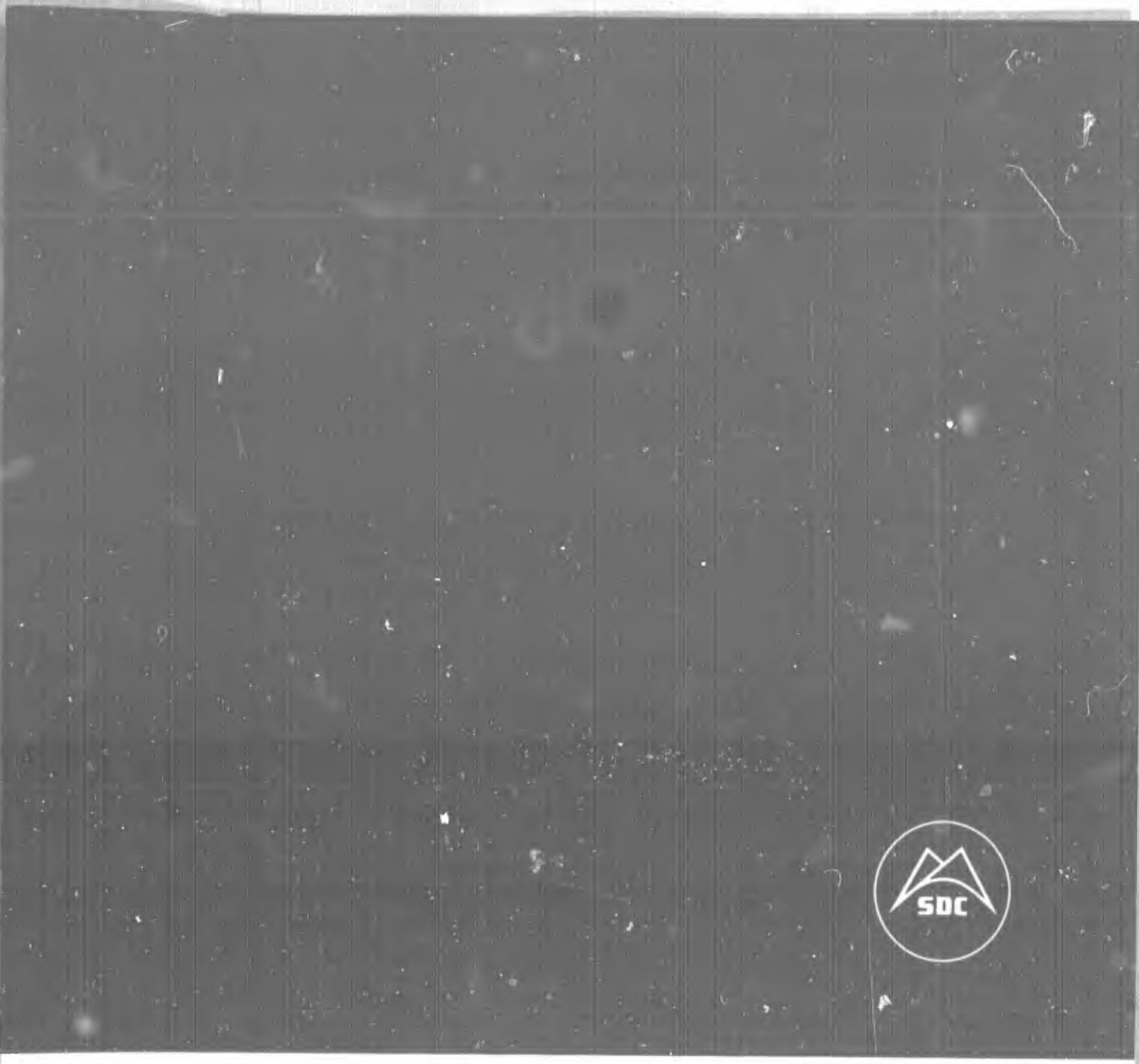
BEST POSSIBLE SCAN



SP-2557

Interactive Displays for Document Retrieval

H. Borko and H. P. Burnaugh

4 August 1966

28

SP-2557

SP *a professional paper*

Interactive Displays for Document Retrieval

by

H. Borko and H. P. Burnaugh

4 August 1966

SYSTEM DEVELOPMENT CORPORATION
2500 COLORADO AVE.
SANTA MONICA
CALIFORNIA
90406

This paper was prepared for submission to Information Display.

(page 2 blank)

ABSTRACT

Interactive computer systems establish a dialog between the man and the machine. By means of displays, the user receives immediate feedback of the results of his actions, and he is able to modify his decisions in order to obtain a system response that is most relevant to his needs.

BOLD (Bibliographic On-Line Display) is an example of a highly automated interactive document storage and retrieval system that is in operation at System Development Corporation. It enables the user to browse through the data base by subject categories or search for specific documents. The response is rapid and the results are satisfying.

1. THE PROBLEM

BOLD (an acronym for Bibliographic On-Line Display) was designed to be a highly automated and versatile document storage and retrieval system. The objective was not simply to automate, but through automation, to achieve a high degree of operational efficiency and user satisfaction.

The procedures followed in most existing retrieval systems are awkward and entail a considerable time delay. The user transmits his request to an information specialist who translates it into a search formula--a set of retrieval terms that are compatible with the index terms used in the information system. Retrieval is usually a two-step process in which document identification numbers are retrieved first and then the documents themselves are obtained. The results of the search are reviewed for reasonableness by the information specialist, and after a delay of hours or days, the material is sent to the requester. More often than not the user is annoyed by the delay he has encountered and is suspicious of the results achieved. He may receive some irrelevant documents and may not receive some relevant documents that he believes are in the system.

In contrast, a person gets a much greater feeling of satisfaction when he uses a traditional library. There, he may browse among the shelves selecting and examining books at random or in a given subject location. He may look through the card catalog under various subject headings seeking documents with promising titles. Finally, if he is not sure of the best way of searching for specific information, he may ask someone for help. It is important to note that a library search procedure is "interactive" in the sense that the user interacts directly with his data and receives immediate feedback on the results of his search. Unquestionably, this sense of direct contact contributes to the user's satisfaction and to the efficiency of the system.

The BOLD system was designed and programmed to provide the user of the computerized retrieval system with the same kinds of capabilities that he finds helpful when using a library. The desired characteristics are as follows:

(a) Language must not be a barrier. Ideally, the user should be able to state his requests in natural English, or failing this, the computer should help him express his needs in an acceptable language. Certainly, the system would fail, if the user first had to learn programming before he could retrieve information.

(b) The user must be provided with information about how documents are organized and classified within the system. He must be allowed to explore the content of these categories just as he would browse among the books on the shelves of a library or cards in a catalog file.

(c) When preparing a search request for specific material, the user must be helped by the system to word this request correctly just as he would be helped by a librarian in a manual system.

In essence, a computerized document storage and retrieval system must make it easy for the human user to interact with the machine in order to retrieve relevant material and screen out irrelevant material. BOLD has been designed with these characteristics in mind.

2. THE SYSTEM ENVIRONMENT

BOLD operates within the System Development Corporation (SDC) Research and Technology Laboratory. It is programmed for use with the IBM Q-32V computer and Time-Sharing System, a very large computer complex with a 64K word core memory and auxiliary storage on drums, discs, and tapes.

In practice, the user has no direct contact with the computer. He sits at one of more than a dozen user stations at SDC or at a remote installation. The inquiry station

consists of a teletypewriter and a cathode-ray-tube display scope with a light pen.

The programming system has two major modules: (1) the data base generator program and (2) the display and retrieval program. Both are written in the JOVIAL language with some of the subroutines in machine code (SCAMP). It is anticipated that these programs will be rewritten for the IBM 360/67 when this computer becomes available in the latter part of this fiscal year.

2.1 The Data Base Generator Subsystem

The BOLD data base generator builds tables of structured information from a Hollerith prestored magnetic tape. The tables are designed with extensive linking between entries referenced by identical key words or phrases to permit rapid retrieval. The data base that is presently being used was obtained from the Defense Documentation Center and consists of abstracts of approximately 6000 documents. For experimental purposes, a subset of these documents is used. The particular tape from which the illustrative examples were derived consists of the first 1745 abstracts and 6883 retrieval terms. The documents are grouped into subject categories organized according to the DDC classification system. However, the program is flexible, and various classification and indexing systems can be used.

The function of the data base generator subsystem is to process prestored data tapes and to establish the linking between the index terms and other descriptive attributes such as authors, titles, contract number, etc. The data tables produced by the data base generator are used by the retrieval program. A technical description of the data base generator and of the display and retrieval subsystems has been prepared by Howard Arnaugh [References 1 and 2], who wrote the programs.

2.2 The Display and Retrieval Subsystem

As the name implies, the display and retrieval subsystem is designed to search the data base that has been structured by the data base generator program and to retrieve and display the requested information. Although this is a difficult enough task in itself, the interactive features make the programs even more complex.

The retrieval program has the following characteristics:

(a) Allows, and therefore monitors, many users who are interrogating the same data base but who - doing simultaneous, independent, real-time ret.

(b) Permits the user to interrogate the dictionary and find out whether a word has been used as an index term--and how often--and whether there are synonyms or related terms.

(c) Allows the user to formulate his requests in natural language with very few computer-oriented restrictions.

(d) Displays retrieved information on a teletype or CRT console.

(e) Copies the retrieved information on a magnetic tape for off-line listing.

The utility and versatility of these programs will be explained by describing the operations of the BOLD system.

3. THE INTERACTIVE ENVIRONMENT

BOLD is an interactive system, which means that a dialog is established between the user and the system so as to enable the user to request and obtain relevant documents from the collection. The requests and the system's responses are stated in as close an approximation to natural language as is possible. Ideally, the user with only a knowledge of the English language and a skill in typing should be able to establish a rapport with the machine. Although this ideal may never be fully achieved, a great deal of human engineering skill has gone into the project so that it could be approximated.

3.1 Interactive Displays and the Browse Mode

The BOLD inquiry station consists of a teletypewriter and cathode-ray-tube display unit (See Figure 1). The user interacts with the system and requests information by

Figure 1. BOLD Inquiry Station

typing on the teletypewriter. He may also make certain requests by using the light pen and the display scope.

After he has logged in and the data base and program tapes are loaded, the system reports this fact by typing

THIS STATION IS NOW UNDER THE CONTROL OF THE BOLD SYSTEM
OPERATION INSTRUCTIONS R OBTAINED BY THE REQUEST:
INSTRUCTIONS/

Simultaneously the display shown in Figure 2 will appear on the scope. This display defines the ten light-pen actions that are available to the user.

The user begins by flashing the "B" character with the light pen or typing BEGIN/ on the teletypewriter. Commands such as BEGIN, SEARCH, BROWSE, CONTINUE, etc., must be followed by a slash. One uses a question mark to ask for help, and for all other interactions no punctuation marks are used.

When the slash is not used and the requester types

BEGIN

the system responds

*NOT FOUND

The system is interpreting the word BEGIN as an index term. Since such a term has not been used, the system responds with *NOT FOUND. The user, recognizing his error, tries again, this time by typing

BEGIN/

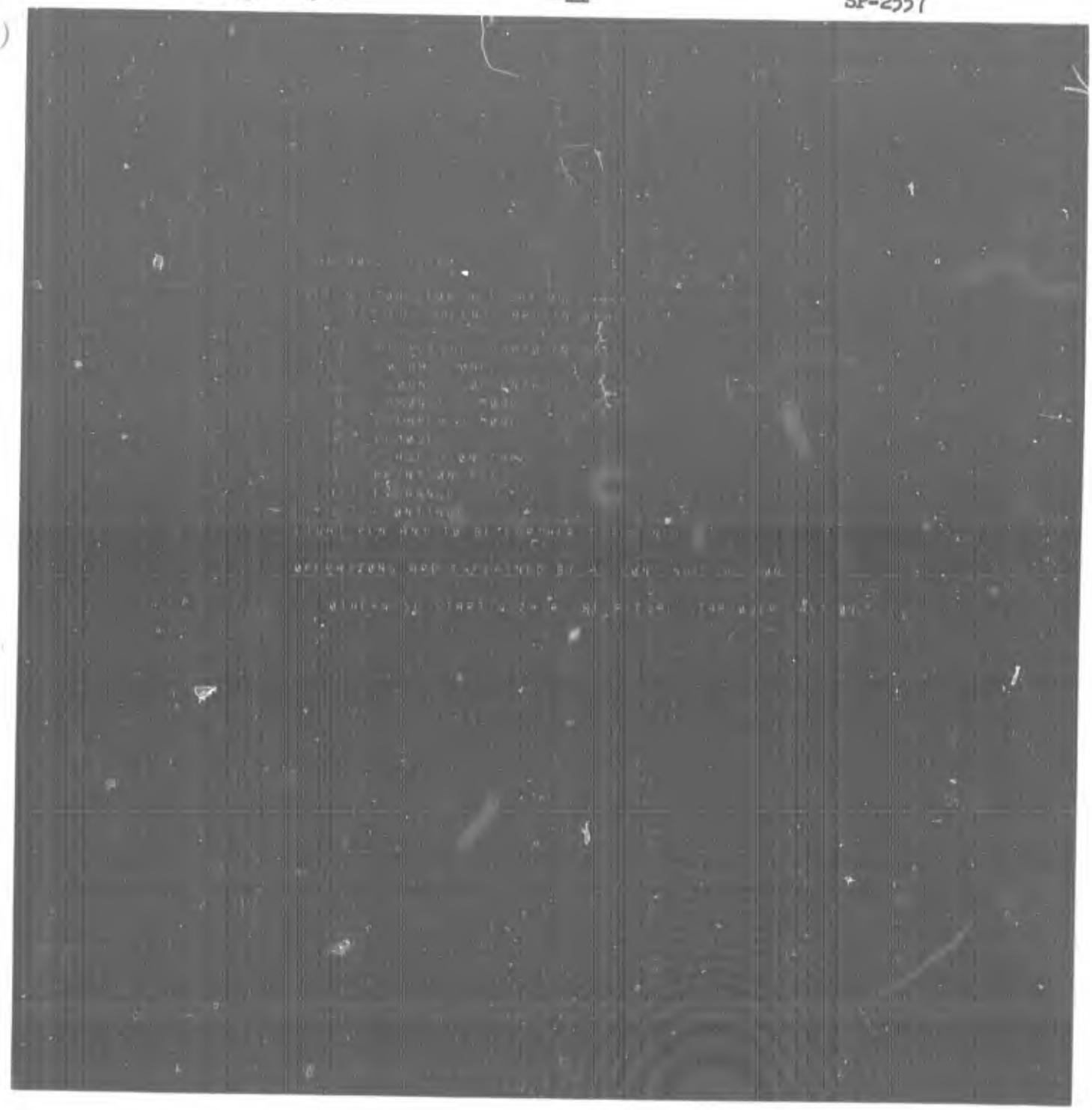


Figure 2. Initial Tutoring Display

This command is accepted (signified by //), and a new display appears (Figure 3) that indicates the 32 divisions or main subject categories into which the data are divided. If the user wishes a further breakdown, he may use his light pen to flash a division. By doing so, he is requesting more information about that category and receives a display of the subdivisions (Figure 4). The display also supplies the number of entries in the category. If he chooses to browse through the items in this category, he may do so by flashing the □ character with his light pen or by typing BROWSE/ on the teletypewriter. The display he would receive is the first abstract in that category (Figure 5).

Usually, when dealing with a large collection of documents, one is not likely to want to browse through all the documents in a category. However, it is conceivable, especially when the document collection is organized in a hierarchical classification schedule such as the Dewey Decimal System, that browsing by category will be useful; this capability is therefore provided.

3.2 Dictionary Interrogation and the Search Mode

A more common method of seeking information is to request documents by subject headings or index terms. Many information centers use a form of coordinate indexing and retrieve information by combining a number of index

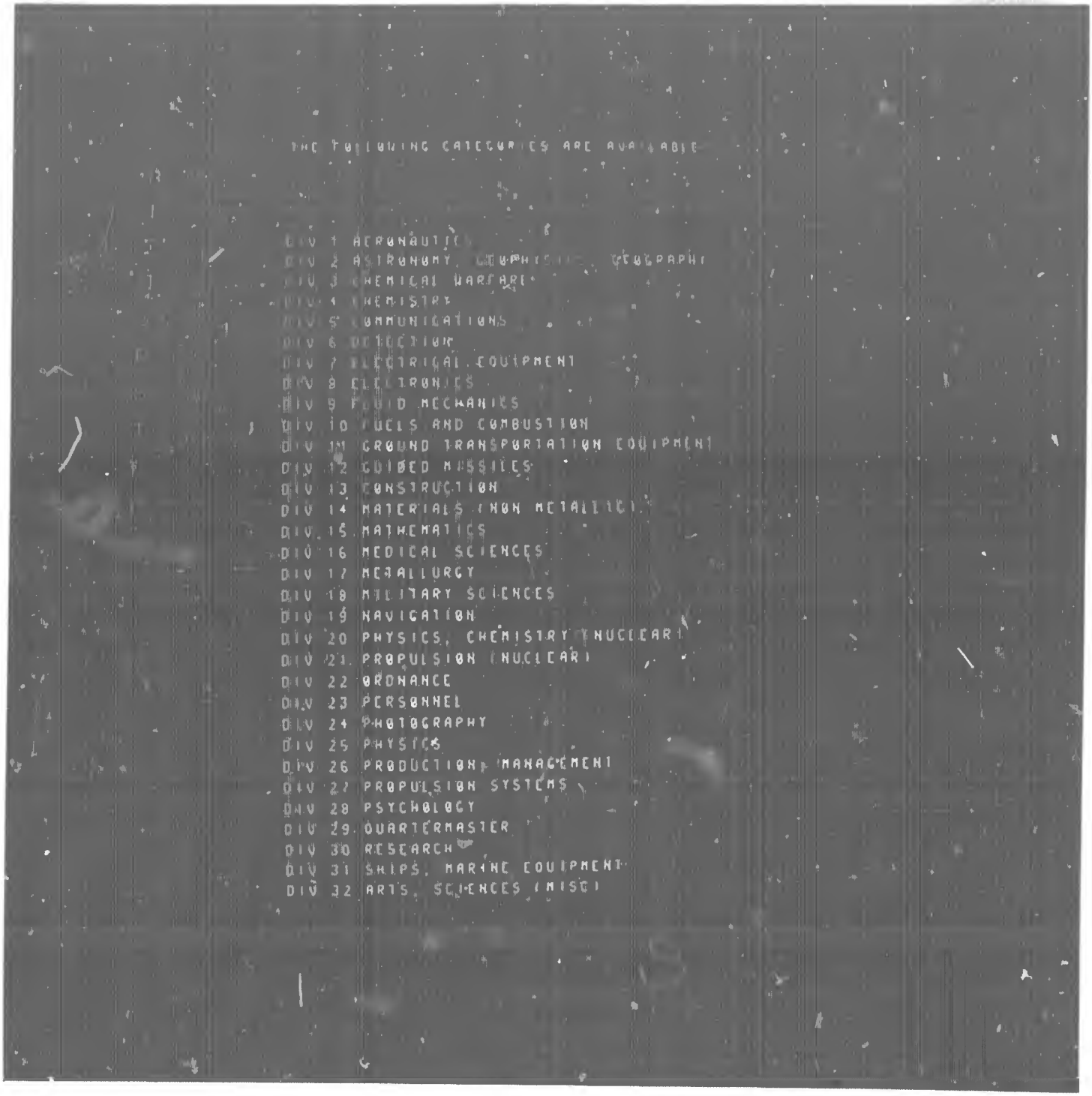


Figure 3. Classification Categories

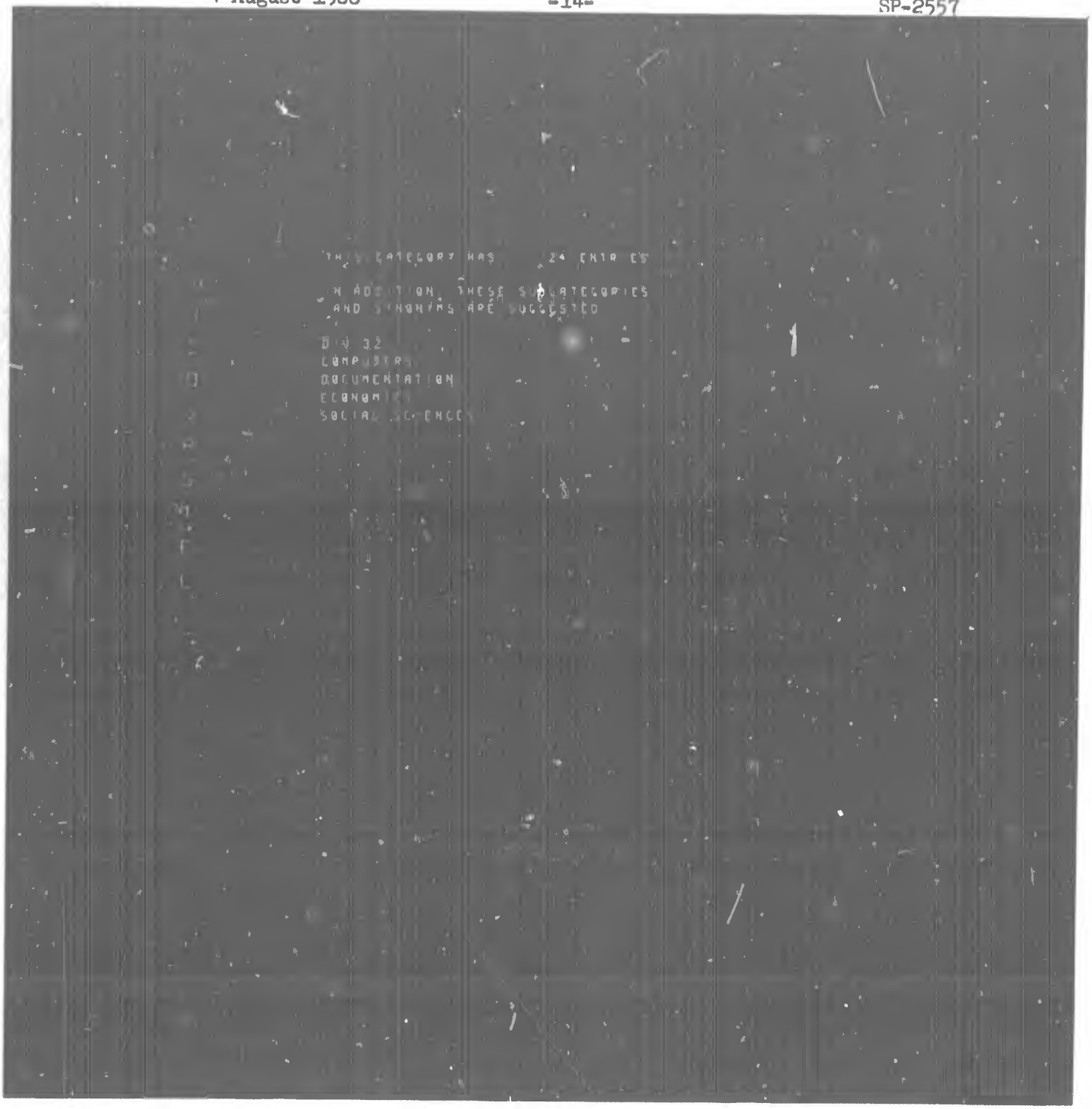


Figure 4. Subdivisions of Category 32

Figure 5. An Abstract Displayed on SCOPE

terms to form a specific request. Usually a trained information specialist must help the user formulate his request for information into a search request made up of approved index terms. In an interactive system, the user requests help by interrogating the dictionary.

By way of illustration, let us suppose the user is doing research in the field of space travel. He is preparing a report on this subject and he wishes to search the collection for relevant articles. He sits at the inquiry station, and after the system is in operation he begins by interrogating the dictionary to determine which words can be used as index terms for retrieval purposes.

The following dialog takes place:

```
SPACESHIPS?
THESE MAY BE RELATED TO SPACESHIPS
SPACESHIP CABINS
SPACESHIPS
SPACESHIPS - POWER SUPPLIES
SPACESHIPS - STABILITY
*END
SPACE?
THESE MAY BE RELATED TO SPACE
SPACE CAPSULES
SPACE CHARGES
SPACE ENVIRONMENTAL CONDITIONS
SPACE FLIGHT
SPACE FLIGHT - CONTROL
SPACE FLIGHT - SURVIVAL
SPACE MEDICINE
*CONTINUE?YES
SPACE MEDICINE - EFFECTIVENESS
SPACE NAVIGATION
SPACE PERCEPTION
SPACE PROBES
SPACE RECOVERY SYSTEMS, INC., EL SEGUNDO, CALIF.
SPACE SCIENCES LAB., GENERAL ELECTRIC CO.,  PHILADELPHIA, PA.
SPACE SHIPS
*CONTINUE?NO
LUNAR FLIGHTS?
*NOT FOUND
MOON FLIGHTS?
*NOT FOUND
MARS FLIGHTS?
*NOT FOUND
MOON?
THESE MAY BE RELATED TO MOON
MOON
MOON - ATMOSPHERE
*END
LUNAR?
THESE MAY BE RELATED TO LUNAR
LUNAR PROBES
*END
MARS?
THESE MAY BE RELATED TO MARS
MARS
MARSH CHARLES A.
MARSHALL JOHN M.
*END
```

He begins by asking whether SPACESHIPS is an index term. He types in the word followed by a question mark. The system responds that in addition to SPACESHIPS there are a number of other similar terms that are also usable index words. The system finds these related terms by dividing the query word in half and locating all index terms that start with the same combination of letters.

The user, now recognizing that the term SPACESHIPS might be too specific, asks for information about the more general term SPACE. Again the system responds with similar terms. Note that the word SPACE by itself is not an index term for it is always used in combination with another word. In response to a dictionary inquiry, the system types seven index terms and then asks the user whether he wishes it to continue. After two such inquiries, the user feels he has enough information on this subject and tries some other terms. Some of the words he tries are not index terms, but in his interaction he finds enough that are.

As a result of this dialog, and with the information he has obtained, he is now in a position to formulate a search request. He selects six terms and formulates these into a search request by indicating that he would like to have displayed the list of document numbers that contains any one of these six terms; that is, he combines these

terms by means of an OR rather than an AND logic, although both the AND and NOT logic are also available.

He makes his requests as follows:

```
SPACESHIPS OR LUNAR PROBES OR MOON OR MARS
      25 ENTRIES ARE REF'D BY      SPACESHIPS
       6 ENTRIES ARE REF'D BY      LUNAR PROBES
      13 ENTRIES ARE REF'D BY      MOON
       3 ENTRIES ARE REF'D BY      MARS
*END
SPACE FLIGHT OR SPACE PROBES
      15 ENTRIES ARE REF'D BY      SPACE FLIGHT
       8 ENTRIES ARE REF'D BY      SPACE PROBES
*END
SEARCH/
      51 ENTRIES
```

Note that when the user types a request, as distinct from interrogating the dictionary, he does not use a question mark. The system tells him how many entries in this data base (1745 abstracts) are referenced by each term.

He now orders the system to

SEARCH/

and the system responds that there are

51 ENTRIES

Since there is a total of 70 documents which have been indexed by these six terms, it is clear that some documents were indexed by more than one.

The system locates these 51 documents and displays the list by identification number and index term. The display appears on the scope (Figure 6). Note that all the documents are not able to be displayed at one time. Of the 51 entries only 37 have been searched. The user may now remove

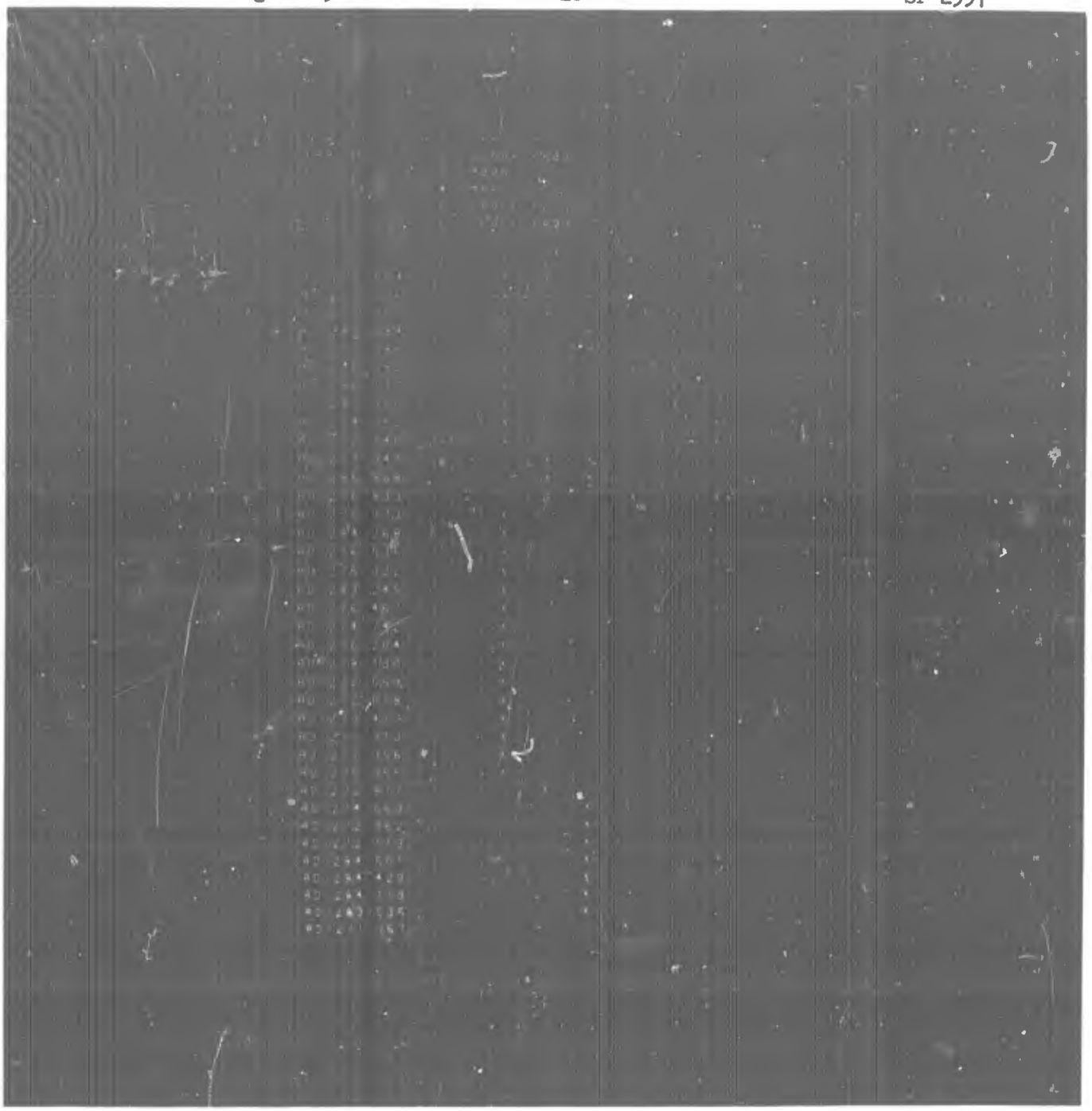


Figure 6. Search Matrix of Retrieved Documents

references to the documents that are of lesser interest. He does this by light-penning "R" and the document number. By light-penning the "C", or continue character, new document references will be displayed. He may also reorder the arrangement of the display by light-penning the "E" character and two document numbers that he wishes to exchange.

Let us suppose that in this case the user decides that he is more interested in space flight to Mars than he is in lunar probes. Probably document # AD-276082, which contains the terms SPACESHIP, MARS, and SPACE FLIGHT, is of most interest. He moves it to the head of the list by exchanging its location with the document that was previously first.

Should the user decide to keep a permanent copy of the display for future reference, he types

TYPE DISPLAY/

and the information is transferred from the scope to the teletype. This information would be identical to Figure 6 except for the subsequent exchange that has been made.

Before requesting copies of the 51 documents that have been indexed by one or more of the six retrieval terms, the user would like to have more information about their contents. He may obtain this information by simply typing BROWSE/ or light-penning the appropriate BROWSE symbol.

The system responds to the BROWSE command with

INPUT ATTRIBUTES WANTED

This is the machine's way of asking whether the user wishes to have displayed the title of the article, the author, the index terms, or the complete abstract. If he doesn't know what choices he has, he types

LIST ATTRIBUTES/

and the system responds with

ALPHA
CORP-AUTHOR
AUTHOR
CONTRACT
XTERM
DIVISION
DATE
TITLE
ABSTRACT
*END

In this instance, the user wishes to see all available information and so he selects ABSTRACT. The system now displays on the scope all available information on each of the documents in descending order. The first abstract to appear on the scope is of document # AD-276082 (Figure 7). Note that the abstract is not complete because of lack of room. The rest of the abstract can be obtained by light-penning the "C" character.

Should an immediate permanent record be wanted, it can be obtained by the command

TYPE DISPLAY/

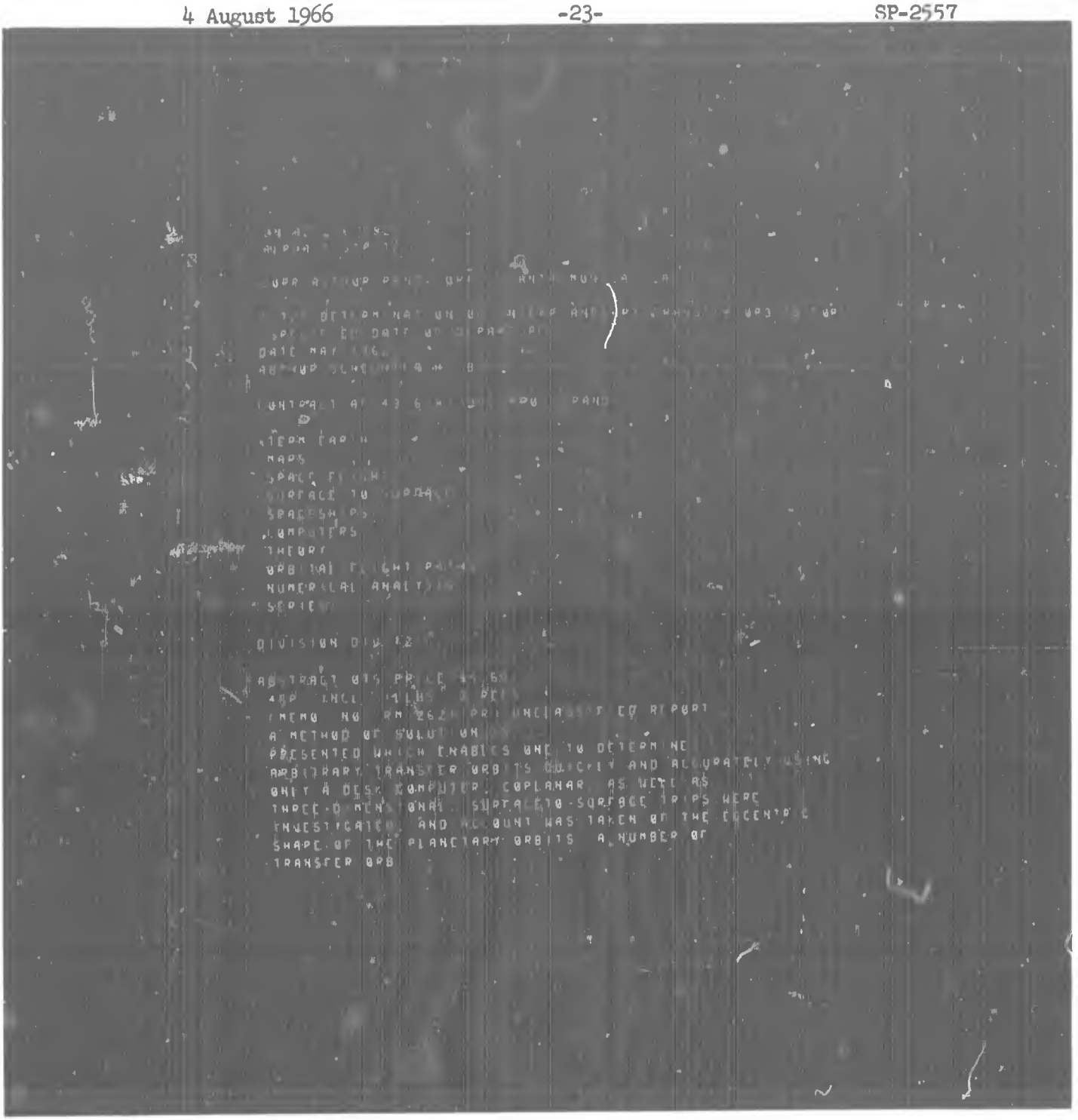


Figure 7. Viewing the Retrieved Abstract

In this manner the user can browse through the entire set of 51 entries that have been retrieved in response to his request or that subset of documents that he has not removed from the display. He may save any information that appears for future reference.

4. USER SATISFACTION

The BOLD information storage and retrieval system enables the user to formulate and modify his requests. He may search through the store of available information. He can, on the spot, determine which documents are relevant and which are not. After seeing the abstracts, he can determine and record the number of those documents that he wishes to read in their entirety. He interacts with the system, and when he leaves the inquiry station he leaves with the feeling that he has obtained most of the relevant material that the system has in store. The response has been rapid, and the experience has been a satisfying one.

REFERENCES

1. Burnaugh, H. P. Data Base Generator for the BOLD System, SDC Document TM-2306/001/01, January 1966.

2. Burnaugh, H. P. Retrieval Program for the BOLD System, SDC Document TM-2306/002/00, January 1966.

BLANK PAGE

Unclassified
Security Classification

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)
System Development Corporation
Santa Monica, California

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
Interactive Displays for Document Retrieval

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

5. AUTHOR(S) (Last name, first name, initial)
Borko, H., Burnaugh, H. P.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
4 August 1966	25	2

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
Independent Research	SP-2557
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. AVAILABILITY/LIMITATION NOTICES
Distribution of this document is unlimited

11 SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT Interactive computer systems establish a dialog between the man and the machine, by means of displays, the user receives immediate feedback of the results of his actions, and he is able to modify his decisions in order to obtain a system response that is most relevant to his needs. BOLD (Bibliographic On-Line Display) is an example of a highly automated interactive document storage and retrieval system that is in operation at System Development Corporation. It enables the user to browse through the data base by subject categories or search for specific documents. The response is rapid and the results are satisfying.

DD FORM 1 JAN 64 1473

Unclassified
Security Classification

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Document Retrieval Interactive Computer Systems Displays BOLD Bibliographic On-Line Display Information Storage & Retrieval						

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.

2a. 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.

2b. 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.

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

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

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

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

9a. 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.

9b. 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.

Unclassified

Security Classification