DIALOG: An operational on-line reference retrieval system

by ROGER K. SUMMIT
Lockheed Palo Alto Research Laboratory
Palo Alto, California

INTRODUCTION

Classification systems in the sciences usually provide an unambiguous structure of mutually exclusive, collectively exhaustive categories. The same formal structuralization, when strictly applied to the classification of technical literature for retrieval purposes, has proved inadequate. At another extreme, approaches to indexing which preclude any hierarchical association are similarly disappointing. The dual dilemma is illustrated in the following quotation:*

> The English language is so rich that [even] many of the most explicit technical and scientific concepts may be represented by several different word symbols (or combinations of word symbols). This apparent literary advantage can become a formidable retrieval disadvantage unless [some means] is developed to enable the user to express his information needs in his vocabulary, and to retrieve relevant information expressed by an originator in his own entirely different vocabulary. Similarly, a user must be able to express his information needs on a generic, or "inclusive class," level with reasonable expectation that the documents retrieved will include not only those which discussed information on the generic class level, but also those which discussed information on the level of the specific members of that generic class.

The differences among information retrieval systems today relate to the manner in which the two problems of ambiguity and specificity are treated. Off-line, batch-processed retrieval suffers an inherent disadvantage of providing no intermediate results for user evaluation and subsequent search redefinition. For these and other reasons, it is felt by many that an on-line computer system, which allows a user to converse directly with the computer in his quest for relevant document citations, can provide a more effective environment for information retrieval than is possible with off-line systems. The on-line system permits information retrieval to be a highly individualized process with respect to time of occurrence, question at hand, and characteristics of user.

Computer technology currently allows the configuration of a real-time, user-directed information storage and retrieval system. Less understood, however, is the problem of directing and controlling such a hardware configuration so as to allow a user who is neither knowledgeable about nor interested in computers to obtain useful results from a large file of document descriptions (citations) in a rapid, convenient, and effective manner.

Related to a previous experiment, CONVERSE, the DIALOG system was developed to investigate the effectiveness of a flexible, user-directed language in accomplishing reference retrieval.

Dialog development

The effectiveness of an on-line, user-directed retrieval system lies in the degree to which it can accomplish the following:

- Provide a variety of "command" functions for communication, search, and display of information from which the user can select those most appropriate to his particular problem.
- Provide the flexibility to include additional commands or other operational modes as new search techniques are developed.
- Assist the user in search definition and in full employment of system capabilities.
- Allow intermediate user evaluation of search results with subsequent request refinement.
- Require a minimum of bookkeeping or remembering on the part of the user in the association of retrieved references with request expressions.
- Minimize elapsed time between query and response.

*"The Engineers Joint Council Action Plan," Appendix II, *Thesaurus of Engineering Terms*, New York, Engineers Joint Council, May 1964.

- Eliminate need for "middle-man" request interpretation by system specialists.
- Allow real-time interaction between user and system for search guidance.

Although a "free-form" language was considered for communication between the user and the computer, it was decided that a better balance between man inconvenience and machine inconvenience was attained through the use of several predefined commands which could be modified by the user according to his own needs. Such a structure allows modular development of the system and also permits the easy incorporation of additional commands if or as the need for them arises.

It was felt that although most users would not be familiar with Boolean algebra, some method of coordinate searching should be allowed. The conclusion was the development of the COMBINE command. If A and B are sets of documents the first of which contains descriptor A and the second of which contains descriptor B, COMBINE A + B results in a set of documents each of which contains either index term A or index term B; COMBINE A*B results in a set of documents each of which contains both index terms A and B; COMBINE A − B results in a set containing term A but not term B. Any set can be used in subsequent COMBINE commands to recursively partition the reference set into successively more relevant subsets. In this manner, a Boolean search strategy evolves in stepwise fashion, and the user is provided information at the conclusion of each step to assist him in defining the next step.

The dialog language

The current operating environment of the DIALOG system in an IBM 360/30 (32 thousand bytes of core) together with two 2311 disk packs (7.5 million bytes each) for programs and intermediate storage, a 2321 Data Cell (415 million bytes) mass storage device for the reference corpus, a 1443 off-line printer, and a 2260 display/1053 printer input/output terminal (Figure 1). The reference file consists of some 300,000 NASA announced citations.

The DIALOG system provides a number of commands which appear as the upper case or shift values of the top row of keys on the display keyboard (Figure 2). The depression of these command keys, together with entry of associated operands, enable the user to instruct the computer in a desired sequence of operations. A search consists of (1) identifying and selecting descriptors (subject of index terms) which reflect the user's interest, (2) combining descriptors into search expressions, and (3) examining retrieved citations and modifying search expressions.

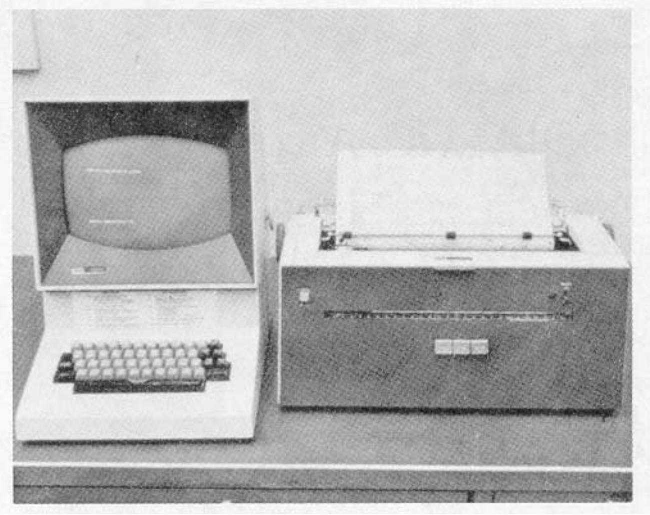


Figure 1.—Remote access terminal

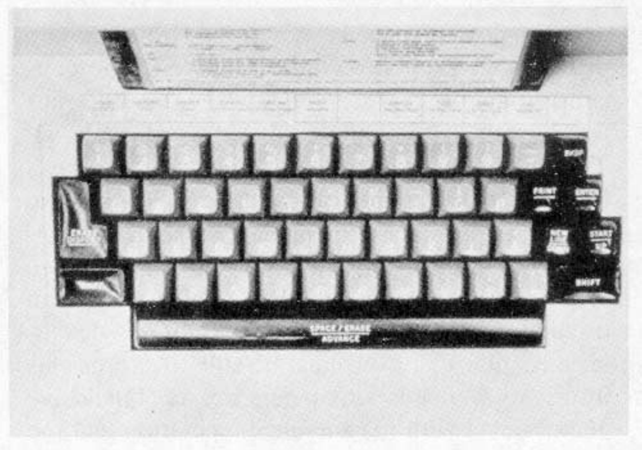


Figure 2.—Remote terminal keyboard

Identification and selection of descriptors

To determine whether a certain term has been used to index documents under consideration, the command EXPAND, entered together with the term, causes a display of a list of actual descriptors alphabetically close to the term entered. Each descriptor is shown with a temporary identification number by which it may be referenced as long as it appears on the display. For each descriptor so displayed, the number of citations to which that descriptor was assigned, as well as the number of terms conceptually related to that descriptor, are also displayed. A display of the conceptually related terms for any displayed descriptor may be obtained by depressing the EXPAND key and entering the descriptor identification number appearing on the display. Any displayed term can be selected by depressing the SELECT command and then entering the descriptor identification number which appears on the display. SELECT causes the citations containing the selected descriptor to be collected for further processing. Each selected descriptor is assigned an identification number and is typed out on the console printer together with the number of citations to which it has been assigned as an index term.

Combination of descriptors

The number of citations associated with any given descriptor is likely to be large (500 to 10,000). Although it is possible to display citations from a single-descriptor set, it is probably more efficient to specify a combination of descriptors which must be present in a citation before it is retrieved. The COMBINE command is used for this purpose. Assume a person is interested in documents pertaining to welding defects in aluminum, and has selected the terms: (1) WELDING (used in 2239 citations), (2) DEFECT (used in 1206 citations), and (3) ALUMINUM (used in 7137 citations). By combining these three terms (i.e., COMBINE 1*2*3 where * stands for "and"), a fourth set of 13 citations results, each of which contains all three terms.

By allowing the repeated use of sets generated by one COMBINE command in the definition of other COMBINE commands, the user can converge in step-wise fashion on čitations of interest. At each step he is provided the size of the resultant set and can either examine individual citations in that set, or modify the set by combining it with other sets (with the COMBINE command).

Examination of retrieved citations and search expression modification

Citations can be displayed wherever desired with the DISPLAY command. This operation will frequently allow the user to discover new descriptors and add them to his search list to further specify his interest.

The KEEP command allows the user to set aside or save selected citations for later printout either on the console printer (TYPE command) or on the off-line printer (PRINT command).

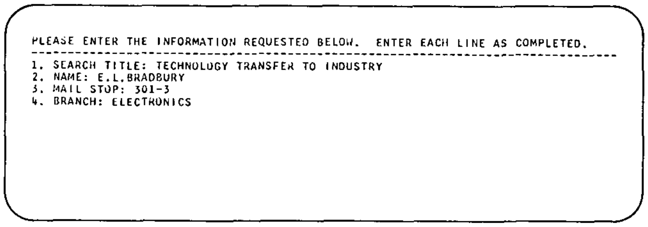
It is possible to further restrict a retrieved set of citations by year of announcement, announcement media (IAA, STAR, CSTAR), or announcement series number. The command LIMIT is provided for this purpose. Although primarily of interest to library personnel, this command allows limiting on any or all of the categories just described.

A search example using dialog

Assume that the user is interested in reports dealing with the transfer of aerospace technology to industry. On the bottom line of the display screen (Figure 2) he will see the message:

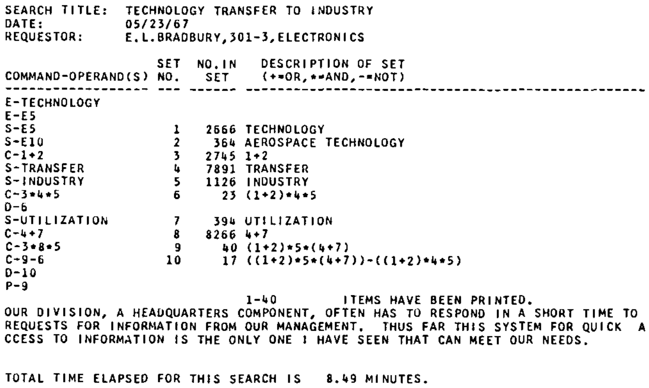
ENTER NEXT COMMAND ►

This is an indication that searching can begin. The user initiates his search by depressing the BEGIN SEARCH command key. This results in an interview display which, when completed, appears as in Display 1.



```
PLEASE ENTER THE INFORMATION REQUESTED BELOW.  ENTER EACH LINE AS COMPLETED.
-------------------------------------------------------------------------------
1. SEARCH TITLE: TECHNOLOGY TRANSFER TO INDUSTRY
2. NAME: E.L.BRADBURY
3. MAIL STOP: 301-3
4. BRANCH: ELECTRONICS
```

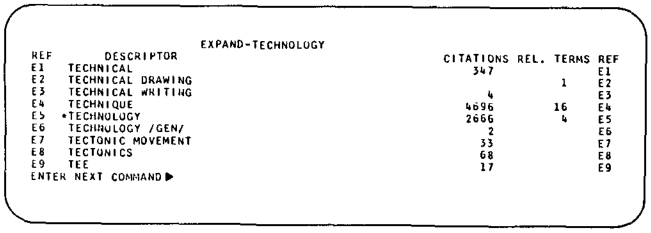
Entry of the final item (the word "electromagnetics" in this case) causes the console printer to print out the search heading at the top of Figure 3. Each subsequent step of the search is typed out on the console printer and appears as Figure 3. (The reader should refer to this figure to follow the search procedure.)



```
SEARCH TITLE:   TECHNOLOGY TRANSFER TO INDUSTRY
DATE:           05/23/67
REQUESTOR:      E.L.BRADBURY,301-3,ELECTRONICS

                   SET  NO.IN   DESCRIPTION OF SET
COMMAND-OPERAND(S) NO.   SET    (+=OR,*=AND,-=NOT)
------------------ ---  -----  ------------------------------------------------
E-TECHNOLOGY
E-E5
S-E5                 1   2666 TECHNOLOGY
S-E10                2    364 AEROSPACE TECHNOLOGY
C-1+2                3   2745 1+2
S-TRANSFER           4   7891 TRANSFER
S-INDUSTRY           5   1126 INDUSTRY
C-3*4*5              6     23 (1+2)*4*5
D-6
S-UTILIZATION        7    394 UTILIZATION
C-4+7                8   8266 4+7
C-3*8*5              9     40 (1+2)*5*(4+7)
C-9-6               10     17 ((1+2)*5*(4+7))-((1+2)*4*5)
D-10
P-9
                              1-40         ITEMS HAVE BEEN PRINTED.
OUR DIVISION, A HEADQUARTERS COMPONENT, OFTEN HAS TO RESPOND IN A SHORT TIME TO
REQUESTS FOR INFORMATION FROM OUR MANAGEMENT.  THUS FAR THIS SYSTEM FOR QUICK  A
CCESS TO INFORMATION IS THE ONLY ONE I HAVE SEEN THAT CAN MEET OUR NEEDS.


TOTAL TIME ELAPSED FOR THIS SEARCH IS   8.49 MINUTES.
```

Figure 3. — Search example (console printer output)

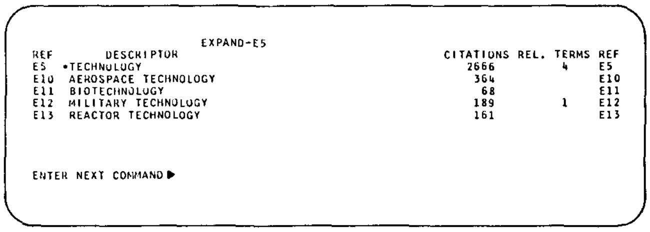
To begin his search, the user wishes to see if "technology" has been used as an index term, and, if so, how many citations contain it as an index item. The user depresses the EXPAND key on the display keyboard and types in "technology." The console printer prints out a command echo (Figure 3) consisting of the first letter of the command (E) together with its operand ("technology"). This provides the user a visual check on what the computer received. The command response then appears on the display screen as shown in Display 2.



```
                         EXPAND-TECHNOLOGY
REF       DESCRIPTOR                                   CITATIONS REL. TERMS REF
E1   TECHNICAL                                            347               E1
E2   TECHNICAL DRAWING                                                1     E2
E3   TECHNICAL WRITING                                      4               E3
E4   TECHNIQUE                                           4696        16     E4
E5  *TECHNOLOGY                                          2666         4     E5
E6   TECHNOLOGY /GEN/                                       2               E6
E7   TECTONIC MOVEMENT                                     33               E7
E8   TECTONICS                                             68               E8
E9   TEE                                                   17               E9
ENTER NEXT COMMAND ▶
```

Notice that the display shows the terms alphabetically near to "technology" (which itself is indicated with an *). By displaying the alphabetically near terms, the user is able to see not only if and to what extent the term he entered has been used as an index term, but also any spelling or ending variations on the term which have been used (e.g., weld versus welding). The user need not even spell the term correctly. "E" numbers are assigned to the displayed index terms for reference purposes.

It can be seen that "technology" is used in an index term in 2,666 citations and has 4 related terms entries in the thesaurus. (Related terms refer to conceptually or hierarchically related terms which are usually associated with a particular term.) To examine the related terms, the user depresses EXPAND and types in E5 which results in Display 3.

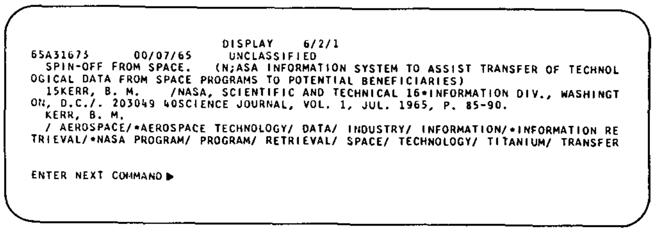


```
                         EXPAND-E5
REF       DESCRIPTOR                                   CITATIONS REL. TERMS REF
E5  *TECHNOLOGY                                          2666         4     E5
E10  AEROSPACE TECHNOLOGY                                 364               E10
E11  BIOTECHNOLOGY                                         68               E11
E12  MILITARY TECHNOLOGY                                  189         1     E12
E13  REACTOR TECHNOLOGY                                   161               E13



ENTER NEXT COMMAND ▶
```

The user notices the descriptor "aerospace technology" and reasons that for his purposes "technology" and "aerospace technology" are equivalent. He thus selects the two terms "technology" and "aerospace technology," and combines the corresponding sets into a third set containing 2,745 citations (in which each citation contains either "technology" *or* "aerospace technology" as index terms). The console typewriter response for these commands is shown in lines 3 through 5 in Figure 3. (Note that OR is coded as "+," whereas AND is coded "*.")

The user now continues his search by selecting "transfer" and "industry" (shown as lines 6 and 7 of Figure 3). He is now ready to combine his selected terms to define his search topic. He wants each citation retrieved to contain either "technology" or "aerospace technology" and "transfer" and "industry." He can effect such a set by depressing the COMBINE key and typing in "3*4*5." This command results in set 6 which contains 23 citations (shown as line 8 of Figure 3).

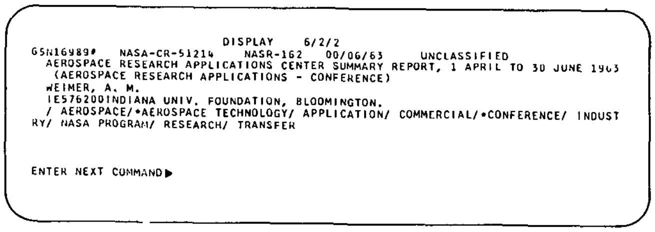
To display these citations, the user depresses the DISPLAY key and types in 6 (the set number), which results in a display of the first citation in the set as shown in Display 4.



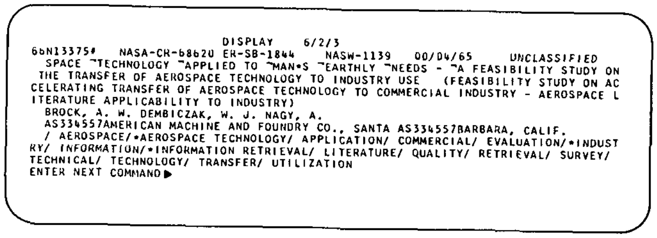
```
                               DISPLAY    6/2/1
65A31673      00/07/65     UNCLASSIFIED
  SPIN-OFF FROM SPACE.   (N;ASA INFORMATION SYSTEM TO ASSIST TRANSFER OF TECHNOL
OGICAL DATA FROM SPACE PROGRAMS TO POTENTIAL BENEFICIARIES)
  15KERR, B. M.    /NASA, SCIENTIFIC AND TECHNICAL 16*INFORMATION DIV., WASHINGT
ON, D.C./. 203049 40SCIENCE JOURNAL, VOL. 1, JUL. 1965, P. 85-90.
  KERR, B. M.
  / AEROSPACE/*AEROSPACE TECHNOLOGY/ DATA/ INDUSTRY/ INFORMATION/*INFORMATION RE
TRIEVAL/*NASA PROGRAM/ PROGRAM/ RETRIEVAL/ SPACE/ TECHNOLOGY/ TITANIUM/ TRANSFER


ENTER NEXT COMMAND ▶
```

Note that the response contains the three specified terms: "transfer," "aerospace technology," and "industry." Successive items can be displayed by depressing ENTER.

Assume the user continues stepping through set 6 to item 3 (Displays 5 and 6). Individual citations can be printed by depressing the PRINT key.



```
                               DISPLAY    6/2/2
65N16989#   NASA-CR-51214    NASR-162   00/06/63     UNCLASSIFIED
  AEROSPACE RESEARCH APPLICATIONS CENTER SUMMARY REPORT, 1 APRIL TO 30 JUNE 1963
   (AEROSPACE RESEARCH APPLICATIONS - CONFERENCE)
  WEIMER, A. M.
  IE576200INDIANA UNIV. FOUNDATION, BLOOMINGTON.
  / AEROSPACE/*AEROSPACE TECHNOLOGY/ APPLICATION/ COMMERCIAL/*CONFERENCE/ INDUST
RY/ NASA PROGRAM/ RESEARCH/ TRANSFER


ENTER NEXT COMMAND ▶
```

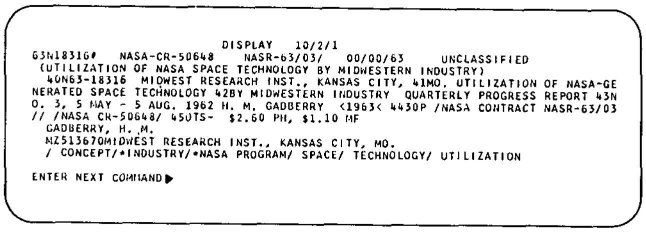


```
                               DISPLAY    6/2/3
66N13375#   NASA-CR-68620 ER-SB-1844    NASW-1139   00/04/65     UNCLASSIFIED
  SPACE ¬TECHNOLOGY ¬APPLIED TO ¬MAN*S ¬EARTHLY ¬NEEDS - ¬A FEASIBILITY STUDY ON
 THE TRANSFER OF AEROSPACE TECHNOLOGY TO INDUSTRY USE   (FEASIBILITY STUDY ON AC
CELERATING TRANSFER OF AEROSPACE TECHNOLOGY TO COMMERCIAL INDUSTRY - AEROSPACE L
ITERATURE APPLICABILITY TO INDUSTRY)
  BROCK, A. W. DEMBICZAK, W. J. NAGY, A.
  AS334557AMERICAN MACHINE AND FOUNDRY CO., SANTA AS334557BARBARA, CALIF.
  / AEROSPACE/*AEROSPACE TECHNOLOGY/ APPLICATION/ COMMERCIAL/ EVALUATION/*INDUST
RY/ INFORMATION/*INFORMATION RETRIEVAL/ LITERATURE/ QUALITY/ RETRIEVAL/ SURVEY/
TECHNICAL/ TECHNOLOGY/ TRANSFER/ UTILIZATION
ENTER NEXT COMMAND ▶
```

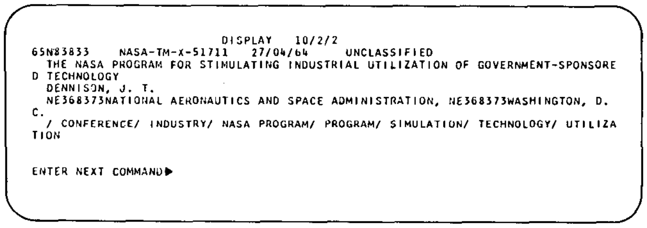
The user continues stepping through the set of citations and notices in item 3 that the term "utilization" is used in the same sense as "transfer" (i.e., "technology transfer to industry" versus "technology utilization by industry"). He thus decides to broaden his search expression to include "transfer" or "utilization." This is accomplished with the following commands:

Command	*Operand(s)*
SELECT	UTILIZATION
COMBINE	4 + 7
COMBINE	3*8*5

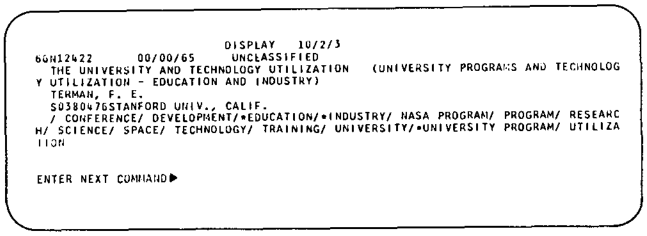
(The results of these commands can be followed in Figure 3.) The final COMBINE command results in set 9 containing 40 citations. To see if the broadened definition returned relevant citations, the user wishes to examine the items in set 9 which do not appear in set 6 (the first search expression). This is accomplished by COMBINE 9-6 (resulting in set 10 containing 17 citations) and DISPLAY 10. A few of the results are shown in Displays 7, 8, and 9.



```
                              DISPLAY   10/2/1
63N18316#   NASA-CR-50648   NASR-63/03/   00/00/63    UNCLASSIFIED
 (UTILIZATION OF NASA SPACE TECHNOLOGY BY MIDWESTERN INDUSTRY)
  40N63-18316  MIDWEST RESEARCH INST., KANSAS CITY, 41MO. UTILIZATION OF NASA-GE
NERATED SPACE TECHNOLOGY 42BY MIDWESTERN INDUSTRY  QUARTERLY PROGRESS REPORT 43N
O. 3, 5 MAY - 5 AUG. 1962 H. M. GADBERRY  <1963< 4430P /NASA CONTRACT NASR-63/03
// /NASA CR-50648/ 450TS-  $2.60 PH, $1.10 MF
  GADBERRY, H. M.
  MZ513670MIDWEST RESEARCH INST., KANSAS CITY, MO.
  / CONCEPT/*INDUSTRY/*NASA PROGRAM/ SPACE/ TECHNOLOGY/ UTILIZATION

ENTER NEXT COMMAND▶
```

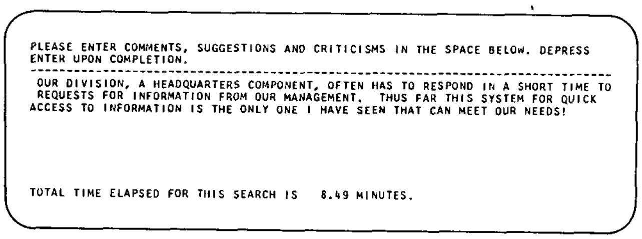


```
                              DISPLAY   10/2/2
65N83833    NASA-TM-X-51711   27/04/64     UNCLASSIFIED
  THE NASA PROGRAM FOR STIMULATING INDUSTRIAL UTILIZATION OF GOVERNMENT-SPONSORE
D TECHNOLOGY
  DENNISON, J. T.
  NE368373NATIONAL AERONAUTICS AND SPACE ADMINISTRATION, NE368373WASHINGTON, D.
C.
  / CONFERENCE/ INDUSTRY/ NASA PROGRAM/ PROGRAM/ SIMULATION/ TECHNOLOGY/ UTILIZA
TION


ENTER NEXT COMMAND▶
```



```
                              DISPLAY   10/2/3
66N12422      00/00/65     UNCLASSIFIED
  THE UNIVERSITY AND TECHNOLOGY UTILIZATION   (UNIVERSITY PROGRAMS AND TECHNOLOG
Y UTILIZATION - EDUCATION AND INDUSTRY)
  TERMAN, F. E.
  S0380476STANFORD UNIV., CALIF.
  / CONFERENCE/ DEVELOPMENT/*EDUCATION/*INDUSTRY/ NASA PROGRAM/ PROGRAM/ RESEARC
H/ SCIENCE/ SPACE/ TECHNOLOGY/ TRAINING/ UNIVERSITY/*UNIVERSITY PROGRAM/ UTILIZA
TION


ENTER NEXT COMMAND▶
```

Any or all of these citations or their accession numbers (identification numbers) can be printed out for future reference (in this example the user prints set 9). The search expression could be further broadened or narrowed by including additional terms. The retrieved sets could be limited by various parameters such as date and publication type. The search is completed when the user depresses END SEARCH. This command results in Display 10 which the user completes. Entry of this information causes elapsed search time and user comments to be printed on the console typewriter, and clears the computer for the next search.



```
PLEASE ENTER COMMENTS, SUGGESTIONS AND CRITICISMS IN THE SPACE BELOW. DEPRESS
ENTER UPON COMPLETION.
--------------------------------------------------------------------------------
 OUR DIVISION, A HEADQUARTERS COMPONENT, OFTEN HAS TO RESPOND IN A SHORT TIME TO
 REQUESTS FOR INFORMATION FROM OUR MANAGEMENT.  THUS FAR THIS SYSTEM FOR QUICK
ACCESS TO INFORMATION IS THE ONLY ONE I HAVE SEEN THAT CAN MEET OUR NEEDS!



TOTAL TIME ELAPSED FOR THIS SEARCH IS   8.49 MINUTES.
```

CONCLUSION

The DIALOG language was developed as a proprietary product by the Information Sciences group at the Lockheed Palo Alto Research Laboratory. It has been implemented with the NASA collection of over 300,000 citations, and is currently being applied to a collection of personnel summaries.

Search topics which have been executed using DIALOG include:

- Interaction of magnetosphere and solar wind
- Vibrational excitation of carbon dioxide
- Molybdenum disulfide as a solid lubricant in spacecraft
- Gas phase reactions of fluorocarbons with oxygen and nitrogen

Search times vary considerably among users, depending on the user's experience and on the complexity of the search. A reasonable average, however, is about 30 minutes of elapsed time and 5 to 10 minutes of computer time per search. At $125 per hour average machine charge, this represents a cost of $10 to $20 per search.

In summary, there are five important characteristics of the DIALOG language:

- The search question is constructed at search time (rather than at index time as is the case with a manual system).
- DIALOG is designed for nonspecialists; i.e., the users themselves, and thus avoids one communication barrier.
- The command language is independent of the particular data it searches.
- As an on-line system, it allows continual redefinition of the search question, based on examination of intermediate results.
- Control of the process lies with the user; the computer merely serves as a data-processing extention of the user.

BIBLIOGRAPHY

D L DREW R K SUMMIT R I TANAKA and R B WHITELEY

An on-line technical library reference retrieval system
American Documentation 17 No 1 3 1966
E HERBERT
Information transfer
International Science and Technology No 51 26 1966
N S PRYWES
Brousing in an automated library through remote access
Computer Augmentation of Human Reasoning Spartan Books Inc Washington D. C. 1965
Information
Scientific American Book W. H. Freeman & Comapny San Francisco Calif. 1966