SP-2338/000/01

SP *a professional paper*

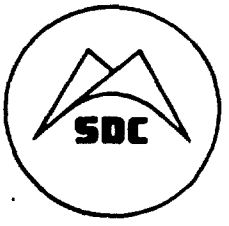
The BOLD (Bibliographic On-Line Display) System

by

Howard P. Burnaugh

6 April 1966

SYSTEM DEVELOPMENT CORPORATION
2500 COLORADO AVE.
SANTA MONICA
CALIFORNIA
90406



The BOLD (Bibliographic On-Line Display) System[1]

ABSTRACT

The BOLD (Bibliographic On-Line Display) system serves as a general purpose vehicle for research on the components of a real-time retrieval system. Specific subjects for investigation are indexing, classification and categorizing schemes, file organization, and user-system communication. The program operates in a "time-sharing" environment doing independent retrieval for multiple simultaneous users. A retrieval station may be any teletype connected to the time-sharing system. A station may be augmented with a CRT console and a light pen for rapid displaying of the retrieval information.

Retrieval is effected by the specification of categories and/or retrieval phrases, using Boolean connectors. There are two modes for retrieval operation: the Browse mode and the Search mode. In the Browse mode the user may specify broad categories and retrieval terms and then browse through the retrieved information entry by entry. The user designates what information is to be returned. This may be anything that is defined in the data base, and may range from a single component (such as author, title, etc., for a bibliographic data set) to a complete body of text (i.e., abstract).

---

[1] Paper to be presented at the Third Annual National Colloquium on Information Retrieval, to be held May 12 and 13, 1966, at the University of Pennsylvania, Philadelphia, Pa.

The Search mode is used to obtain a list of entries, any of which is referenced by one or more of the retrieval terms designated. The retrieved set of entries is sorted and presented as a matrix, correlating the entries (by name or number identification) with the set of retrieval phrases. The sort causes those entries referenced by the greatest number of retrieval designators to be listed first. With the light pen, entries may be removed from the CRT display and/or transferred to a listing tape for report generation.

The program permits an approach to natural language in the indexing and retrieval queries. There is extensive capability for interrogation and modification of the dictionary by the user. The modification techniques allow hierarchical structures to be specified and changed (in addition to adding, deleting and changing dictionary entries).

The data components for an entry are defined for each data file. Therefore, although the system was designed for library retrieval, the programs are capable of handling a wide range of non-bibliographic data sets. The system is being checked with a data set of over 6000 documents, indexed and abstracted.

Introduction

A retrieval system has the basic objective of getting those pieces of stored information that best satisfy a retrieval request. To accomplish this objective merely requires the combining of various functions (such as input interpretation, searching, etc.) into a program. But the selection of the functions to be included and the capabilities allotted to each function result in quite different retrieval systems. The functions comprising the BOLD system (Bibliographic On-Line Display) were selected to produce a system which not only is a usable retrieval system but also affords a means for research on the components of such a system.

BOLD is one of the studies of the Language Processing and Retrieval Staff at System Development Corporation. The projects under study in this group fill a spectrum from studying sentence grammar (machine parsing), through machine indexing and abstracting, to information retrieval (BOLD). This shows BOLD as a part of an evolving system for complete machine processing of English text.

Several factors determined the design of the BOLD system. One was the philosophy for the retrieval function. The retrieval would be in one of two modes for the user. In the browsing option the program would return the requested information for a retrieval set established by categories and/or indexing terms designated by the user. The user would then "browse" through each returned entry of the set. In the search mode option the indexing terms are designated with each entry to aid the user in selecting entries.

Another design determinant was to have the capability for the program to serve as a system for research. One of the building blocks for a retrieval program is some concept of indexing the data. In order to study the retrieval effectiveness of indexing, BOLD would need the capability for manipulating entries in the dictionary by suppressing, adding, changing, and deleting indexing terms. Also, its use of category designators would have to be flexible. The components under study include the structure of the data base, the use of natural language, the use of a biased dictionary, and the comparative retrieval effectiveness of various indexing and classification systems. In addition to the retrieval capabilities and research possibilities, the system would have to allow for, and take advantage of, a "time-shared" environment.

The following design features governed the combining of selected functions into the BOLD system:

1. It operates in a time-shared environment.
2. It is general purpose in the type of data to be accommodated, with no restriction on the quantity.
3. It monitors multiple users under its control who are doing simultaneous, independent, real-time retrieval.
4. It aids the "lay-user" in developing the retrieval request by allowing extensive interrogation of the dictionary for synonyms and related indexing terms (either by spelling or hierarcy).
5. It permits natural-language queries of the dictionary.

6. It retrieves according to two operational modes, Search and Browse.
7. It presents the retrieved information on a teletype or a CRT display console.
8. It produces copies of the retrieved information on the teletype or on a magnetic tape for off-line listing.

Descriptions of these features will show how they have been incorporated into the BOLD system and how they can be used to effect retrieval.

The System Environment

The system uses the computer facilities of the Research and Technology Laboratory at SDC. This includes teletypes, CRT display consoles, and light pens connected to the SDC Time-Sharing system for the IBM Q-32V computer. This is a large-scale computer with a 64K word core memory and auxiliary storage on drums and disc. The system is written in the JOVIAL language with some of the subroutines in machine code (SCAMP).

The Data Base

For efficiency, the BOLD system was split into two programs--the data base generator and the retrieval program. The original concept was as a library retrieval program with a data base of millions of documents. As the design evolved it became possible to make the program general purpose by converting the list of alphabetical retrieval terms into a dictionary, i.e., entries are defined by other entries. Several processing functions specifically for handling documents are included, as specified below.

Information to be processed is organized into sequential data entries (where, for a document retrieval system, a data entry is an abstract, with title, author, retrieval terms, etc.). Included in each data entry are all terms by which the entry is to be referenced. These are defined by first designating "descriptors" for the entry. A descriptor name must have an asterisk as its first letter and is required to be one word, where a space designates an end to a word. Descriptors for a document data-entry might be *author, *publisher, *retrieval-terms, etc. The hyphen in *retrieval-terms is to form "*retrieval terms" into one word. Retrieval designators follow each descriptor name, with each designator followed by a slash. The last retrieval designator for a descriptor is followed by two slashes. Each data entry is concluded with the limiter "*end /".

Several descriptors are specifically defined in the program to handle document data. These are *title, *dn (16-character-maximum accession number, unique for each entry), and *abstract. Each is followed by the described information and concluded by two slashes.

An example[1] of a document data entry is as follows:

*dn AD-276 285 //

*corp-author General Electric Co., Cincinnati, Ohio //

---

[1] All examples are from a data base of 1121 document abstracts obtained from ASTIA. The descriptor names are in some cases pseudonyms, to contract the descriptor name into one word, i.e., xterm for indexing term.

```
*title Research on the effects of hydrocarbon fuels on the operation
   of vapor phase fuel systems //
*date 1 Oct 1961 //
*author Colley W. C. / Kutzko G. G. //
*contract AF 33(616)8224 //
*xterm jet engines - cooling / jet engine fuels / fuel systems /
   heat transfer / heat exchanger / coolants / hydrocarbons /
   vapors / enthalpy / decomposition / test methods / test equipment //
*division div 27 / div 10 / div 25 / div 30 //
*abstract (abstract text) //
*end /
```

As the table of retrieval terms is compiled, each entry generates a list of pointers to the stored information referenced by that entry. The program does not produce a sequential list of pointers; instead, it creates closed loops for each retrieval term. The loops are embedded in the data entries.

Each retrieval term in the dictionary has an associated address that locates a data entry referenced by that term. Within the data entry is the address of the next data entry referenced by the term. This linking eventually closes on the first data entry to form a closed loop.

Several factors prompted this design for the tables. The merits and faults of tables with such extensive associative linking are being

evaluated. Since the data are not totally indexed, this structure is feasible. Once the data are processed in this manner, it is easy to convert the loops to the more conventional lists (if desired).

The User-System Interface

One of the design features allows the program to function with multiple users who are doing simultaneous, independent retrieval. The BOLD user can interrogate the dictionary to ascertain relevant retrieval terms. The retrieval is on-line, in real-time, which allows the user to accept and reject alternatives and to modify, or restart, a request at any time. During these interactions in which the user establishes a retrieval request, the program functions faster than the user. This enables the system to monitor multiple users without affecting the response time.

A user's station consists of a teletype connected to the SDC Time-Sharing system. This may be augmented with a CRT display console and light pen. The display console affords a more rapid means for presenting requested information than the teletype.

Stations may be added and dropped from the system at any time by directives, which may be from any currently tied station:

JOIN A AND SCOPE A'

DROP B

(where A and B are teletype station numbers and A' is a display console number).

When a station is added without a display console the program translates the displays to the teletype. Only the teletype number is designated when a station is removed from the system; any associated display console is automatically dropped.

Dictionary Interrogation

The dictionary is the means by which the user is able to transpose his request into the language used to index the data. The BOLD dictionary uses the basic entity of a "descriptor" as explained in describing the data base. All dictionary entries, other than descriptors, are defined for these descriptors. Each entry can designate an equivalent entry (synonym), a subordinate entry, and a superordinate entry. The hierarchical relationships may be specified and modified at any time, as explained under "Other System Features and Miscellaneous Actions," page 21.

In addition, each entry can have an associated comment. This may be any message--such as a referral statement, a clarification, etc. The design of the dictionary, as to the means and extent for creating entry associations, is one of the areas for extensive study.

A request for information from the dictionary must be preceded by a period. The program searches the dictionary for all root forms of the word. These are typed back with the count of the referenced items for each term. In the following examples capitalized words are designations by the user and the remainder of the content of an example is the response from the program.

```
.HEAT
     6 entries are ref'd by heat
     1 entries are ref'd by heaters
     2 entries are ref'd by heating
*end
```

When a single term is found, the program automatically includes the term as a retrieval designator. When more than one form for a retrieval term is found, as above, the program does not include any of the forms for retrieval. A particular form of a word is designated by following the word with a colon.

```
.HEATERS:
     1 entries are ref'd by heaters
*end
```

Another instance that will cause multiple references for a term, and thus require qualification, is when two entries are the same but define different descriptors. A hypothetical example for a library system might be:

```
.WILEY, JOHN
     1000 entries are ref'd by Wiley, John  *publisher
        2 entries are ref'd by Wiley, John  *author
*end
```

The qualification would be:

```
.*AUTHOR = WILEY, JOHN
     2 entries are ref'd by Wiley, John
*end
```

The dictionary is examined for alternatives by following the input phrase with a question mark. The program communicates back all entries that have similar spelling or that are defined as synonyms or as superordinate to the input phrase. These are put on the display scope or on the teletype according to a D or T following the question mark. If neither is specified, the teletype is used. The degree of comparison for near-spellings is qualified by a C following the question mark with the number of characters for comparison. If no comparison qualifier is specified, the program

requires comparison through one half of the input term,but a minimum of the first four characters.

.HEAT?

The following may be similar to heat

heat
thermodynamics
enthalpy
heat exchangers
heat of formation
heat of fusion
heat of sublimation
heat production
heat resistant alloys
heat resistant polymers
heat transfer
heat treatment
heaters
heating
*end

In this example, thermodynamics and enthalpy are defined as a superordinate word and a synonym, respectively. The rest qualify because of spelling. When the terms are put on the teletype, the program asks "continue?" after every sixth term. A yes response causes the listing to continue. Any other response terminates the listing and effects the new request.

Retrieval terms are separated with the Boolean connectors: and, or, and not. Absence of a connector results in an or relationship. Examples of possible arrangements are the following:

.$term_1$
.$term_1$ or $term_2$ or $term_3$
.$term_1$ and $term_2$
.not $term_1$
.$term_1$ and not $term_2$
.and $term_1$

The last example ands $term_1$ with all previous terms.

The count of entries returned by the program is for the complete data base. At any time the request COUNT will cause the program to relate back the exact number of entries established, taking into account all terms and the Boolean connectors.

Retrieval

The initial presentation to the user is a display (see Figure 1) providing the data have been categorized. Categories may be established by the data base generator or through the retrieval program. Categories are treated as other retrieval terms except that when a category is designated, retrieval is restricted to entries in the category.

Action buttons are available as a column of characters for light-pen activation. The characters are for the following actions:

B retrieval startover

↑ return to last higher category, or go back one entry

Σ count of current retrieval set

□ Browse mode

◇ Search mode

D delete

S save on tape

T print on teletype

R restore

C continue

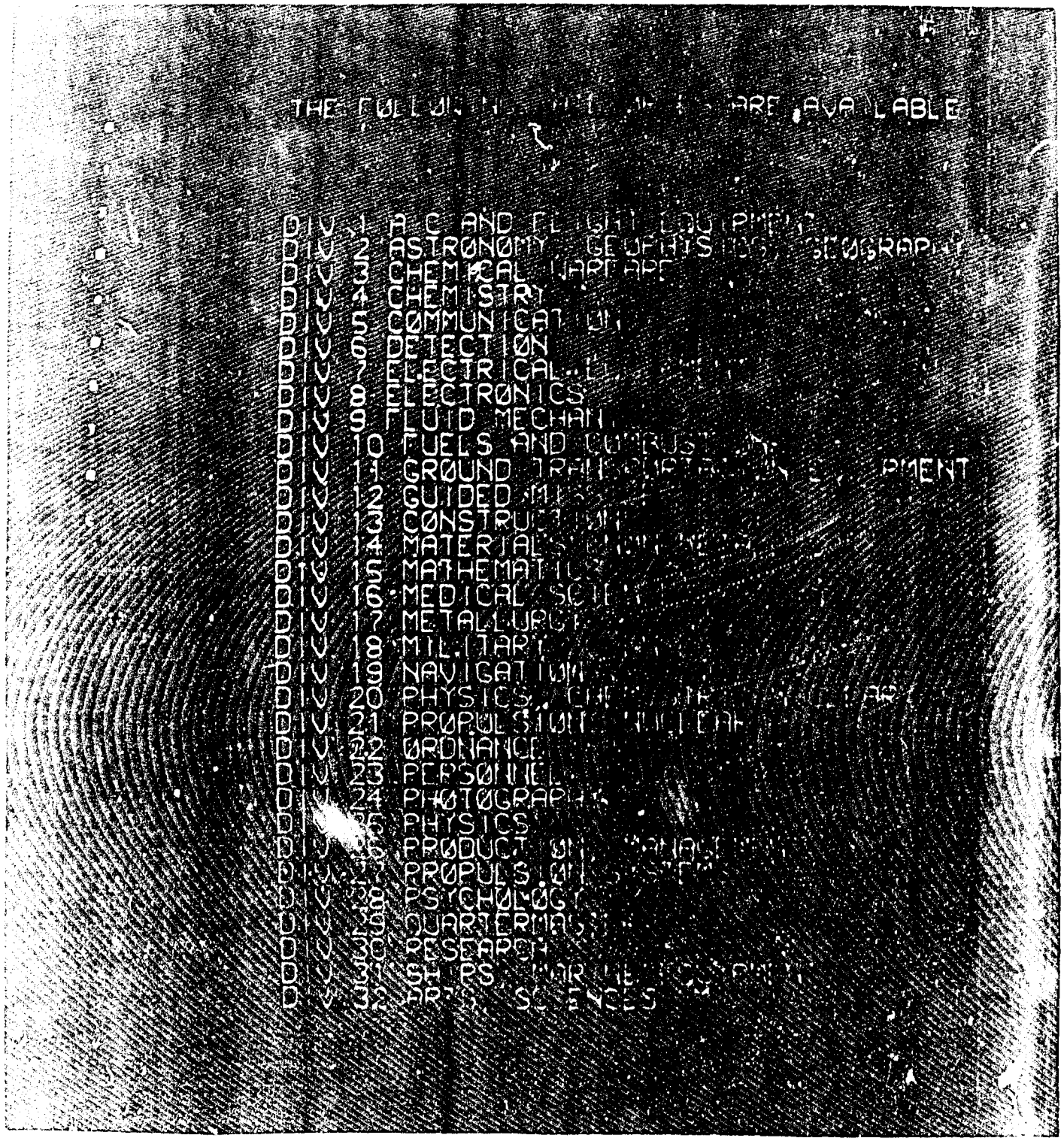


Figure 1. Display of Category Titles

A category is selected (for learning associated retrieval terms) by firing the light pen on any part of the category definition. When a category is selected the display will be changed to present the category and its subcategories and/or synonyms (see Figure 2).

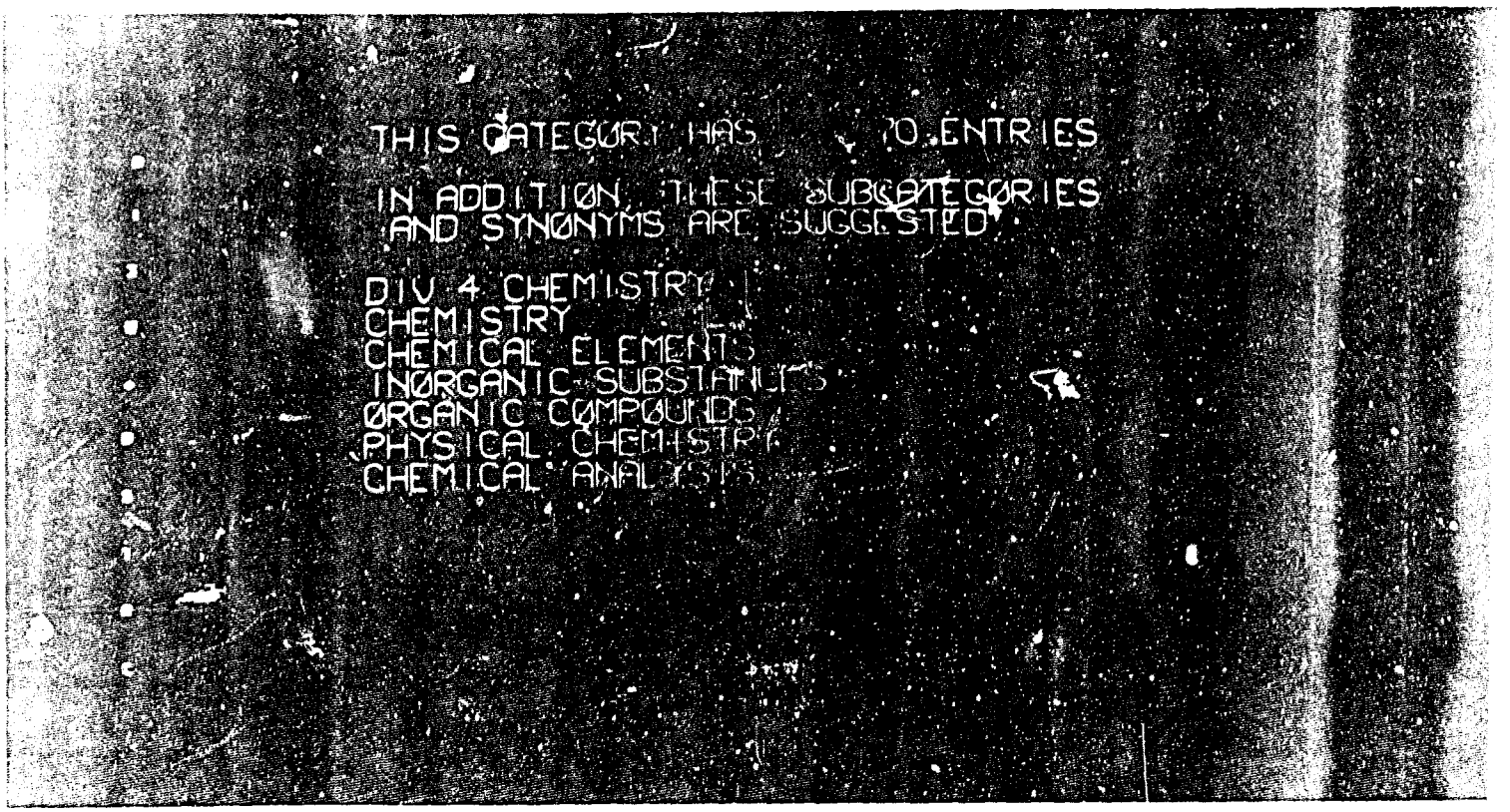


Figure 2. Display of Major Terms Within Div 4 Chemistry

The light pen may be used to select any of the displayed synonyms for a further search through the dictionary definitions (see Figure 3).

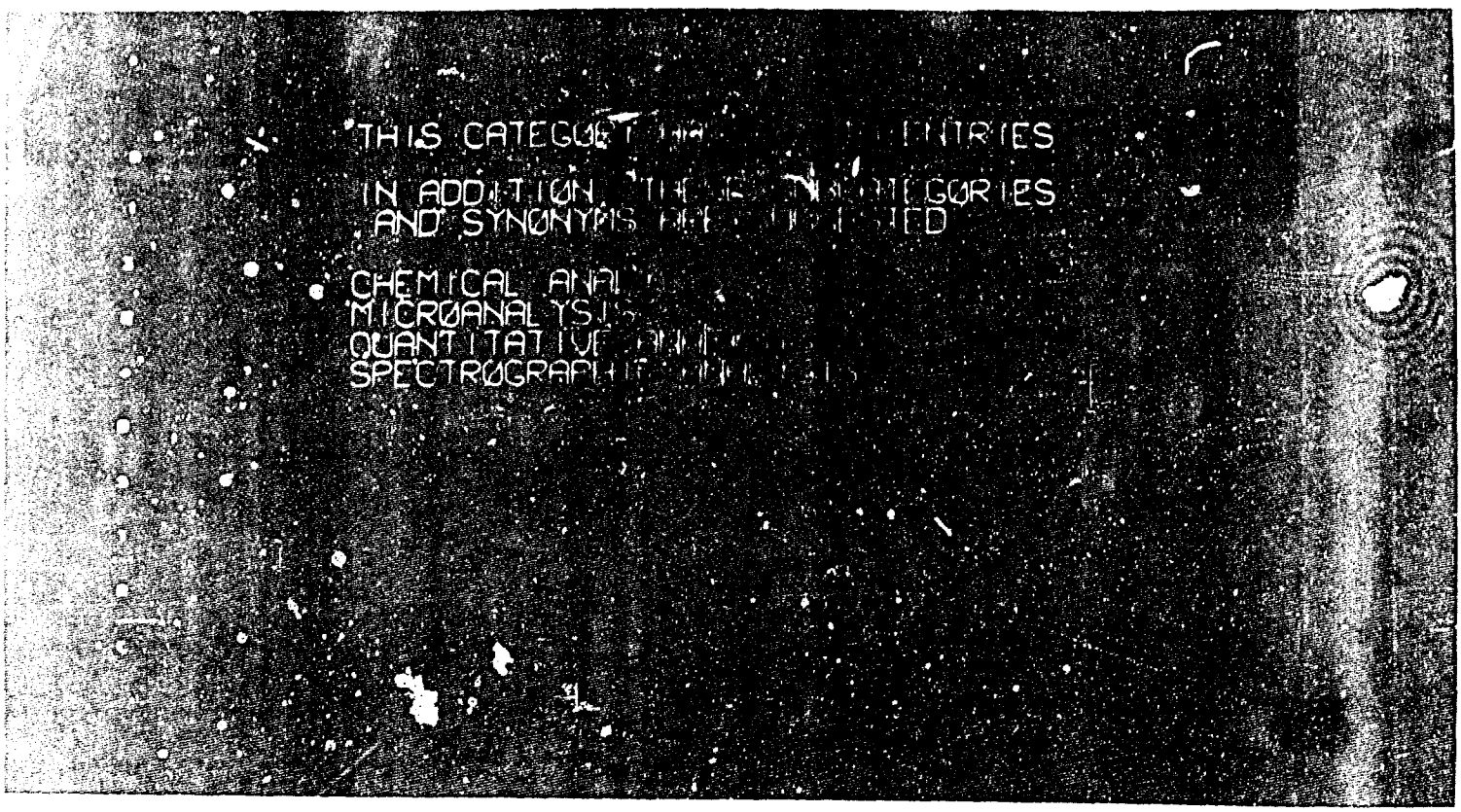


Figure 3. Display of Next Lower-Level Term to Chemical Analysis

The display also shows the count of entries in the data base for the activated term. The selection of a category or retrieval term with the light pen does not cause the term to be considered for the retrieval set.

The two modes for viewing the requested information are the Browse mode and the Search mode. The retrieval mode is designated by selecting the respective activation character with the light pen or by specifying the mode on the teletype, i.e., SEARCH MODE or BROWSE MODE. Retrieval terms must be specified before either mode is designated.

The Search mode displays a correlation between the entries and the retrieval terms. When the Search mode is requested, the program assigns columns in the display to the retrieval terms and designates the assignment at the top of the display area. The retrieved entries are sorted and displayed, one entry per line on the CRT, starting with the entries including the most retrieval terms. The column for the appropriate term is ticked if it references the entry. Figure 4 shows a request in the Search mode.

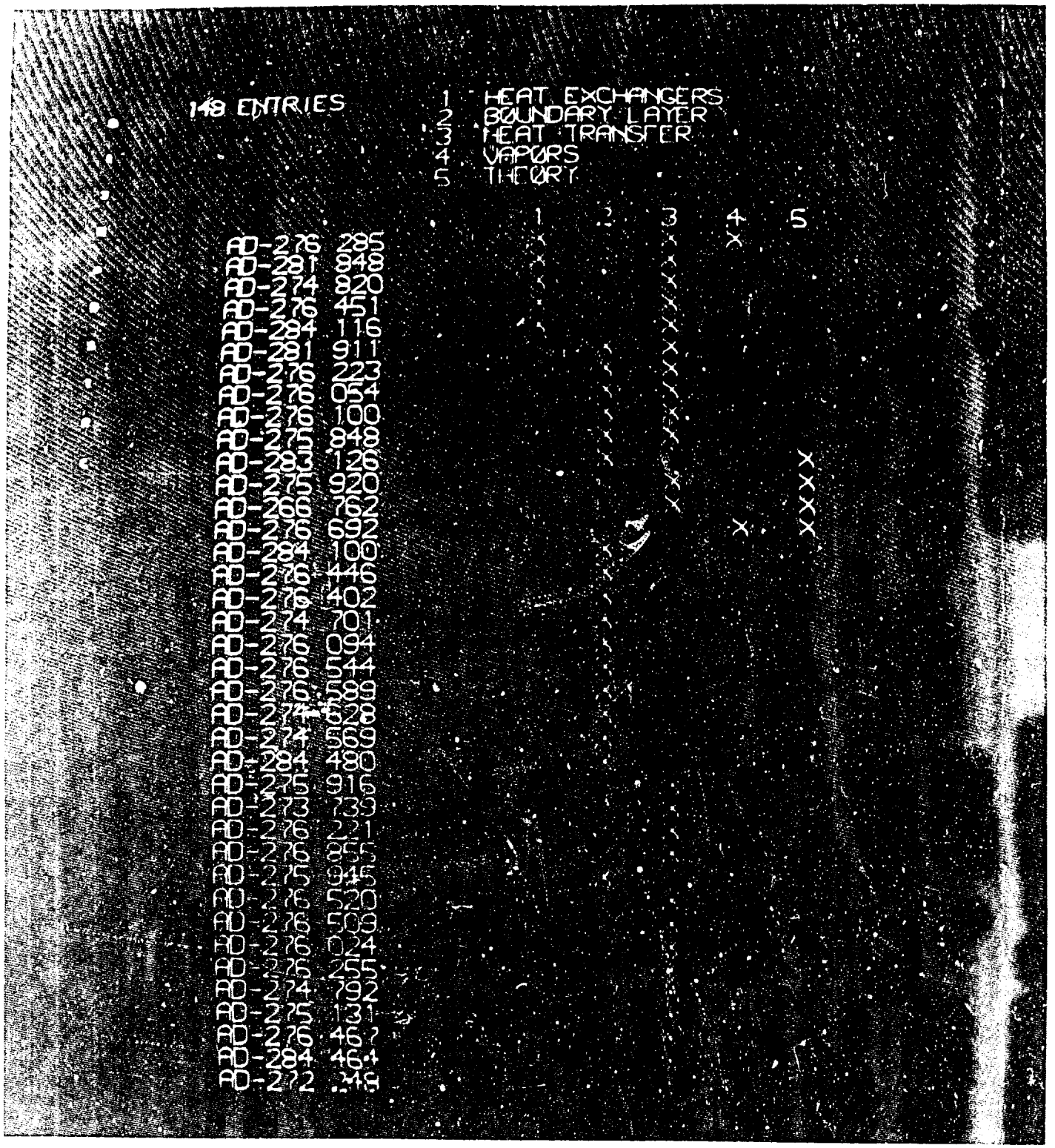


Figure 4. Search Mode Display

Activation on the delete character, followed by activation on some entry, causes that entry to be removed and the next undisplayed entry to be included. Activation on the print on teletype character, followed by activation on a line of the display, causes the contents of that line to be transferred to the teletype for a permanent copy. The abstract for any entry is put on a DLO tape by activation on the save on tape character followed by the activation on the desired entry.

For the Browse mode, the program prints a message on the teletype requesting the information to be retrieved. Any information that was defined as descriptors when the data base was generated may be specified. The command LIST DESCRIPTORS will result in a listing on the teletype of the available descriptors for the data. A maximum of three descriptors may be requested (the document number is always displayed.) If the abstracts are requested, all the stored information for the entry is displayed. Figure 5 shows a display in the Browse mode with the authors and titles requested.

Figure 5. Browse Mode Display for Requested Authors and Titles

Figure 6 shows the Browse mode request for the abstracts. Information is transferred to the teletype and permanent copies of abstracts obtained in the same manner as in the Search mode.

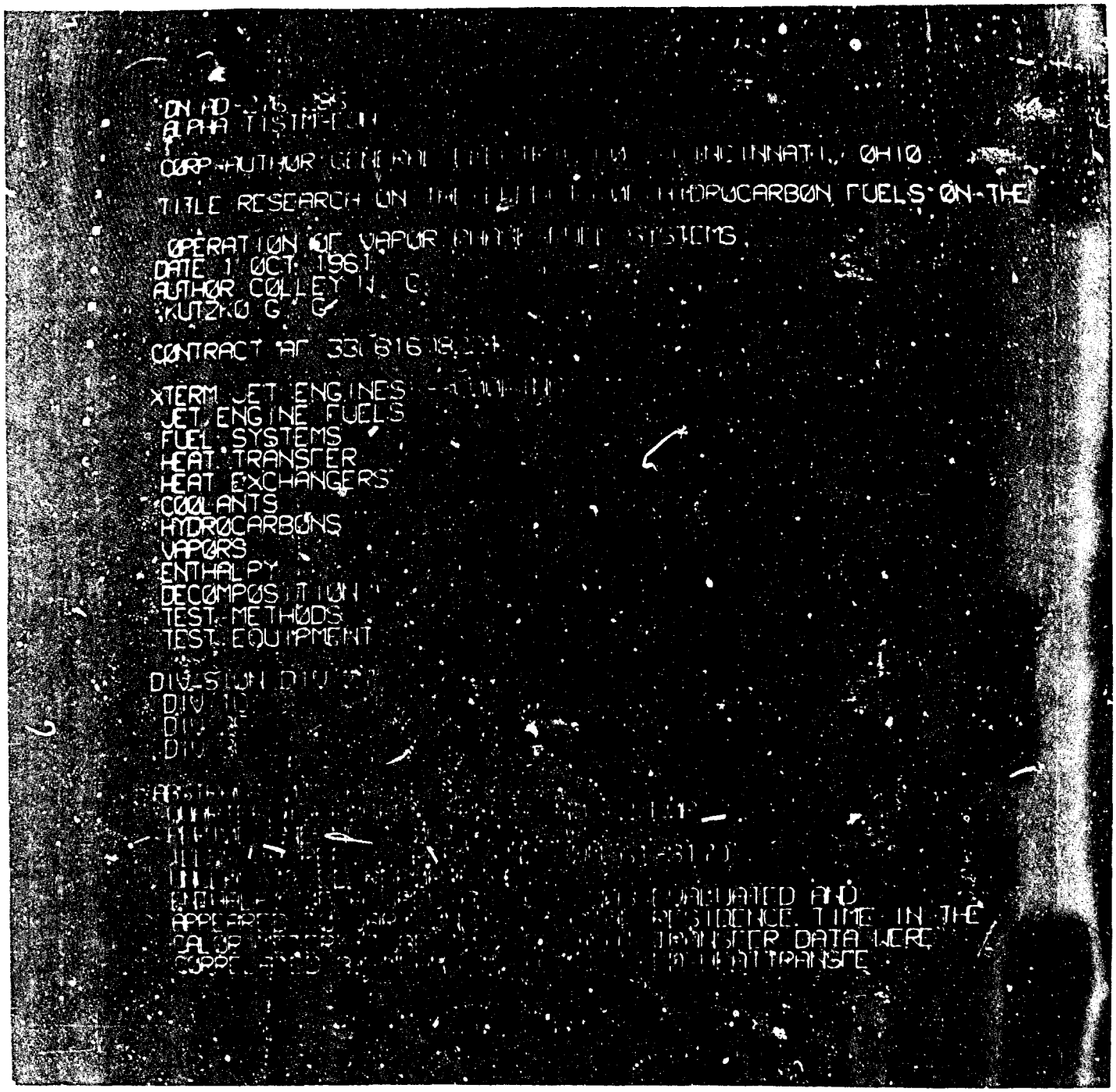


Figure 6. Browse Mode Display for an Abstract Request

Other System Features and Miscellaneous Actions

A single abstract is requested by the command GET followed by the entry identification. It is displayed as in the abstract request for the Browse mode.

A term is deleted from retrieval consideration by the command CANCEL followed by the term.

The contents of a CRT are transferred to the teletype by the command TYPE DISPLAY.

Categories and dictionary hierarchies are designated by the following procedures. The preparatory command *DICTIONARY CHANGE is necessary to make changes to the dictionary. After the program responds *ready the following options are available:

To cause an entry to be treated as a category:

.$term_1$ = category

To add a new entry to the dictionary, the message is the word ADD followed by the phrase to be inserted:

ADD $term_1$

To change a dictionary entry, the message has the following format:

CHANGE (present term) = (desired term)

To remove a dictionary entry, the message is the word DELETE followed by the term to be expurgated:

DELETE $term_1$

Note: When an entry is deleted from the dictionary, it is also deleted as a reference for all data entries.

To define terms as synonyms:

$term_1 = term_2 = term_3$

To define a hierarchy between two terms:

$term_1 < term_2$ ($term_1$ is subordinate)
$term_3 > term_4$ ($term_4$ is subordinate)

To cause an entry to be treated as though it were not in the dictionary:

SUPPRESS $term_1$

To return a suppressed entry to availability:

RESTORE $term_1$

It is necessary to conclude the changing mode with an END to return to retrieval operation. The capability for on-line additions and changes to the data entries will be added in the near future.

In building the hierarchies in the dictionary the terms may not be available (or stated in the best form) for a logical structure. The capability to add new terms allows the dictionary entries to be interrelated without being restricted to any one set of retrieval terms for a data base. The ability to change a dictionary entry allows the program to compensate for some aspects of poor indexing term phrasings. One of these aspects that has been encountered is the use of various forms of an individual term, e.g., weld, welding, weldings (these could all be equated to welding). Another aspect is that a choice of the form of a retrieval term can be

different for different research facilities using the same data base. These capabilities enable the dictionaries for a common data base to be biased, so that retrieval is made more efficient in specific subject areas for the different locations.

The research in using the system is just beginning. It is hoped that the studies made with BOLD will help to establish the most valuable characteristics to be incorporated in future retrieval systems.

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
The BOLD (Bibliographic On-Line Display) System

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

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

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
6 April 1966	23	

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
Independent Research b. PROJECT NO.	SP-2338/000/01
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. AVAILABILITY/LIMITATION NOTICES
This document has been cleared for open publication and may be disseminated by the Clearing House for Federal Scientific & Technical Information

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13 ABSTRACT

The BOLD (Bibliographic On-Line Display) System serves as a general purpose vehicle for research on the components of a real-time retrieval system. Specific subjects for investigation are indexing, classification and categorizing schemes, file organization, and user-system communication. The program operates in a "time-sharing" environment doing independent retrieval for multiple simultaneous users. A retrieval station may be any teletype connected to the time-sharing system. A station may be augmented with a CRT console and a light pen for rapid displaying of the retrieval information. Retrieval is effected by the specification of categories and/or retrieval phrases, using Boolean connectors. There are two modes for retrieval operation: the Browse mode and the Search mode. In the Browse mode the user may specify broad categories and retrieval terms and then browse through the retrieval information entry by entry. The user designates what information is to be returned. This may be anything that is defined in the data base, and may range from a single component (such as author, title, etc., for a bibliographic data set) to a complete body of text (i.e., abstract).

DD FORM 1 JAN 64 1473

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
BOLD Bibliographic On-Line Display Real Time Retrieval Indexing File Organization Time-Sharing						

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, roles, and weights is optional.