descriptors. In essence, search strategy and results are concerned with three set theoretical relationship, intersection, union and negation.

Prior to programming the computer for a search, the Librarian has a choice of the following three conditions:

(a) The intersection of two sets, A and B, is the set of all elements common to both set A and set B. We use a plus sign to symbolize intersection: A + B

(b) The union of two sets, A and B, is the set of all elements which belong to either set A, or to set B or to both sets A and B. We employ a slash to symbolize union; A/B

(c) The negation of set A by set B is the set of all elements which belong to set A but not to set B. We symbolize negation by a minus sign: A-B

Essentially, the aforementioned series of possible configuration is the key for information retrieval on the GE-225 Computer as designed for our Library.

PHASE III - RANDOM ACCESS - INFORMATION RETRIEVAL

In January, 1964 the MSD Library magnetic tape holdings were transferred to the GE disc storage subsystem and an A.T.&T. teletypewriter was installed in the Library - thereby eliminating the prior requirement to keypunch a search program directly into the GE 225 computer memory system.

The disc storage unit subsystem is a large capacity, fast random access storage device for information processing systems. It is a new and vastly flexible filing medium that stores information so that data so recorded in a random location can be immediately returned for further processing.

The DSU subsystem makes it feasible to maintain virtually any business or information file on a current basis, transaction by transaction. The method of random filing and updating records has advantages over the necessity of sequency or batching in order to update reco·ds.

The Library can make unbelievably rapid retrieval and while the engineer waits the search query is typed on the teletypewriter - within 30 to 45 seconds - the reply is received, citing the following basic data:

(1) Our library computer code number

(2) Total number of documents satisfying the request and which will meet the search parameters

(3) A chronological listing by document accession number of data desired in the search query.

Management can make expedient and rapid decisions from current reports and statements that are easily produced by a GE data processing system featuring DSU subsystem. The most economical factors that can be determined for thousands of different items that can be obtained with controlling production lot sizes, keeping inventory at optimum levels and being able to set re-order points automatically-all of the accurate and timely information that management needs for rapid decision making in today's fast moving information climate.

The GE-MSD Library disc file can be queried from remote locations (Florida, Texas or California) whose offices or libraries might happen to be part of the data network system.

The General Electric DATANET-30 was designed specifically to enable GE-235 Computer to automatically receive and process information originated at locations remote from the computer center and also to automatically send information (replies, results, etc.) to the remote locations.

The DATANET-30 serves as the primary control and connecting link between the GE-235 information processing system and the transmission line and remote data-originating and receiving equipment. Broadly speaking, the unit functions as a buffer and conversion device; that is, it accepts information in bit-serial form from the transmission lines and converts it into the computer. It also receives information from the computer and conditions it for release on the outgoing transmission lines.

Remote stations may be connected to the DATANET-30 through a variety of transmission facilities. These include leased or public telephone and telegraph lines and privately owned, two-wire cables. The DATANET-30 can also be used to connect the GE-235 computer to public message networks, such as AT&T's teletypewriter exchange service (TWX), or Western Union's TELEX service. A number of other special communications services now being offered by the telecommunication companies may also be used.

The sequential scan counter of the DATANET-30 is designed for connection to up to 128 transmission facilities, thus enabling it to accommodate 128 directly connected remote stations. However, the number of remote users that can be satisfied is dependent upon the memory speed of the GE-235 and also on the application being performed. We feel the system can handle thirty (30) users before the saturation point is reached. The remote terminal devices can be located within a single building, within a given city, or they can be scattered throughout the nation.

USER ACTION AND INTERACTIONS WITH THE RANDOM ACCESS SYSTEM

In the early part of 1966, random access consoles were strategically assigned to many of the Engineering/Scientific operations throughout the Division. This was now an ideal opportunity to test the system from the users' viewpoint.

The Library thesaurus was distributed on a pilot test basis to two engineering groups for their education and subsequent use as the medium to query the system. The language problem between the engineer and the library

had to be overcome. Once this barrier in semantics was under control, our staff taught the library users how to phrase questions. This experiment proved very interesting for the library personnel, in that we were not too far off in the aerospace vocabulary and jargon codification; but most important was the role of the user in recommending additions, removal of terms and modifications to make this system more useable and meaningful.

In May, 1966, we assigned a console to our Branch Library, 3198 Chestnut Street, Philadelphia. Fortunately, the Librarian is also a trained scientist with an empirical approach, and was able to make maximum utilization of random access for the almost 4,000 people whom she serves.

By November, 1966, eleven Engineering/Scientific operations were doing their own search strategy and querying the system - independent of Library Staff supervision or interdiction.

The positive aspects of remote access are already cited in my paper. The critique of the eleven users can best be capsuled as:

(a) Output is only in accession number of document number format.

(b) Intrinsic limitations on the disc storage unit makes it extremely expensive to add title, author corporate author, etc.

(c) A desk catalog to be published at least every month to include above is needed for wide dissemination.

In the meantime, because of the cost limitations, we are considering the print-out of all the titles in numeric sequence, adding it to our thesaurus distribution and by this process will save considerable amount of disc storage space. These are the titles of documents which is part of our

mechanized document circulation and recall - which we can "marry" with our RANDOM ACCESS PROGRAM.

This hardware problem, hopefully, will be resolved when the Library will switch to the GE/600 system sometime in 1967. The cost factors, as to whether the GE/MSD Library will go all-out on random access, is a top management decision.

FUTURE PLANE AND PROGRAM

Current plans for the next two years include conversion of the NASA search tapes for the GE-235 computer program, and incorporation of this file on the random access equipment.

This would enable the GE-MSD Library to interrogate virtually the entire aerospace literature. As a recipient of NASA microfiche collection, the search tapes would result in a speedy and dynamic marriage of both bibliographic retrieval and the end product -- the needed data packaged and available to the Library's customer.

The open literature would be solved by subscribing to the AIAA microfiche collection, which utilizes the same indexing classification as NASA.

The "information explosion" need not be explosive, if we make optimal use and plan ahead, with the computer to solve retrieval and the microformat as the logical adjunct.

BIBLIOGRAPHY

Kodroff, B. and Johnson, D. M., "A Set Theoretical Approach to Mechanized Information Retrieval," General Electric Company Report 62SD130, Oct. 31 (1962).

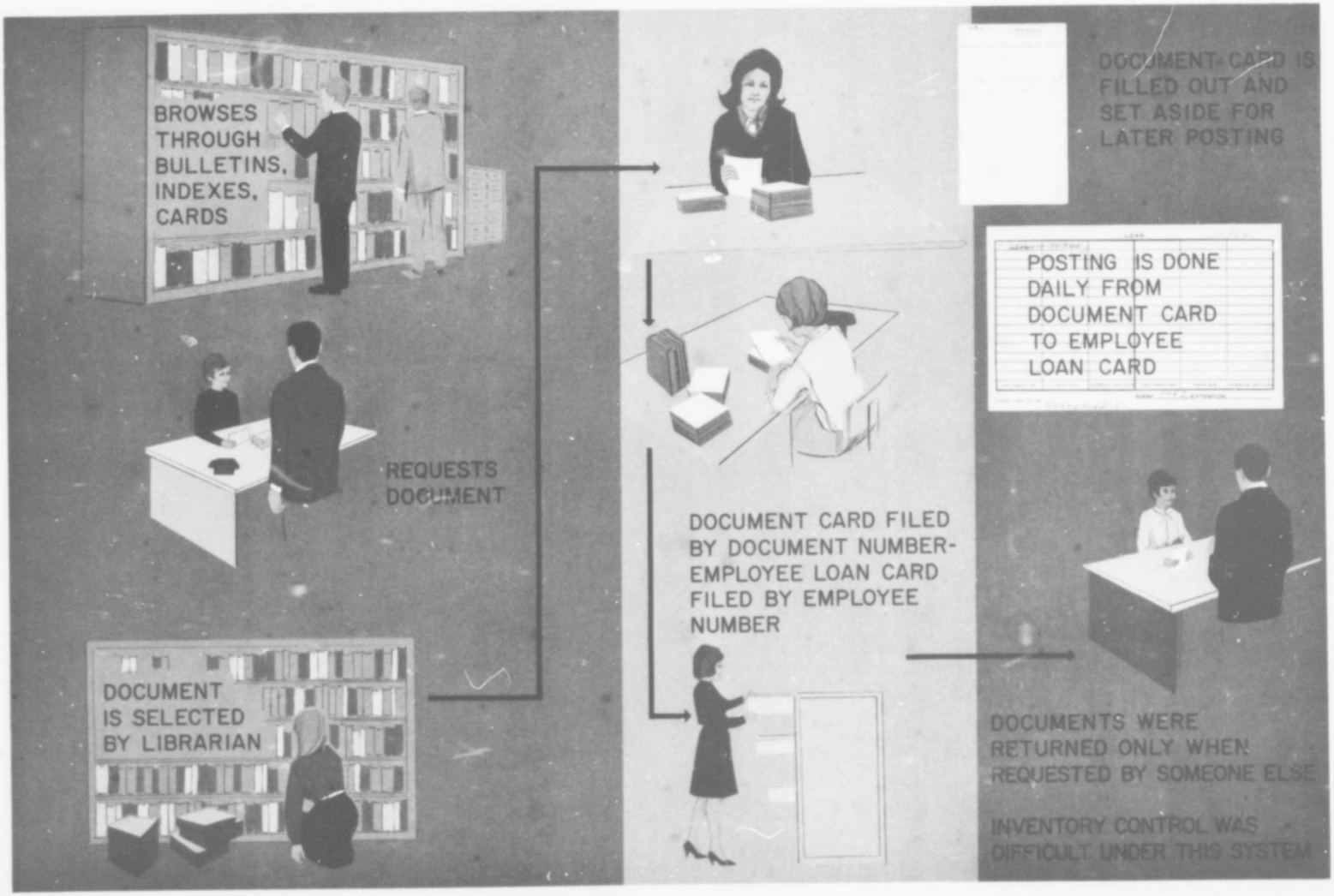
Chasen, L. I. and Kodroff, B., "The Automation of a Document Library," Bull. Sp. Libraries Council of Phila., 28, No. 3, February (1962).

Chasen, K. I., "Planning, Organizing and Implementing Mechanized Systems in a Space Technology Library," Am. Doc. Inst. Paper No. 117, Oct. (1963).

"Compatibles/200 Disc Storage Unit Subsystems," General Electric Company Computer Department Bulletin CPB-323, April (1964).

"DATANET-30 System and Operating Manual," General Electric Company Computer Department Bulletin CPB-289A, January (1964).

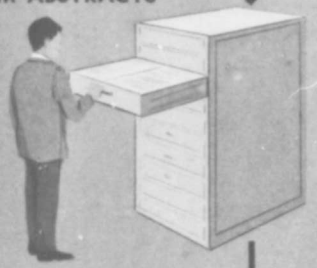
ORIGINAL MANUAL SYSTEM



MECHANIZED SYSTEM

CUSTOMER REQUESTS COMPUTER SEARCH

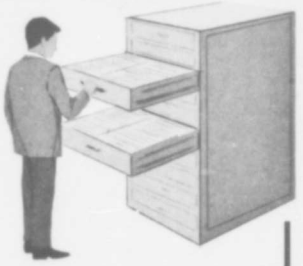
HE SELECTS DOCUMENT FROM ABSTRACTS



DOCUMENT IS SELECTED BY THE LIBRARIAN

TWO DUPLICATE LOAN CARDS ARE PUNCHED

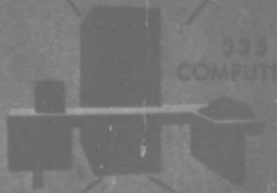
ONE CARD FILED BY DOC NUMBER
ONE CARD FILED BY EMPLOYEE NUMBER



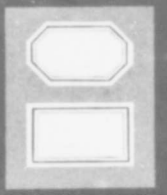
MONTHLY - OVERDUE DOCUMENT CARDS ARE SELECTED OUT BY COLLATOR. THESE ARE THEN MATCHED WITH TITLE FILE AND FORWARDED TO COMPUTER.

INPUT FROM LIBRARY

MASTER PERSONNEL FILE



OVERDUE NOTICES



MAILING LABELS

TO DOCUMENT HOLDERS

INFORMATION RETRIEVAL - SET THEORY

OPERATORS

+ PLUS (INTERSECTION)
/ OR (UNION)
− BUT NOT (NEGATION)
( ) INCLUSIVE

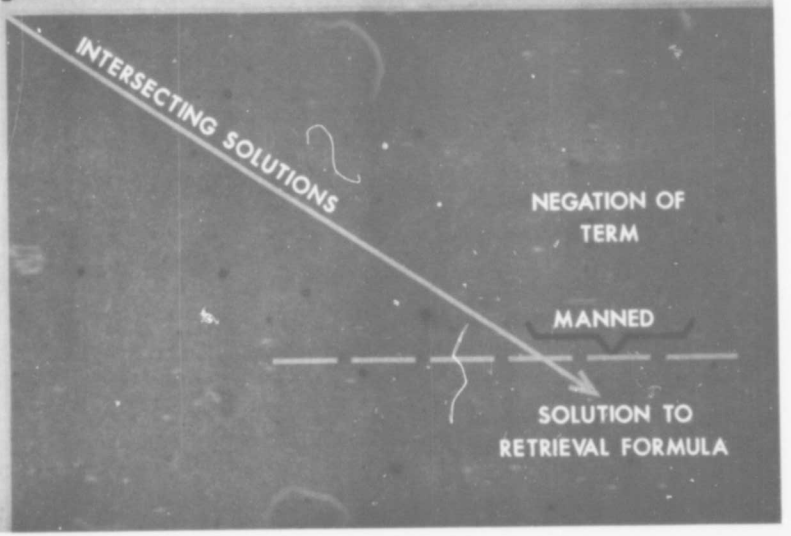
RETRIEVAL FORMULA

(SPACE * VEHICLES / SATELLITE * VEHICLES)+
(NAVIGATION / NAVIGATION * SYSTEMS / GUIDANCE / CONTROLS) − MANNED

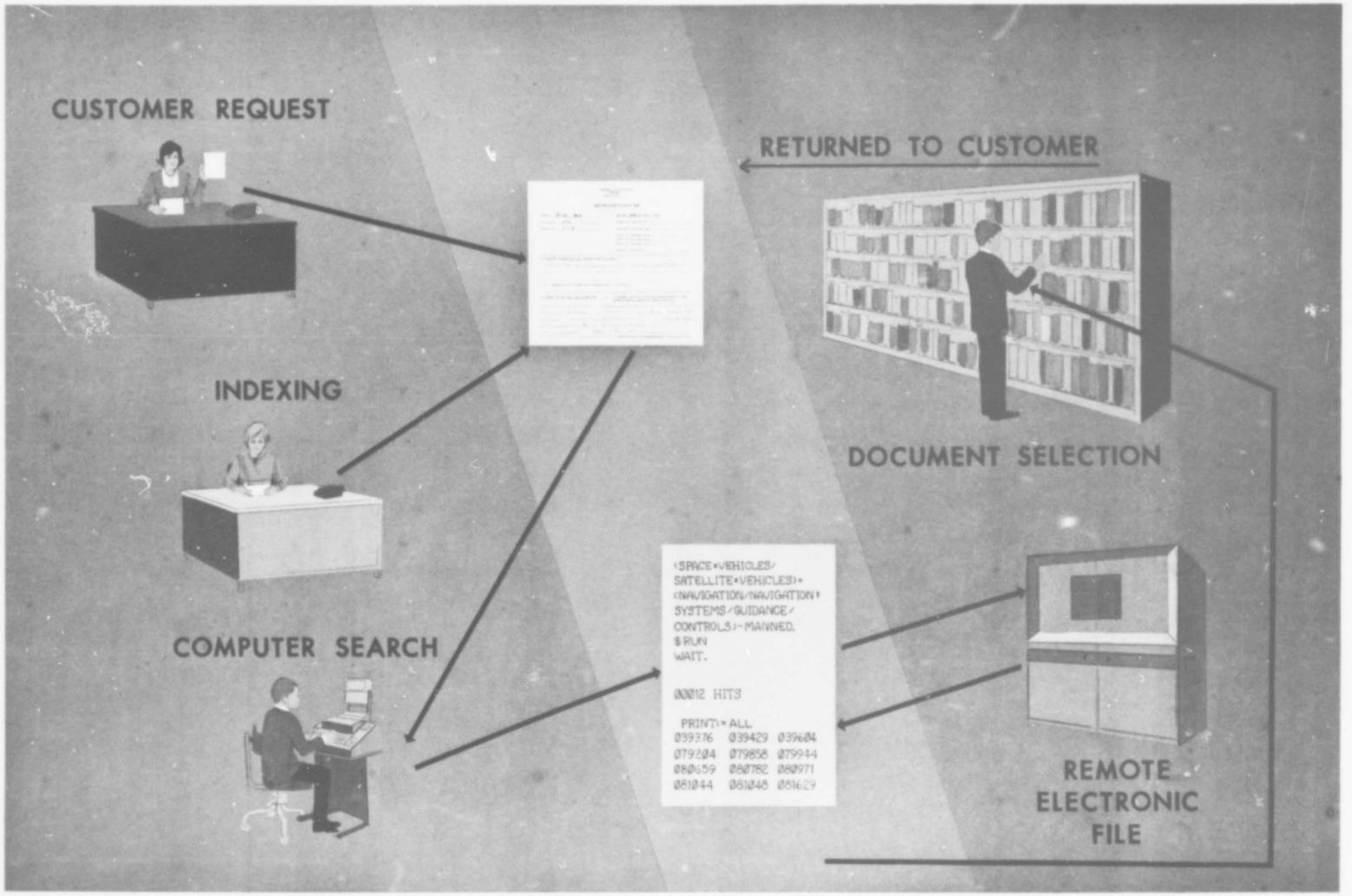
UNION SOLUTION
- NAVIGATION
- NAVIGATION * SYSTEMS
- GUIDANCE
- CONTROLS

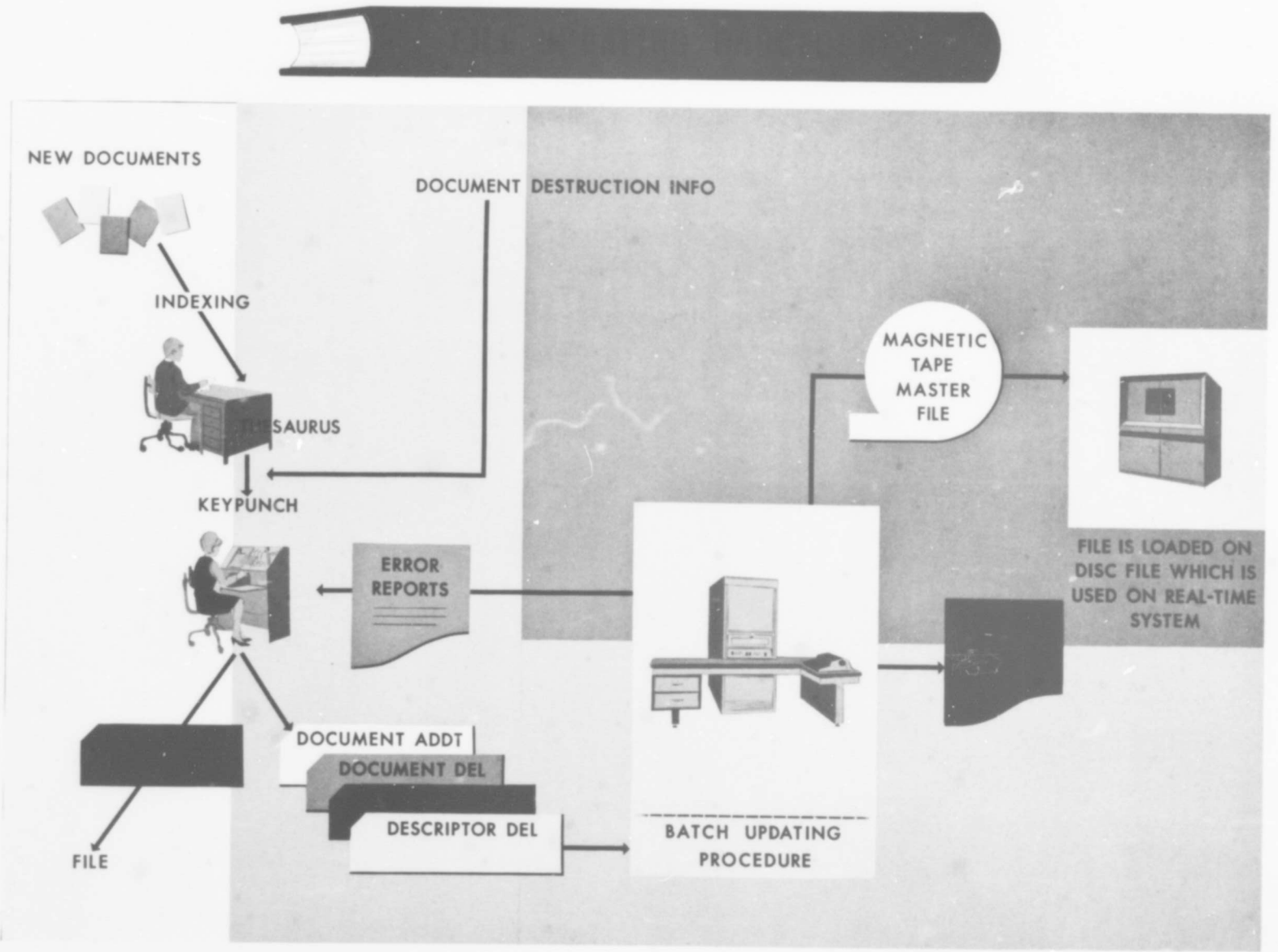
SPACE * VEHICLES
SATELLITE * VEHICLES

UNION SOLUTION



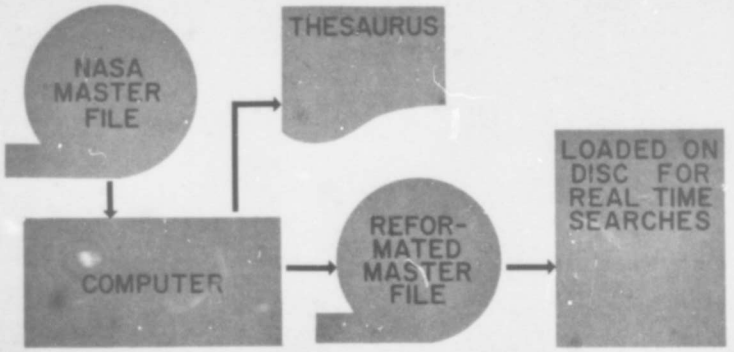
INFORMATION FLOW – CURRENT REAL TIME SYSTEM



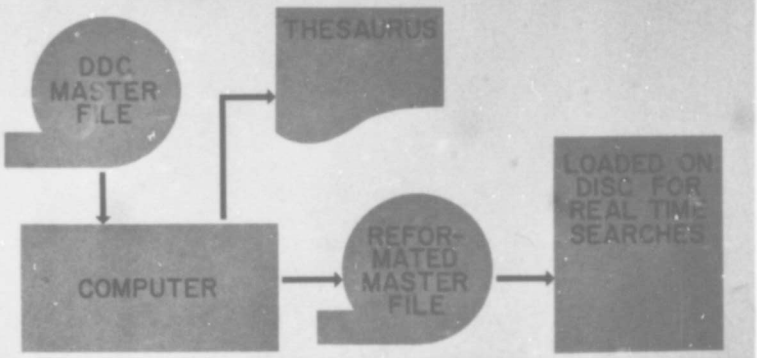

NEW DOCUMENTS
DOCUMENT DESTRUCTION INFO
INDEXING
THESAURUS
KEYPUNCH
MAGNETIC TAPE MASTER FILE
FILE IS LOADED ON DISC FILE WHICH IS USED ON REAL-TIME SYSTEM
ERROR REPORTS
DOCUMENT ADDT
DOCUMENT DEL
DESCRIPTOR DEL
BATCH UPDATING PROCEDURE
FILE

PROPOSED SYSTEM EXPANSION

NASA FILE



DDC FILE



TITLES

```
 $TDOC
READY.

(SPACE*VEHICLES/SATELLITE*VEHICLES)*(NAVIGATION/NAVIGATION*SYSTEMS/
GUIDANCE/CONTROLS)-MANNED.
$RUN     .
WAIT.

00012 HITS

 PRINT=ALL
039376 039429 039684 079284 079858 079944 080659 080782 080971 081044
081048 081629

 SEND ID
081048
ANALYSIS OF HOMING PHASE RENDEZVOUS MANEUVERS
SEND ID
081629
ADVANCED DATA CORRELATOR STUDY AND DEVELOPMENT
SEND ID
NO
```

SELECTIVE DOCUMENT DISSEMINATION