DEFENSE TECHNICAL INFORMATION CENTER
8725 JOHN J. KINGMAN ROAD
FORT BELVOIR, VIRGINIA 22060-6218

IN REPLY REFER TO: DTIC-R (FOIA 2019-153)

JUL 25 2019

John Greenewald
27305 W Live Oak Rd., Suite 1203
Castaic, CA 91384

Dear Mr. Greenewald:

This is in response to your request dated July 20, 2019, requesting information under the Freedom of Information Act (FOIA) (enclosure 1). Under Department of Defense rules implementing the FOIA, published at 32 CFR 286, your request was categorized as "other".

The document that you have requested AD0776299, entitled "Optimization of Retrieval Techniques and File Structures for the CIRCOL (Central Information Reference and Control Online System)", is approved for public release. It is enclosed.

To date, there are no assessable fees for services from DTIC. Please understand that other members of the public may submit a FOIA request for copies of FOIA requests received by this office, or the names of those who have submitted requests. Should such occur, your name and, if asked for, a copy of your request will be released; however, your home address and home telephone number will not be released. Other private citizens who have obtained your name by using such a request may contact you; however, correspondence from the DoD about your request will be on official letterhead. Please contact me at (703) 767-9204 if you have any questions. Thank you for your interest in obtaining information from DTIC.

Sincerely,

Michael Hamilton
FOIA Program Manager

Enclosure

AD-776 299

OPTIMIZATION OF RETRIEVAL TECHNIQUES AND FILE STRUCTURES FOR THE CIRCOL (CENTRAL INFORMATION REFERENCE AND CONTROL ON-LINE) SYSTEM

Frederic L. Scheffler

Dayton University

Prepared for:

Rome Air Development Center

February 1974

DISTRIBUTED BY:

NTIS
National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-73-420	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) OPTIMIZATION OF RETRIEVAL TECHNIQUES AND FILE STRUCTURES FOR THE CIRCOL SYSTEM		5. TYPE OF REPORT & PERIOD COVERED Final Report 1 Jul 72 - 31 Aug 73
		6. PERFORMING ORG. REPORT NUMBER UDRI-TR-73-53
7. AUTHOR(s) Frederic L. Scheffler		8. CONTRACT OR GRANT NUMBER(s) F30602-72-C-0505
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Dayton Research Institute 300 College Park Avenue Dayton, Ohio 45469		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 31025F IDHS 0409
11. CONTROLLING OFFICE NAME AND ADDRESS Rome Air Development Center (IRDT) Griffiss Air Force Base, New York 13441		12. REPORT DATE February 1974
		13. NUMBER OF PAGES 148
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Same		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Search Response Time	Information Retrieval
Document Retrieval	CIRCOL
On-Line Systems	Evaluation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Central Information Reference and Control On-Line (CIRCOL) System is a computer-based document reference retrieval system maintained by the Foreign Technology Division (FTD) in support of Scientific/Technical intelligence production. CIRCOL System performance was analyzed by systematically studying the effects of twelve primary factors on search response time. The primary factors included: the specification of positional logic, the use of country codes as a search term vs. use as a qualifier, the use of LRANGE specification,

20. ABSTRACT (continued)

the use of the truncation feature, the use of the DATE field as a qualifier, the use of the LRANGE command vs. use of the DATE field as a qualifier, the use of labelled vs. unlabelled Boolean statements, the running of one complex search vs. formulating and running several logically equivalent search strategies, posting density, the number of documents retrieved, the number of users simultaneously on-line, and the order in which search lines are entered. The findings were used to produce a CIRCOL User's Guide Supplement (included in this report as Appendix A) which describes techniques that can be used to manipulate the appropriate factors to optimize the formulation of search strategies. Also included in this report is a study of the use of CIRCOL by four types of users, recommendations on how to optimize file structures and system processing, and an investigation of the IBM Storage and Information Retrieval System (STAIRS) was made with respect to its suitability to the CIRCOL System.

PREFACE

This report covers research performed by the University of Dayton Research Institute, Information Systems Section, between 1 July 1972 and 31 August 1973. Rome Air Development Center, Griffiss Air Force Base, New York, sponsored the effort under Contract F30602-72-C-0505, Job Order Number IDHS0409. Mr. Nicholas M. DiFondi (IRDT) was the RADC Project Engineer.

The report includes the analysis of CIRCOL system performance as a function of various factors which influence retrievals and response time; it also provides an analysis of user on-line searching differentiated into four distinct types of users.

The author acknowledges the efforts and contributions of Ms. Jacqueline F. March, Ms. Marie E. Shanley, Mr. Eugene R. Egan, and Mr. Howard H. Schumacher, all of the University of Dayton Research Institute, and Mr. Donald Quigley of the Foreign Technology Division, FTD/NIIB.

This technical report has been reviewed by the RADC Information Office (OI) and is releasable to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved.

APPROVED: Nicholas M. Difondi

NICHOLAS M. DIFONDI
Technical Evaluator

APPROVED: Franz H Dettmer

FRANZ H. DETTMER
Colonel, USAF
Chief, Intel & Recon Div

FOR THE COMMANDER:

CARLO P. CROCETTI
Chief, Plans Office

EVALUATION

The objective of this study was to determine from an in-depth analysis of user behavior system performance of the Central Information Reference and Control On-Line (CIRCOL) document retrieval system at FTD. System performance was measured in terms of the amount of time it took the retrieval mechanism to search the data base (search response time). Significant conclusions are:

a. The number of documents retrieved was the most significant factor in causing long search response times.

b. Due to computational requirements placed upon the computer when positional logic is specified in a request, search response time was expected to be long. However, since positional logic reduces the number of documents retrieved a trade-off situation is created resulting in short retrieval times.

c. The use of country codes as search terms vs. their use as qualifiers significantly reduced search time.

d. The use of the LRANGE command is highly effective in reducing search response time.

e. Qualifications of the request by date causes a reduction in search time only when the date range is quite small. The most efficient way of qualifying by date is to combine the DATE field specification with the LRANGE command.

f. Labelled Boolean logic permits more flexibility in search strategy formulation and usually results in faster searches.

g. The posting density of individual index terms significantly effects search response time. This is in agreement with results listed in "a" above because heavily posted terms cause large numbers of documents to be retrieved. Only if two heavily posted terms are combined with a logical AND could smaller retrievals occur and respon time be shortened.

These findings have resulted in the production of the CIRCOL USER's GUIDE SUPPLEMENT to aid the user in formulating searches for efficient retrievals. This effort is in support of the written word exploitation mission as stated in TPO #4.

Nicholas M. Difondi

NICHOLAS M. DIFONDI

SUMMARY

An investigation was made of the CIRCOL system, a natural language text processing information storage and retrieval system maintained by the Air Force Foreign Technology Division for the retrieval of scientific and technical research and intelligence information. CIRCOL is used by various intelligence and research and development organizations to serve their information needs. Retrieval by a number of subject and bibliographic retrieval parameters and combinations thereof is possible. Retrieval is accomplished both by specifying items in the Search Mode, which utilizes inverted files, and in the Qualification Mode which utilizes a checking routine for the occurrence or nonocurrence of a specified item in a designated fixed format field.

The investigation encompassed four interrelated phases. Phase 1 was concerned with those system and search strategy characteristics which affect retrieval results and search response times. Phase 2 involved studying actual use of the CIRCOL system by four categories of types of users. Phase 3 incorporated the findings of the first two phases to suggest techniques which could be applied by the user, and system processing and file structure modifications to optimize retrieval from CIRCOL. Phase 4 involved the investigation of the Storage and Information Retrieval System (STAIRS) with a model CIRC data base to determine its suitability as a possible software package for the FTD application.

PHASE 1 - FACTORS AFFECTING SEARCH RESPONSE TIMES AND RETRIEVALS

In Phase 1, twelve factors were identified which influence the search response time. These are:

1. The specification of positional logic.
2. The use of the country code as a search term vs. use as a qualifier.
3. The use of the LRANGE specification command. (The LRANGE command limits the extent of the file searched by a specified range of accession numbers).
4. The use of the truncation feature.
5. The use of the DATE field as a qualifier.
6. The use of the LRANGE command vs. the use of the DATE field as a qualifier.
7. The use of labelled Boolean statements vs. the use of unlabelled Boolean statements.

8. The running of one complex search strategy vs. formulating and running several logically equivalent search strategies.
9. The posting density.
10. The number of documents retrieved.
11. The number of users simultaneously on-line.
12. The order in which search lines are entered.

Results of the Phase 1 investigation showed that the number of documents retrieved (Point 10 above) was the most significant factor in causing long search times. Apparently, the internal system mechanisms for addressing the Text File represent the major time-consuming step in search processing. The Text File must be referenced on retrieval to permit on-line display or off-line printing of document records.

The specification of positional logic (Point 1 above) requires that the Master File be accessed to establish the search term relationships specified. This requirement should increase search response time. In practice, however, there is a trade-off between the increased time to apply the positional logic and the reduced time effected by retrieving a correspondingly fewer number of documents.

The use of country codes (Point 2 above) as search terms vs. their use as qualifiers indicated that the use of country codes as search terms significantly reduces search response times, especially when no other qualifiers are present. When the number of country codes exceeds three or four and other qualifiers are present in the search strategy, there is little difference between the use of country codes as search terms and the use of country codes as qualifiers.

The LRANGE command (Point 3 above) is highly effective in reducing search response times, both due to the fewer number of document postings and the fewer number of documents actually retrieved.

The truncation feature (Point 4 above) results in a logical OR series of terms whose first 'n' characters match the 'n' characters specified by the truncated form. No differences in search response times were found between specifying the truncated form and specifying the equivalent logical OR series in the search strategy. The user must be cautious in using the truncated form, not only in terms of the logical OR series generated, but also regarding intermixing of logical OR's and AND's in the same search line.

The DATE field used as a qualifier (Point 5 above) causes a reduction in search response time only when the date range is quite small. The most efficient way of qualifying by date is by combining the DATE field specification with the LRANGE command (Point 6 above).

Labelled Boolean logic (Point 7 above) permits more flexibility in search strategy formulation and usually results in faster response times. Experiments with the running of a long search vs. running of several equivalent shorter searches (Point 8 above) showed that longer searches actually consume less overall time; the overall time consists of user time at the terminal plus the summation of search execution times. Increasing posting densities of individual terms (Point 9 above) increases search response times. However, if the number of retrievals resulting from a logical AND statement of two heavily posted terms is smaller than the number of retrievals corresponding to a less heavily posted individual term, the search response time will be less for the former condition, i.e., for the smaller number of retrievals.

The number of documents retrieved (Point 10 above) is highly significant in increasing search response times as indicated earlier. The number of simultaneous on-line users (Point 11 above) affects search response times erratically, especially with many users on-line. The response time generally increases with more users on-line. The order in which search lines is entered is important (Point 12 above). Search lines should be entered such that the number of documents anticipated on retrieval corresponding to each line increases from the first line to the final line of the search.

PHASE 2 - STUDY OF THE USE OF CIRCOL BY FOUR TYPES OF USERS

Users were categorized as follows:

1. FTD information specialists
2. FTD intelligence analysts
3. Outside R&D organizations
4. Outside Intelligence Agencies

Studies of the various types of users indicated that there are no startling differences in the use of the CIRCOL system by the different types of users. R&D users tend to make more subject-oriented requests than intelligence users. Intelligence users make extensive use of personalities and authors in their searching. Only limited searching of facilities, locations and nomenclature was performed, in part because no convenient and reliable means of performing searches of this type are available. Almost no instances of searches occurred in which both subject terms and personalities/authors, facilities, locations, and nomenclature were incorporated into one search strategy. Over a period of time, all users tended to make more use of the features available to them, especially the LRANGE command and labelling. Another observable trend over a period of time was the increase in the number of author/personality searches.

Analysis of on-line terminal records indicated that few problems were encountered by the users in interacting with the system. FTD information specialists are most proficient with the system and make fewest errors.

PHASE 3 - CIRCOL USER'S GUIDE SUPPLEMENT AND SYSTEM RECOMMENDATIONS

A number of factors studied in Phase 1 can be controlled by the user in formulating search strategies. A supplement to the CIRCOL User's Guide was prepared and distributed to CIRCOL users. The purpose of the supplement was to indicate to the user techniques for optimizing search strategies.

Recommendations were made to optimize file structures and system processing in order to improve search response times. Among these recommendations were suggestions for the implementation of several new commands. These suggested commands were:

1. A LABEL command to effect automatic labelling of search statements,
2. A TERMS command to permit on-line viewing of available terms
3. An ENTER command to permit the entering of terms selected from the displayed list of terms into a strategy.

System prompting of the user to enter known qualification statements initially in his strategy was recommended to reduce overall search response time. The establishment of a supplementary Dictionary File containing high frequency search terms was suggested as a means of reducing search response times. Removal of never-used data from the fixed format fields in the Master File would improve the system. The system processing steps which effect referencing of the Text File corresponding to the DPS numbers retrieved, were found to represent the most time-consuming aspect of CIRCOL. Thus, this factor is the primary cause of long search response times. It was recommended that these processes be reviewed and modified appropriately to make them more efficient.

PHASE 4 - INVESTIGATION OF (STAIRS) THE STORAGE AND INFORMATION RETRIEVAL SYSTEM

STAIRS was investigated to determine if it would provide features and capabilities which would serve the needs of CIRC users. A pilot data base of CIRC documents was available with STAIRS, and limited investigation indicated that STAIRS with certain modifications would be suitable for the FTD/CIRC application. It is basically easy to use and provides excellent response times.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont'd)

TABLE OF CONTENTS (cont'd)

TABLE OF CONTENTS (cont'd)

TABLE OF CONTENTS (cont'd)

LIST OF TABLES

LIST OF TABLES (cont'd)

TABLE PAGE

SECTION 1

INTRODUCTION

1.1 SYNOPSIS

The Centralized Information Reference and Control On-Line (CIRCOL) system is an on-line computer-based information storage and retrieval system which has been established and is maintained by the Foreign Technology Division (FTD) at Wright-Patterson Air Force Base (W-PAFB). The purpose of CIRCOL is to serve the needs of Department of Defense intelligence agencies, plus research and development organizations, both within the Department of Defense and in other sectors of the Federal Government, as well as qualified contractors to the Federal Government. CIRCOL consists of document reference records of foreign scientific and technical and intelligence information derived from a large number of sources. CIRCOL provides the means by which users around the country can search information in the CIRCOL data base by direct access through remote computer communications terminals.

CIRCOL is a natural language text processing system which provides a number of search parameters for the user. Because of the large number of search parameters or search fields available, the system is extremely flexible, and the user can exercise many search options for retrieving documents from the system.

CIRCOL has been in active use since 1969, and its success has been proven by the facts that the user community has grown in size and the use by individual user groups has continually increased. However, no comprehensive systematic study of CIRCOL had been made up to 1972. The purpose of the work performed by the University of Dayton Research Institute and reported herein was to study CIRCOL system performance characteristics and to investigate the actual use made of CIRCOL by various types of users. Our investigation culminated in a set of recommendations for optimization of retrieval techniques and file structures in order to make CIRCOL more efficient and effective for the CIRCOL users.

The work program and results are presented in Sections 2 through 7 of this report. A literature survey was made to determine the state of the art of natural language text processing systems, with particular regard to performance and cost effectiveness factors. This survey is presented later in this Section. The final part of this Section provides an outline of the program plan followed during our investigation.

Section 2 gives a more detailed description of CIRCOL. Section 3 discusses the Document Processing System (DPS) which is the basic software package for CIRCOL. Section 4 presents our findings on the primary identifiable factors which affect CIRCOL search response time, and the relative effect of each factor. Section 5 covers our study of the actual use of CIRCOL by the user community. In this part of the study, the user population was categorized into four groups and the differences and similarities in the use of CIRCOL by the various groups was determined.

Section 6 consists of observations and recommendations for the CIRCOL system based on our findings from the work described in Sections 4 and 5. Section 7 discusses a brief evaluation made of the Storage and Information Retrieval System (STAIRS) software package which first became available during the course of the CIRCOL study. The STAIRS system was evaluated in terms of its suitability for the FTD CIRC applications, particularly from the user standpoint.

1.2 NATURAL LANGUAGE TEXT PROCESSING AND RETRIEVAL

In order to assess the state of the art of natural language text processing indexing, storage, and retrieval systems, a literature review was undertaken. Thereby, a proper perspective of the CIRCOL system could be obtained, especially with regard to the appropriateness of the CIRCOL system for on-line retrieval of scientific and technical intelligence information.

There is extensive literature available on automatic indexing from natural language or free text. There is still considerable controversy over natural language indexing and retrieval vs manual indexing using indexer-assigned keywords from a controlled vocabulary. Stevens[1,2,3,4] has prepared some excellent state-of-the-art reviews of automatic indexing.

Historically, CIRCOL developed from an assigned-keyword indexing approach. A CIRC Thesaurus of Topic Tags was used for manually indexing documents. For the FTD application, it appeared that automatic

[1] M.E. Stevens, Automatic Indexing: A State-of-the-Art Report, NBS Monograph 91, National Bureau of Standards, U.S. Department of Commerce, Washington, D.C., March 1965.

[2] M.E. Stevens, Research and Development in the Computer and Information Sciences, Volume 1: "Information Acquisition, Sensing, and Input - A Selective Literature Review," NBS Monograph 113, Volume 1. National Bureau of Standards, U.S. Department of Commerce, Washington, D.C., March 1970.

[3] M.E. Stevens, Research and Development in the Computer and Information Sciences, Volume 2: "Processing, Storage and Output Requirements in Information Processing Systems - A Selective Literature Review," NBS Monograph 113, Volume 2, National Bureau of Standards, U.S. Department of Commerce, Washington, D.C., May 1970.

[4] M.E. Stevens, Research and Development in the Computer and Information Sciences, Volume 3: "Overall System Design Considerations - A Selective Literature Review," NBS Monograph 113, Volume 3, National Bureau of Standards, U.S. Department of Commerce, Washington, D.C., June 1970.

indexing might prove advantageous. Smith, Hoffman, and Cornell[5] performed a study referred to as the CIRC ON-LINE EXPERIMENT (COLEX). This study indicated the feasibility of an automatic indexing and retrieval system for the FTD application.

According to Cuadra[6], the issue of automatic versus manual information systems seems to be resolved. Automation is inevitable. He states, "the question is less 'to computerize or not to computerize' than what subsystem to computerize first, or what computerized services to take advantage of - when or if the money is available. The benefits and value of information along with pricing and marketing present new and thorny problems. And the thorniest is to decide which values to attach a dollar sign to."

King, Neel, and Wood[7] conducted a comparative study on the retrieval effectiveness of two alternate input and search systems. They used such measures as recall, fallout, precision, and total retrieval. One system used manually-indexed document files searched by controlled vocabulary, while the other employed full text input using natural language statements for searching. Both systems were applied to a common data base. Results indicate that "the two systems perform at approximately the same level of effectiveness, although estimated total average retrieval was found to be slightly greater for free-text searching than for descriptor searching at all levels of recall."

Information science appears to be in a transition period when emphasis is shifting from performance and performance testing to an increasing interest in costs and cost effectiveness. Many reports on comparative studies can be found in the literature, but the results are not conclusively in favor of one system over another (manual versus automatic). The emphasis seems to be shifting toward cost effectiveness evaluations in an effort

[5] J.L. Smith, J.D. Hoffman, and J.C. Cornell, COLEX (CIRC ON-LINE EXPERIMENT), RADC-TR-68-332, Rome Air Development Center, Griffiss Air Force Base, N.Y., November 1968.

[6] C.A. Cuadra, in Annual Review of Information Science and Technology, Volume 7, p. 57, Washington, D.C.: American Society for Information Science, 1972.

[7] D.W. King, P.W. Neel, and B.L. Wood, Comparative Evaluation of the Retrieval Effectiveness of Descriptor and Free-Text Search Systems Using CIRCOL, Report #0199, RADC-TR-71-311, January 1972.

to determine not the performance per se, but the performance per dollar. Vickery[8] states "the computer does not make retrieval more automatic. On the contrary, it gives us the opportunity to make it more imaginative." Flexibility and speed seem to be the primary factors favoring an automatic information system.

Viable evaluation techniques have been developed in many areas of information science and these techniques are being improved upon. The fact remains, however, that in a time when economic factors are so important, major effort must be directed toward designing the most cost effective information retrieval system.

Cleverdon[9] states, "All the experimental evidence shows that there is no technique, no method of carrying out any particular operation that will, in all circumstances or under all situations, prove to be the most satisfactory; equally so, in any given situation, it is possible that there will be several different designs that, from the viewpoint of performance, would produce strictly comparable results. In such circumstances, cost must be the determining criterion." Future information retrieval system evaluations must include the users and their specific requirements. Again from Cleverdon[9], "...there has been the viewpoint of users that information services are for free; ...that when any economic difficulties arise, information services are expendable." The information industry must be able to show "...that investment in information services can be economically justified."

Standardization of information retrieval system software and hardware, and user training are indicated by Knox[10] as being essential to future development and cost effectiveness. He argues, "The mechanisms for transferring technological information have not changed appreciably in the past 30 to 40 years. They are not actually a system; rather, they are a mix of publicly-supported and privately-owned organizations, each pursuing

[8] B.C. Vickery, in Annual Review of Information Science and Technology, Volume 6, p. 138, Chicago, Illinois: Encyclopedia Britannica, Inc., 1971.

[9] C.W. Cleverdon, in Annual Review of Information Science and Technology, Volume 6, pp. 68-69, Chicago, Illinois: Encyclopedia Britannica, Inc., 1971.

[10] W.T. Knox, "Systems for Technological Information Transfer," Science, Volume 181, No. 4098, pp. 418-419, August 1973.

its own discipline or problem-oriented objectives." What changes have taken place (advent of technical reports, federal support, and computerization), while greatly improving performance, "...have also vastly complicated its structure and mode of use," Engineers and scientists have no more time to interact with an information retrieval system now than 25 to 30 years ago - and new users seem to have even less time. Yet the volume of information available to current users has increased roughly 16 times since World War II. Nothing short of a highly computerized and automated system even has the potential of matching users' needs with this growth in information. Knox[10] recommends that to improve a user's ability to interact with an information retrieval system, "...the following steps are desirable:

1) Specific training for all persons in high school, college and adult school in the technological information system...

2) A broader, more intensive training of information professionals...

3) Much greater standardization of components...

4) Greater reliance on pricing for full cost recovery..."

One of the points made by Kochen[11], in summarizing comments made by panelists during a meeting of the Special Interest Group of the ASIS on Behavioral Sciences, was that fund priorities should be changed. "More money should be devoted to make data more usable...even...at the expense of decreased funding for collecting and disseminating data. Collection, organization..., and maintainence of data bases is...necessary... but not sufficient." His summary of the panel's discussion goes on to say that "While we do not yet know as much as we need to about structuring, updating, and accessing very large files, great expenditures toward this end will yield but little gain relative to the gains that the same investment would produce in better ways of synthesizing data, of making it more directly usable in problem solving."

Information retrieval systems, manual or automatic, to be effective must be readily accessible, user oriented, and capable of satisfactory recall with precision and relevance. Automatic, computerized systems with portable terminals and dial-up access provide a very convenient inteface

[11] M. Kochen, "On the Economics of Information", Journal of the American Society for Information Science, Volume 23, No. 4, pp. 281-283, July-August 1972.

between user and information sources to the extent that, with the exception of limited or even personal information files, computerization of information retrieval systems is inevitable. Two interesting studies on small information retrieval systems point out the need for limited-scale manual systems. Graves and Helander[12] in a comparative study between manual and automatic information retrieval systems found manual systems to be favorable over automatic for a limited application data base using Petroleum Abstracts. Jahoda[13] cities similar results for small (10,000 documents or less) information retrieval systems.

The criterion for choice between manual and automatic indexing and retrieval seems to boil down to cost effectiveness. Lancaster[14] concludes that after considering the variables and "trade-off comparisons" of cost between the two types of systems, final judgments must still be "tempered with common sense." He cites the words of Congressman Melvin Laird[15]: "We should not allow cost effectiveness to cost us our effectiveness to cost us our effectiveness."

In conclusion, the state of the art indicates that computer-based natural language text processing systems are not only effective performance-wise but also in terms of cost effectiveness for very large data bases. Further advances in increasing user and user-system effectiveness are seen as an area requiring further work. Considering the FTD application, the CIRCOL system is appropriate, and the effort undertaken to discover ways to optimize retrieval efficiency and system efficiency is well worthwhile.

[12] R. W. Graves, Jr. and D. P. Helander, "A Feasibility Study of Automatic Indexing and Retrieval", IEEE Transactions on Engineering and Speech, Volume EWS-13, No. 2, pp. 58-59. September 1970.

[13] G. Jahoda, Information Storage and Retrieval Systems for Individual Researchers, New York, N. Y.: Wiley-Interscience, 1970.

[14] F. W. Lancaster, "The Cost Effectiveness Analysis of Information Retrieval and Dissemination Systems", Journal of the American Society for Information Science, Volume 22, No. 1, pp. 12-27, January-February 1971.

[15] M. Laird, quoted in Missile/Space Daily, p. 161, 7 April 1964.

1.3 PLAN FOR STUDYING THE CIRCOL SYSTEM

The plan for studying the CIRCOL system was formulated to incorporate four interrelated phases. Specifically, Phase 1 was concerned with those system and search strategy characteristics which affect search response times. It was desired to isolate and identify those factors which cause long response times and to establish techniques for improving response times.

Phase 2 involved studying actual use of the CIRCOL system by four user categories. In this phase the objective was to determine any significant differences in user searches of the system, both in terms of search content and in terms of variations in patterns of search strategy formulation techniques. Also, it was desired to determine if significant changes in user behavior were occurring over a period of time.

Phase 3 incorporated the findings of the first two phases to suggest possible file structure modifications to optimize the retrieval from CIRCOL, especially with respect to improved search response times. Also, the results of the first two phases led to a Supplement to the CIRCOL Users' Guide. This supplement suggests search strategy formulation techniques within the user's control by which he can optimize search results from CIRCOL. A copy of the CIRCOL Users' Guide supplement is presented as Appendix A to this report.

Phase 4 involved the investigation of the IBM Storage and Information Retrieval System (STAIRS) software package with a model data base. The purpose of this phase of the work was to consider the suitability of STAIRS for the FTD application and to compare STAIRS with DPS in terms of system capabilities and features, response times, and ease of use.

Throughout the work, the underlying philosophy was to study the system and the user population under actual operating conditions, so that realistic data would be obtained. It was recognized that under actual operating conditions, certain uncontrolled variables, for example, queuing of searches to access various files, would occur. However, since this situation does occur normally, typical operating conditions should have been encountered over the duration of the experimentation with CIRCOL.

SECTION 2

THE CIRCOL SYSTEM

2.1 CONTENTS

The Centralized Information Reference and Control On-Line system was established to meet the intelligence information needs of the scientific and technical intelligence community and Government research and development scientists and engineers. The data base currently contains approximately 850,000 scientific and technical documents with new documents being added at a rate of 15,000 to 20,000 per month. These documents consist of references of scientific and technical research and intelligence interest derived from open literature and intelligence documents.

Open literature includes documentation from sources which have significant scientific and technical content. The information originates outside the United States, and includes material from the USSR, East Europe, and Pacific area communist countries, as well as Western World countries. Open literature sources include periodicals, irregular serials, dissertations, monographs, textbooks, reviews, newspapers, and trade literature.

Intelligence information reports are produced throughout the intelligence community and contain information of scientific and technical significance. Semi-finished and finished intelligence reports consist of working papers or intelligence contractor reports, and other finished intelligence documents which have scientific and technical significance.

All documents entered into the CIRCOL system are converted to a standardized format consistent with the Data Base Description (DBD) which has been established for CIRCOL. The DBD provides both bibliographic and textual elements. Bibliographic elements provide identification and characterization data about each document. They are fixed-length fields. Textual elements provide information on the subject content of the document, the author and/or personalities concerned with the document, the document source, translation availability, releasability restrictions, and complete security classification data. Textual elements are composed of variable-length fields.

2.2 RETRIEVAL

Document references can be retrieved from the data base by specifying text derived search terms, and, at the user's option, by bibliographic or descriptive data in a number of different combinations. Generally searches are performed by specifying the subject content and/or author or personality desired. CIRCOL searching is accomplished by two modes of operation -- the search mode and the qualification mode. In the search mode, the variable length fields (textual elements) are searched.

Searches can be qualified by specifying such items as the country of information, date of information, subject code, and classification, which are derived from fixed length fields (bibliographic elements); these operations are referred to as the qualification mode. Qualification can only be specified subsequent to specifying search terms. Qualification specifications have the effect of further limiting and refining the search output. For example, a search on steel manufacturing could be restricted to include only those documents more recent than the beginning of the calendar year 1972.

Retrieval fields which are available for searching (search mode) are as follows:

1. Subject matter content (derived from title and text).
2. Authors/personalities.
3. Country codes.
4. Releasability restrictions.

Retrieval fields which are available for qualification (qualification mode) are as follows:

1. Country of information origin (except for USSR)
2. USSR as the country of information origin
3. COSATI Subject Code[*]
4. Date of Information
5. Type of Document
6. Security classification
7. CIRC Accession Number
8. Publication Country
9. DPS Number

2.3 THE COMPUTER SYSTEM

The basic information storage and retrieval system in use by the Foreign Technology Division (FTD) is the Document Processing System (DPS), a proprietary IBM software package designed to operate under

[*] COSATI Subject Codes are a standard set of subject classification codes established by the Committee On Scientific And Technical Information (COSATI)

Operating System (OS) 360. Several modifications have been made to the basic software to accommodate specific features desirable for the FTD application. The system is currently operating on the IBM 360/65 computer located at the Foreign Technology Division, Wright-Patterson Air Force Base. IBM System 360 Document Processing is a natural language or free-text indexing system. Because of the importance of the operational characteristics of DPS for the understanding of the subsequent sections of this report, DPS is described in detail in Section 3. The reader is also referred to the CIRCOL USER'S GUIDE, publication FTD-MP-22-14-73, which describes the CIRCOL application from the user's viewpoint.[16]

[16] Anonymous, Central Information Reference and Control (CIRC) On-Line (CIRCOL) User's Guide, FTD-MP-22-14-73, Foreign Technology Division, Wright-Patterson Air Force Base, Ohio, July 1973.

SECTION 3

DESCRIPTION OF THE DOCUMENT PROCESSING SYSTEM

The Operating System/360 (OS/360) Document Processing System (DPS) works with the original documentary information and extracts text-contained keywords, at machine speeds, without human interpretation. This mechanical extraction process preserves the author's terminology in its original context. DPS notes the location of each text word with respect to its neighbors, each sentence in its paragraph, and each paragraph within its document. In addition, the system notes the number of times a given word appears in the entire file of documents, and identifies each word with every document in which it has appeared. The document requester can set up a search whose terms and the arrangement thereof are significant to him for retrieval. The system consists of a set of interrelated files which accpts natural language text in machine-readable form and processes it in accordance with the data base description and with certain text processing criteria.

DPS performs two major processing functions -- input document processing, and searching. In addition to these, there are several utility procedures which essentially aid in the maintenance of the principal processing functions. Associated with these functions of DPS, as well as with any information storage and retrieval system, are a host of decisions which the designer of the specific application must make. Specifically, he must decide which DPS options should be utilized. These selected options will characterize the final processing configurations.

3.1 INPUT DOCUMENT PROCESSING

Phase I of DPS input processing performs the initial processing of machine-readable text. This initial processing transforms the input text into intermediate data sets holding bibliographic information and extracted keywords, while exercising optional data editing. This preliminary editing is invoked by the system, at the option of the user, and results in: the removal or changing of special characters within or surrounding words in the input text; the deletion of words shorter than a user-specified length; the truncation of words exceeding a user-specified length; and the deletion of words in the text which also appear on a user-provided 'common word' list. The common word list removes semantically nonsignificant words such as prepositions, articles, and auxiliary verbs.

After the editing functions have been performed, the remaining words are compared with a controlled Dictionary File of words previously determined as acceptable to the system. Again, at the user's option, nonfound words that were neither edited out of the list nor found in the Dictionary are printed out. This completes Phase I of input document processing and presents to a lexicographer off-line decisions as to the disposition of these nonfound words.

Phase II of input document processing releases the intermediate data sets created by Phase I for automatic updating into the DPS. Three specific files are effected by Phase II: the Master File, the Dictionary File, and the Vocabulary File.

The Master File contains the bibliographic record data (fixed format fields) and the word position data and word identification codes corresponding to the variable length text data entered. In DPS, each allowable keyword has a corresponding word identification number which is used for internal processing. The Dictionary File contains the allowable alphanumeric keywords. In conjunction with each keyword, statistical data and an internal word identification code number are maintained in the Dictionary File.

The Vocabulary File consists of all the word identification codes corresponding to the Dictionary terms. For each word identification code, a list of document control numbers is maintained. Each document containing an allowable keyword in its text is recorded and identified with the appropriate word identification number in the Vocabulary File. Thus, the Vocabulary File is the heart of the retrieval system. It is an inverted index; that is, each index term is stored with its associated document identification (document control) numbers.

3.2 SEARCH PROCESSING

3.2.1 System Identification of Documents which Qualify in Response to a Search Strategy

The search processing facility of DPS enables a system user to locate, identify, and print out a listing of documents in the system which meet the specifications established by user-written search strategies. The strategies are written in near-English lanugage and can consist of either labelled Boolean or unlabelled Boolean logical statements.

The simplest search strategies are written so that the occurrence of search terms in the text of documents is sufficient to effect retrieval. This is called "document level" logic. An example is a search requiring documents on 'CORROSION AND ALUMINUM'.

In addition to strategies which select documents on the occurrence of a term or set of terms, DPS provides the ability to retrieve a document by the positional relationship of terms. That is, the searcher may require two (or more) terms to be within the same paragraph, the same sentence, within a specified number of words of each other, or adjacent. This is a strong and narrowing search capability, but requires that the search process address the Master File positional identifiers. An example of such a search is 'STRESS CORROSION' of 'ALUMINUM ALLOYS' requiring adjacent positions.

An example of a labelled Boolean search request follows. System reactions to the line entries are explained below the example.

```
$1  stress & corrosion(+1)
$2  aluminum & alloys(+1)
$3  $1 & $2
    if date ge 72
```

The line labelled $1 requires that the word STRESS occur adjacent to the word CORROSION in the same sentence. Line $2 requires that the word ALUMINUM occur adjacent to the word ALLOYS in the same sentence. Line $3 requires that the conditions of Lines $1 and $2 both be satisfied. These keyword conditions are specified using the search mode. After these keyword conditions are satisfied, the further restriction date is specified using the qualification mode. For the entire request, only those documents which meet the specified keyword conditions and which also meet the specified bibliographic data requirements (in this case, the date of publication) will be retrieved.

System processing of a search begins with an edit of the input statements. The keywords are located in the Dictionary File, from which pointers to the Vocabulary File and the word frequency are extracted. The word frequency is used to reorder the words in a given line to minimize data handling.

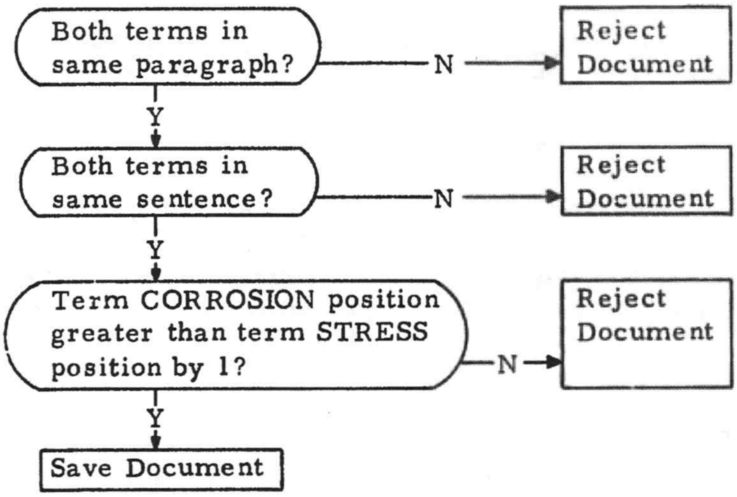
When all pointers to the Vocabulary File have been retrieved, a merging of the strings (or series) of document numbers for STRESS and CORROSION will result in a new string that contains only document numbers common to both the original strings (a logical AND operation). This new string now consists of a set of documents which contain both words but not necessarily in the correct position relative to each other. The logical AND process is also performed for ALUMINUM and ALLOYS. A final merging operation occurs to accommodate the statement $3 $1&$2; a new string results which consists only of those documents containing: STRESS, CORROSION, ALUMINUM, and ALLOYS. Up to this point, the word positions have not yet been considered by the system.

In the previous sample search, the merging of document-number strings produces a series of document numbers representing all documents that were retrieved simply on the basis of "document level" logic. The system then references the Master File and quickly narrows the search (and selection) by checking bibliographic data for a field named DATE having a value equal to or greater than 1972. If DATE in any retrieved document meets the specification (GE 72), the search programs then check the encoded text for the following context or positional conditions:

1. The document contains the word STRESS adjacent to the word CORROSION in the same sentence.

2. The document contains the word ALUMINUM adjacent to the word ALLOY in the same sentence.

In essence, the system checks for positional requirements in order of increasing specificity. It is already known that the words STRESS, CORROSION, ALUMINUM, and ALLOYS must appear in the same document. The next step is to check the positional relationships of STRESS and CORROSION, and ALUMINUM and ALLOYS. Therefore the system checks as follows:


Both terms in same paragraph?
N
Reject Document
Y
Both terms in same sentence?
N
Reject Document
Y
Term CORROSION position greater than term STRESS position by 1?
N
Reject Document
Y
Save Document

If the candidate document still qualifies, it is referenced for subsequent on-line or off-line display. This process continues for each document number in the final qualified list. The requester is advised of the number of documents satisfying the request.

3.2.2 Truncation Feature and Ancillary Utility Routines

An important feature of the search processing of DPS is the truncation or word stem function. A truncated form of the search term, denoted by suffixing the term with a dollar sign ($), will allow all terms in the Dictionary beginning with the same letters as the truncated search term to be used in the search. FTD has also instituted an automatic depuralization feature which treats plurals and singular forms as exact equivalents.

A series of ancillary utility routines are available with DPS which support and help to maintain the central functions of the system. These are essential to the effective operation of the system. A detailed description of these ancillary routines is available from the IBM Program Description and Operations Manual.

SECTION 4

CIRCOL SEARCH RESPONSE TIME

4.1 INTRODUCTION

Experience with CIRCOL by various user groups had shown that search response time can vary considerably. Search response times can be attributed to a number of factors, but up to the present study, no systematic approach had been taken to isolate, identify, and quantify these factors in terms of search response time. In accordance with the philosophy of performing the study within normal operating conditions, experiments were designed to test the effect on search response time of the various factors identified. One objective beyond merely discovering the effects of factors on search response time was to establish user-controllable techniques in search strategy formulation which could be employed to reduce search response times. The factors affecting search response time which were isolated and identified are as follows:

1. The specification of positional logic.
2. The use of the country code as a search mode term vs. use as a qualifier.
3. The use of the LRANGE specification command. (The LRANGE command is a special command which limits the extent of the file searched, by a specified range of DPS accession numbers).
4. The use of the truncation feature.
5. The use of DATE as a qualifier.
6. The use of the LRANGE command vs. the use of DATE as a qualifier.
7. The use of labelled Boolean statements vs. the use of unlabelled Boolean statements.
8. The running of one complex search strategy vs. formulating and running several logically equivalent search strategies.
9. The posting density.
10. The number of documents retrieved.

11. The number of users simultaneously on-line.

12. The order in which search lines are entered.

4.2 EXPERIMENTAL PROGRAM

In order to determine the effects of the various factors on search response times under actual operational conditions, an experimental program was devised. The experiments were designed with actual search strategies, usually under two basic conditions; namely, search strategies using only the search mode for retrieval and the same search strategies but modified with a standard set of qualification statements. With the set of search strategies employed for a given experiment, all user-controllable conditions were maintained constant except for the factor being studied. The period of time between the computer response message GOIN' SEARCHIN' and the computer response message //// (which occurs immediately prior to the message 'XXX DOCS SATISFY') was defined as the search execution time. The search execution time is the elapsed period of time required for DPS to carry out those functions required in response to the search strategy. The search execution time is a precisely defined term which is a measure of the search response time.

In order to describe the structure of the experiments, it is helpful to define some terms. The factor or variable being studied is one of the twelve factors described above. The value is the particular quantity assumed by the factor for a specific step of the experiment. The search strategy is that combination of specific search terms used to test the effect of the factor. The standard set of qualifiers is an arbitrarily selected set of qualification statements used to qualify the searches formulated to test the factors. The search execution time is the elapsed time from the GOIN' SEARCHIN' computer message and the //// computer message.

For example, in testing positional logic, positional logic is the factor. Its values are: (+1); (sen); (par); and logical AND (document level). A given search strategy is subjected to the various values of the factor being studied. The process is repeated for a number of search strategies. An entire experiment consists of all the selected search strategies being subjected iteratively to all values of the factor being studied.

In general, the structure of the experiments is as follows:

Experiment (Factor 1)

	Condition	Search Execution Time (min)
Search strategy #1		
Search mode only	Factor 1=Value 1	xx.x
Search + standard set of qualifiers	Factor 1=Value 1	xx.x
Search strategy #1		
Search mode only	Factor 1=Value 2	xx.x
Search + standard set of qualifiers	Factor 1=Value 2	xx.x
Search strategy #1		
Search mode only	Factor 1=Value n	xx.x
Search + standard set of qualifiers	Factor 1=Value n	xx.x
Search strategy #2		
Search mode only	Factor 1=Value 1	xx.x
Search + standard set of qualifiers	Factor 1=Value 1	xx.x
Search strategy #2		
Search mode only	Factor 1=Value 2	xx.x
Search + standard set of qualifiers	Factor 1=Value 2	xx.x
Search strategy #2		
Search mode only	Factor 1=Value n	xx.x
Search + standard set of qualifiers	Factor 1=Value n	xx.x
+Search strategy #n		
Search mode only	Factor 1=Value 1	xx.x
Search + standard set of qualifiers	Factor 1=Value 1	xx.x
Search mode only	Factor 1=Value n	xx.x
Search + standard set of qualifiers	Factor 1=Value n	xx.x

In formulating search strategies, the system imposes certain restrictions. Specifically, the specifying of the search mode parameters must precede the specifying of the qualification mode parameters; and the qualification mode cannot be used independently. Any given search line or string must consist either of logical AND or of logical OR expressions; logical AND's and OR's cannot be intermixed within a search line.

By using the LRANGE command, the user can specify the portion of the file to be searched. The STATIONS command permits the user to determine the number of on-line users. An example of a labelled search strategy with qualifiers is given as follows:

```
     $1 electronic($)                          (single term)
     $2 industr($), manufactur($),             (logical OR series)
        plant, facilit($)
     $3 $1 & $2 (SEN)                          (logical AND with
                                                positional logic)
     $4 if cntyussr eq y
     $5 if infocnty sc al, bu, cz, ge,
        hu, po, ru, yu
     $6 if date ge 60                          qualification
     $7 if classif lt 1                        statements
     $8 if subjcode sc 09,13,14
     $9 if $4 or $5
    $10 if $6 and $7 and $8 and $9
        !lrange 500000,749828                  lrange command
 GOIN' SEARCHIN'
                                               search execution
                                               time = xx.x min.
 ////
     480 DOCS SATISFY

 !stations(don)                                stations command to
                                               determine the number of
                                               on-line users

DATE = 24 OCT 1973 **** TIME =  13:30:51

   12 STATIONS SIGNED ON

S=32 L=3    S=13 L=4     S=20 L=5     S=02 L=6     S=01 L=7    S=03 L=8

IRC         WILSON -PR   GARNER       TUT          BRAWN       IRC

S=47 L=9    S=48 L=12    S=24 L=13    S=48 L=16    S=11 L=18   S=14 L=21

HOLLEB      MIRKIN       LASL-MYERS KRONK          CRISLIP     J. E. T.

   22 OFFLINE REQUESTS

      532 OFFLINE HITS
```

S = station number
L = telephone line number or hardwire line number

4.3 SPECIFICATION OF POSITIONAL LOGIC

4.3.1 Description

Positional logic refers to the relative positions of words as they occur in incoming text. As was discussed in Section 3, word positions are assigned by DPS; they are designated by: paragraph number; sentence number within the paragraph, and word number within the sentence. Word positions are assigned after deletion of the common words. The user can specify on retrieval the relative positions of words within the text by indicating the positional relationship appropriately. Following a logical AND requirement for two words, e.g., MISSILE & SYSTEM, the positional relationship is indicated by the designation (PAR), (SEN), (+n), (-n), or (±n), where n is a numeric value. These designations require the words to appear within the same paragraph, within the same sentence, within +n words in the same sentence, within -n words in the same sentence or within ±n words in the same sentence.

When the user imposes a positional requirement on the word pair, the system must access not only the Vocabulary File, which establishes the co-occurrence of the two words in the same document, but also the Master File, which contains the word positional data. One can readily infer that specification of positional logic requires more computer system manipulation and, hence, longer search response times. Other factors also are present which necessarily affect search response times. In particular, the posting density of the search terms, i.e., the number of documents associated with each of the terms, and the number of documents ultimately retrieved by the search are significant. Also, of course, the overall system activity with the inherent aspects of queuing, etc. certainly affect search response time.

4.3.2 Results

In the experimental design for positional logic, the specificity of positional relationships of the search words in the document were varied and the corresponding search execution times were measured. Both independent word pair searches and actual searches requiring positional logic statements were run with CIRCOL. Forty-five word pairs were tested with increasing specificity of positional logic A & B; A&B(PAR); A&B(SEN); A&B(±1) A&B(+1). As the specificity increased, the number of documents retrieved decreased. Search execution times for document level logic (Boolean AND) were about 25% less than for positional logic requirements. However, there was no detectable difference in search execution times as the degree of specificity increased from paragraph logic (PAR) to word adjacency logic (+1). Average search execution times were

1.00 minutes with positional logic and 0.75 minutes with simple AND logic. In other words, search execution times did increase somewhat for word pairs when positional logic specification required the system to access the Master File, but once this file was in use, increasing the specificity, which requires additional checking steps within the Master File, did not affect search response time.

Experiments were conducted with seven actual searches for which the positional logic was varied with increasing specificity: AND; SEN; ±1; +1. The set of seven searches contained an average of five positional logic statements per search. The searches were run as labelled searches both in the search mode only and with a standard set of qualifiers. (Labelled searches consist of search statements to which the user assigns a label, for example, in '$1 corrosion,' the designator $1 is the label.) The standard set of qualifiers applied was as follows:

```
$mm    if cntyussr eq y
$nn    if infocnty sc al, bu, cz, ge, hu, po, ru, yu
$oo    if subjcode sc (3 most probable for search subject)
$pp    if date ge 60
$qq    if classif lt 1
$rr    if $mm or $nn
$ss    if $oo and $pp and $qq and $rr
```

Composite results of these searches are shown in Table 1.

TABLE 1

Search Execution Time and Document Retrievals as a Function of Specificity of Positional Logic (Qualified and Unqualified)

Positional Logic Specified	Search Mode Only		Qualified	
	Search Execution Time (in minutes)	Number of Docs	Search Execution Time (in minutes)	Number of Docs
+1	5.1	350	5.1	236
±1	5.4	358	5.8	242
SEN	5.3	435	5.0	257
AND	5.0	533	5.4	309

4.3.3 Conclusions

From the results it can be seen that the specification of positional logic apparently does not greatly affect the average search execution time. There is a slight increase in search response time when positional logic is specified as contrasted with simple AND logic. Search response differences due to increasing specificity after positional logic is required are negligible. The results indicate that search response behavior is more erratic for qualified searches than for search-mode-only searches. Of course, qualified searches also require the Master File to be accessed to check the qualification requirements. Thus, qualified searches in which positional logic is also specified should result in the longest search response times. As the results show, however, search response times are essentially the same for all conditions.

It should be noted that as the positional specificity decreases, the number of documents retrieved increases, as would be expected. It is known from later experiments that the number of documents actually retrieved significantly affects search response time. Therefore, there is a trade-off phenomenon which provides some explanation for the near constancy of search response time when comparing the positional logic requirements with simple AND logic. Specifically, the search response time 'saved' as a result of formulating unqualified simple AND searches (not requiring access to the Master File) is offset by the additional time required for the system to deal with more documents retrieved.

This effect became particularly noticeable when the results for individual search strategies were examined. The situation which resulted by far in the longest search response time involved the retrieval of a large number of documents, even with the specification of qualifiers and highly specific (+1) positional logic.

Referring to Section 3 on the operation of DPS, one can understand why this phenomenon should be. Since the number of documents initially retrieved at the document level (simple AND logic) is large, there is a large number of documents which must be checked for qualification and/or positional restrictions when these requirements are specified.

In conclusion, it can be stated that search response time is little affected by qualification and/or positional logic because the system time required for additional processing is offset by the system time saved in having to deal with fewer documents retrieved. The longest search response times occur when many documents are retrieved in conjunction with positional logic and/or qualification specifications in the search strategy.

4.4 USE OF COUNTRY CODES IN THE SEARCH OR QUALIFICATION MODE

4.4.1 Description

The data base description and input rules for CIRCOL allow for the country of publication or the country of information to be available on retrieval either in the search mode or the qualification mode. Of course, the country of information or publication is very important to intelligence users of CIRCOL within the Foreign Technology Division and to users within other intelligence agencies. Since the majority of documents in the data base are derived from the USSR, a separate fixed field qualifier is allocated for Russian documents; the corresponding country code '-ur-' is not available as a searchable item in the search mode.

On retrieval by country code, the user can indicate that the country code must be in the search mode, that it must be a qualifier, or that retrievals must include Russian documents (or must exclude certain country codes or Russian documents). The search strategy specifications are given as follows:

```
$1 -al-, -bu-, -cz-, -ge-, -hu-, -po-, -ru-, -yu- (search mode)
if infocnty sc al, bu, cz, ge, hu, po, ru, yu  (qualification mode)
if cntyussr eq y  (inclusion of Russian documents)
```

4.4.2 Results

It was desired to determine experimentally the effect of specifying country codes in the search mode vs. the qualification mode on the search response time. Accordingly, experiments were designed so that the primary factors varied were the mode of country code specification (search or qualification mode) and the number of country codes. Country codes were specified as a logical OR series. Interactions of country code specifications and search complexity, number of retrievals, and term posting densities were also taken into account. Both qualified and unqualified searches were run. The country codes selected were East European codes as follows:

8 country codes

Albania	al	55 docs
Bulgaria	bu	3846 docs
Czechoslovakia	cz	9373 docs
East Germany	ge	9373 docs
Hungary	hu	6348 docs
Poland	po	9516 docs
Rumania	ru	5076 docs
Yugoslavia	yu	2969 docs
		45546 docs

4 country codes

Albania	al	55 docs
Bulgaria	bu	2836 docs
Czechoslovakia	cz	9373 docs
East Germany	ge	9373 docs
		21637 docs

3 country codes

China	ch	2915 docs
Czechoslovakia	cz	9373 docs
East Germany	ge	9373 docs
		21660 docs

2 country codes

China	ch	2915 docs
Czechoslovakia	cz	9373 docs
		12287 docs

1 country code

Poland	po	9516 docs

The composite results for four actual search requests are shown in Table 2.

TABLE 2

Search Execution Time and Document Retrievals as a Function of Country Code Specification (Search Mode vs. Qualification Mode)

Number of Country Codes	Search Mode		Standard Set of Qualifiers	
	Search Exec. Time (minutes)	Number of Docs	Search Exec. Time (minutes)	Number of Docs
Country Code in Search Mode				
8	4.3	115	4.6	16
4	3.5	64	3.5	6
3	3.7	93	2.6	11
2	2.3	65	2.5	8
1	2.5	27	2.3	3
Country Code in Qualification Mode				
8	6.9	121	5.8	16
4	5.7	67	5.5	6
3	7.5	93	4.6	11
2	5.7	66	5.1	8
1	4.2	29	2.6	3

8 country codes	al, bu, cz, ge, hu, po, ru, yu
4 country codes	al, bu, cz, ge
3 country codes	ch, cz, ge
2 country codes	ch, cz
1 country code	po

4.4.3 Conclusions

The results indicate that the number of country codes specified in a search affects the average search execution time. Especially when subject search terms are specified in the search strategy (no qualifiers specified) and the number of country codes is small, there is a decided advantage to specifying the country codes in the search mode. As the number of country codes (and, correspondingly, the number of postings) increases, there appears to be less and less advantage to using country codes in the search mode. For eight countries, it makes little difference whether the countries are specified in the search mode or the qualification mode. As the number of countries decreases, search execution times are significantly less for the search mode compared to the qualification mode.

When the search contains other qualifiers, there is less advantage to specifying country codes in the search mode. Nonetheless, if the number of country codes is small, search execution times are less, even if other qualifiers are present. It appears that if other qualifiers are present and the number of country codes is fairly large, it is actually more advantageous in terms of reducing search response time to specify country codes in the qualification mode.

The factors which control the response times in the specification of country codes in the qualification mode vs. in the search mode are the occurrence of other qualification requirements, positional logic, and the posting density. It should be noted that for searches which consist only of keywords with no positional logic, there is no need for the system to access the Master File. Thus, if country codes are specified in the search mode, the entire search can be run without reference to the Master File. However, the number of documents retrieved serves to increase the search response time, so the number of documents retrieved may tend to offset the advantage of not having to access the Master File.

Summarizing from the experiment with country codes used as search terms or as qualifiers, the following observations are made:

(1) For a single country code, even if fairly heavily posted, the search will run significantly faster if the country code is used as a search term rather than as a qualifier, whether or not other qualifiers are present.

(2) For several country codes (2-4), it seems advantageous to use the country codes as search terms if the search is otherwise unqualified. If other qualifiers are present, there seems to be little difference between country codes as qualifiers and country codes as search terms.

(3) It is generally a good practice to use country codes as search terms rather than as qualifiers unless many country codes (8 or more) are involved.

4.5 USE OF THE LRANGE COMMAND

4.5.1 Description

The LRANGE (Limit Range) command was introduced as a feature for CIRCOL by FTD to offer to the user the option of restricting his search to portions of the file. This capability permits the user to scan only recent material or various ranges within the file depending on the user's desires. By using the LRANGE command, the search execution time can be significantly reduced compared to searching the entire file. Also, the LRANGE command can be used to correlate with certain CIRCOL updates as will be explained in Section 4.6.

4.5.2 Results

The LRANGE experiment was performed to determine the effect of LRANGE on the search response time. Five searches, both qualified and unqualified, were run with LRANGE as the independent variable. The LRANGE command is used in conjunction with DPS numbers. The format of the command is:

```
!lrange xxxxxx,xxxxxx (x's = numeric characters of DPS
                       document accession numbers.)
```

If only one number is specified, it is the lower bound; the upper bound is the most recent (highest) number in the system. The actual size of the document file searched is obtained from the difference between the specified or default upper bound and the specified or default lower bound. It should be noted that since CIRCOL is based on inverted files, it is not completely accurate to refer to "size" of the document file searched; rather, the values of the DPS numbers in the inverted Vocabulary File are checked by the system and only those falling in the specified range are processed.

The composite results of the LRANGE experiment are given in Table 3.

TABLE 3

Search Execution Time and Document Retrievals as a Function of LRANGE value

LRANGE	Number of Docs Searched	Search Exec. Time		Number of Docs Retrieved	
		Search Mode Only(min)	Qualified (min)	Search Mode Only	Qualified
700000, 750000	50000	2.5	2.3	32	9
600000, 750000	150000	4.2	3.5	130	40
500000, 750000	250000	3.5	4.6	217	66
400000, 750000	350000	5.6	4.4	279	95
365000, 750000	385000	4.2	4.6	298	103

4.5.3 Conclusions

The search execution times show a significant decrease as the number of documents specified by LRANGE decreases. The functional relationship is not linear, however. The effect of LRANGE is definitely more pronounced in those ranges having the smaller numbers of documents. At about the point at which the document file size is greater than 200,000, the effect of LRANGE becomes much less. Again, one must recognize the competing factor of the number of documents retrieved. As the LRANGE size increases, the number of documents retrieved also increases. For those document ranges representing greater than 200,000 documents, the number of retrievals is probably the predominant factor contributing to search response time. Also, with larger LRANGE's, the number of documents is approaching the size of the entire file. Hence, LRANGE is not as effective when the range specified is fairly large.

In summary, we conclude the following:

(1) LRANGE is especially effective in reducing search time and the number of documents retrieved when the LRANGE specified is small. As the LRANGE

increases in size, other variables tend to offset the beneficial effect of LRANGE specification.

(2) LRANGE has a greater beneficial effect on the execution times of searches which tend to have long search times than on those which run quickly.

(3) LRANGE reduces the search execution times of unqualified searches more than the response times of qualified searches.

4.6 USE OF THE TRUNCATION FEATURE

The truncation feature of CIRCOL permits the user to specify a word stem for searching rather than having to enter an entire set of terms with a logical OR statement. Truncation requires that all Dictionary terms whose first n characters match a user-specified character string be searched as a logical OR group. The truncation designation is accomplished by appending a ($) symbol to the word stem specified. Thus, a user can enter work($) instead of: work, workability, workable, worked, working. It was desired to determine the effect, if any, of specifying a truncated form of a word in place of the equivalent logical OR series and in place of the most heavily posted single term of the series.

Ten actual searches containing truncated terms were run under three conditions: single term; truncated term; and logical OR series. These searches were run both with and without the standard set of qualifiers. An LRANGE of 500000, 749828 was specified. The composite average search times and document retrievals for the ten searches are summarized in Table 4.

TABLE 4

Average Search Execution Times and Document Retrievals for Various Truncation Specifications for Ten Searches

Specification	Search Time (minutes)		No. of Documents	
	Unqual.	Qual.	Unqual.	Qual.
Single term	0.8	1.0	64	38
Truncated term	1.7	1.5	100	46
Logical OR series	1.7	1.7	100	46

From these results it can be seen that there is essentially no difference in search execution time between the logical OR series and the truncated term, either in the qualified or the search mode. Apparently, the DPS software treats the truncation exactly as a logical OR series. The chief advantage of truncation is for the convenience of the user in formulating his search strategy.

In addition to searching by truncation and equivalent logical OR series, single Dictionary terms were run to determine the extent to which truncation affects search execution times and document retrievals. The single term represented the most heavily posted or the most likely term in place of the truncation. Referring to Table 4, it can be seen that truncation definitely increases search time and document retrievals compared to a single search term as would be expected. The search response time penalty imposed by specifying truncation is relatively small. In individual cases, however, the number of postings and number of terms in a logical OR series resulting from truncation may cause a significant increase in search execution time. For example, if one were to use the term CHEMISTRY, the number of document postings would be substantial. If, however, one were to specify CHEMI($), the document postings would increase dramatically, primarily due to the influence of the word CHEMICAL. Thus, the user must be duly cautious in using the truncation feature, both from the standpoint of search execution time and the possibility of obtaining numerous nonrelevant retrievals due to the truncation.

In summary, we conclude that the truncation is a useful feature enabling the user to specify a word stem search. The truncated form and the equivalent logical OR series result in the same search response time, but the truncated form is much easier for the user. Due care must be exercised by the user in specifying a truncated form to preclude inappropriate retrievals and excessive search run times.

4.7 USE OF DATE AS A QUALIFIER

4.7.1 Description

The CIRCOL system provides for qualification of searches by date. By date specification, the user can retrieve documents which are limited by the date of information. Thus, he can enter a qualification statement to his search strategy which will restrict the output to those documents which not only are retrieved by the search mode requirements but which also qualify according to the user-specified date or range of dates. The date of information command is given by the following formats:

if date eq 72	eq	:	equals
if date gt 70	gt	:	greater than
if date ge 71	ge	:	greater than or equal
if date lt 72	lt	:	less than

4.7.2 Results

The effect of specifying date ranges on the search execution time was determined over a number of specific date range values. Three actual searches were run with date values from 1960-1972. Searches were run both without additional qualifiers and with the standard set of qualifiers being applied. The composite results are shown in Table 5.

TABLE 5

Search Execution Time and Document Retrievals as a Function of the Value of the Date Qualifier

Date Value	Search Exec. Time		No. of Docs Retrieved	
	Date as Only Qualifier	Qualified by Std. Set	Date as Only Qualifier	Qualified by Std. Set
If Date GE 60	6.7	3.5	410	158
If Date GE 61	9.3	3.4	409	158
If Date GE 62	8.9	4.0	400	157
If Date GE 63	7.8	5.3	396	157
If Date GE 64	8.3	5.8	389	154
If Date GE 65	5.9	6.7	368	149
If Date GE 66	5.7	5.6	352	131
If Date GE 67	6.8	6.4	293	121
If Date GE 68	6.5	5.5	267	113
If Date GE 69	6.1	4.1	253	96
If Date GE 70	5.8	4.4	186	72
If Date GE 71	5.4	2.9	75	18
If Date GE 72	4.0	2.0	5	0

From the results it can be seen that the search execution times and the number of documents retrieved are higher when only the date is specified as a qualifier, than they are when the standard set of qualifiers is used. As with earlier experiments, this finding reflects the fact that the additional time required for the system to check the fixed format qualification fields in addition to the date field in the Master File is more than offset by the time saved in retrieving and handling fewer documents. The effect is particularly noticeable for high date values, which result in significantly fewer retrievals.

4.7.3 Conclusions

The results show that as the date value specified increases, the number of documents and the search execution times decrease. In other experiments it was learned that if the date is not specified at all and no other qualifiers are present, the search will run faster than if the date is specified. It should also be noted that, similar to the LRANGE, as the date value specified becomes lower, the number of documents which qualifies tends to approach the number which would have been retrieved from the entire file without any date qualification.

In conclusion, the use of the date as a qualifier from the standpoint of search response time is not very effective until the date range value specified is small. As the date value thus becomes more restrictive, the search response time is significantly reduced, primarily because the number of retrievals becomes correspondingly less.

4.8 COMPARISON OF THE DATE QUALIFIER AND THE LRANGE COMMAND

4.8.1 Description

As can be inferred from Sections 4.6 and 4.7, there is a certain correlation between the LRANGE command and the specification of date as a qualifier. The LRANGE command permits the user to limit the range of documents by DPS number, whereas the date permits the user to specify the date of information. If the user knows the dates on which updates were made and he knows the corresponding DPS numbers within the update periods, he can use the LRANGE command in effect to specify a date range. For example, if he knows that the DPS number corresponding to January 1972 is 650,000, then a search with the command: !lrange 650000 would include 1972 and 1973 documents, since 1972 documents obviously could not have been entered prior to the January 1972 update. It is possible, however, for documents whose date of information precedes 1972 to be entered subsequent to January 1972. To ensure that only 1972 documents would be retrieved, the user can specify both the LRANGE command and the date command in his search as in the following example.

```
OPTION CIRCOL MV, TEXT
surface & air(sen) & missiles(sen)
!lrange 650000
if date ge 72
end
```

Such a search represents the most efficient search of documents whose date of information is subsequent to 1 January 1972.

It is necessary to follow the mechanism of system operation to understand the usefulness of the LRANGE command and date qualifier used together. First, the LRANGE command initially limits the number of documents retrieved by DPS number in the inverted Vocabulary File. Then the date specification is checked in the Master File, but only for the limited number retrieved from the Vocabulary File. The same retrievals would occur by specifying the date without the LRANGE command, but then all documents retrieved in the search mode must be checked in the Master File to select only those which qualify by date. Hence, the search execution time necessarily would be greater for the latter search. This phenomenon was borne out by experiments.

4.8.2 Results

Experiments were run for four actual searches using date specification and using the equivalent LRANGE command. Both unqualified searches and the standard set of qualifiers were applied. The results are shown in Table 6.

TABLE 6

Search Execution Time and the Number of Documents
Retrieved for various dates and equivalent LRANGES

!LRANGE/ Date value	Search Execution Time		Number of Docs Retrieved	
	Search Mode Only (minutes)	Qualified (minutes)	Search Mode Only	Qualified
LRANGE 365000 if date ge 70	3.9 6.0	6.5 9.3	583 410	360 225
LRANGE 470000 if date ge 71	3.6 5.7	4.5 5.5	434 175	255 75
LRANGE 650000 if date ge 72	1.1 4.8	2.0 2.9	149 19	78 13

4.8.3 Conclusions

The LRANGE command is very efficient in reducing search response times. Qualification of a search by date can be most efficiently accomplished by specifying the date qualification value in conjunction with the use of the appropriate LRANGE command.

4.9 THE USE OF LABELLING IN SEARCH STRATEGIES

4.9.1 Description

The CIRCOL system provides for both labelled and unlabelled search strategies. Labelling can be applied by the user to the search lines which comprise his search strategy. The user assigns a label or line designation with a '$' notation followed by a number. The advantage of labelling is that line references can then be combined in various logical combinations just as search terms can be combined. The contents of each labelled line are manipulated simply by indicating the line reference number. Labels can be applied both in the search mode and the qualification mode. An example of an unlabelled strategy and its labelled equivalent are shown as follows:

UNLABELLED

```
 1 OPTION CIRCOLMV, TEXT
 2 radiation & shielding(+1)
 3 or heat & shield(+1)
 4 or supersonic & aerodynamics
   (+1)
 5 or laminar & boundary(+1)
   & layer(+1)
 6 or ablative
 7 if infocnty sc al, bu, cz, ge, hu,
   po, ru, yu
10 and classif lt 1
11 and subjcode sc 01, 11, 22
12 and date ge 60
13 or cntyussr eq y
14 and classif lt 1
15 and subjcode sc 01, 11, 22
16 and date ge 60
end
```

LABELLED

```
 1 OPTION CIRCOLMV, TEXT
 2 $1 radiation & shielding(+1)
 3 $2 heat & shield(+1)
 4 $3 supersonic & aerodynamics
   (+1)
 5 $4 laminar & boundary(+1) &
   layer(+1)
 6 $5 ablative
 7 $6 $1, $2, $3, $4, $5
 8 $7 if infocnty sc al, bu, cz, ge, hu,
   po, ru, yu
11 $8 if cntyussr eq y
12 $9 if classif lt 1
13 $10 if subjcode sc 01, 11, 22
14 $11 if date ge 60
15 $12 if $7 or $8
16 $13 if $9 and $10 and $11 and $12
end
```

4.9.2 Results

The effect of labelling on the search execution time was determined by running equivalent labelled and unlabelled searches, both in the search mode only and with the standard set of qualifiers. The composite results for five searches are shown in Table 7.

TABLE 7

Search Execution Time and Document Retrievals for Labelled and Unlabelled Searches

Status	Search Execution Time		Number of Docs Retrieved	
	Search Mode Only	Qualified	Search Mode Only	Qualified
Unlabelled	6.1	7.9	984	595
Labelled	6.1	7.2	984	595

The results for the average of five searches tend to indicate little difference between labelled and unlabelled searches. Examination of individual search data, however, show that labelling is distinctly advantageous. For searches which run quickly with a small number of retrievals, the labelled technique was clearly better, especially for qualified searches. Also, it is usually much easier for the user to formulate fairly complex search strategies using the labelled method. Examples of simple and complex strategies are shown as follows:

Simple

```
$1  weld($), join($)
$2  powder & alloy(sen)
$3  $1 & $2
end
```

Complex

```
 $1  aluminum & alloy(sen)
 $2  titanium & alloy(sen)
 $3  light & metal(sen) & alloy(sen)
 $4  $1, $2, $3
 $5  forming, extru($)
 $6  heat & treatment(sen)
 $7  grain
 $8  size, growth
 $9  $7 & $8(sen)
$10  $5, $6
$11  $4 & $9 & $10
 end
```

4.9.3 Conclusions

The use of labelling in search strategy formulation reduces search response time, especially for complex search strategies and for a relatively small number of retrievals. Use of labelling both for the search mode and the qualification mode generally results in shorter search response times.

4.10 COMPARISON OF ONE LONG SEARCH STRATEGY VS. SEVERAL EQUIVALENT SHORTER SEARCHES

Certain complex searches require considerable search execution times. It was desired to compare the running of one long search with equivalent multiple shorter searches. Accordingly, several searches were selected and run as follows: one search; two equivalent searches; three equivalent searches. The times for carrying out the searches were calculated as the summation of the search execution times. Also the required user times were added to the search execution times, since additional user time is required when more than one search is entered. Five labelled searches were performed with this methodology. Table 8 shows the composite results.

TABLE 8

Overall Search Execution Time as a Function of the Number of Equivalent Searches

No. of Equivalent Searches	Overall Search Execution Time (s.e.t.)	Search and User Time (u.t.)	Docs. Retrieved*
Single search	s.e.t. = 9.1	s.e.t. + u.t. = 12.0	588
Two searches	s.e.t. = 10.3	s.e.t. + u.t. = 14.1	596*
Three searches	s.e.t. = 9.4	s.e.t. + u.t. = 15.0	603*

* - Including duplications

The results indicate that the overall search and user time increases as the number of equivalent searches increases, whereas the summation of the actual search execution times remains essentially constant. With nearly constant search execution times, it is expected that the overall user and search time should increase as the number of equivalent searches increases, since a certain amount of time is required to enter the searches through the terminal. We conclude that one long search takes less overall time than equivalent shorter searches.

4.11 THE EFFECT OF POSTING DENSITY

4.11.1 Description

A limited number of experiments was performed to determine the effect of posting density, i.e., the number of documents associated with a given term, on the search execution time. One could logically infer that more heavily posted terms would result in longer search times, simply because the probability of obtaining a higher number of retrieved documents is higher. In the experiments performed, word phrases were used such that the document postings for the word phrase components in an experiment occurred in approximately the same ratio. For example, for a given experimental series, if the word phrase (A)THERMAL (B)CONDUCTIVITY were used, another word phrase, (C)LASER (D)COMMUNICATIONS would be selected such that the following relationships would hold true:

$$\frac{A}{C} = \frac{B}{D}$$

$$\frac{A}{B} = \frac{C}{D} = K \text{ (Constant)}$$

A, B, C, and D represent the posting densities for the respective terms. Similarly, other word phrases would be selected such that:

$$\frac{A}{C} = \frac{B}{D} = \frac{E}{F} \;\cdots\cdots\; \frac{M}{N} = \frac{O}{P} \cdots\cdots$$

$$\frac{A}{B} = \frac{C}{D} \;\cdots\cdots\cdots\; \frac{M}{O} = \frac{N}{P} \;\cdots\cdots\cdots\cdots\cdots\cdots\; = K$$

4.11.2 Results

The data for these experiments including the documents retrieved and the search times obtained are given in Table 9. Actual posting densities were obtained by running single word searches.

The results of the experiment indicate that the posting density on any given term is much less significant than the actual number of documents retrieved. For example, in one experiment the individual postings on both RADIATION and SIMULATION were high, but only 12 documents were retrieved, and the search execution time was only 0.5 min. On the other hand, the individual postings on OPTICAL and SIGHT are fairly low, but 59 documents were retrieved resulting in a search time of 0.6 min. In another experiment the individual terms of the phrases 'MISSILE TACTICS' and 'MAGNETIC TAPE' are posted within the same relative order of magnitude. However the documents retrieved amounted to 6 and 484 respectively for the two phrases. Corresponding search execution times were 0.9 and 3.2 min.

We conclude that the posting density of individual terms has some effect on the search execution time. However, the number of documents actually retrieved is a far more significant factor in determining search execution time.

4.12 THE EFFECT OF THE NUMBER OF DOCUMENTS RETRIEVED

4.12.1 Description

Considerable evidence had been accumulated to indicate that a major factor affecting search execution time was the number of documents retrieved on a search. In order to examine this factor, all of the data accumulated from our preceding experiments were extracted and ordered in terms of the number of documents retrieved, without regard to the specific experiment (positional logic,LRANGE, information country as a search term or as qualifier, etc.) from which the data was derived. Because of the variation in the number of documents retrieved for the individual experiments, the compilation of documents retrieved should provide a representative sampling of the various types of experiments run.

TABLE 9

Effect of Posting Density on Search Execution Time

Exp. No.	Word Phrase/Posting Density			No. of Docs Retrieved	Search Time (min)
1	optical & sight(+1)	8673	858	59	0.6
	laser & window(+1)	9707	992	8	0.3
	radiation & simulation(+1)	28937	2988	12	0.5
2	sun & visor(+1)	1750	55	15	0.4
	heat & vaporization(+1)	27559	962	57	0.6
	heat & shield(+1)	27559	988	71	0.6
3	beryllium & fluoride(+1)	1702	3682	14	0.3
	shock & tube(+1)	5480	11591	184	1.2
4	exhaust & gas(+1)	973	1516	271	1.7
	hypersonic & flow(+1)	1516	32878	108	1.1
5	missile & tactics(+1)	16234	1404	6	0.9
	magnetic & tape(+1)	24322	2252	484	3.2
6	microwave & amplifier(+1)	3031	5833	24	0.6
	grain & growth(+1)	5054	9679	132	1.6
	exhaust & value(+1)	1516	2934	38	0.5
7	discharge & tube(+1)	7157	11591	262	4.2
	power & plant(+1)	23388	38473	2520	10.8
	sensory & perception(+1)	417	696	3	0.3

4.12.2 Results

The search execution times were averaged corresponding to the number of retrievals; the number of document retrievals was expressed by ranges. The data were extracted with a differentiation between unqualified, qualified, and overall search results. The results are presented in Table 10.

The results definitely confirm the trend of longer search times as a function of the number of documents retrieved. Particularly as the number of documents retrieved exceeds 200, there appears to be an accelerating trend toward increasing search time as a function of the number of documents retrieved. For more than 200 documents the degree of scatter of data is considerable, although the upward trend of search execution time is clearly evident. There is no discernable difference in terms of search execution time between qualified and unqualified searches, probably because the additional system time required for checking the qualification statements is offset by the retrieval of fewer documents.

4.12.3 Conclusions

We conclude that the effect of the number of retrievals on search execution time is highly significant, particularly when the number of retrievals exceeds 200. In fact, the effect of a large number of retrievals apparently more than offsets the effect of qualification requirements and the complexity of search logic. Therefore, it is probably better to formulate search strategies initially to as specific a level as might be desired, regardless of the complexity of the strategy. Full use of applicable qualification statements should be made initially.

Appropriate use of positional logic does not affect search execution time deleteriously. The user should recognize, however, that if his positional logic requirements are too restrictive, he may fail to retrieve relevant documents. Sentence positional logic may be better for retrieval than adjacent word logic, as in the following example:

TABLE 10

Search Execution Time as a Function of the Number of Documents Retrieved

Number of Documents Retrieved		Search Execution Time (min)		
Range	Total No. of Searches	Search Mode only (A)	Qualified (B)	Weighted Average* of (A) & (B)
0-10	(66)	2.8	2.3	2.5
11-20	(55)	2.9	2.5	2.8
21-30	(37)	2.6	4.0	3.5
31-40	(27)	3.4	2.8	3.0
41-50	(18)	2.5	3.0	2.8
51-60	(24)	2.5	4.0	3.3
61-70	(14)	4.0	3.8	3.8
71-80	(13)	4.5	4.0	4.3
81-90	(18)	4.0	3.0	3.7
91-100	(11)	2.0	4.2	3.0
101-200	(71)	4.2	4.3	4.3
201-300	(54)	5.5	6.6	5.8
301-400	(37)	5.9	6.8	6.3
401-500	(22)	6.1	3.5	5.6
501-600	(15)	6.0	-	6.0
601-700	(8)	7.1	8.0	7.2
701-800	(5)	5.3	15.5	9.4
801-900	(0)	-	-	-
901-1000	(5)	5.8	3.2	4.8
1001-2000	(11)	6.6	9.6	7.3
2001-3000	(10)	7.0	12.0	9.0
3001-4000	(4)	10.0	-	10.0

* Average of (A) & (B) = $\frac{\Sigma A + \Sigma B}{\text{total no. of searches}}$

$1 adhesive & bonds(+1)

(The phrase 'adhesive bonds' must appear in the document to effect retrieval.)

$1 adhesive & bonds(sen)

(Less restrictive strategy permitting documents to be retrieved which express the concept differently, e.g., "testing of bonds cured at 300°C using Narmco adhesives.")

4.13 THE EFFECT OF THE NUMBER OF SIMULTANEOUS ON-LINE USERS

4.13.1 Description

In order to examine the effect of the number of users on-line simultaneously, the data accumulated from our previous experiments were extracted and compiled. The data were arranged in order according to the number of simultaneous on-line users. A representative sampling of experiment types was achieved in the same manner as was described in Paragraph 4.12.

In addition, a special experiment was conducted in which CIRCOL users were instructed to sign on and use the system at various time periods while we ran the same search with a progressive incremental increase in the number of users.

4.13.2 Results

The results of these experiments are shown in Tables 11 and 12. The results are quite interesting. It was somewhat surprising to observe that, on the average, the number of users signed on affects the search execution time to a relatively small degree. There was only about a one-minute increase in search time between six users and twelve users. Also, there was no observable difference between unqualified and qualified searches as far as the effect of the number of users signed on. A pattern of search execution times was observed; namely, that the longer the search tended to run, the greater was the effect of the number of users signed on.

TABLE 11

Search Execution Time as a Function of the Number of Users Signed on

Number of Users Signed On	Search Execution Time (min.)		
	Search Mode Only (A)	Qualified (B)	Weighted Average* of (A) & (B)
3	1.0	1.4	1.2
4	0.9	3.5	2.2
5	2.0	3.2	2.6
6	3.5	3.9	3.7
7	3.6	3.4	3.5
8	4.4	3.6	4.0
9	3.9	4.6	4.2
10	4.4	4.6	4.4
11	5.5	4.8	5.1
12	4.5	4.3	4.4
13	4.3	4.7	4.5
14	4.5	4.2	4.4
15	4.7	5.2	5.0
16	12.0	6.0	9.0
17	5.3	2.8	4.0

* Average of (A) & (B) = $\frac{\sum A + \sum B}{\text{total no. of searches}}$

The operational configuration of CIRCOL is such that five identical copies of key files and programs are maintained in the computer system. The effect of this operational configuration is that for up to five simultaneous users, each user can access CIRCOL without any queuing, since he is switched to a copy of the file or program which is not in use. When more than five users are on the system, the system must hold the search commands and search specifications in queue until the system completes processing of the searches already in progress and a copy of the program or file needed for the new search becomes available.

Obviously this queuing results in longer search execution times. However, because of the five copies of files and programs, the degree of queuing is nominal for up to about twelve users. The explanation for this can be seen by considering an analogy with teller windows at a bank controlled by a single waiting line. If there are five teller windows and only six people, the sixth person can take the next available window from five possibilities. If there are nine people in queue, however, there are four people "competing" for the next available position. Hence the overall waiting and transaction time of the ninth person is longer than that of the sixth person. Similarly, with twelve persons, the five windows are queued "two deep" with only two additional persons competing for the "third deep" queue position. Beyond twelve patrons, the queuing "competition" becomes greater, and the last person's total time in the bank tends to increase greatly.

In the controlled experiment, additional users entered and became active with system in increments of ten users and five users per time period as the day progressed. The University ran the same search under different conditions of numbers of active users. As can be seen from Table 12, it is apparent that the efficiency of the system decreases as the number of users increases. Although reasonable execution times can still be obtained with a high number of users, system response becomes much more erratic and less predictable, even for simple searches. For a simple search which runs rapidly, the search execution time was still less than two minutes with up to 15 users signed on. However, for 16 or more users, execution times ranged from around 4 minutes to over 13 minutes. Search times for 16 to 25 users seem rather random. One of the longest execution times (12.2 min.) was recorded for 15-18 users signed on, but the same search was able to run in approximately half the time (6.2 min.) with 25 users on-line.

TABLE 12

Search Execution Time as a Function of the Number of Simultaneous On-Line Users

Search #1

```
$1 sea & floor
$2 sea & bottom
$3 structure, installation
$4 $1,$2
$5 $3 & $4
if cntyussr eq y
and date ge 69
end     (70 docs retrieved)
```

Search #2

```
!lrange 500000,749828
$1 microstructure
$2 grain & growth(+1)
$3 $1 & $2
end          (15 docs retrieved)
```

No. of Users Signed on	Search Execution Time (min.)	
	Search #1	Search #2
1	-	0.3
2	-	-
3	2.0	0.4
4	-	-
5	-	0.7
6	-	0.5
7	4.3	0.4
8	5.9	2.0
10	7.7	-
11	5.0	0.8
12	5.3	1.3
13	10.0	-
14	5.6	4.5
15	10.0	0.9
16	11.1	9.3
17	15.1	12.2
18	7.6	11.1
19	14.0	4.1
20	9.3	4.9
21	5.9	4.1
22	8.6	-
23	12.2	13.5
24	12.3	6.8
25-29	24.5	7.1

- = No data available

Similar effects were noticed for another search, which was somewhat more complicated. This search always required at least 4 minutes, except for one instance where only 3 users were signed on. Again, search times tended to stay within reasonable bounds for up to around 14-16 users; then the search execution times became unpredictable.

The erratic response of the system with a high number of users is not surprising. Search time depends on many factors, such as queuing time for accessing certain files. When many users are signed on, these factors have a great effect on search execution time, but the effects are less predictable.

An explanation for the erratic search response times can be derived by referring once again to the bank teller window analogy. Anyone who has been in a bank knows that the actual transaction time can vary from a few seconds to many minutes. The transaction time of the people in front of a person in the queue greatly affects his own waiting and transaction time, irrespective of how short his own transaction time may be. Some days people seem to be making deposits in established accounts, and many people are accommodated quickly. Other days it seems that everybody ahead of you is performing time-consuming multiple account transactions or opening new accounts, and your overall waiting time seems interminable, even though your transaction can be accomplished quickly.

Another factor which must be considered, regarding search response time with a large number of users, is the system resources which must be utilized simply to keep track of the queuing and switching. Those resources required for serving as a "traffic cop" cannot be used for processing. Hence, system efficiency degrades rapidly and search response times increase exponentially beyond a certain critical number of on-line users.

From these data we conclude that it is best to avoid running searches, especially complex ones, when many users are signed on. Random variables seem to become much more significant when 15 or more users are active.

We conclude that the number of users signed on affects search execution time in an exponential relationship as described above; the effect is rather small for a small number of users, but becomes important beyond a critical number of users. Of significance is the fact that the longer the search run-time tends to be, the more deleterious is the effect of the number of users logged on. For relatively short, simple searches, the effect is quite small. As the number of users becomes quite large, the system performs very erratically, generally with quite long search execution times.

4.14 THE EFFECT OF THE ORDER OF SEARCH STATEMENT INPUT

4.14.1 Description

A factor which had not been addressed previously was whether the order of entering search statements would affect search execution times. It has been shown that the number of documents ultimately retrieved has a definite effect on the search execution time. An unknown factor was the effect, if any, of the number of documents retrieved by the individual search statements and the location of the search statement within the strategy. To test the possible effect of search statement order in the strategy, an experimental series of searches was designed. A search strategy generally consisted of some number of individual search statements such that some statements alone would result in many retrievals, whereas other individual search statements would result in relatively few retrievals. A logical OR series of heavily posted terms would result in many retrievals:

$1 measur$, tests, testing, analy$

A logical AND search would normally result in rather few retrievals:

$1 radar & station(+1)

Our hypothesis was that for a particular search strategy, the individual search lines should be entered in the order of increasing number of anticipated retrievals.

Other factors explored were the effects on search execution time of the posting density of individual terms, and variations in formulation of equivalent search strategies with different phrases and logical combinations.

4.14.2 Results

A set of searches was run according to the following pattern:

A. Increasing Posting Density:

$1 Least densely posted term
$2 Next least densely posted term
$3 Most densely posted term
$4 $1 & $2 & $3

B. Decreasing Posting Density

$1 Most densely posted term
$2 Next most densely posted term
$3 Least densely posted term
$4 $1 & $2 & $3

A variation of the search patterns indicated above included the use of two terms with intrastring OR logic, within a labelled search line, such that Pattern A and Pattern B were run with ORed term pairs instead of individual terms. Table 13 shows the average results for six searches.

TABLE 13

Search Execution Time as a Function of Posting Density Order of Search Statements within the Search Strategy

Condition	Posting Density in LRANGE	Search Execution Time	Documents Retrieved
Increasing Posting Density	Term(s) 1 = 2869 Term(s) n = 12861 (n = 2, 3, 4)	1.5	258
Decreasing Posting Density	Term(s) n = 12861 Term(s) 1 = 2869 (n = 2, 3, 4)	1.3	258

The results of this experiment indicate that the order of the search terms within the search strategy does not affect the search execution time, whether the terms are entered in order of increasing or decreasing posting density. This finding holds true for individual terms and for small logical OR groups of terms.

A second experimental series was run to determine if the order of a heavily posted large logical OR group within the search strategy would affect search execution time. Search patterns were run such that the posting density of the search statement with an individual term (I. T.) was less than the posting density of the search statement containing the OR series (Σ OR); also searches were run with the posting density of the I. T. greater than Σ OR.

	Type 1 (I. T. first line)		Type 2 (Σ OR first line)
	$1 Individual term (I. T.)		$1 Logical OR series
A	$2 Logical OR series (Σ OR>I. T.)	B	$2 Individual term
	$3 $1 & $2		$3 $1 & $2
	$1 I. T. (I. T. > Σ OR)		$1 Σ OR
X	$2 Σ OR	Y	$2 I. T. (I. T. > Σ OR)
	$3 $1 & $2		$3 $1 & $2

The results are presented in Table 14.

From these results we conclude that the Σ OR series does affect search execution time significantly, depending on its order within the search strategy, regardless of whether the posting density of the Σ OR series is greater or less than the posting density of the individual term (or small OR series). It is definitely advantageous to place the long OR series (Σ OR) as near to the end of the search as possible. An explanation for this is that for a long OR series, the Dictionary file must be accessed a significant number of times (once per term). Even when the actual total number of postings in the Vocabulary file is small for a long OR series, the requirement for many points of access to the Dictionary file more than offsets the advantage of low posting density. The generation of a subset of documents from the file based on a Σ OR series appears to be one of the least efficient DPS internal processes.

It should be noted that if the individual term is listed first, an initial subset of documents is created corresponding to the document postings for that term against which the Σ OR series is matched, thus drastically reducing the number of postings derived from the Σ OR series which must be maintained for further search processing. To illustrate, let us consider I. T. = A and Σ OR = B, C, D, E with the corresponding posting densities a, b, c, d, e; used in the following strategy:

$1 = A
$2 = B, C, D, E
$3 = $1 & $2

A is entered as the first search statement in the search; it has a posting density of 'a'. B, C, D, and E are entered as the second search statement. Posting densities b, c, d, and e are then

TABLE 14

Search Execution Time as a Function of Posting Density and Order of Search Statements within the Search Strategy

Condition	Posting Density	Search Time	Documents Retrieved
Type 1 1. T. = first line;I. T. Σ OR (avg. of 8 searches)	I. T. = 2498; Σ OR = 15061	2. 8	944
Type 2 I. T. = second line; I. T. <Σ OR (avg. of 8 searches)	I. T. = 2498; Σ OR = 15061	4. 0	944
Type 1 I. T. = first line;I. T. >Σ OR (avg. of 5 searches)	I. T. = 5480; Σ OR = 1564	1. 7	93
Type 2 I. T. = second line; I. T. >Σ OR (avg. of 5 searches)	I. T. = 5480; Σ OR = 1564	2. 2	93

carried in the search only to the extent that 'a' intersects with b, c, d, or e symbolized as follows: $a \cap b$, $a \cap c$, $a \cap d$, and $a \cap e$. If B, C, D and E are entered first, b, c, d, and e must be carried in full until the matching operation with 'a' occurs.

A third experimental series was run to determine the effect of the order in a search strategy of a phrase having a lower posting density than other elements of the search strategy, even though the individual phrase components themselves are the most heavily posted terms in the search. Stating this situation symbolically, the following search strategy is representative:

Strategy		Conditions
\$1 A & B(+1)	\$1 C, D	a & b(+1) < c, d
\$2 C, D	\$2 A & B(+1)	a > c, d
\$3 \$1 & \$2	\$3 \$1 & \$2	b > c, d

Table 15 shows the results for five searches.

For comparison purposes, the search execution times corresponding to the individual words and word phrases were also determined. The great effect of the number of retrievals on search execution time is dramatically evident; this result confirms the phenomenon reported previously. Regarding the effect on search execution time of the order of a word phrase whose posting density is less than that of individual terms or a short OR series, even though the posting density of the phrase components is high, there is virtually no effect.

A fourth experimental series was run to confirm our findings on posting density, the order of OR series within the search strategy, and the use of positional logic. Specifically search patterns of the following type were run:

TABLE 15

Search Execution Time as a Function of the Order of Search Statements within an 'a&b(+1)' Search Strategy vs. a 'c,d' Search Strategy

Condition	Posting Density	Search Time	Documents Retrieved

a&b(+1) = first line
a&b(+1) < c, d
a > c, d; b > c, d;
a > b
	a&b(+1) = 200
	c, d = 1811
	a = 7847
	b = 1958
		1.0	23

c, d = first line
a&b(+1) < c, d
a > c, d; b > c, d;
a > b
	a&b(+1) = 200
	c, d = 1811
	a = 7847
	b = 1958
		1.2	23

Individual terms, phrases			
a&b(+1)	200	2.3	200
c, d	1811	3.9	1811
a	7847	10.1	7847
b	1958	4.3	1958

Type 1

```
$1  A
$2  B, C, D, E
$3  $1 & $2
```

Type 2

```
$1  B, C, D, E
$2  A
$3  $1 & $2
```

Type 3

```
$1  A & B
$2  A & C
$3  A & D
$4  A & E
$5  $1, $2, $3, $4
```

From our previous results we predicted that search efficiency should occur in the order: Type 1/Type 3/Type 2. Experimental results are shown in Table 16.

TABLE 16

Effect of Various Search Strategy Formulation Techniques on Search Execution Time (Average for 13 searches)

Type	Search Time	Docs Retrieved
Type 1	2.4	617
Type 2	3.7	617
Type 3	2.6	617

As can be seen, the experimental results do indeed corroborate our prediction, and thus confirm our previous findings. It is interesting to note that the difference between Type 1 and Type 3 is small. If OR consists of very many terms, the search efficiency order possibly could be changed to Type 3/Type 1/Type 2.

4.14.3 Conclusions

The results indicate that search statements should be entered in the following order:

(1) Search statements with long OR series are entered last.

(2) Search statements resulting in few document retrievals should be entered first.

(3) Search statements with positional logic specifications should be entered early in the search strategy.

Ordering a search strategy in increasing posting density order for individual terms or for short OR series is slightly helpful in improving search processing efficiency. However, when a long OR series is introduced, a highly significant phenomenon occurs which overshadows the effect of posting density. The efficiency of the search strategy can be dramatically improved by placing the longest OR series as far down in the search statements as possible. In considering the effect of various factors on search execution time, the two most significant found are (1) the number of terms in an OR series and (2) the overall posting density. Obviously the two factors are interrelated, since longer OR series would tend to have higher overall posting densities. By appropriate positioning of search lines within a strategy, significant reductions in search execution times can be effected.

4.15 SUMMARY OF FACTORS AFFECTING SEARCH RESPONSE TIME

A number of factors are present in the CIRCOL system operation which can affect the search execution time. These factors are dependent on the CIRCOL system itself, on the number of active users of the system at any given time, and on the types of commands and search operations which the system must process. The CIRCOL system is composed of a number of files and processing steps. A detailed description of DPS/CIRCOL is given in Section 3.

Many factors affecting search response time are "system bound", i.e., the computer programs and file structures control the specific mechanisms by which the system processes the searches which are submitted to it. The number of users actively utilizing the system is also a factor over which any given user has no control. However, the user can control the manner in which he enters searches into the system, and efficiency in searching can thus be enhanced. A summary of CIRCOL system factors and their effect on search response time is given as follows:

1. Positional logic - The specification of positional logic (word adjacency, co-occurrence of words in the same sentence or paragraph) has almost no noticeable effect on search response time. The reason is that specifying positional logic reduces the number of retrievals, thus establishing a tradeoff in terms of response time, since the additional time required by the system to check the Master File is offset by the system not having so many retrievals to handle.

2. Country codes - The entering of country codes in the search mode generally results in reduced search execution time. Country codes should be entered in the qualification mode only if other qualifiers are present and the number of country codes and associated postings is high.

3. !LRANGE - The LRANGE command, which limits the range of document numbers searched, is highly effective in reducing search response time. The LRANGE command takes effect early in the search process in the Vocabulary File, thus saving search time in subsequent processing.

4. Truncation - The use of the truncation feature has the effect of creating a logical OR series. There is no difference in search response time between entering the truncated form of a word stem and the equivalent logical OR series. The user should recognize, however, the fact that he is actually entering a logical OR series and he should be sure that he does not include undesired terms in the logical OR expression.

5. DATE as a qualifier - The date of information can be specified as a qualification statement. As the date range becomes narrower, the search execution time is reduced at a faster rate, primarily due to the retrieval of fewer documents.

6. !LRANGE and DATE - The use of the LRANGE command in conjunction with the date, when the date is an important retrieval specification, represents the most efficient means of searching resulting in lowest search response times. A table of updates and the date when each update was made is given in the Supplement to the CIRCOL Users' Guide (See Appendix A).

7. Labelling - Labelling of search statements in a search strategy generally results in more efficient searching with lower response times.

8. One long strategy vs equivalent shorter strategies - The running of equivalent shorter strategies does not offer any advantages to the user. More user time at the terminal is required, and if one of the shorter strategies caused a large number of retrievals, overall search execution time would also be greater.

9. Posting density - More heavily posted terms cause longer search execution times, both because more document numbers must undergo processing, and because the number of retrievals tends to be greater.

10. Number of documents retrieved - The number of documents ultimately retrieved by a search greatly affects the search response time. The user should exercise care to avoid strategies for which it could be anticipated that many retrievals could occur. Judicious use of the LRANGE feature and consideration of the word frequency and document frequency listings in the CIRCOL Dictionary should be helpful.

11. Number of users on-line simultaneously - As the number of users actively using the system increases, search response times increase rapidly. With many users (more than 15) search response times become quite erratic.

12. Order of entering search lines - Greatest efficiency is achieved by relegating long logical OR series towards the end of the search. The number of documents anticipated to be retrieved should govern the order of the other search statements in the strategy. Search statements should be placed in order from the least number of documents anticipated to the largest. Also, search statements should be placed in order of term posting density. It should be noted, however, that the result

of a logical AND statement of two fairly heavily posted terms may result in few documents retrieved. Thus such a statement should precede a single term which, by itself, would result in more documents retrieved.

4.16 PREPARATION OF THE SUPPLEMENT TO THE CIRCOL USERS' GUIDE

Based on our findings of the effect of various factors on the search execution time, it was evident that a number of these factors can be controlled by the user to reduce search execution time. Also, certain interactive techniques are available to aid the user in ensuring that the search strategy employed is appropriate for obtaining relevant retrievals. In order to disseminate this information to users of the CIRCOL system, a supplement to the CIRCOL Users' Guide was prepared. This supplement was entitled "Procedures for Optimizing Search Results from CIRCOL". The supplement is presented in this report as Appendix A.

SECTION 5

ANALYSIS OF CIRCOL SEARCHING BY USER TYPE

5.1 INTRODUCTION

The CIRCOL system is used extensively every day by a number of organizations. Since a primary mission of the Foreign Technology Division is to fulfill the intelligence and research information needs of the Department of Defense, many military agencies and organizations all over the country perform searches of the CIRCOL data base. CIRCOL covers foreign open literature sources as well as classified materials. CIRCOL contains in its data base the most comprehensive coverage of foreign scientific and technical literature available in this country. In order to make this resource of technical information available for nonmilitary applications, CIRCOL is also available to certain civilian Government agencies and to Government contractors.

An important aspect of the study performed by the University of Dayton, was the use made of CIRCOL by various types of users as well as by the entire user community. By analyzing the ways in which CIRCOL is used by various types of users, certain implications for optimizing CIRCOL should be revealed, both in terms of user-controllable factors and in terms of changes which could be made in file structures and internal system operations.

After the organizations using CIRCOL were reviewed, four major categories were selected for analysis:

Type 1: FTD Information Specialists
Type 2: FTD Intelligence Analysts
Type 3: Outside R&D Organizations
Type 4: Outside Intelligence Agencies

5.2 PROGRAM FOR ANALYZING SEARCHES BY USER TYPE

The program for the analysis of searches by user type took into account the following factors:

1. Significant differences (if any) in searching by user type in terms of search content and search strategy patterns.

2. Changes (if any) in searching over a period of time by user type.

3. Specific retrieval parameters actually used in search strategies.

4. Problems encountered and search results obtained by users in interacting with CIRCOL.

The source data for analysis was derived from copies of off-line printouts of actual CIRCOL searches run; some on-line terminal records were also provided for analysis. In analyzing the data, only the search strategy itself was examined. We did not look at the actual documents retrieved to determine relevance or recall, since such an analysis would have been beyond the scope of our study.

The analysis of user search behavior was primarily statistical in nature. Search strategy patterns were characterized by categories as follows:

1. Searches using interstring OR logic to connect lines.

   These searches usually employ (+1) logic between individual terms, creating multi-word phrases which are then ORed together. Sometimes, however, a line can consist of a single term. Searches of this pattern are usually labelled.

   Example:

   ```
   $1  automatic & machining(+1)
   $2  bearing & manufacturing(+1) & equipment(+1)
   $3  synthetic & diamond(+1) & manufacturing(+1)
   $4  electronic & manufacturing(+1) & machinery(+1)
   $5  laser & machining(+1)
   $6  laser & metal(+1) & cutting(+1)
   $7  $1, $2, $3, $4, $5, $6
   end
   ```

2. Searches with lines connected by interstring OR, then combined to one line by interstring AND.

   This type of search requires the presence of the elements of one line (either single term, country code, or intrastring AND terms) with any of another group of lines.

Example:

```
$1  field & artillery(+1)
$2  target & acquisition(+1)
$3  command & control
$4  ammunition, system
$5  $2, $3, $4
$6  $1 & $5
end
```

3. Search lines connected in a logical OR series, and OR series combined with logical AND to one search line containing country codes.

Example:

```
$1  electronic & countermeasure(+1)
$2  ecm
$3  jamming, confus($)
$4  $1, $2, $3
$5  -ch-, -cz-, -ge-, -hu-, -po-
$6  $4 & $5
$7  if cntyussr eq n
$8  if subjcode sc 01, 09, 17
$9  if $7 and $8
end
```

4. One-line keyword searches.

This type of search may be a single word, or several words joined by AND, +1, SEN, or PAR logic. These searches are seldom labelled, but are frequently qualified.

Example:

```
japan & atomic(+1) & energy(+1) & research(+1)
       & institute(+1)
if infocnty sc ja
end
```

5. Searches using interstring AND to connect intrastring OR lines.

These searches are frequently labelled, but positional logic is seldom used. The following search is typical of this category:

```
$1  epoxy, polyimide
$2  resin, polymer
$3  composite, laminate, reinforce($)
$4  $1 & $2 & $3
end
```

6. Searches using interstring AND to combine single terms or intrastring AND lines.

These searches are usually short, seldom having more than 3 or 4 words. Sometimes positional logic is used, but most of the searches use simple AND logic.

Example:

```
$1  carbon & fiber
$2  metal
$3  $1 & $2
end
```

7. Any search using a logical NOT on a search line containing subject keywords.

Example:

```
$1  waste & disposal(sen)
$2  waste & treatment(sen)
$3  $1, $2
$4  radioactiv($), nuclear
$5  $3 & $4(not)
end
```

8. Other searches.

Some searches fail to fall into any of the above categories, especially those searches which combine terms on several logic levels.

9. Author searches.

The only keywords used in this type of search are the names of authors. Sometimes only one author is requested, but often several names are connected by interstring or intrastring OR logic.

Example:

```
a alekseyev, a. i. a, a brodskiy, i. m. a
end
```

These searches are seldom labelled or qualified, and frequently contain only one author's name.

In addition to the search pattern categories, searches were characterized by the actual retrieval categories specified, both in the search mode and in the qualification mode.

1. Search Mode

   a. Search category
   b. Number of search terms
   c. Number of lines in search mode
   d. Format of output requested
   e. Type of search terms used (keyword, author, country code)
   f. Use of labelling
   g. Use of LRANGE to limit returns
   h. Use of positional logic(+1, SEN, other)
   i. Use of truncation
   j. Use of Boolean NOT logic
   k. Use of NOT logic with distribution control markings (:1:, :2:, etc.) to eliminate restricted documents
   l. Specific terms used.

2. Qualifying Mode

   a. date
   b. cntyussr
   c. infocnty
   d. classif
   e. subjcode
   f. datatype
   g. accessnr
   h. filmnr

i. publicnty
j. dpsnr
k. Use of labelled qualifiers

A limited number of searches were also examined to determine the extent to which subject search terms could be further categorized as follows:

1. Subject search terms
2. Personalities/authors
3. Facilities, e.g., Deutsche Forschungsanstalt für Luft und Raumfahrt (German Research Establishment for Air and Space Travel)
4. Locations, e.g., Moscow, USSR
5. Nomenclature, e.g., MIG-21

The frequency of occurrence of individual subject search terms was also determined, both for the overall user community and by user type.

CIRCOL provides the user with various options for displaying the actual document records retrieved. Both on-line display and offline printouts are controlled by the user-selected output option. The user can display various fields or elements of each document record, depending on the particular option selected. The available display or output options are as follows:

Output Codes

Cod	Element
D	Accession number
E	Film number (microfiche/microfilm)
F	Classification
G	Security downgrading code
H	Date of publication
I	Day/month of publication
J	Country/countries of information
K	Publication country
L	COSATI subject code
M	Does CNTYUSSR EQ Y or N?
A	Equivalent to Codes D, E and F
C	Equivalent to Codes D through M
B	Text elements
N	Signifies no output desired.

As part of the analysis of the use of the system by various user types, statistics were compiled on the users' specification of the various output options.

5.3 RESULTS

Results of the analysis of the off-line printout records are presented for a three-month period covering August-October 1972 and for the one-month period of January 1973. One reason for covering two discrete periods was to determine if any significant changes were occurring in user searching behavior. The results are presented by user type.

5.3.1 Analysis of Searches by User Type and by Retrieval Categories (August-October 1972).

Table 17 shows the frequency distribution by user type of the 3076 individual search strategies examined for the August-October 1972 period. Table 18 provides a further breakdown of the data in Table 17, showing statistics by user type for the search mode. Table 19 indicates a more detailed breakdown of the use of the qualification mode by user type.

From Table 18, the percentage of searches using subject keywords is highest for research and development users; it decreases in order for non-FTD intelligence agencies, FTD information specialists and FTD intelligence analysts. This result was expected, since one would expect R&D-oriented personnel to be most interested in retrieval by subject, and intelligence analysts to be greatly interested in retrieval by author. On an overall basis, it is somewhat surprising that the percentage of searches requiring at least one subject keyword is as low as it is. Retrieval by author only, at 46%, represents a surprisingly high use of this retrieval parameter for all user types. Apparently FTD intelligence analysts tend to submit their subject-oriented requests to FTD information specialists, as inferred from the higher percentage of subject requests run by FTD information specialists. Outside (non-FTD) intelligence agencies tend to use CIRCOL more frequently for subject requests than FTD as indicated by the higher percentage of subject requests.

Regarding complexity of search strategy, by far the majority of search strategies are performed with 1-4 terms and 1-4 lines. Searches by author only certainly account for many of these short, simple searches. However, it would be expected that subject-oriented searches would be more complex. From the results, it appears that rather simple strategies are being prepared, even for subject searches. Positional logic is being applied for the majority of keyword searches. Generally, the less complex a subject search strategy is, the broader it is in scope, and retrievals are less precise. There appears to be relatively little differentiation between user

TABLE 17

Frequency Distribution of CIRCOL Search Characteristics by User Type

SEARCH MODE	User Type 1 FTD Information Specialist	User Type 2 FTD Intelligence Analyst	User Type 3 Research & Development	User Type 4 Non-FTD Intelligence Agencies	Overall %
1. No. Searches	729 (24%)	712 (23%)	295 (10%)	1340 (44%)	3076
2. Search Pattern*					
1	19	30	7	180	236 (8%)
2	72	40	39	81	232 (8%)
3	6	4	1	17	28 (1%)
4	82	99	108	348	637 (21%)
5	58	49	33	128	268 (9%)
6	101	36	34	47	218 (7%)
7	4	3	5	3	15 (0%)
8	18	7	1	11	37 (1%)
9	369	444	67	525	1405 (46%)
*Search patterns are described on pp. 73-75.					
3. No. of Search Terms					
1	311	344	36	442	1133 (37%)
2	129	130	69	236	564 (18%)
3	71	70	59	143	343 (11%)
4	57	42	29	76	204 (7%)
5	49	21	19	63	152 (5%)
6	20	21	19	62	122 (4%)
7	16	12	6	47	81 (3%)
8	16	18	4	40	78 (3%)
≥ 9	60	54	54	231	394 (13%)

TABLE 17 (Cont'd)

SEARCH MODE	User Type 1 FTD Information Specialist	User Type 2 FTD Intelligence Analyst	User Type 3 Research & Development	User Type 4 Non-FTD Intelligence Agencies	Overall %
4. No. of Search Lines					
1	417	486	134	787	1824 (59%)
2	10	71	28	42	151 (5%)
3	161	56	87	218	522 (17%)
4	41	19	19	64	143 (5%)
5	50	33	15	86	184 (6%)
6	24	10	4	43	81 (3%)
7	11	9	5	38	63 (2%)
8	9	8	1	19	37 (1%)
≥ 9	6	20	2	43	71 (2%)
5. Type of Search Term					
Subject Keyword(s) Only	297 (41%)	247 (35%)	230 (78%)	777 (58%)	1548 (50%)
Author Only	369 (51%)	444 (62%)	64 (22%)	519 (39%)	1396 (46%)
Country Code Only	16	6	0	10	32 (1%)
Subject Keyword(s) Author	24	7	0	2	33 (1%)
Subject Keyword(s)/ Country Code	23	8	1	32	64 (2%)
Author/Country Code	0	0	0	0	0 (0%)
Subject Keyword(s)/ Author/Country Code	0	0	0	0	0 (0%)
6. Use of Labelling/LRANGE Features					
Labelled	232 (32%)	107 (15%)	100 (34%)	736 (55%)	1175 (38%)
LRANGE'd	18 (2%)	90 (13%)	8 (3%)	173 (13%)	289 (9%)
Labelled/LRANGE'd	64 (9%)	13 (2%)	0 (0%)	154 (11%)	231 (8%)
Unlabelled/No LRANGE Used	415 (57%)	502 (71%)	187 (63%)	277 (21%)	1381 (45%)

TABLE 17 (Cont'd)

SEARCH MODE	User Type1 FTD Information Specialist		User Type 2 FTD Intelligence Analyst		User Type 3 Research & Development		User Type 4 Non-FTD Intelligence Agencies		Overall	%	
7. Positional Logic											
(+1)	95	20%	126	19%	73	48%	553	43%	847	(27%)	32%
SEN	65		10		35		11		121	(4%)	
Other	7		8		43		26		84	(3%)	
None	585	80%	578	81%	153	52%	764	57%	2080	(68%)	
8. Truncation (at least one truncated term)	49		33		24		31		137	(4%)	
9. Boolean NOT logic	4		20		5		6		35	(1%)	
10. Control Designation Specified	9		4		63		589		665	(22%)	

TABLE 17 (Cont'd)

QUALIFICATION MODE	User Type 1 FTD Information Specialist	User Type 2 FTD Intelligence Analyst	User Type 3 Research & Development	User Type 4 Non-FTD Intelligence Agencies	Overall %
1. No. Qualified Searches	352 (48%)	257 (36%)	153 (52%)	1006 (75%)	1768 (57%)
2. Date	148 (20%)	109 (15%)	95 (32%)	600 (45%)	952 (31%)
3. IF CNTYUSSR EQ Y	127 (17%)	102 (14%)	96 (33%)	380 (28%)	705 (23%)
4. INFOCNTY	55 (8%)	31 (4%)	16 (5%)	175 (13%)	277 (9%)
5. CLASSIF	62 (9%)	52 (7%)	29 (10%)	237 (18%)	380 (12%)
6. SUBJCODE	69 (9%)	43 (6%)	17 (6%)	261 (19%)	390 (13%)
7. DATATYPE	12 (2%)	21 (3%)	13 (4%)	125 (9%)	171 (6%)
8. ACCESSNR	5	4	2	3	14
9. FILMNR	0	0	0	0	0
10. PUBLICNTY	0	0	0	0	0
11. DPSNR	7	2	0	2	11
12. Labelling of Qualifiers	0 (0%)	2 (1%)	0 (0%)	19 (1%)	21 (1%)
OUTPUT FORMAT					
1. A	0	2	3	3	8 (0%)
2. B	18	257	14	636	925 (30%)
3. C	1	0	2	0	3 (0%)
4. BC	710	449	276	701	2136 (69%)
5. Other	0	4	0	0	4 (0%)

TABLE 18

Characteristics of Search Mode Specifications by User Type

	User Type 1	User Type 2	User Type 3	User Type 4
Total Searches	729	712	295	1340
Searches with keywords (KW; KW/Author; KW/Cnty Code) (KW/Author/Cnty Code)	344 (47%)	262 (37%)	231 (77%)	811 (60%)
No. of terms/search (Percent based on total searches)				
1-4	568 (78%)	586 (82%)	193 (65%)	897 (67%)
5-8	101 (14%)	72 (11%)	48 (16%)	212 (16%)
≥ 9	60 (8%)	54 (8%)	54 (18%)	231 (17%)
No. of lines/search (Percent based on total searches)				
1-4	629 (86%)	632 (89%)	268 (91%)	1111 (82%)
5-8	94 (13%)	60 (8%)	25 (8%)	186 (14%)
≥ 9	6 (1%)	20 (3%)	2 (1%)	43 (3%)
Positional logic (Percent based on subject keyword searches)				
+1	95	126	73	553
SEN	65	10	35	11
Other	7	8	43	26
Overall	167 (49%)	144 (55%)	151 (55%)	590 (73%)
LRANGE (Percent based on total searches)	82 (11%)	103 (14%)	8 (3%)	327 (24%)
Labelling (Percent based on total searches)	296 (41%)	120 (17%)	100 (34%)	890 (66%)
Truncation (Percent based on subject keyword searches)	49 (14%)	33 (13%)	24 (11%)	31 (4%)

TABLE 19

Use of Qualification Factors in Search Strategies by User Type

	User Type 1	User Type 2	User Type 3	User Type 4
Qualified Searches	352 (48%)	257 (36%)	153 (52%)	1006 (75%)
Qualification factor (Percent based on qualified searches)				
1. DATE	148 (42%)	109 (42%)	95 (62%)	600 (60%)
2. IF CNTYUSSR EQ Y	127 (36%)	102 (40%)	96 (63%)	380 (38%)
3. CLASSIF	62 (18%)	52 (20%)	29 (18%)	237 (24%)
4. SUBJCODE	69 (20%)	43 (17%)	17 (11%)	261 (26%)
5. INFOCNTY	55 (16%)	31 (12%)	16 (10%)	175 (17%)
6. DATATYPE	12 (3%)	21 (8%)	13 (9%)	125 (12%)
7. ACCESSNR	5 (1%)	4 (2%)	2 (1%)	3 (0%)
8. DPSNR	7 (2%)	2 (1%)	0 (0%)	2 (0%)
9. PUBLCNTY	0 (0%)	0 (0%)	0 (0%)	0 (0%)
10. FILMNR	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Labelling of Qualifiers	0 (0%)	2 (1%)	0 (0%)	19 (2%)

types regarding search strategy complexity, although both R&D and non-FTD intelligence agencies tend toward more complex strategies. Considering the greater subject orientation of these two user types, this result is not unexpected.

The use of the LRANGE and truncation features is not great for any of the user types. It is not certain whether these features are being little used because the user is not accustomed to applying them, or whether they recognize these features but consider them inappropriate for the strategies they are formulating. On the other hand, labelling is used quite extensively. Considering the searches which used keywords, 71% of these were formulated with labelled Boolean logic. One can infer from this that labelling is found to be useful for subject-oriented searching. As our previous studies have shown, labelling also helps to reduce the search execution time, particularly for more complex searches.

Table 19 shows the aspects of the search qualification mode used by user type. The general pattern suggests relatively little differentiation among user types. All users tend to rely on the date qualifier most heavily as the primary qualification factor. As has been learned from our work on system response time, greater use of the ! LRANGE feature in conjunction with date as a qualifier would probably improve system response time significantly.

The first six search qualification factors listed in Table 19 are used rather heavily, although the frequency of use declines rather rapidly; interestingly this occurs almost in parallel for all user types. Labelling of qualifiers is almost never employed. Film number and publication country were never used over the one- and three-month periods over which statistics were compiled. Sufficient use is made of at least the first six qualification factors to indicate that they are important in the retrieval process for all user types.

Because of the unexpectedly high percentage of author-only searches, it was decided to examine author/personality-only searches and subject keyword searches independently in terms of search complexity. Since no differentiation had been made up to this point, the author-only searches were probably influencing the statistics to make it appear that search strategies were less sophisticated than was acutally the case. By differentiating the searches into author/personality-only searches and subject keyword searches, more accurate statistics could be derived. Table 20 shows the results for subject keyword and author/personality-only searches by user type.

The analysis of these data suggests that there are essentially no differences between user groups regarding the number of search terms

TABLE 20

Distribution of Author-Only and Subject Keyword Searches by User Type

	User Type 1 FTD Information Specialist	User Type 2 FTD Intelligence Analyst	User Type 3 Research & Development	User Type 4 Non-FTD Intelligence Analyst	Overall
No. of Searches					
Keyword	360	268	228	815	1671
Author	369	444	67	525	1405
Total	729	712	295	1340	3076
No. of Terms per Search					
Keyword					
1	40 (11%)	25 (10%)	16 (7%)	110 (13%)	191 (11%)
2	81 (22%)	59 (22%)	64 (28%)	134 (16%)	338 (20%)
3	50 (14%)	40 (15%)	57 (25%)	138 (17%)	235 (17%)
4	50 (14%)	35 (13%)	25 (11%)	64 (8%)	174 (10%)
5	41 (11%)	21 (8%)	18 (8%)	53 (6%)	133 (8%)
6	16 (4%)	19 (7%)	14 (6%)	54 (7%)	103 (6%)
7	15 (4%)	12 (4%)	5 (2%)	45 (5%)	77 (5%)
8	15 (4%)	12 (4%)	2 (1%)	35 (4%)	64 (4%)
≥ 9	52 (14%)	45 (17%)	27 (12%)	182 (21%)	306 (18%)
Author Only					
1	271 (73%)	319 (72%)	20 (30%)	332 (63%)	942 (67%)
2	48 (13%)	71 (16%)	5 (7%)	102 (19%)	226 (16%)
3	21 (6%)	30 (7%)	2 (3%)	5 (1%)	58 (4%)
4	7 (2%)	7 (2%)	4 (6%)	12 (2%)	30 (2%)
5	8 (2%)	0 (0%)	1 (2%)	10 (2%)	19 (1%)
6	4 (1%)	2 (0%)	5 (7%)	8 (1%)	19 (1%)
7	1 (0%)	0 (0%)	1 (2%)	2 (0%)	4 (0%)
8	1 (0%)	6 (1%)	2 (3%)	5 (1%)	14 (1%)
≥ 9	8 (2%)	9 (2%)	27 (40%)	49 (9%)	93 (7%)

per keyword search. As might be expected, a significant number of keyword searches exceeded nine terms per search. Somewhat unexpected is the fairly large number of single term, two-term, and three-term keyword searches, comprising about 48% of all keyword searches. Alternative modes of expression of concepts (synonyms and synonymous phrases) apparently are not frequently considered in search strategy formulation. In those few cases where the narrative statement of request was available this tendency was confirmed, since the search strategy was usually merely a direct transcription of the phrases in the request to a search strategy format. Considering the possibilities with CIRCOL for specifying positional logic and Boolean NOT logic, it would appear that the capabilities of the system for sharpening search strategies to obtain better precision on retrieval are not utilized extensively. It is interesting to observe that the average number of keywords per search does not differ significantly among user types. FTD information specialists average 4.3 keywords per search; FTD intelligence analysts, 4.5 keywords, R&D personnel, 3.9 keywords, and outside intelligence agencies, 4.7 keywords per search.

With regard to author/personality-only searches, by far the majority (67%) are single-term searches. Here the R&D user is a notable exception, since most of his author/personality-only searches are for a series of authors/personalities. For intelligence agencies, nearly all author/personality-only searches are for one or two authors/personalities per search.

5.3.2 Changes in Searching over a Period of Time

Data for January 1973 are presented in Table 21. For comparison purposes, the August-October data are presented adjacent to the January 1973 data. Also for easy comparison, the data are given in terms of percentages instead of the actual frequencies. The number of searches for the January period was 1375, whereas for the three-month period the monthly average was only 1025, showing that the overall search activity was increasing significantly. The largest gain was accounted for by research and development users. A significant increase also occurred for outside intelligence agencies. Use by FTD remained nearly constant. However, since FTD had been actively using the system for some time, their use had stabilized by this time. The use of CIRCOL by outside intelligence agencies and research and development users was increasing as more agencies were added to the community of users, and as useful results started to be achieved.

It was desired to determine if overall changes in search behavior were occurring, and if significant changes among user types could be seen between the two periods. Major groups of search characteristics were categorized by search mode, qualification mode, and output format.

TABLE 21

Frequency Distribution of CIRCOL Search Characteristics by User Type

SEARCH MODE	User Type 1 Aug-Oct 1972	User Type 1 Jan 1973	User Type 2 Aug-Oct 1972	User Type 2 Jan 1973	User Type 3 Aug-Oct 1972	User Type 3 Jan 1973	User Type 4 Aug-Oct 1972	User Type 4 Jan 1973	Overall Aug-Oct 1972	Overall Jan 1973
1 No. Searches	729 (24%)	322 (23%)	712 (23%)	275 (20%)	295 (10%)	160 (12%)	1340 (44%)	618 (45%)	3076	1375
2. Search Pattern										
1	3%	2%	4%	4%	2%	9%	13%	15%	8%	9%
2	10%	2%	6%	2%	14%	15%	6%	6%	8%	5%
3	1%	1%	1%	1%	0%	0%	1%	0%	1%	0%
4	11%	9%	14%	10%	37%	30%	26%	15%	21%	14%
5	8%	9%	7%	10%	11%	13%	10%	11%	9%	10%
6	14%	10%	5%	3%	12%	14%	4%	5%	7%	6%
7	1%	1%	0%	3%	2%	0%	0%	1%	0%	1%
8	2%	2%	1%	0%	0%	1%	1%	3%	1%	2%
9 (author/ personality only)	51%	68%	62%	68%	23%	19%	39%	44%	46%	51%
3. No. of Search Terms										
1	43%	59%	48%	53%	12%	9%	33%	25%	37%	37%
2	18%	11%	18%	15%	23%	19%	18%	17%	18%	16%
3	10%	8%	10%	8%	20%	9%	11%	7%	11%	8%
4	8%	7%	6%	8%	10%	15%	6%	7%	7%	8%
5	7%	5%	3%	2%	6%	11%	5%	7%	5%	6%
6	3%	1%	3%	2%	6%	9%	5%	4%	4%	3%
7	2%	2%	2%	1%	2%	1%	4%	3%	3%	2%
8	2%	1%	3%	3%	1%	1%	3%	4%	3%	3%
≥9	8%	7%	8%	10%	18%	24%	17%	25%	13%	18%
4. No. of Search Lines										
1	57%	73%	68%	71%	45%	32%	59%	46%	59%	56%
2	1%	2%	10%	9%	9%	8%	3%	3%	5%	5%
3	22%	12%	8%	7%	29%	19%	16%	19%	17%	16%
4	6%	3%	3%	1%	6%	21%	6%	6%	5%	7%
5	7%	5%	5%	7%	5%	2%	5%	13%	6%	9%
6	3%	1%	1%	0%	1%	7%	3%	5%	3%	3%
7	2%	2%	1%	1%	2%	3%	3%	2%	2%	2%
8	1%	1%	1%	0%	0%	5%	1%	2%	1%	2%
≥9	1%	2%	3%	1%	1%	4%	3%	4%	2%	3%

TABLE 21 (Cont'd)

Frequency Distribution of CIRCOL Search Characteristics by User Type

SEARCH MODE (cont'd)	User Type 1		User Type 2		User Type 3		User Type 4		Overall	
	Aug-Oct 1972	Jan 1973	Aug-Oct 1972	Jan 1973	Aug-Oct 1972	Jan 1973	Aug-Oct 1972	Jan 1973	Aug-Oct 1972	Jan 1973
5. Type of Search Term										
Subject Keywords Only	41%	28%	35%	28%	78%	81%	58%	55%	50%	46%
Author/Personality Only*	51%	68%	62%	68%	22%	19%	39%	44%	46%	51%
Country Code Only	2%	0%	1%	1%	0%	0%	2%	0%	1%	0%
Subject Keyword(s)/A-P	3%	1%	1%	1%	0%	0%	0%	0%	1%	1%
Subject Keyword(s)/Country Code	3%	3%	1%	1%	0%	0%	5%	1%	2%	2%
A-P/Country Code	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Subject Keyword(s)/A-P/Country Code	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
6. Use of Labelling/LRANGE Features										
Labelled	32%	23%	15%	11%	34%	43%	55%	42%	38%	30%
LRANGED	2%	39%	13%	20%	3%	4%	13%	5%	9%	15%
Labelled/LRANGED	9%	10%	0%	3%	0%	4%	11%	34%	8%	19%
Unlabelled, No LRANGE Used	57%	28%	71%	66%	63%	49%	21%	20%	45%	36%
7. Positional Logic										
(+1)	13%	9%	18%	12%	25%	29%	41%	33%	25%	23%
SEN	9%	7%	1%	3%	11%	10%	1%	3%	4%	4%
Other	0%	1%	1%	1%	15%	6%	2%	1%	3%	2%
None	80%	83%	81%	84%	50%	55%	55%	62%	68%	71%
8. Truncation (at least one truncated term)	7%	4%	5%	4%	8%	2%	2%	4%	4%	4%

* Author/Personality is abbreviated as A-P

TABLE 21 (Cont'd)

Frequency Distribution of CIRCOL Search Characteristics by User Type

SEARCH MODE (cont'd)	User Type 1 Aug-Oct 1972	User Type 1 Jan 1973	User Type 2 Aug-Oct 1972	User Type 2 Jan 1973	User Type 3 Aug-Oct 1972	User Type 3 Jan 1973	User Type 4 Aug-Oct 1972	User Type 4 Jan 1973	Overall Aug-Oct 1972	Overall Jan 1973
9. Boolean NOT logic	1%	1%	3%	5%	2%	0%	0%	2%	1%	2%
10. Control Designation	1%	3%	1%	0%	21%	36%	44%	58%	22%	31%
QUALIFICATION MODE										
1. % Qualified Searches	48%	34%	36%	29%	52%	66%	75%	74%	57%	55%
2. Date	20%	9%	15%	4%	32%	43%	45%	35%	31%	24%
3. IF CNTYUSSR EQ Y	17%	12%	14%	12%	33%	34%	28%	44%	23%	30%
4. INFOCNTY	8%	3%	4%	5%	5%	3%	13%	13%	9%	8%
5. CLASSIF	9%	19%	7%	10%	10%	14%	18%	28%	12%	21%
6. SUBJCODE	9%	4%	6%	6%	6%	7%	19%	22%	13%	13%
7. DATATYPE	2%	1%	3%	1%	4%	6%	9%	6%	6%	4%
8. ACCESSNR	1%	0%	1%	0%	1%	5%	0%	0%	0%	1%
9. FILMNR	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
10. PUBLCNTY	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
11. DPSNR	1%	0%	0%	4%	0%	0%	0%	0%	0%	1%
12. Labelling of Qualifiers	0%	0%	1%	0%	0%	0%	1%	0%	1%	0%

TABLE 21 (Cont'd)

Frequency Distribution of CIRCOL Search Characteristics by User Type

OUTPUT FORMAT	User Type 1		User Type 2		User Type 3		User Type 4		Overall	
	Aug-Oct 1972	Jan 1973	Aug-Oct 1972	Jan 1973	Aug-Oct 1972	Jan 1973	Aug-Oct 1972	Jan 1973	Aug-Oct 1972	Jan 1973
1. A	0%	0%	1%	0%	1%	1%	2%	0%	0%	0%
2. B	3%	13%	36%	48%	5%	13%	47%	67%	30%	44%
3. C	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
4. BC	97%	87%	63%	52%	94%	86%	52%	33%	69%	56%
5. Other	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%

The most significant change in the search pattern from the August-October 1972 to the January 1973 time period is the trend toward more searches by author/personality-only among in elligence users. The research and development type user tended slightly to use more subject keyword-oriented searches. More than half of the searches run in January were author/personality-only type searches. Author/personality-only searches by the FTD information specialists and intelligence analysts amounted to 68% of all searches run by these two groups, an increase of 17%. Author/personality-only searches by outside intelligence agencies amounted to 44%, an increase of 5%.

The number of search terms used in searching shows some change. Among the FTD intelligence groups, the percentage of single-term searches increased considerably. This correlates with the fact that the percentage of author/personality-only type searches increased markedly. Most FTD searches for author/personality are performed for one author/personality per search. For both R&D and outside intelligence agencies, the number of searches having nine or more terms increased significantly. Such searches are usually keyword searches, and the use of additional terms probably represents greater use of synonyms and alternative means of expressing concepts in keyword searches. We infer that with increasing experience, users are able to formulate better searches.

The number of search lines used in searching indicates interesting changes. In conjunction with the tendency toward author/personality searches, the two FTD user types show a tendency toward single-line searches, as would be expected. However, among the R&D and outside intelligence agencies there is a greater trend toward more lines in the searches. It is expected that the number of lines per search should increase with the same basic trend as indicated by an increase in the number of terms. The number of lines per search, which reflects the inclusion of alternative means of expressing concepts and more complex strategies, probably also suggests more sophistication on the part of users performing keyword subject searches with increasing experience.

Item 5 under the Search Mode in Table 21 indicates that nearly all searches are either exclusively author/personality-only searches or subject keyword only. Very few searches use combinations of the three types of search terms available - subject keyword, author/personality, and country code. As of January 1973, there is no trend toward using country codes in the search mode instead of in the qualification mode.

Data on the use of the labelling feature indicate a decrease in the use of labelling, except for the R&D user type. However, considering the great increase in author/personality-only searches, the decreased use of labelling is expected, since labelling is primarily beneficial only for more complex searches. The use of labelling by the R&D user type, who tends to make more keyword searches, increased from 34% to 43%.

The use of the LRANGE feature was much greater for both the FTD information specialist and FTD intelligence analyst groups. The use of the LRANGE and labelled features together increased very significantly for the non-FTD intelligence analyst groups from 11% to 34%. We infer from these results that the users are generally becoming more aware of these features, especially the LRANGE feature, and therefore the use of these features is more widespread.

The application of positional logic shows little change among user groups for the two periods. Most keyword searches are formulated with positional logic of +1 or SEN logic, with the +1 (phrase) logic being by far the most frequently used logical operator. Positional operators are significant only for subject keyword searches, so it is expected that the use of positional logic for all searches would decrease with the increase in author/personality-only searches.

The use of Boolean NOT logic shows essentially no change. The use of control designations in searches definitely shows an increase, especially among R&D and outside intelligence agency users.

In considering the qualification mode for searches, the extent of qualification decreased considerably for FTD information specialists and intelligence analysts, whereas the extent of qualification increased significantly for research and development type users and remained about the same for outside intelligence agencies. The decrease in the use of the qualification mode by FTD users is expected, because of the large increase in the number of author/personality-only searches. Author/personality-only searches rarely require qualification, either for limiting the number of returns, which is usually small, or for designating other qualification parameters, such as country code.

The use of the qualification mode by research and development users demonstrates significant increases in the date, the classification, and the data type as qualifiers. The use of qualification parameters by outside intelligence agencies shows a decrease in date as a qualifier, but an increase in the following qualifiers: IF CNTY USSR EQ Y, classification, and subject code.

For both time periods there is practically no use of film number, publication country, or DPS number as qualifiers by any of the user types. Interestingly, labelling of qualifier statements is almost never performed, even though labelling of lines in the search mode, especially with the use of the LRANGE feature, tended to be greater for the January 1973 period.

Output formats indicate that Formats B and BC are requested almost exclusively. However, all user groups are relying more on Output Format B. In the August-October 1972 period 69% of output requested was in Format BC, but in January Format BC declined to 56% with a corresponding increase in Format B from 30% to 44%.

5.3.3 Analysis of Subject Keywords Used in Searching

Actual subject keywords used in searching CIRCOL were analyzed in two different ways. The frequency with which subject keywords are used on retrieval was a matter of some interest. The data from the August-October period were examined to determine the frequency of use of various terms. For high frequencies the distribution by user type was also indicated. Below 10 occurrences, only the terms and frequencies are given. The data are presented in Appendix B.

A further analysis of search terms was performed by categorizing the search terms as subject keywords (KW), authors/personalities (P), facilities (F), locations (L), and nomenclature (N). Country codes were not considered as locations for this analysis. In addition to the use of the various categories of search terms, combinations of categories actually used were also determined. The following categories were established:

Keywords (KW)

Personalities (including authors) (P)

Facilities (F)

Nomenclature (N)

Keywords with personalities (KW/P)

Keywords with facilities (KW/F)

Keywords with nomenclature (KW/N)

Facilities with locations (F/L)

Personalities with facilities,
locations, or nomenclature (P/FLN)

The composite results for all user types are shown in Table 22. This group of searches totals 4398, a sufficient number to indicate statistical trends. As can be seen from the table, searches by author/personality-only are run most often; searches by subject keywords make up the second largest category. The other categories show that searching is performed by order of frequency as N, F, L. Searches involving combinations of P, F, L, and N type data in conjunction with KW's are quite infrequent.

A further analysis was made to determine if there were any differentiation by user type. The results of this analysis are also shown in Table 22. The percentages of occurrences are given for each category of

TABLE 22

Distribution of Types of Search Terms by User Type (Percentage)

Search Term	Type 1	Type 2	Type 3	Type 4	Overall
KW	26.8%	31.5%	74.0%	47.1%	46.5%
P	54.7%	64.0%	21.5%	41.4%	39.9%
F	2.8%	0.3%	1.3%	4.3%	5.0%
L	3.4%	0.2%	0.7%	0.9%	4.3%
N	7.2%	2.1%	1.1%	3.1%	1.4%
KW/P	2.1%	1.0%	0%	0.1%	0.7%
KW/F	0.6%	0%	0.4%	0.8%	0.7%
KW/L	1.5%	0.1%	0.2%	0.7%	0.5%
KW/N	0.4%	0.7%	1.3%	0.3%	0.5%
F/L	0.4%	0.1%	0%	0.9%	0.5%
P/FLN	0.1%	0%	0%	0%	0%

search terms. It is interesting to note that KW/P searches are made almost exclusively by User Types 1 and 2. KW/F and KW/L searches are run primarily by both FTD and non-FTD intelligence users. R&D users tend to use KW/N more frequently than other groups. As with KW/F and KW/L searches, F/L searches are almost exclusively used by FTD and non-FTD intelligence users. The analysis shows that there are definite trends by user type regarding the types of search terms used in combinations. However, as shown by the overall composite, the frequencies of subject keywords used in combination with P, F, L, and N data are very small, even within specific user types.

5.3.4 Problems Encountered and Results Obtained by CIRCOL Users

Although the relevance and recall of results obtained by CIRCOL users was beyond the scope of this study, it was desired to determine the number of documents actually retrieved and ordered off-line by users. Specifically, we wanted to find out if there were some relatively low "threshold number" of documents which users tended to achieve by their search strategy prior to ordering off-line printouts. Tables 23 and 24 show the data for the number of documents ordered off-line for author/personality-only searches and for subject keyword searches as shown by document quantity ranges. The distribution by percentages is indicated. Table 25 shows the composite data for all searches.

From Table 25 we can see that the range of 10-19 documents represents the quantity of documents which were retrieved and ordered off-line most frequently. For small quantities of documents, it is interesting to note the high frequency with which only one document was retrieved and ordered. Similarly, the frequencies for 2, 3, 4, 5, 6, 7, 8, 9 documents are fairly high. Beyond 150 documents, the frequencies drop off rapidly.

In comparing keyword and author/personality-only searches, the trend is similar for both types. For keyword searches, however, the range of frequency distribution is much greater, with some instances in which fairly large numbers of retrievals (more than 200) occurred. For author/personality-only searches, only rarely were more than 100 documents retrieved; of course, author/personality-only searches would tend to result in fewer documents retrieved. Similarly, the frequencies with which low numbers of retrievals occurred was high, especially for author/personality-only searches.

Comparison of user types shows that, on an overall basis, User Types 1, 2, and 4 show a generally similar trend. For these user types, the number of documents retrieved and ordered tends toward the 1-20 range. For User Type 3, the number of documents retrieved and ordered tends more toward the 10-50 range. Looking at the differences between keyword and author/personality searches among the various user

TABLE 23

AUTHOR/PERSONALITY SEARCHES

Frequency of Searches Corresponding to No. of Documents Retrieved by Percent

Retrieved No. of Docs.	User Type 1	User Type 2	User Type 3	User Type 4	Overall
1	15.5%	7.6%	3.1%	12.2%	11.3%
2	11.2%	9.2%	4.1%	8.3%	9.1%
3	7.7%	6.3%	1.0%	8.2%	7.2%
4	6.2%	3.2%	3.1%	6.8%	5.4%
5	6.4%	7.1%	2.1%	4.6%	5.7%
6	5.8%	4.6%	0%	4.3%	4.6%
7	3.1%	3.8%	1.0%	3.4%	3.4%
8	2.9%	2.9%	0%	2.6%	2.6%
9	3.4%	3.8%	1.0%	3.4%	3.4%
10-19	17.5%	20.8%	9.3%	20.4%	19.2%
20-29	11.2%	11.3%	10.3%	8.4%	10.1%
30-39	4.8%	7.3%	6.6%	5.5%	5.9%
40-49	2.9%	3.8%	2.1%	2.1%	2.8%
50-59	0%	2.1%	8.3%	1.9%	1.7%
60-69	0.7%	1.3%	3.1%	1.6%	1.3%
70-79	0.7%	1.1%	7.2%	0.9%	1.2%
80-89	0%	1.0%	8.2%	0.8%	1.0%
90-99	0%	0.2%	3.0%	0.5%	0.4%
100-149	0%	2.1%	13.4%	1.3%	1.7%
150-199	0.2%	0.6%	4.1%	0.6%	0.7%
200-299	0%	0.2%	4.1%	0.4%	0.4%
300-399	0%	0%	2.1%	0.4%	0.2%
400-499	0%	0%	3.1%	0%	0.1%
500-999	0%	0%	1.0%	0.1%	0.1%
≥ 1000	0%	0%	0%	0%	0%

TABLE 24

KEYWORD SEARCHES

Frequency of Searches Corresponding to No.
of Documents Retrieved by Percent

Retrieved No. of Docs.	User Type 1	User Type 2	User Type 3	User Type 4	Overall
1	7.3%	4.8%	6.4%	9.8%	5.0%
2	6.5%	3.1%	2.8%	5.0%	4.7%
3	4.7%	3.6%	3.9%	3.7%	3.9%
4	1.9%	2.2%	2.2%	3.3%	2.7%
5	4.5%	1.1%	4.2%	2.7%	3.0%
6	2.6%	1.1%	1.9%	2.0%	2.0%
7	2.4%	1.4%	2.2%	2.6%	2.3%
8	2.4%	3.6%	1.7%	1.3%	1.9%
9	1.5%	2.2%	1.7%	1.5%	1.7%
10-19	12.5%	15.7%	14.4%	10.1%	12.1%
20-29	10.3%	10.7%	10.0%	8.8%	9.5%
30-39	7.1%	11.8%	5.6%	4.4%	6.2%
40-49	5.2%	5.4%	5.8%	5.0%	5.2%
50-59	4.1%	6.4%	6.9%	4.8%	5.2%
60-69	5.2%	3.4%	5.3%	3.2%	3.9%
70-79	2.2%	2.0%	1.4%	3.2%	2.5%
80-89	3.0%	2.5%	2.5%	2.0%	2.3%
90-99	0.9%	2.8%	2.2%	2.4%	2.1%
100-149	7.3%	5.4%	6.1%	6.8%	6.6%
150-199	1.7%	3.9%	5.0%	4.8%	4.1%
200-299	2.8%	4.5%	3.3%	4.5%	4.0%
300-399	2.4%	1.1%	1.4%	3.1%	2.4%
400-499	0.9%	1.1%	1.1%	2.0%	1.5%
500-999	0.9%	1.1%	1.1%	2.5%	1.6%
≥ 1000	0.5%	0%	0%	0.7%	4.3%

TABLE 25

Frequency of Searches Corresponding to No.
of Documents Retrieved by Percent

Retrieved No. of Docs.	User Type 1	User Type 2	User Type 3	User Type 4	Overall
1	11.8%	6.5%	5.7%	10.8%	9.4%
2	9.0%	6.8%	3.0%	6.4%	6.7%
3	6.4%	5.3%	3.2%	5.5%	5.4%
4	4.3%	2.9%	2.4%	4.7%	4.0%
5	5.5%	4.8%	3.8%.	3.5%	4.3%
6	4.4%	3.3%	1.3%	2.9%	3.2%
7	2.8%	3.8%	1.8%	3.4%	3.2%
8	2.6%	3.0%	0.3%	1.8%	2.3%
9	2.6%	3.3%	0.5%	2.3%	2.5%
10-19	15.3%	18.6%	13.3%	14.3%	15.1%
20-29	10.8%	10.9%	10.0%	8.7%	9.6%
30-39	5.8%	8.9%	5.9%	4.8%	6.0%
40-49	3.9%	4.3%	5.0%	3.9%	4.1%
50-59	1.8%	3.6%	7.2%	3.6%	3.6%
60-69	2.6%	1.9%	4.8%	2.5%	2.7%
70-79	1.4%	1.8%	2.6%	2.2%	2.0%
80-89	1.4%	1.6%	3.9%	1.5%	1.7%
90-99	0.4%	1.0%	2.4%	1.6%	1.3%
100-149	3.2%	3.1%	7.7%	4.6%	4.3%
150-199	0.9%	1.8%	4.8%	3.1%	2.5%
200-299	1.2%	1.8%	3.5%	2.9%	2.8%
300-399	1.0%	0.4%	1.5%	2.0%	1.4%
400-499	0.4%	0.4%	1.5%	1.2%	0.9%
500-999	0.2%	0.4%	1.8%	1.3%	0.9%
≥ 1000	0.2%	0%	0%	0.4%	0.2%

types, the overall tend carries through, i. e., R&D users tend to retrieve and order more documents than intelligence users, both for author/personality-only searches and for keyword searches.

From the analysis of the number of off-line retrievals, it is difficult to observe any strong tendency to achieve some typical number of retrieved documents being ordered off-line. It is surprising, in fact, that although most searches result in retrievals of less than 100, there are still instances of users ordering retrieved documents off-line of up to and beyond 1000 documents, especially by FTD information specialists and outside intelligence agencies. In general it can be stated that the number of retrieved documents ordered off-line usually amounts to fifty or fewer.

From an analysis of on-line terminal records, an indication of recurring problems was derived. Most problems seemed to be occurring with research and development users followed by outside intelligence agencies. Because of extensive experience with CIRCOL, FTD intelligence analysts and information specialists encountered very few problems with CIRCOL. The problems were largely of a mechanical nature, for example: failure to specify positional logic properly; failure to combine lines of a search together properly at the conclusion of the search, particularly for labelled searches; failure to label the line in which labelled lines were combined; improper specification of entries in the qualification line; intermixing of logical AND and OR in the same line, often when truncation was specified; and format errors, particularly with author/personality searches.

Rather few instances of the use the QUERY command to check search strategies were noted. The QUERY command would have precluded the occurrence of many of the errors found. Considering the time wasted when an improper strategy is transmitted to the computer for searching, this command should be used routinely to check searches of any complexity. Examples of errors are listed in decreasing order of occurrence as follows:

Example 1: Label not applied to line composed of
back-reference labelled lines

```
      INCORRECT                         CORRECT

$1  aluminum & alloys(+1)       $1  aluminum & alloys(+1)
$2  titanium & alloys(+1)       $2  titanium & alloys(+1)
$3  $1, $2                      $3  $1, $2
$4  fatigue                     $4  fatigue
    $3 & $4                     $5  $3 & $4
end                             end
```

Example 2: Labelled lines not combined at conclusion of search

INCORRECT	CORRECT
$1 missiles	$1 missiles
$2 heat & seeking(sen)	$2 heat & seeking(sen)
$3 heat & sensing(sen)	$3 heat & sensing(sen)
$4 $2, $3	$4 $2, $3
end	$5 $1 & $4
	end

Example 3: Intermixing logical AND and OR in the same search line

INCORRECT	CORRECT
$1 ignit($) & properties(+1)	$1 ignit($)
$2 burn($) & rate(+1)	$2 properties
$3 solid & propellants(+1)	$3 $1 & $2(+1)
$4 $1, $2	$4 burn($)
$5 $3 & $4	$5 rate
end	$6 $4 & $5(+1)
	$7 solid & propellants(+1)
	$8 $3, $6
	$9 $7 & $8
	end

Example 4: Positional logic improperly specified

INCORRECT	CORRECT
$1 steels	$1 steels
$2 hot & working(+1)	$2 hot & working(+1)
$3 hot & rolling	$3 hot & rolling(+1)
$4 $2, $3	$4 $2, $3
$5 $1 & $4	$5 $1 & $4
end	end

Example 5: Format errors

INCORRECT	CORRECT
@Bolushenko; Ya. K.@	@Bolushenko, Ya. K.@

5.4 CONCLUSIONS

The analysis of the off-line and terminal search records categorized by the four user types shows that there are definite differences in the ways that CIRCOL is used, primarily in terms of search content. The primary differences occur between the intelligence users and the research and development user. Intelligence users make extensive use of authors and personalities as retrieval terms, often making searches of only one author/ personality per search. Research and development users search primarily by subject keywords, but surprisingly a significant number of author/ personality-only searches are run by this group of users, also.

On examining the search mode portion of the searches, it was found that the number of terms and number of lines used in searching was low. The large number of author/personality searches accounts to some extent for the low number of search terms and lines, but even the subject keyword searches demonstrated very little use of synonyms or logical combinations which could have been used advantageously. In the later time period, users tended to formulate better search strategies.

In the qualification mode, IF CNTYUSSR EQ Y, the date, the information country(ies), the classification, and the subject code qualifiers were used extensively. Data type was used to some extent; access number, film number, publication country, and DPS number were used almost never. The LRANGE feature for limiting searches was infrequently used. Labelling, especially of subject keyword searches, was usually employed in the search mode, but labelling of qualification mode statements was very rarely used.

Comparison of searches taken over two discrete periods of time showed greater use of CIRCOL by research and development users and by outside intelligence users in the later period. Even greater use of authors/ personalities as search terms was evident. More use was being made of CIRCOL features such as truncation and LRANGE, probably as a result of more experience and greater familiarity with the system.

Analysis of search terms as subject keywords, personalities, facilities, locations, and nomenclature (P, F, L, N) indicated that the majority of PFLN searching is done by intelligence users. Very few searches used combinations of PFLN type search terms and subject keywords.

An examination of document retrievals indicated a wide range of documents retrieved per search and ordered off-line. Most frequently the number of documents fell within the range of 10-20, and overall, the number of documents ordered was usually fewer than 50.

Review of on-line terminal records showed that relatively few problems were being experienced by CIRCOL users; research and development users were most likely to make errors and FTD information specialists rarely made errors. Nearly all problems encountered were due to errors with the mechanics of entering searches. The QUERY command, which can be used effectively to review search strategies before transmitting the search to the computer, was used very infrequently. Many errors could have been avoided by using the QUERY command; its use is recommended.

SECTION 6

OBSERVATIONS AND RECOMMENDATIONS FOR THE CIRCOL SYSTEM

Based on our analysis of the CIRCOL system, both from the standpoint of factors affecting search response time and actual usage of the system by various types of users, certain observations and recommendations can be made. It should be noted that CIRCOL represents a specific application of DPS. In order to meet certain requirements of the FTD environment, several modifications of the basic DPS software have been made by FTD to accommodate these needs. Therefore the recommendations and observations made apply to the CIRCOL application of DPS.

6.1 CIRCOL COMMANDS AND DIALOGUE

The command language of CIRCOL is quite straightforward and easy to use. A particularly desirable feature is the option of using a short form of interactive dialogue for experienced users. The default dialogue is semi-tutorial and is useful for less-experienced CIRCOL users. The short form of the dialogue is invoked by entering !SHORT. This command can be entered immediately prior to the sign-on procedure. Reversion to the long form is effected by entering !LONG.

The QUERY command is very useful for reviewing a search before entering an END command to transmit the search to the computer. Especially when a number of corrections have been made, !QUERY is useful to display the actual status of the search with the corrections made. Review of the search using the QUERY command can avoid many errors.

Several commands are available for corrections. The '!' used by itself at any point in specifying a search line (or search statement) will remove that entire line. Character-by-character corrections can generally be made by the reverse arrow or underscore key. The key required for correction is dependent on the specific terminal being used. The REGRESS command permits the user to return to any line in the search strategy, also removing the search lines subsequent to the line specified. The CORRECT command permits the user to correct a previous line without removing intervening lines between the corrected line and the point at which the CORRECT command is invoked. The CANCEL command eliminates the search and logs the user off the system. The DELETE command permits the user to delete a search at any point and return to the beginning line of a new search. The commands available for corrections allow for a convenient and quick means of making corrections in search strategy specifications for all situations requiring corrective action. The use of the QUERY command in conjunction with the appropriate correction command(s) enables the user to enter searches properly with minimal effort.

Labelled searches were observed to cause some problems, because the user occasionally neglected to enter labels consistently, especially when tying previous labelled statements together. A LABEL command is recommended. This command would cause the system to label search lines automatically, once the user had selected this option. An UNLABEL command should also be provided to permit the user to return to the unlabelled logic mode as desired.

It is recommended that a command be provided which would enable the user to scan the Dictionary for actual search terms available. The user should be able to see, on-line, those terms which are alphabetically close to a given term or word stem. It is recommended that a command such as !TERMS followed by the term or word stem for which the user wanted to see a list of terms be provided. A RETURN command could permit the user to return to his search strategy. An example of the use of this command would be as follows:

```
     User:          !terms      work#
                                          Document            Word
     Computer:                            Frequency           Frequency

                    WOOL                     456                 747
                    WOOLEN                    20                  25
                    WORD                    2013                3510
                 #  WORK  #                34949               65535
                    WORKABILITY              168                 221
                    WORKABLE                 117                 130
                    WORKBOAT                   5                   5

TO EXPAND TERM LIST ENTER 'UP(N)' OR 'DOWN(N)'

TO RETURN TO SEARCH STRATEGY ENTER !RETURN

     User:          down 5
                                          Document            Word
                                          Frequency           Frequency

                    WORKER                  6405               11212
                    WORKFORCE                200                 205
                    WORKHARDENING             12                  15
                    WORKING                19995               27702
                    WORKMEN                   31                  38
```

A further refinement would be for the system to code the term list and then permit the user to select those terms he wanted to string together in a logical OR series. The logical OR series would then be inserted in the search strategy as the next search line. An ENTER command could be provided to permit the user to enter the logical OR series, Use of the ENTER command would automatically return the user to the search strategy.

```
                                  Document        Word
                                  Frequency       Frequency

          A.     WOOL                456             747
          B.     WOOLEN               20              25
          C.     WORD               2013            3510
          D.   # WORK  #           34949           65535
          E.     WORKABILITY         168             221
          F.     WORKABLE            117             130
          G.     WORKBOAT              5               5

TO EXPAND TERM LIST ENTER 'UP(N)' OR 'DOWN(N)'

TO SELECT TERMS ENTER !ENTER FOLLOWED BY TERM CODES
(A,B,.....etc.)

TO RETURN TO SEARCH STRATEGY ENTER !RETURN

   down 5                         Document        Word
                                  Frequency       Frequency

          A.     WOOL                456             747
          B.     WOOLEN               20              25
          C.     WORD               2013            3510
          D.  #  WORK  #           34949           65535
          E.     WORKABILITY         168             221
          F.     WORKABLE            117             130
          G.     WORKBOAT              5               5
          H.     WORKER             6405           11212
          I.     WORKFORCE           200             205
          J.     WORKHARDENING        12              15
          K.     WORKING           19995           27702
          L.     WORKMEN              31              38

TO EXPAND TERM LIST ENTER 'UP(N)' OR 'DOWN(N)'

TO SELECT TERMS ENTER !ENTER FOLLOWED BY TERM CODES
(A,B,....etc.)

TO RETURN TO SEARCH STRATEGY ENTER !RETURN

!enter d,e,f,j,k
```

The LRANGE command limits the extent of the data base searched. It is effective in reducing search response times. Its use has been covered in detail in Sections 3 and 4.

At the conclusion of a successful search the user is advised of the number of documents retrieved. The dialogue then prompts the user to qualify the search, if desired. Generally qualification will reduce the number of documents ultimately retrieved. Often, however, the user may wish to qualify the search at the outset. It is recommended that the user be prompted to enter appropriate qualification statements initially. As our studies have shown, if qualification statements are entered at the beginning, document retrievals will be fewer and search response time will be faster. The user would still have the option of further qualification to reduce the number of documents at the conclusion of his original search.

It is also recommended that a wider choice of output options be made available for on-line display. Currently, to obtain titles, for example, the user must also obtain much auxiliary information as well. The user should have the capability of displaying on-line those user-selected portions of the document record which he desires to see, e.g., titles, authors/personalities, information country, date, source, etc.

6.2 CIRCOL FILES

CIRCOL files are discussed in detail in Section 2. Based on our studies of factors affecting response times and of actual use of the system by the users, several recommendations are in order.

6.2.1 Dictionary File

The Dictionary File is ordered in alphabetic order. Access to the Dictionary File during the processing of a search occurs essentially in alphabetic sequence. From our studies of the frequency with which certain terms are used in searching, we found that certain terms are used with quite high frequencies, whereas the great majority of terms actually in the Dictionary are rarely or never used. It is recommended that a means be found for high frequency terms to be addressed early during the processing of the search. One means of accomplishing rapid access to high frequency terms could be a supplementary high frequency Dictionary file or table which would be ordered by the frequency with which such Dictionary terms are entered into the system. The incorporation of such a file should significantly reduce search execution time, especially if all search terms entered were in that file. However, such a table or auxiliary file would involve more storage. The table should probably be derived by obtaining statistical data from off-line search records and implementing the table manually. However, the CIRCOL system could be programmed to monitor searches entered to obtain frequency data for search terms.

6.2.2 Master File

The Master File contains fixed format field data as well as search term positional data. There are currently certain fixed field data which are provided but which are never used in the qualification mode. Since such data are not used, it would be possible to eliminate such fixed fieid data from the Master File. The data would still be available in the Text File. Storage requirements for Master File data would be reduced, and some improvement in search response time could be anticipated. Specifically the publication country and film number fields were never used in searching and could be eliminated from the Master File.

6.3 CIRCOL PROCESSING

There are certain aspects of CIRCOL processing which are significantly inefficient. One such area is in processing the qualification mode. If the user adds qualification statements to a search following the original running of the search, the entire search must be processed again through both the search mode and the qualification mode. Efficiency could be increased if the system would save the results of the original search. Then, if the user decided to enter additional qualification statements, the saved group of documents originally retrieved would have to be checked only against the Master File in order to select from that group those documents which also met the additional qualification requirements.

Our studies have shown that the most time-consuming step in the search process is the compilation of "pointers" (pointers are special internal codes) which establish the correlation between the DPS identification number retrieved and the actual document record in the Text File. It is the time-consuming process of compiling "pointers" which accounts for the fact that the number of documents retrieved is the most significant factor in causing long search execution times. The reason for the pointers is to permit the user either to view the document records on-line or to order the documents to be printed off-line. The system must not only identify the retrieved documents by DPS number, but the cross reference to the actual document record in the Text File must be made. Thus, when the system informs the user how many documents were retrieved, it has also located the document records and is ready to display or print the records as indicated by the user.

It is recommended that the computer programs which establish the pointers for locating the document records in the Text File be reviewed. Modifications in the system mechanics for establishing pointers definitely are indicated.

Our studies showed that the order in which various types of search statements were entered can significantly affect the search response time. Although the user can control the order of specifying search statements, he may not necessarily take advantage of the most efficient method. A possibility which could be explored would be for the system to order the search lines automatically to optimize the efficiency of search processing. The optimum order is: single term; intrastring logical AND; intrastring logical AND with positional logic; intrastring logical OR. A system routine to accomplish the ordering of search statements would also have to check labelled searches to ensure that back-referenced statements were entered prior to the subsequent processing.

SECTION 7

THE STORAGE AND INFORMATION RETRIEVAL SYSTEM (STAIRS)

The IBM Company has developed a new complete software package for information storage and retrieval systems called Storage and Information Retrieval System (STAIRS). This system became available toward the end of our evaluation work with CIRCOL/DPS. STAIRS was designed as an on-line system and has many capabilities which could prove useful for the CIRC application. It was desired to determine if STAIRS could provide features and capabilities which would serve the needs of the CIRC users. Many of the features and program changes suggested in Section 6 are already incorporated in STAIRS. However, it is important to note that a DPS data base cannot be directly converted to a STAIRS data base. If conversion of a DPS-based system were to be made to STAIRS, the entire DPS data base would have to be reloaded into the STAIRS system. Thus, any DPS system manager wishing to upgrade the system capabilities would be faced with an important decision of whether to modify DPS or to convert to STAIRS or another storage and retrieval software package. As part of our evaluation program, we undertook a short-term evaluation of STAIRS in comparison with CIRCOL/DPS.

7.1 DESCRIPTION OF STAIRS

STAIRS is a natural language processing storage and retrieval system which extracts terms from natural language text for later retrieval. As specified by the system designer, STAIRS also stores data in fixed length fields which can be addressed by the user to qualify his searches. Conceptually, STAIRS is very similar to DPS, but the mechanics of system operation are quite different.

For retrieval, STAIRS operates in different modes. The mode of operation is established by the user uy entering the appropriate commands. The Search Mode is invoked by the command ..SEARCH. Once the mode-determining command is entered, the system informs the user that the system is operating in the mode specified by the user. For document retrieval there are basically three operating modes - the Search Mode, the Select Mode, and the Browse Mode. The Search Mode permits the user to search variable-length text-derived search terms, the Select Mode permits the user to qualify his search by specifying fixed field length data, and the Browse Mode permits the user to display document records or portions of document records on-line. The Browse Mode also permits the user to print document records off-line.

In the Search Mode the user can specify both single terms and term co-occurrences. He can apply "nesting" of search terms and logical operators within parentheses to insure that the appropriate groups are treated properly by the search. In addition, he can specify the document record paragraph or field in which the term(s) must appear. For example, retrieval can be specified so that the terms must appear in the title. In contrast to DPS, the user can intermix logical AND's and OR's within the same search line. An important difference between the two systems is that STAIRS provides line-by-line retrievals, whereas with DPS the entire search must be entered before searching can begin. Thus, the user is continually advised of the number of retrievals effected by each line. Also, STAIRS automatically labels each search line with a search line number so that back-referencing of search statements can be specified in subsequent search statements.

Examples of search strategies using the STAIRS system and CIRCOL are shown below:

CIRCOL

```
  1  OPTION CIRCOLMV, TEXT
  2  $1 surface & air(sen) & missile
     (sen)
  3  $2 design, specification
  4  $3 $1 & $2
  5  !lrange 600000
  6  end
GOIN' SEARCHIN'

////
399 DOCS SATISFY
```

STAIRS

```
..search
AQUARIUS-SEARCH MODE-
BEGIN YOUR QUERY AFTER THE
STATEMENT NUMBER
00001  :  surface with air with
          missile
RESULT    94 OCCURRENCES
          67 DOCUMENTS
00002  :  1 and (design or specifi-
          cation)
RESULT    13 OCCURRENCES
          13 DOCUMENTS
00003  :  1.title.
RESULT    5 OCCURRENCES
          5 DOCUMENTS
```

Referring to the STAIRS strategy, note that STAIRS provides line by line results. Other STAIRS features mentioned above are shown as follows:

```
        00001              surface with air with missile
automatic labelling

        00002              1  and  (design or specification)
                                   nesting parentheses

        00002              1  and  (design or specification)
                             intermixing logical 'and' and
                             'or' in same line

        00001              surface with air with missile
        RESULT                          94 OCCURRENCES
                                        67 DOCUMENTS
          line-by-line advisement of retrievals

        00003             1. title.
                     back-referencing

        00003             1. title.
                               restriction of search phrase to the
                               title field
```

In the Select Mode the user can specify certain data which are available in the fixed-length formatted fields. He specifies the search line label number, the field name, the operator, and the value within the field, e. g., 5 INFDATE GT 72. The meaning of this statement is: for those documents retrieved corresponding to search line 5, select therefrom those documents whose dates of information exceed 1972. In contrast to the DPS Qualification Mode, the Select Mode operates only on those documents initially retrieved in the Search Mode.

In the Browse Mode the user can display on-line the document records or portions thereof which he wishes to see. The system prompts the user to specify those document record portions desired. Thus he can browse only document titles, originating organizations, authors, and countries of information, or any other combination of document record elements. Thus, in contrast to CIRCOL/DPS, he has many output options available. Off-line printing can also be accomplished using the ..PRINT or ..MAIL command; entire records or only selected portions can be printed.

7.2 EVALUATION

STAIRS has many other features and capabilities which make the STAIRS system very attractive, but a detailed description of STAIRS and its potential application for CIRC is beyond the scope of the work reported herein. We did some limited experimentation with the STAIRS system using a pilot data base of CIRC documents. In using the STAIRS system with the pilot data base we found that the system is very easy to use and it provides excellent response times, even when many documents are retrieved. However, when heavily posted terms are specified in a search, especially in a logical OR series, the search response time is affected adversely. Still, even with attempts to create search strategies which we knew would require long run times, e.g., chemical adj properties, the longest search response times were on the order of five minutes. It should be noted that the pilot data base consisted of only about 24,000 documents compared to over 800,000 documents in CIRCOL. The governing factor, of course, is the number of postings and the number of retrievals which result.

Our experience with STAIRS demonstrates that the system processing of "pointers" to the document records is much more efficient than with DPS. Also the CIRCOL application of DPS uses data cells for the storage of the Text File, whereas STAIRS uses disc storage, which permits faster access times. It should be recognized that STAIRS was specifically designed for a multi-user time-sharing environment, whereas DPS was intended primarily for batch mode operation. Multi-user access to CIRCOL/DPS is possible only through a "hybrid" system which combines the teleprocessing executive program with the DPS retrieval system. Many of the features which we recommended in Section 6 for DPS are already available with STAIRS. However, as pointed out earlier in this Section, a decision to convert to STAIRS involves not only the acquisition of STAIRS software, but also the reloading of data already committed to DPS.

The purpose of our evaluation of STAIRS was to determine its suitability for the CIRC application. We conclude that STAIRS with some modifications is entirely suitable for FTD CIRC, from the standpoints of both system updating and the users' ability to interact effectively with the system.

REFERENCES

1. M.E. Stevens, Automatic Indexing: A State-of-the-Art Report, NBS Monograph 91, National Bureau of Standards, U.S. Department of Commerce, Washington, D.C., March 1965.

2. M.E. Stevens, Research and Development in the Computer and Information Sciences, Volume 1: "Information Acquisition, Sensing and Input - A Selective Literature Review", NBS Monograph 113, Volume 1, National Bureau of Standards, U.S. Department of Commerce, Washington, D.C., March 1970.

3. M.E. Stevens, Research and Development in the Computer and Information Sciences, Volume 2: "Processing, Storage, and Output Requirements in Information Processing Systems - A Selective Literature Review", NBS Monograph 113, Volume 2, National Bureau of Standards, U.S. Department of Commerce, Washington, D.C., May 1970.

4. M.E. Stevens, Research and Development in the Computer and Information Sciences, Volume 3: "Overall System Design Considerations - A Selective Literature Review", NBS Monograph 113, Volume 3, National Bureau of Standards, U.S. Department of Commerce, Washington, D.C., June 1970.

5. J.L. Smith, J.D. Hoffman, and J.C. Cornell, Colex (CIRC-On-Line Experiment), RADC-TR-68-332, Rome Air Development Center, Griffiss Air Force Base, N.Y., November 1968.

6. C. A. Cuadra, in Annual Review of Information Science and Technology: Volume 7, p. 52, Washington, D.C.: American Society for Information Science 1972.

7. D.W. King, P.W. Neel, and B. L. Wood, Comparative Evaluation of the Retrieval Effectiveness of Description and Free-Text Search Systems Using CIRCOL, Report #0199, RADC-TR-71-311, January 1972.

8. B.C. Vickerey, in Annual Review of Information Science and Technology: Volume 6, p. 138, Chicago, Illinois, Encyclopedia Britannica, Inc., 1971.

9. C. W. Cleverdon, in Annual Review of Information Science and Technology: Volume 6, pp. 68-69, Chicago, Illinois, Encyclopedia Britannica, Inc., 1971.

10. W. T. Knox, "Systems for Technological Information Transfer", Science, Volume 181, #4098, pp. 418-419, August 1973.

11. M. Kochen, "On The Economics of Information", Journal of the American Society for Information Science, Volume 23, No. 4, pp. 281-283, July-August 1972.

12. R. W. Graves Jr., and D. P. Helander, "A Feasibility Study of Automatic Indexing and Retrieval", IEEE Transactions on Engineering and Speech, Volume EWS-13, No. 2, pp. 58-59, September 1970.

13. G. Jahoda, Information Storage and Retrieval Systems for Individual Researchers, New York, N. Y.: Wiley-Inter-Science, 1970.

14. F. W. Lancaster, "The Cost Effectiveness Analysis of Information Retrieval and Dissemination Systems", Journal of the American Society for Information Science, Volume 22, No. 1, pp. 12-27, January-February 1971.

15. M. Laird, quoted in Missile/Space Daily, p. 161, 7 April 1964.

16. Anonymous, Central Information Reference and Control (CIRC) On-Line (CIRCOL) User's Guide, FTD-MP-22-14-73, Foreign Technology Division, Wright-Patterson Air Force Base, Ohio, July 1973.

APPENDIX A

PROCEDURES FOR OPTIMIZING SEARCH RESULTS FROM CIRCOL

SUPPLEMENT TO THE CIRCOL USERS' GUIDE

FOREWORD

The CIRCOL Users' Guide describes the procedures by which the user can interact with the CIRCOL system. A complete description of CIRCOL and the data base is provided, and the methods and options for performing searches of CIRCOL are given, including examples. It is assumed that the user is already familiar with the CIRCOL Users' Guide. This supplementary manual is intended to suggest means of applying basic operational procedures in ways to optimize searching from the CIRCOL data base, both in terms of reduced turnaround time and in quality of results.

The University of Dayton Research Institute (UDRI) has undertaken a study of the CIRCOL system under actual operating conditions to determine certain operational characteristics of CIRCOL, especially from the standpoint of the user, and to establish from our findings certain techniques which can be utilized by users to optimize search results from the CIRCOL data base. Use of the procedures suggested in this supplementary manual should result in faster and more satisfactory search results.

TABLE OF CONTENTS

SECTION PAGE

1.0 FACTORS AFFECTING SEARCH RESPONSE TIME

The search execution time or search response time is defined as the time between the user command END, until the computer responds with XXX DOCUMENTS SATISFY YOUR REQUEST. Depending on the complexity of the search request, the number of active users on-line at a given time, and other factors, the search response time can vary considerably.

From studies performed by UDRI, it was shown that certain factors affect search response time more adversely or beneficially than others. By applying search strategy optimization techniques which the user can control, search response times often can be reduced significantly. The significant factors and means of optimizing search strategies are presented in the order in which these features affect search response time.

1.1 Posting Density

1.1.1 Description

The posting density or word frequency/document frequency greatly affects the search response time. The more heavily posted the term is, the longer the search response time. In addition to the posting density of individual terms, the effective posting density is increased by truncation and/or by specifying a logical OR series. It should also be noted that there is a cumulative effect induced by high posting density, i.e., other factors, such as specifying positional logic, result in increasingly longer search times than would be the case with lightly posted search terms.

The posting density (document/word frequency) can be ascertained from the CIRCOL Dictionary. The user should recognize that a heavily posted term or logical OR series is not a very effective discriminator in improving the precision of search results. For example, the phrase CHEMICAL COMPOUNDS would require an excessively long search run time due to the heavy postings on each term. Similarly, a truncated term such as MEASUR($) would tend to cause a long search response time, and may not provide particularly useful as a discriminating feature of the search. However, there are situations which require a heavily posted search term(s) to be used. A means of handling heavily posted terms is indicated below.

1.1.2 Ordering of Lines in the Search

One means of minimizing the effect of heavy posting density is to place search lines in the order from the least heavily posted term(s) to the most heavily posted terms. If differences in posting density are not great, the effect on search response time is minimal. However, if there is a considerable difference, response time will be substantially reduced by placing the most heavily posted terms at the end of the search.

Example 1:

```
        1  OPTION CIRCOLMV, TEXT
        2  $1 boron                              (low posting density)
        3  $2 fiber, filament, reinforc($)       (high posting density)
        4  $3 $1 & $2(sen)
        5  end
GOIN' SEARCHIN'
```

Example 2:

```
        1  OPTION CIRCOLMV, TEXT
        2  $1  polyimide                         (low posting density)
        3  $2  electrical                        (moderate posting density)
        4  $3 resistance, resistivity            (high posting density)
        5  $4 $2 & $3(+1)
        6  $5 $1 & $4
        7  end
GOIN' SEARCHIN'
```

1.1.3 Use of Personalities as Search Terms

If personalities alone are used as search terms, the search will proceed with a very rapid response time, since the posting density for any given personality or combination of personalities is very small compared to most search terms. Furthermore, the personality file is separate from the subject Dictionary file.

When personalities are used in combination with subject search terms, the search will run with about the same search response time as when the subject search terms alone are used. For combination personality/subject searches, it is a good practice to list personalities first. (Example 1) By listing personalities first, the user ensures that they are included. Subsequent modification (Example 2), if any, is usually concerned with the subject keyword portion of the search, so it is more convenient to list this portion toward the end.

Example 1

Personality/Subject Search (Original Strategy)

```
        1  OPTION CIRCOLMV, TEXT
        2  $1 @dovlatyan,  v. v.@,  @volf,  l. a.  @,  @vitulskaya,  n. v. @
        3  $2 halogenation
        4  $3 $1 & $2
        5  end
GOIN' SEARCHIN'

////

        4 DOCS SATISFY
```

Example 2

Modification of Original Strategy

```
        1  OPTION CIRCOLMV, TEXT
        2  $1 @dovlatyan,  v. v. @,  @volf,  l. a. @,  @vitulskaya,  n. v. @
        3  $2 halogenation, bromination, fluorination, chlorination
        4  $3 $1 & $2
        5  end
GOIN' SEARCHIN'

////

        8 DOCS SATISFY
```

1.2 Truncation

Truncation permits the user to search for word stems and is a very convenient shorthand notation for entering terms. With regard to search response time, there is no difference between truncating and entering the equivalent logical OR series of all words to be gained by truncating. The user is cautioned, however, to check the CIRCOL Word Listing to ensure that the truncation does not cause search words which are not intended. For example, TRANSFER($) would cause the search to include TRANSFERASE (an enzyme) and TRANSFERRIN (a globulin). Also, the user must be careful that very long logical OR series (especially resulting in heavy postings) are not created. It should be recalled that there is a 40-word limit for the maximum number of search terms. As an example, the truncated word stem MAGNET($) would result in a logical OR series of 76 terms.

1. 3 Country Code

The country(ies) of interest other than USSR can be specified in the search strategy either in the search mode or in the qualification mode. From the standpoint of search response time, faster responses usually occur when the countries are listed in the search mode. As indicated in Paragraph 1. 1. 2, posting density effects can be minimized by ordering search lines properly. If the overall posting density corresponding to the country codes is high, the country code logical OR series should be placed at the end of the search; conversely, for low posting density, the country code(s) should be placed at the beginning. If other qualifiers are used, it is still more efficient to use country codes as search terms instead of qualifiers, unless the number of country codes and/or overall postings is fairly large. Then countries should be specified in the qualification mode.

Example 1:

```
        1  OPTION CIRCOLMV, TEXT
        2  $1 -uk-
        3  $2 ceramic, glass
        4  $3 industry, manufactur($)
        5  $4 $2 & $3(sen)
        6  $5 $1 & $4
        7  end
GOIN' SEARCHIN'
```

Example 2:

```
        1  OPTION CIRCOLMV, TEXT
        2  $1 ceramic, glass
        3  $2 industry, manufactur($)
        4  $3 -cz-, -ge-, -po-
        5  $4 $1 & $2(sen)
        6  $5 $3 & $4
        7  end
GOIN' SEARCHIN'
```

Example 3:

```
        1  OPTION CIRCOLMV, TEXT
        2  $1 ceramic, glass
        3  $2 industry, manufactur($)
        4  $3 $1 & $2(sen)
        5  $4 if cntyussr eq y
        9  $5 if infocnty sc al, bu, cz, ge, hu, po, ru, yu
       10  $6 if datatype eq e
       11  $7 if classif lt 2
       12  $8 if $4 or $5
       13  $9 if $6 and $7 and $8
       14  end
GOIN' SEARCHIN'
```

1.4 LRANGE

The LRANGE feature limits the portion of the file being searched. The LRANGE is particularly effective in minimizing the search response when the LRANGE is specified at a high DPS number, thus corresponding to a small portion of the file. As the LRANGE is specified to cover larger portions of the file, the effect of LRANGE on reducing response times becomes less and less noticeable. The use of LRANGE to take advantage of interactive feedback is discussed in Section 2.

If the user is interested in quite recent material, the LRANGE can be useful in limiting the portion of the file which has to be searched. It is only necessary to know the DPS number corresponding to the calendar year date in order to restrict the search to the recent years of interest. The DPS numbers corresponding to the years are approximately as follows:

Year	DPS No.
1970	365000
1971	470720
1972	642802
1973	823395

The year indicated is the year in which documents started to be accessed into the CIRCOL Data Base. Some older documents may be entered into the system in a particular year, but it would be impossible, for example, for 1972 documents to have a DPS number lower than the lowest number for that year. Therefore search response for recent documents can be greatly expedited by specifying an LRANGE to accommodate only that portion of the file which is sure to include the year of interest (see Appendix A-1).

In subsequent examples, the first example represents the optimum strategy. The number of retrievals, search response times and number of stations are given simply as illustrations of the approximate degree of optimization which can be effected.

Example 1

```
    1  OPTION CIRCOLMV,TEXT
    2  $1 semiconductor,transistor
    3  $2 if date ge 73
       !lrange 823395
OK
    7  end
GOIN' SEARCHIN'
////
    1 DOCS SATISFY
```

5 min. 18 sec. (5:18)   16 stations

Example 2

```
    1  OPTION CIRCOLMV,TEXT
    2  $1 semiconductor,transistor
    3  $2 if date ge 73
    7  end

GOIN' SEARCHIN'
////
    1 DOCS SATISFY
```

15 min. 50 sec. (15:50)   16 stations

Note that the same documents would be obtained if the user had specified IF DATE GE 73 and the LRANGE had not been included, but the response time would be significantly longer as illustrated in Example 2.

1.5 Qualifying

The qualification mode is used to provide further restrictions on the search. According to the CIRCOL interactive dialog, the user is asked if he wishes to qualify his search after he is informed of the number of retrieved documents. Often the number of documents retrieved serves as the basis by which the user qualifies his search. However, it has been found that if the user knows ahead of time that he wishes to qualify, the overall search response time will be reduced by specifying the qualification requirements initially.

Example 1

```
     1  OPTION CIRCOLMV, TEXT
     2  $1 heart & valve(sen)
     3  $2 repair, surgery
     4  $3 $1 & $2
     5  $4 if cntyussr eq y
     9  $5 if infocnty sc ch
    10  $6 if subjcode sc 06
    11  $7 if $4 or $5
    12  $8 if $6 and $7
    13  end
GOIN' SEARCHIN'
////
    31  DOCS SATISFY

1:24   13 stations
```

Example 2

```
    1  OPTION CIRCOLMV, TEXT
    2  $1 heart & valve(sen)
    3  $2 repair, surgery
    4  $3 $1 & $2
    5  end
GOIN' SEARCHIN'
////
   36 DOCS SATISFY

1:18  10 stations

QUALIFY?
       y
    8  $4 if cntyussr eq y
    9  $5 if infocnty sc ch
   10  $6 if subjcode sc 06
   11  $7 if $4 or $5
   12  $8 if $6 and $7
   13  end
GOIN' SEARCHIN'
////
    31  DOCS SATISFY

3:00    12 stations
```

1.6 Labelling

1.6.1 Search Mode

The use of labelling permits the construction of rather complex strategies by logically tying together labelled strings. It is important to understand the meaning of each labelled line when using labels. Studies have shown that labelled searches tend to run with shorter search response times than equivalent unlabelled searches. When tying together labelled lines in logical series, response time will improve if labelled lines are logically combined in as few steps as possible.

Examples of alternative methods of preparing search strategies achieving the same result are shown as follows, with strategies listed in order of increasing response time:

Example 1

```
     1  OPTION CIRCOLMV, TEXT
     2  $1 weathering, exposure, climate
     3  $2 plastics, polymers
     4  $3 strength
     5  $4 $1 & $2 & $3
     6  end
GOIN' SEARCHIN'
////
     56  DOCS SATISFY
```

3:48   15 stations

Example 2

```
     1  OPTION CIRCOLMV, TEXT
     2  $1 weathering
     3  $2 exposure
     4  $3 climate
     5  $4 $1, $2, $3
     6  $5 plastics
     7  $6 polymers
     8  $7 $5, $6
     9  $8 $4 & $7
    10  $9 strength
    11 $10 $8 & $9
    12  end
GOIN' SEARCHIN'
////
     56  DOCS SATISFY
```

4:26   13 stations

Example 3

```
    1  OPTION CIRCOLMV, TEXT
    2  $1 weathering, exposure, climate
    3  $2 plastics, polymers
    4  $3 $1 & $2
    5  $4 strength
    6  $5 $3 & $4
    7  end
GOIN' SEARCHIN'
////
    56  DOCS SATISFY

5:04    13 stations
```

Example 4

```
    1  OPTION CIRCOLMV, TEXT
    2  weathering, exposure, climate
    3  and plastics, polymers
    4  and strength
    5  end
GOIN' SEARCHIN'
////
    56 DOCS SATISFY
11:36     13 stations
```

1.6.2 Qualification Mode

Labelling in the qualification mode is also advantageous, both regarding search response time and in constructing the search. As has been mentioned previously, qualifiers should be provided in the original strategy if it is known ahead of time that qualifiers are to be used. When tying together labelled qualification lines, it must be remembered that it is necessary to spell out AND and OR. As in the search mode, AND and OR logical lines must be separated.

Example:

LABELLED

```
 1  OPTION CIRCOLMV, TEXT
 2  $1  pollution, contamination
 3  $2  air
 4  $3  $1 & $2
 5  $4  if cntyussr eq y
 9  $5  if infocr.y sc ge, po, cz, hu
10  $6  if classif lt 2
11  $7  if date gt 70
12  $8  if $4 or $5
13  $9 if $6 and $7 and $8
14  !lrange 500000
OK
 14 end
GOIN' SEARCHIN'
////
  363 DOCS SATISFY

3:59  10 stations
```

UNLABELLED

```
 1  OPTION CIRCOLMV, TEXT
 2  pollution, contamination
 3  and air
 4  if cntyussr eq y
 8  and classif lt 2
 9  and date gt 70
10  or infocnty sc ge, po, cz, hu
11  and classif lt 2
12  and date gt 70
13  !lrange 500000
OK
 13 end
GOIN' SEARCHIN'
////
  363 DOCS SATISFY

5:20  10 stations
```

1.7 Positional Logic

The specification of positional logic generally has very little effect on search response time, regardless of the particular position specified, e.g., +1, +2, SEN, PAR. However, if one is searching heavily posted terms, the imposition of positional logic magnifies the search response time noticeably. The use of +1 positional logic has the effect of specifying a word phrase and therefore tends to be restrictive, sometimes actually failing to retrieve useful documents. It should be remembered that CIRCOL is a free text system, and therefore authors can state the same concept in different ways. For example, requiring the phrase MARAGING STEELS in a search could eliminate a document on FORMING OF HIGH NICKEL ALLOY STEELS BY THE MARAGING PROCESS. It has been found that SEN logic is very useful as a positional indicator. Other positional indicators (+2, -1, etc.) do not offer any apparent advantage over SEN positional logic.

Example: (+1 logic)

```
 1  OPTION CIRCOLMV, TEXT
 2  $1  maraging & steels (+1)
 3  end
GOIN' SEARCHIN'
////
```

(SEN logic)

```
 1  OPTION CIRCOLMV, TEXT
 2  $1  maraging & steels (sen)
 3  end
GOIN' SEARCHIN'
////
```

```
141 DOCS SATISFY                    147 DOCS SATISFY
```

There are situations in which word phrases (+1 logic) should be specified in searching. Particularly when the terms being specified are heavily posted and/or if the phrase expresses a concept and is required for precision on retrieval, the +1 positional logic should be specified.

Example:

```
 1  OPTION CIRCOLMV, TEXT
 2  $1  solar & radiation (+1)
 3  $2  ultraviolet & radiation (+1)
 4  $3  $1, $2
 5  $4  reflect ($)
 6  $5  spaceship, spacecraft
 7  $6  $3 & $4 & $5
 8  end
GOIN' SEARCHIN'
```

1.8 Number of Users On-Line

The number of users signed on and using the CIRCOL system affects search response time very significantly. This is one factor, however, over which the user has no control. It can be stated that when more than 15 users are signed on, the search response time increases at an ever-increasing rate. Studies have shown that search response times become very erratic, especially for searches having heavily posted terms; such searches already tend to have long run times. By determining the number of stations with the !STATIONS(N) command, the user can avoid long-running searches at peak use periods. The effect of the number of users on short searches, e.g., author searches, is not nearly as great.

1.9 Summary

The findings of the effect of various factors on search response time are summarized and stated in the form of guidelines for searching to achieve minimal search response times as follows:

a. Minimize the use of heavily posted search terms as much as possible; logical OR series have the effect of increasing posting density.

b. When heavily posted terms are necessary in the search, arrange the order of the search lines such that more heavily posted terms occur towards the end of the search.

c. Searches for authors only run very quickly. When combining authors and search terms in a search, it is a good practice to list authors first.

d. Truncation is a useful tool for the input of a logical OR series of words with a common word stem. Care must be exercised to preclude the occurrence of too many terms and/or the introduction of terms having the same word stem but different meanings.

e. Country codes normally should be entered as search terms. Only if a number of country codes is desired in conjunction with several other qualifying statements should country codes be entered as qualifiers.

f. LRANGE should be applied whenever appropriate. To search recent years only, the use of LRANGE in conjunction with a corresponding date range greatly reduces search response time.

g. Apply qualifiers as part of the original search if it is known previous to running the search that qualifiers are to be included.

h. Use labelled search lines; when logically combining labelled search lines use as few lines as possible for this purpose. Labelled qualifiers help reduce search response time.

i. Use positional logic carefully. Response times with heavily posted terms in conjunction with positional logic are quite long. The use of (+1) logic to specify a word phrase is more restrictive than the use of (SEN) logic. The choice of (+1) or (SEN) positional logic depends on the posting density and precision required on retrieval.

j. When many users are on-line, avoid running searches in which heavily posted terms are used. Such searches already tend to have long run times, and with many users the search response time is made even longer.

2.0 USING INTERACTIVE FEEDBACK TO IMPROVE SEARCH STRATEGIES

Because of the interactive features of CIRCOL, including on-line display, it is desirable to take advantage of these capabilities to improve the search strategies from the standpoint of the content of the search strategies. Part 1 dealt with means of improving search response time by taking into account various search factors and means of applying them. Part 2 suggests methods for enhancing the relevance of the documents retrieved.

2.1 Use of the LRANGE Feature to Estimate the Number of Retrievals

The specification of LRANGE restricts the portion of the file searches, as was indicated in 1.4. Also, the higher the LRANGE (hence the smaller the file), the more rapidly the search can be completed. When a search strategy is formulated, even though it is desired ultimately to search the entire file, a high LRANGE can be specified, and the search will run very quickly. The number of documents which satisfy the search provides some indication of the number of documents which would have been retrieved, if the entire file had been searched.

For example, if an LRANGE of 800,000 had been selected and the data base contains 837,098 documents, the document base being searched is about 40,000 compared to a total file size of about 800,000, a factor of 20. Therefore, the number of documents from the entire file could be expected to be about 20 times as great as the number retrieved using the LRANGE specified. If the number of returns is 20, about 400 could be expected from the entire file. If 0 documents satisfy, a lower LRANGE specification can be made, e.g., 700,000. With this LRANGE, the multiplication factor would change to about 6, e.g., if 5 documents satisfy with an LRANGE of 700,000 the same search of the entire file will result in about 30 retrievals.

2.2 On-line Display of Retrieved Documents

Once documents have been retrieved, the user has the option of obtaining on-line display. With Display Option B, the unclassified text elements can be displayed on-line. Classified text elements may be reviewed by referring to the microfiche library. Only the first five documents are displayed on-line in response to the on-line display command. By scanning the text elements of the documents displayed, the user can get a reasonably good idea of whether the search is actually retrieving appropriate documents. Often this feedback will provide to the user alternative terms for expressing concepts or the scanning of the items may suggest recurring terms which are causing inappropriate retrievals.

Example 1: Methods for solid waste disposal and reclamation

```
  1 OPTION CIRCOLMV, TEXT
  2 $1  solid & waste (sen)
  3 $2  disposal, reclamation, recycl($)
  4 $3  $1 & $2
  5 $4  if classif lt 1
  6 !lrange 500000
OK
  6  end
GOIN' SEARCHIN'
////
  16  DOCS SATISFY
QUALIFY?
  n
OUTPUT FORMAT?
  b
ON LINE?
  y
STANDBY
```

From these returns it was seen that documents on radioactive waste disposal were being recovered. Solid radioactive wastes were definitely not desired, so a search strategy modification was in order.

2.3 Modification of the Initial Search Strategy

2.3.1 Modification Based on Text Elements

Once the initial results have been retrieved and displayed on-line, it is then possible to modify the search strategy and rerun the search. By utilizing the feedback provided by on-line display of text elements, the results achieved in the modified strategy should improve significantly over the initial strategy. The technique can actually be used iteratively until the user is satisfied that an optimum strategy has been formulated. Then the search can be run against the entire file with confidence that the relevance of retrievals will be good. From the previous example, the search strategy was modified as follows:

```
1 OPTION CIRCOLMV, TEXT
2 $1  solid & waste (sen)
3 $2  disposal, reclamation, recycl($)
4 $3  radioactiv($)
5 $4  nuclear
6 $5  $3, $4
```

```
    7  $6  $1 & $2 & $5(not)
    8  $7 if classif lt 1
    9  end
GOIN' SEARCHIN'
```

2.3.2 Modification Based on Personalities

Another effective method of utilizing interactive feedback from on-line display is to make note of the personalities. Generally, specific personalities tend to be involved in certain subject areas. Therefore by making personality search for those personalities identified by the initial search, the user can retrieve additional documents with a high probability that these documents will be related in subject matter to the topic of interest. By requesting these document references on-line and scanning them, the user can also derive keywords which may be helpful in modifying a search to reflect more accurately the subject matter actually desired. A major advantage of this technique is that personality searches run very quickly compared to subject keyword searches. Therefore a number of interactive iterations can be performed with a very short search response time.

2.4 Summary

The use of interactive feedback to improve the relevance of documents retrieved is summarized and stated in the form of guidelines as follows:

a. Specify a high LRANGE for an initial search.

b. Estimate the retrievals from the entire file by multiplying by a factor corresponding to the portion of the file searched, e. g., LRANGE = 700000; DOC FILE = 837, 098: MULTIPLICATION FACTOR = 6.

c. Ask for on-line display of documents with Display Option B.

d. Review output of titles and topic tags and authors by scanning on-line display.

e. Look for terms which may serve as alternate means of expression of basic concepts in the search or which may be causing inappropriate retrievals.

f. Run an "author only" search on the primary authors of those particularly relevant documents retrieved from the initial search strategy. Look for terms as in step e.

g. Modify initial search strategy based on findings from the preceding steps.

h. Repeat steps a-g as desired.

i. Run optimized search against entire file or that portion of the file which was to be searched originally.

APPENDIX A-1

UPDATE INFORMATION

DATE	DPSNR	CIRC PRODUCTION RUN
Oct 2, 1969	333256	-
Mar 8, 1970	378395	-
Apr 17, 1970	391296	-
May 7, 1970	396567	-
May 21, 1970	398474	-
June 14, 1970	411383	-
July 9, 1970	414282	-
July 15, 1970	415278	-
Aug 23, 1970	429192	-
Sept 1, 1970	430651	-
Oct 6, 1970	442520	-
Oct 15, 1970	443913	-
Nov 5, 1970	451219	-
Nov 17, 1970	453610	-
Dec 7, 1970	462034	-
Dec 11, 1970	462960	-
Jan 11, 1971	470720	-
Jan 18, 1971	471603	-
Feb 3, 1971	480226	-
Feb 26, 1971	497478	Dec 25, 1970
May 4, 1971	505749	Jan 5, 1971
May 18, 1971	513982	Feb 5, 1971
June 3, 1971	523411	Feb 19, 1971

UPDATE INFORMATION

DATE	DPSNR	CIRC PRODUCTION RUN
June 6, 1971	533215	Mar 12, 1971
June 9. 1971	541726	Mar 26, 1971
June 13, 1971	550790	Apr 23, 1971
June 16, 1971	558267	May 14, 1971
June 20, 1971	568236	May 28, 1971
July 2, 1971	575477	June 21, 1971
July 23, 1971	584351	July 9, 1971
Aug 16, 1971	595852	Aug 6, 1971
Sept 3, 1971	601286	Aug 20, 1971
Sept 22. 1971	611008	Sept 10, 1971
Oct 20, 1971	617705	Oct 1, 1971
Nov 21, 1971	624927	Oct 15, 1971
Dec 2, 1971	633243	Nov 5, 1971
Dec 29, 1971	642802	Dec 3, 1971
Jan 19, 1972	652393	Dec 24, 1971
Feb 1, 1972	661412	Jan 14, 1972
Feb 28, 1972	669667	Jan 28, 1972
Mar 13, 1972	679476	Feb 18, 1972
Apr 25, 1972	688280	Mar 3, 1972
May 11, 1972	697336	Mar 16, 1972
June 1, 1972	705901	Apr 7, 1972
June 14, 1972	718795	Apr 28, 1972
June 16, 1972	728750	May 15, 1972

UPDATE INFORMATION

DATE	DPSNR	CIRC PRODUCTION RUN
June 23, 1972	739816	June 9, 1972
July 26, 1972	749828	June 30, 1972
Aug 16, 1972	760417	July 28, 1972
Sept 18, 1972	769792	Aug 11, 1972
Sept 27, 1972	781163	Aug 25, 1972
Oct 18, 1972	793881	Sept 15, 1972
Oct 31, 1972	802982	Sept 29, 1972
Nov 13, 1972	814354	Oct 13, 1972
Dec 22, 1972	823395	Nov 24, 1972
Feb 26, 1973	830579	Jan 5, 1973
Apr 2, 1973	837098	Feb 16, 1973
May 11, 1973	841892	Apr 6, 1973

APPENDIX B

FREQUENCY OF USE OF SUBJECT KEYWORDS

Subject Keyword	User Type 1	User Type 2	User Type 3	User Type 4	Overall
Missile	19	10	22	52	103
Institute	9	1	1	88	99
Laser	15	16	11	30	72
Radar	1	24	8	36	69
Research	12	0	7	39	58
Communication	9	7	5	31	52
Aircraft	7	15	2	27	51
Control	11	5	12	17	45
Metal	12	2	3	27	44
Equipment	5	1	4	31	41
Dispersion	0	1	0	38	39
Test	5	6	5	16	32
Combustion	11	1	0	19	31
Detection	0	1	0	30	31
Infrared	2	17	1	11	31
IR	2	3	1	24	30
Air	5	7	6	11	29
High	3	2	12	11	28
Chemical	0	13	3	10	26
Electronic	4	3	3	16	26
Steel	7	0	5	14	26
System	2	3	3	18	26

APPENDIX B (cont'd)

Subject Keyword	User Type 1	User Type 2	User Type 3	User Type 4	Overall
Alloy	10	1	1	13	25
Guidance	0	3	1	21	25
Material	1	2	0	22	25
Plastic	0	1	0	24	25
Propellant	12	4	1	8	25
Plant	4	1	0	19	24
Radiation	0	7	1	16	24
Effect	1	4	2	16	23
Temperature	1	3	5	14	23
Engine	10	1	3	8	22
Production	3	0	0	19	22
Gas	9	5	0	7	21
Solid	8	1	2	10	21
Carbide	9	3	4	4	20
Command	4	0	9	7	20
Reconnaissance	1	1	1	17	20
Conference	16	0	0	3	19
Facility	5	0	1	13	19
Ground	3	4	0	12	19
Space	7	4	0	8	19
Systems	0	6	5	8	19
Tracking	0	9	2	8	19
Warfare	1	13	1	4	19
Analysis	1	2	0	15	18
Antenna	0	1	8	9	18
Explosive	0	0	5	13	18
Moscow	6	0	0	12	18
Telecommunication	6	0	0	12	18
Titanium	6	8	1	3	18
Vehicle	2	1	1	14	18
Weapon	1	0	3	14	18

APPENDIX B (cont'd)

Subject Keyword	User Type 1	User Type 2	User Type 3	User Type 4	Overall
Computer	8	0	0	9	17
Liquid	5	3	3	6	17
Passive	0	0	0	17	17
Power	3	0	1	13	17
Silicon	1	12	0	4	17
Technology	1	4	1	11	17
CW	0	2	0	14	16
Development	5	0	1	10	16
Intercept	0	0	1	15	16
Optic	0	0	0	16	16
Antitank	0	0	11	4	15
Hardening	0	2	0	13	15
Heat	7	3	0	5	15
Theory	1	0	0	14	15
-CH-	2	4	0	8	14
Airfield	0	1	3	10	14
Center	9	0	0	5	14
ECM	1	0	1	12	14
Electric	0	1	1	12	14
Magnetic	0	1	4	9	14
Microwave	6	1	3	4	14
Nickel	10	0	1	3	14
Plasma	0	5	3	6	14
Strengthened	0	0	0	14	14
Telephone	6	0	0	8	14
Compound	1	0	1	11	13
Construction	0	3	0	10	13
Counter-measure	1	4	0	8	13
Electron	3	5	3	2	13
Fiber	1	2	2	9	13
Propulsion	7	1	0	5	13
Pulse	0	4	1	8	13
Site	1	0	0	12	13
Submarine	1	0	2	10	13
ultrasonic	1	1	1	10	13
University	0	0	0	13	13
Warhead	0	4	2	7	13

APPENDIX B (cont'd)

Subject Keyword	User Type 1	User Type 2	User Type 3	User Type 4	Overall
Base	10	0	0	2	12
Carbon	1	2	0	9	12
Engineering	1	0	0	11	12
Facilit($)	4	0	0	8	12
Fuel	4	2	1	5	12
Hardened	0	0	0	12	12
Machine	3	0	0	9	12
Military	4	0	1	7	12
Science	3	0	0	9	12
Storage	0	0	2	10	12
Strengthening	0	0	0	12	12
Surveillance	0	0	0	12	12
Tank	0	1	1	10	12
Carrier	0	1	0	10	11
Cave	0	0	0	11	11
Ceramic	3	5	0	3	11
Design	2	2	0	7	11
Electromagnetic	1	2	0	8	11
Extrusion	1	0	0	10	11
Line	0	3	1	7	11
Method	0	0	0	11	11
Mortar	0	1	0	10	11
Naval	1	0	2	8	11
Physics	2	0	0	9	11
Pollution	0	2	7	2	11
Range	1	0	2	8	11
Satellite	1	0	5	5	11
Superalloys	11	0	0	0	11
Tunnel	0	2	0	9	11
Zinc	4	1	0	6	11

APPENDIX B (cont'd)

Subject Keyword	User Type 1	User Type 2	User Type 3	User Type 4	Overall
A	1	2	0	7	10
Acoustic	2	1	0	7	10
Armored	0	0	0	10	10
Automatic	3	0	2	5	10
Balloon	1	9	0	0	10
D	2	0	0	8	10
Field	1	0	0	9	10
Frequency	0	1	2	7	10
Glass	0	2	2	6	10
Homing	0	2	0	8	10
hydrogen	5	1	3	1	10
Industry	4	0	0	6	10
Lightning	0	2	1	7	10
Low	0	0	2	8	10
Materials	0	3	0	7	10
Powder	3	0	0	7	10
Protection	0	4	0	6	10
R	2	0	0	8	10
SAM	0	0	1	9	10
Self	5	0	0	5	10
Semi-conductor	0	8	1	1	10
Simulation	2	4	0	4	10
Surface	5	0	0	5	10
Thermal	2	2	2	4	10
Vacuum	2	5	1	2	10

APPENDIX B (cont'd)

9 Occurrences

Aluminum
Defence
Electrical
Electronics
Energy
Flow
Helicopter
Horizon
Installation
International
Jamming
Management
Medical
Metallurgy
Motor
Neutron
Oxide
Radio
Reactor
State
Strength
Structure
Training
Transfer
Turbine
Underground
Viral
Wave

8 Occurrences

-CZ-
-FR-
ABM
Academy
Agent
Armor
Ballistic
Cold
Complex
Controlled
Cutting
Economic
Fire
Flight
Forming
Gun
ICBM
Laborator($)
Leningrad
Lithium
Mathematical
Natural
Network
Oxygen
Planning
Selenide
Ship
Silo
Solar
SSM
Superconductivity
Technical
Union
Warning
Water
Zero

7 Occurrences

-PK-
-PO-
Acceleration
Aerosol
Artillery
B
Beam
Central
Climate
Cobalt
Composite
Conjugate
Countermeasures
Data
Develop
Early
Earth
Graphite
Guided
Hangarette
Hybrid
Ionosphere
IRBM
Lasers
Launch
Metallic
Navigation
Nitrogen
Nondestructive
Optical
Personnel
Polar
Pressure
Resistance
Resistant
Shielding
Statistical
Strategic
Sub
Synthesis
Synthetic
Tantalum
Telegraph
Television
Thrust
VLF
Wear

APPENDIX B (cont'd)

6 Occurrences

- -JA-
- Absorption
- Adhesive
- Agents
- Ammunition
- Aviation
- Awacs
- Battery
- Bomb
- Bonding
- Boron.
- Carbamate
- Cesium
- Chemistry
- Chromium
- Circuit
- Coefficient
- Counter
- Defensive
- Density
- Device
- DF
- Elint
- Experimental
- Filament
- Fiscal
- Fuz($)
- GCI
- Generator
- Impulse
- Linear
- Manufactur($)
- Maser
- Minesweeping
- mm
- Model
- MRBM
- Munition
- Particle
- Pay
- Performance
- Photo
- Potassium
- Protective

6 Occurrences

- Reentry
- Refueling
- Remote
- Runway
- Security
- Selenium
- Sensor
- Shook
- SS11
- Station
- Takeoff
- Target
- Telegraphy
- Teleprinter
- Testing
- Transmission
- Transporter
- Treatment
- Tungsten
- TV
- Underwater
- Vapor
- Welding
- ZR

5 Occurrences

- -IT-
- -UK-
- Adaptive
- Administration
- Aerial
- Airborne
- Artic
- Arm($)
- Arsenic
- Atmospheric
- Atomic
- Badger
- Ball
- Bearing
- Cartridge
- Cholera
- Coating
- Creep
- Cross
- Cryogenic
- Decision
- Decoy
- Detector
- Devices
- Discharge
- Doppler
- Dynamic
- Factory
- Fabrication
- Fighter
- Fracture
- Friction
- Games
- Gamma
- Hazard
- Hydride
- Incendiary
- Instrument
- Instrumentation
- Integrated
- Intelligence
- Internal
- Irradiation
- Jammer

APPENDIX B (cont'd)

5 Occurrences

Laboratory
Lanthanum
Locking
Machinery
Matter
Meteoroid
Mine
Molybdenum
Noise
Nonlinear
Numerical
Offensive
Orbita
Organization
Over
Pol
Polytechnical
Press
Problem
Programming
Projectile
Property
Ramjet
Receiver
Red
Reliability
Rolling
Saf($)
Scan
Scanning
Sheet
Simulator
Smoke
Spaceborne
Spectrometer
Spray($)
SS5
Stress
Superconductor
Superconducting
Supersonic
Technological
Time

5 Occurrences

Tools
Transformer
Transister
Two
V
Weapons
X-ray
Zirconium
2