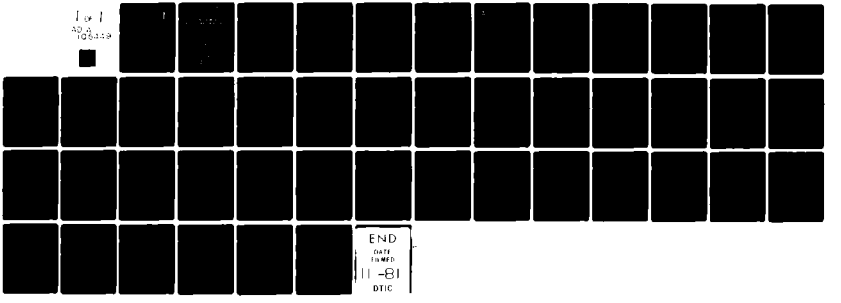
AD-A105 449 DEFENSE TECHNICAL INFORMATION CENTER ALEXANDRIA VA O--ETC F/G 5/2
DEFENSE TECHNICAL INFORMATION CENTER FREE TEXT EXPERIMENT - TEC--ETC(U)
OCT 81 J L CARNEY, C J THOMPSON

UNCLASSIFIED NL



PHOTOGRAPH THIS SHEET

AD A105449

DTIC ACCESSION NUMBER

II

LEVEL

I

INVENTORY

Defense Technical Information Center
Cameron Station, Alex, VA. Office of Information Systems and Technology

Defense Technical Information Center
Free Text Experiment - Technical Report File

DOCUMENT IDENTIFICATION Final Rept. Oct. '81

Approved for public release;
Distribution Unlimited

DISTRIBUTION STATEMENT

ACCESSION FOR	
NTIS	GRA&I
DTIC	TAB
UNANNOUNCED	
JUSTIFICATION	
BY	
DISTRIBUTION /	
AVAILABILITY CODES	
DIST	AVAIL AND/OR SPECIAL
A	

DISTRIBUTION STAMP

SELECTED
OCT 14 1981
D

DATE ACCESSIONED

81 10 14

DATE RECEIVED IN DTIC

PHOTOGRAPH THIS SHEET AND RETURN TO DTIC-DDA-2

DTIC FORM OCT 79 70A

DOCUMENT PROCESSING SHEET

DEFENSE TECHNICAL INFORMATION CENTER FREE TEXT EXPERIMENT - TECHNICAL REPORT FILE

John L. Carney, Project Officer
Carlynn J. Thompson
Charles D. Edmondson

October 1981
Final Report

Office of Information Systems and Technology
Defense Technical Information Center
Cameron Station
Alexandria, Virginia 22314

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) DEFENSE TECHNICAL INFORMATION CENTER FREE TEXT EXPERIMENT - TECHNICAL REPORT FILE		5. TYPE OF REPORT & PERIOD COVERED Information Retrieval Dec 1979 - Jan 1981 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) John L. Carney Carlynn J. Thompson Charles D. Edmondson		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Office of Information Systems and Technology Defense Technical Information Center Cameron Station, Alexandria, Virginia 22314		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE October 1981 13. NUMBER OF PAGES 39
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified - Unlimited 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
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)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Information Retrieval, Free Text Inversion, Defense RDT&E On-Line System, DROLS		

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

DTIC has conducted an experiment that utilizes the free text inversion technique to retrieve records from the Technical Reports data base. It was felt that free text inversion as a form of retrieval would be beneficial in supporting DTIC's revitalized role in the transfer of technical information.

The intent of this project was to determine if the free text inversion technique could be implemented on the DROLS on-line system without severely impacting the on-line system's storage capacity and to determine whether the

added capability could improve upon or replace the present practice of using open-ended/identifier indexing terminology.

Based on the results of the experiment, it was determinded that searching of DTIC's Technical Reports data base is a viable retrieval technique that will provide searchers with an additional enhancement that can be used to augment DTIC's present retrieval methods. However, evidence that will support the use of free text in lieu of the use of identifiers and open-ended terminology remains inconculsive. Additional data collected in a "live environment" will have to be analyzed before a conclusion can be made concerning the use of open-ended terminology.

Although the results of the experiment indicate that it would be feasible for DTIC to implement free text searching of the Technical Reports data base as an alternate way of retrieving documents, there is concern that the critical factors of storage and response time may be adversely affected in a "live environment." Because of this, the study recommends a delay in scheduling this task until the implementation of free text inversion of the management data bases (WUIS and P&DPP) is complete and the results analyzed. Work on this later effort is currently in progress. If there is little evidence of system degradation, it is recommended that DTIC proceed with the free text inversion of the complete Technical Reports file. If there should be an adverse system response; the study proposes several options for further consideration.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)*		5. TYPE OF REPORT & PERIOD COVERED
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR*(s)*		8. CONTRACT OR GRANT NUMBER*(s)*
9. PERFORMING ORGANIZATION NAME AND ADDRESS		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE
		13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)*
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT *(of this Report)*		
17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*		
20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*		

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

TABLE OF CONTENTS

DEFENSE LOGISTICS AGENCY
DEFENSE TECHNICAL INFORMATION CENTER
CAMERON STATION
ALEXANDRIA, VIRGINIA 22314

PREFACE

DTIC-J
Oct 81

As part of DTIC's developmental program, the Office of Information Systems and Technology, in cooperation with the Directorate of Data Systems and Data Base Services, conducted this experiment that utilizes free text to retrieve records from the Technical Report data base. It was felt that this technique as a form of retrieval would be beneficial to our user community and at the same time support DTIC's revitalized role in the DoD technical information program.

The major focus of the experiment was to determine if free text could be implemented on the Defense RDT&E On-Line System (DROLS) without severely impacting computer storage and to determine whether the added capability could improve upon or replace the preseant practice of using open-ended/ identifier terminology.

The results of the experiment showed that free text searching of the Technical Report data base is a viable technique that will provide searchers with an additional enhancement to augment present retrieval methods. However, evidence supporting the use of free text in lieu of open-ended identifiers remains inconclusive.

Although the experiment demonstrated that it would be feasible to implement free text searching of the Technical Report data base, there is some concern that the critical factors of storage and response time may be adversely affected in a "live environment." Therefore, it is recommended that we delay implementation until additional experience is gained from the use of free text in the management data bases (WUIS, R&DPP, IR&D) which is presently operational.

Prepared by:

John L. Carney
Technical Information Specialist

Approved by:

Cecil A. Myatt, JR.
Director, Office of Information Systems and Technology

INTRODUCTION:

This paper describes the methodology, results, and conclusions of an experiment at the Defense Technical Information Center (DTIC) utilizing the free text inversion technique to retrieve records from the Technical Report Data Base. The project was conducted by in-house personnel during the period December 1979 through January 1981.

DISCUSSION:

As part of a new thrust to upgrade the DoD Scientific and Technical Information Program, DTIC has been given the opportunity to undertake a significantly improved effort that provides for new and enhanced technical information services to support the Defense research and development community. These new services are expected to contribute to greater efficiency, productivity, and user satisfaction throughout the DoD community. To achieve this objective it is essential that DTIC enhance its capabilities to employ modern electronic information storage and retrieval technology. DTIC is also to carry out internal research and development to support experimentation and studies directed toward advanced and innovative methods of processing and transferring information. It is within this framework that the experiment with free text was proposed.

Free text inversion as a form of retrieval is one area of information science that is considered to have significant promise in support of DTIC's revitalized mission; especially in the transfer of information. Evidence demonstrating the acceptance and utilization of this form of retrieval by major scientific and technical information on-line systems continues to grow throughout the information community. Examples of major systems using free text supplementing controlled vocabulary include NASA/RECON, NLM/MEDLARS, LOCKHEED/DIALOG, and SDC/ORBIT. Other systems, such as INFOCEN, LEXIS, NEXIS and CIRC II utilize free text inversion as the principal form of retrieval.

Using the free text inversion technique, every word, except common high-frequency words, is extracted from the narrative portion of the document and is posted to an inverted file. This file is then searched using the same techniques DTIC now uses in the Defense RDT&E On-Line System (DROLS).

The intent of this project was to determine if the free text inversion technique could be implemented on the DROLS without severely impacting systems storage and to determine whether the added capability could, in fact, improve upon or replace the present practice of using open-ended/identifier indexing terminology.

APPROACH:

A text inversion and retrieval technique similar to that employed by SDC/ORBIT and NLM/MEDLARS was selected for this experiment. This involved the inversion of all terms not on a "stop word list" and a two-step retrieval technique. The first step being a simple boolean search of single words to isolate all text (records) containing the desired terms. The second step includes a string search (text search) of the direct file for the desired terms in a prescribed sequence or format. The advantage of this approach is that the capability described in the second step has already been implemented by DTIC. Also, the experimental inverted file structure is similar in format to our existing inverted file.

Utilizing as many existing programs as possible the Directorate of Data Systems created a test free text inverted file and collected statistical data concerning that file. The most important question to be answered was one concerning the growth of the inverted file. The Technical Reports were processed in small portions to prevent processing errors and to gain easier access to adequate computer time. For the purposes of this experiment, the abstract and unclassified title fields were processed by the free text programs.

At the outset of the experiment, an inverted file without terms was created. The first group of 15,000 technical reports was inverted via a COBOL program written for this experiment. The output was written in the form of our existing inverted file transactions. As mentioned before, it was planned to use as many of the existing inverted file programs as possible (obviously to save the resources required to write new programs). Transactions were formatted exactly to our production program specifications. The free text terms were assigned role code 55 (TITLE SEARCH KEY). It was recognized that although this role indicator was in use for title search keys, there would be no conflict since the test inverted file contained only free text index terms. Role code 60 will be used if free text searching is implemented.

The inverting program generated seven or eight reels of inverted file transactions for 15,000 documents. These transactions were then sorted by term and accession number. The sort contains its own coding that eliminates duplicate transactions (generated when the same word appears more than once in a document). At first an attempt was made to sort the entire group of transactions in one large sort, but the disk storage requirements and run time were too great. Therefore, pairs of reels were sorted together and then merged to produce the final output in one sequence. The inverted file (on tape) created from the previous group of 15,000 documents was then updated, accumulating more terms.

Each time an updated inverted file was produced, the number of new terms added to that file was noted. The most recent technical reports were processed (AD-A's - unclassified/unlimited and AD-B's - unclassified/limited). No classified documents were included in this experiment, since it simplified the handling of printouts.

The inverted file frequency statistics program (LDA1R) was modified to process the free text inverted files. This program generates a page of statistics showing how many terms are on the inverted file in different frequency ranges. This program was run against each successive inverted file. The following table presents frequency statistics of the test file.

INVERTED FILE FREQUENCY STATISTICS
(For 128,278 Reports there were a total of 5,451,737 postings.)

TERM FREQUENCIES	FREE TEXT	
	Number	Postings
0	0	0
1*	65654	65654
2-3**	24165	56284
4-7	12743	65355
8-15	8090	87414
16-31	5492	121226
32-63	3852	173273
64-127	2789	250582
128-255	1976	357601
256-511	1421	514325
512-1023	983	703593
1024-2947	626	900528
2048-4095	322	914979
4096-8191	118	660152
8192-16383	39	414306
16384-32767***	8	166465
32768-65535	0	0
65536-131071	0	0
Over 131071	0	0

* 65,654 terms had 1 posting for a total of 65,654 postings.
** 24,165 terms had 1 posting for a total of 56,284 postings.
*** 8 terms had 16,384 - 32,767 postings for a total of 166,465 postings.

For the purpose of this experiment, special characters were treated as spaces when inverting text. This means that hyphenated forms such as "state-of-the-art" were broken apart into separate terms and posted separately. "AN/TNH-20" would be posted under "TNH" and "20," since "an" is on the stop word list. This method was selected since it avoids the use of complicated rules (which those searching the files must also remember and use). When using the elimination of special characters the information is stored on the inverted file in a disaggregated form. If hyphenated words were not separated, it would mean that the individual terms could not be searched. Only the entire "bound term" would be searchable. For example, the phase "open-ended" would have to be searched two ways: "open-ended" and "open and ended."

Once the free text inverted file was created, it was loaded to mass storage and made available for testing. The Directorate of Data Base Services and the Office of Information Science and Technology independently developed a strategy to test the retrieval effectiveness of the new inverted file.

Concurrent with building the inverted file, a free text stop word list was developed (appendix A). The list consists of approximately 250 common words which were thought to be of no use in retrieving documents. Articles, prepositions, and common verb forms make up the bulk of this list. U, C, S, SRD, etc. were also added so that the classification indicators would not appear on the inverted file. The objective here was to keep the list as short as possible not only for the convenience of the retriever but also to minimize "no hits." For the purpose of this test the National Library of Medicine (NLM) Stop Word List (appendix D) was used as a basis for the final DTIC list. During the process of creating the inverted test file, frequency counts were collected of all words that were encountered in the 145,000 technical reports sample. The frequency counts were then used to refine the stop word list to fit DTIC's needs more closely. The following is an example of nine words from the inverted test file showing their occurrences:

TR FREQUENCY STATISTICS

No. of occurences	Term
33,191	report
28,244	system
27,907	data
24,516	test
24,019	1
21,868	study
21,751	two
21,225	program
19,396	2

The test stop words and high-frequency words were circulated among the Directorate of Data Base Services retrievers who participated in the Technical Report Data Base free text test and the Vocabulary Control Group. They were asked to mark terms on both lists (test stop words, high frequency words) that they felt should not be included in DTIC's stop word list. These are marked by asterisks (*), on appendixes A and B.

Low frequency words and words the retrievers felt were important were removed from the test stop word list and high frequency words were added to create the final DTIC stop word list (appendix C). In order to present an effective list, and at the same time minimize the burden on the user, we selected a middle of the road approach and constructed a stop word list that is less than half the size of the NLM list (255 words) but more than the System Development Corporation/ORBIT list (9 words).

Instructions for the retrievers were developed which described the procedures for activating the dial-up on-line system for the free text test along with examples to assist in searching the Technical Report test file (appendix E).

Two different experiments were then conducted simultaneously by different DTIC organizational elements. One experiment described in appendix F was conducted in a controlled environment and focused on assessing the retrieval effectiveness of specific terms searched in both the free text system and as identifiers or open-ended terms. The second (appendix G) was performed by retrieval analysts in searching actual and simulated data base queries using the same free text inverted file. This second experiment was primarily concerned with (1) determining the technical capabilities of free text retrieval, (2) limitations of the use of free text retrieval and (3) whether free text could be used as a substitute for or supplement present indexing procedures.

In implementing the test, six retrieval analysts participated on a rotating basis and performed test queries. Other analysts then contributed additional strategies and test questions. No overall statistics are available, but it was estimated that the group put in a total of 28 hours at the terminal. During the test period, approximately 55 bibliographies were batched and examined by retrieval analysts. Some of the tests, with a small number of hits, did not require the batching of bibliographic products and were

examined at the terminal. The bibliographies that were batched for the test involved both user requests and simulated search questions. Although the resulting products were not mailed directly to users, some of them were used for supplementary information for bibliographies that were derived from the regular retrieval process.

FINDINGS:

For this experiment, the title and abstract fields of 145,000 Technical Report records were inverted. The terms were then used to build a test free text inverted file (this file contained only free text terms--no terms from the existing inverted file were used). The inverting program processed documents at the rate of four per second.

The sample free text inverted file was 4000 tracks long. If this inverted file grew linearly, the entire file Technical Report data base would generate a free text inverted file of 28,000 tracks or the equivalent of almost two full disk drives. A graph showing the term growth in the TR Data Base is included as appendix H.

The large increase in the inverted file (it is now 17,000 tracks) will not slow the search routines, since the inverted file is searched by looking at two tables which point to the appropriate entry on the inverted file (two-level index file). The real impact on the system will not be in speed in searching the larger inverted file, but rather in the use of mass storage and the time required to merge larger results.

Free text searching in the controlled experiment produced a greater number of hits than the use of identifiers/open-ended terminology, with a slight decrease (2.5%) in relevancy of the items retrieved. For actual statistics refer to the report (appendix F).

Experimenting with free text searching using customer requests, the analysts reported that there were no technical problems experienced during this limited trial. However, it was found that the use of free text searching cannot adequately replace current methods of indexing and retrieval. Awareness of stop words was found to be an important factor since it is not uncommon that these words occur in searches.

CONCLUSIONS:

Based on the results of both experiments, the use of free text searching of DTIC's Technical Report data base is a viable retrieval technique that will provide searchers with an additional capability to augment DTIC's present retrieval methods. However, evidence that will support the use of free text in lieu of the use of identifiers and open-ended terminology remains inconclusive. Divergent views on this aspect of the project are expressed in each of the experimentation results.

From the data processing point of view it was concluded that the following problems must be addressed before free text is implemented in the Technical Report data base: (1) It is unlikely that the current open-ended/identifiers and the new free text terms should both be stored on the inverted file; maintaining both may make it too large. Therefore, the modification of procedures using open-ended/identifier terminologies will have to be considered. (2) File maintenance programs would have to be modified to cancel any free text terms when a document is cancelled or replaced. (3) Another problem surfaces due to the fact that a term cannot appear on the file both as a subject term and a nonsubject term. If a free text term matches a subject term, (fields 1, 49, or 50) a notation is made on the error list and the term is posted as a subject item. (4) The limitations on numbers of postings in the inverted file will have to be raised.

RECOMMENDATION:

It is recommended that DTIC implement free text searching of the Technical Report data base (both the Direct and Current files) as an alternate way of retrieving documents. However, the critical factors of storage and response time are still of sufficient concern to dictate a cautious approach. Therefore, the scheduling of this task should await the implementation of the free text inversion of the management data bases (WUIS, IR&D, and R&DPP) which is scheduled for May 1981. If after studying the impact this effort has on DROLS, there appears to be little concern about system degradation, DTIC should proceed with the free text inversion of the complete Technical Report file. If, however, there was evidence of more than expected degradation, it is recommended that DTIC consider implementing a limited free text search capability of the unclassified title, field #6. This approach will enable DTIC to test the water in an operational mode with a minimal impact on response time and storage. At the same time this capability will support another development task currently under way which is designed to provide an effective on-line duplicate checking of the Technical Report data base and an enhanced reference capability.

Needless to say, if implementation of free text inversion of the management data base shows considerable system degradation, the Technical Report file implementation should be postponed until further investigation into the causes and remedies of the degradation is conducted. At that time a decision will have to be made whether to continue the project.

Estimates have been made of the resource impact, based on the test of 145,000 technical reports, for three alternative approaches for implementing free text in the technical report system. This assumes an estimated file size of one million records.

Alternative 1: 30 Years of Titles and 30 Years of Abstracts

Mass Storage : 2 Disk packs
Time required to convert : 150 hours processing time (one time)
Programing changes : 1 month (one time)

Alternative 2: 30 Years of Titles and 10 Years of Abstracts

Mass Storate : 1 Disk pack
Time required to convert : 120 hours processing time (one time)
Programing changes : 1 month (one time)

Alternative 3: 30 Years of Titles only

Mass Storage : 1/2 Disk pack
Time required to convert : 90 hours processing time (one time)
Programing changes : 1 month (one time)

ACKNOWLEDGEMENTS:

Other individuals in the agency who have contributed to this experiment include: Ellen V. McCauley, Fuller E. Murfree, Alvin W. Miller, and Thomas F. Lahr.

APPENDIX A: TEST STOP WORDS IN THE TECHNICAL REPORT DATA BASE

A
ABOUT
ACCORDINGLY
AFFECT
AFFECTED
AFFECTION
AFFECTS
AFTER
AGAIN
AGAINST
ALL
*ALMOST
ALREADY
ALSO
ALTHOUGH
ALWAYS
AMONG
AN
AND
ANOTHER
ANY
ANYONE
APPARENTLY
ARE
ARISE
AS
ASIDE
AT
AUTHOR
AWAY
BE
BECAME
BECAUSE
BECOME
BECOMES
BEEN
BEFORE
BEING
BETWEEN
BOTH
BRIEFLY
BUT
BY
C
CAME
CAN
CANNOT
CERTAIN
CERTAINLY
CFRD
COPYRIGHT
COULD
CRD
DID
DIFFERENT
DO
DOES
DONE
DUE
DURING
EACH
*EFFECT
*EFFECTS
EITHER
ELSE
ENOUGH
ESPECIALLY
ETC
EVER
EVERY
FOLLOW
*FOLLOWING
FOR
FOUND
FROM
FURTHER
FY
GAVE
GETS
GIVE
GIVEN
GIVING
GONE
GOT
HAD
HARDLY
HAS
HAVE
HAVING
HERE
HOW
HOWEVER
IF
IMMEDIATELY
IMPORTANCE
IMPORTANT
IN
INTO
IS
IT
ITS
ITSELF
JUST
KEEP
KEPT
*KG
*KM
*KNOWLEDGE
LARGELY
LIKE
MADE
MAINLY
MAKE
MANY
MAY
*MG
MIGHT
*ML
MORE
MOST
MOSTLY
MUCH
MUG
MUST
NEARLY
NEARLY
NECESSARILY
NEITHER
NEXT
NO
NONE
NOR
NORMALLY
NOS
NOT

* = words retrievers noted as being important for retrieval

APPENDIX A (CONT'D)

NOTED
NOW
OBTAIN
OBTAINED
OF
OFTEN
ON
ONLY
OR
OTHER
OUGHT
OUR
OUT
OVERALL
OWING
PARTICULARLY
PAST
PERHAPS
PLEASE
POORLY
POSSIBLE
POSSIBLY
POTENTIALLY
PREDOMINANTLY
PRESENT
PREVIOUSLY
PRIMARILY
PROBABLY
PROMPT
*PROMPTLY
QUICKLY
QUITE
RATHER
READILY
REALLY
RECENTLY
REGARDING
REGARDLESS
RELATIVELY
RESPECTIVELY
RESULTED
RESULTING
RESULTS
S
SAID
SAME
SEEM
SEEN
SEVERAL
SFRD
SHALL
SHOULD
*SHOW
SHOWED
SHOWN
SHOWS
SIGNIFICANTLY
SIMIALR
SIMILARLY
SINCE
SLIGHTLY
SO
SOME
SOMETIME
SOMEWHAT
SOON
SPECIFICALLY
SRD
*STATE
*STATES
STRONGLY
SUBSTANTIALLY
SUCCESSFULLY
SUCH
SUFFICIENTLY
THAN
THAT
THE
THEIR
THEIRS
THEM
THEN
THERE
THEREFORE
THESE
THEY
THIS
THOSE
THOUGH
THROUGH
THROUGHOUT
TO
TOO
TOWARD
U
UNDER
UNLESS
UNTIL
UP
UPON
USE
USED
USEFULLY
USEFULNESS
USING
USUALLY
VARIOUS
VERY
WAS
WERE
WHAT
WHEN
WHERE
WHETHER
WHICH
WHILE
WHO
WHOSE
WHY
WIDELY
WILL
WITH
WITHIN
WITHOUT
WOULD
YES
YET

* = words retrievers noted as being important for retrieval

APPENDIX B: HIGH FREQUENCY POTENTIAL STOP WORDS

ABILITY
ABOVE
ACCOMPLISHED
ACCORDANCE
ACCORDING
ACHIEVED
ADDITIONAL
ADEQUATE
ALLOW
ALLOWS
ALONG
AMOUNT
ANALYZED
APPARENT
APPEAR
APPEARS
APPENDIX
APPROPRIATE
*APPROXIMATELY
ASSIGNED
ASSIST
ASSUMED
ATTEMPT
ATTEMPTS
ATTENTION
AVAILABLE
BEST
BETTER
BRIEF
CALCULATE
CALCULATED
CALLED
CAUSE
CAUSED
CAUSES
CHARACTERIZED
CHOSEN
COLLECTED
COMPARABLE
*COMPARATIVE
COMPARE
COMPLETE
COMPLETED
COMPLETELY
CONCERNED
CONCERNING
CONCLUDED
CONCLUDES
*CONCLUSIONS
CONDUCTED
CONSIDERABLE
CONSIDERED
CONSISTED
CONSISTENT
DESCRIBE
DESCRIBE
DESCRIBES
DESIRED
DESIRED
DETAILED
DETAILS
*DETERMINATION
DETERMINDED
DETERMINE
DETERMINING
*DEVELOPED
*DEVELOPING
DIRECTED
DISCUSSED
DISCUSSES
DISCUSSION
EARLIER
EFFORTS
EMPHASIS
ENCOUNTERED
ENTIRE
ESTABLISH
ESTABLISHED
EVALUATE
EVALUATED
EXAMINED
EXAMINES
EXCEPT
EXIST
EXISTING
EXPECTED
EXTENT
FEATURES
FEW
FINALLY
FOLLOWED
FULLY
*FUNDAMENTAL
GENERAL
GENERALIZED
GENERALLY
GENERATED
GIVES
GOOD
GREATER
HIS
IDENTIFIED
IMPLEMENTED
*IMPORVED
IMPROVING
INCLUDE
INCLUDED
INCLUDES

* = words retrievers noted as being important for retrieval

APPENDIX B (CONT'D)

INCLUDING
INCREASED
INCREASING
INDICATE
INDICATED
INDICATES
INPROVE
INSTALLED
INTENDED
*INTEREST
INTRODUCED
*INTRODUCTION
INVESTIGATE
INVESTIGATED
INVOLVED
INVOLVING
KNOWN
LATTER
LIMITED
*LOCATED
MEANS
*MEASURES
MEET
MET
NECESSARY
NEED
NEEDED
*OBJECTIVES
*OBSERVED
OBTAINING
OCCUR
OCCURRED
OCCURS
OPERATED
OPTIMUM
PARTICULAR
PERFORM
PERFORMED
PERMIT
PLACED
PORTION
POSSIBILITY
*PREDICTED
*PREPARED
PRESENTED
PRESENTLY
PRESENTS
PREVIOUS
PRIOR
*PRODUCE
PRODUCED
*PROPOSED
PROVIDE
PROVIDED
PROVIDES
PROVIDING
*PROVING
PURPOSES
RECEIVED
RECENT
*RECOMMENDATIONS
*RECOMMENDED
*REDUCED
RELATED
REPORT
REPORTED
REQUIRE
REQUIRED
REQUIRES
RESULTS
REVEALED
REVIEW
REVIEWED
SELECTED
SIGNIFICANT
SPECIFIED
STILL
STUDIED
*STUDIES
*STUDY
SUBJECTED
SUBSEQUENT
SUCCESSFUL
SUGGEST
SUGGESTED
SUITABLE
*SUMMARIES
SUMMARIZED
TAKEN
TESTED
THUS
TOGETHER
TYPES
TYPICAL
UNDERTAKEN
USEFUL
UTILIZED
UTILIZING
VARIETY
WAY
WE
WELL
WRITTEN

* = words retrievers noted as being important for retrieval

APPENDIX C: STOP WORD LIST

A
AFTER
ALSO
AN
AND
ANY
ARE
AS
AT
AUTHOR
AVAILABLE

BE
BEEN
BEING
BETWEEN
BOTH
BUT
BY

C
CAN
CFRD
CONDUCTED
CONSIDERED
COULD
CRD

DESCRIBED
DESCRIBES
DESIGNED
DETERMINE
DETERMINED
DIFFERENT
DISCUSSED
DUE
DURING

EACH

FOR
FOUND
FROM
FURTHER

GENERAL
GIVEN

HAS
HAVE
HOWEVER

IF
IN
INCLUDED
INTO
INVESTIGATED
IS
IT
ITS

MADE
MAY
MORE
MOST

NO
NOT

OBTAINED
OF
ON
ONLY
OR
OTHER
OUT

PERFORMED
POSSIBLE
PRESENT
PRESENTED
PRESENTS
PROVIDE
PROVIDED
PROVIDES

RELATED
REPORT
REQUIRED
RESULTS

S
SEE
SELECTED
SEVERAL
SFRD
SHOULD
SHOWN
SIGNIFICANT
SOME
SRD
STUDIES
SUCH

TESTED
THAN
THAT
THE
THEIR
THERE
THESE
THEY
THIS
THOSE
THROUGH
TO
TYPES

U
UNDER
UP
USE
USED
USING

VARIOUS
VERY

WAS
WERE
WELL
WHEN
WHERE
WHICH
WHILE
WILL
WITH
WITHIN
WITHOUT
WOULD

APPENDIX D: ELLHILL STOPWORD LIST

A
ABS
ABOUT
ACCORDINGLY
AFFECT
AFFECTED
AFFECTING
AFFECTS
AFTER
AGAIN
AGAINST
ALL
ALMOST
ALREADY
ALSO
ALTHOUGH
ALWAYS
AMONG
AN
AND
ANOTHER
ANY
ANYONE
APPARENTLY
ARE
ARISE
AS
ASIDE
AT
AWAY
BE
BECOME
BECAUSE
BECOME
BECOMES
BEEN
BEFORE
BEING
BETWEEN
BIOL
BOTH
BRIEFLY
BUT
BY
CAME
CAN
CANNOT
CERTAIN
CERTAINLY
CHEM
COPYRIGHT
COULD
DID
DIFFERENT
DO
DOES
DONE
DUE
DURING
EACH
EFFECT
EFFECTS
EITHER
ELSE
ENOUGH
ESPECIALLY
ETC
EVER
EVERY
FOLLOWING
FOR
FOUND
FROM
FURTHER
GAVE
GETS
GIVE
GIVEN
GIVING
GONE
GOT
HAD
HAS
HARDLY
HAVE
HAVING
HERE
HOW
HOWEVER
IF
IMMEDIATELY
IMPORTANCE
IMPORTANT
IN
INTO
IS
IT
ITS
ITSELF
JUST
KEEP
KEPT
KG
KM
KNOWLEDGE
LARGELY
LIKE
MADE
MAINLY
MAKE
MANY
MAY
MG
MIGHT
ML
MORE
MOST
MOSTLY
MUCH
MUG
MUST
NEARLY
NECESSARILY
NEITHER
NEXT
NO
NONE
NOR
NORMALLY
NOS
NOT
NOTED
NOW
OBTAIN
OBTAINED
OF
OFTEN
ON
ONLY
OR
OTHER
OUGHT
OUR
OUT
OVERALL
OWING
PARTICULARLY
PAST
PERHAPS
PLEASE
POORLY
POSSIBLE
POSSIBLY
POTENTIALLY
PREDOMINANTLY
PRESENT
PREVIOUSLY
PRIMARILY
PROBABLY
PROMPT
PROMPTLY
QUICKLY
QUITE
RATHER
READILY
REALLY
RECENTLY
REFS
REGARDING
REGARDLESS
RELATIVELY
RESPECTIVELY
RESULTED
RESULTING
RESULTS
SAID
SAME
SEEM
SEEN
SEVERAL
SHALL
SHOULD
SHOW
SHOWED
SHOWN
SHOWS
SIGNIFICANTLY
SIMILAR
SIMILARLY
SINCE
SLIGHTLY
SO
SOME
SOMETIME
SOMEWHAT
SOON
SPECIFICALLY
STATE
STATES
STRONGLY
SUBSTANTIALLY
SUCCESSFULLY
SUCH
SUFFICIENTLY
THAN
THAT
THE
THEIR
THEIRS
THEM
THEN
THERE
THEREFORE
THESE
THEY
THIS
THOSE
THOUGH
THROUGH
THROUGHOUT
TO
TOO
TOWARD
UNDER
UNLESS
UNTIL
UP
UPON
USE
USED
USEFULLY
USEFULNESS
USING
USUALLY
VARIOUS
VERY
WAS
WERE
WHAT
WHEN
WHERE
WHETHER
WHICH
WHILE
WHO
WHOSE
WHY
WIDELY
WILL
WITH
WITHIN
WITHOUT
WOULD
YET

APPENDIX E: TEXT WORD SEARCHING IN THE TECHNICAL REPORT TEST FILE

1. Special Characters are treated as blanks.

phrase	search statement	
AN/800-1	800 and 1	(AN is on the stop list)
F-15	F and 15	
TF30-P-3	TF30 and P and 3	
Cobalt-alloyed silicon	Cobalt and alloyed and silicon	

2. The inverted file is constructed of single words from the title and abstract.

3. A Stop Word list is enclosed.

4. 150,000 Technical Reports were inverted in the test file.
The ranges are:

ADA-000001 - ADA-075000
ADB-000001 - ADB-045000
AD 900000 - AD 924000

5. Free text allows the searcher to find very narrow highly relevant documents if used properly. Do not search free text for dogs if you are looking for documents on German Shepherds.

6. Some types of searches that free text is appropriate for include:

a. Variant Spellings - color or colour

b. Variant Word Forms - mine, mines, mined, mining

c. Misspellings - retrieval or retreival

d. Alpha Numerics - M -16

e. Chemical Terminology - Tetracycline

f. Names - Laplace transform, Einstein Equations, or Hodgkins Disease

g. Foreign Words - Radiatsii (Radiation)

h. Other Controlled Vocabulary

7. Free text inverted file searching can be used in conjunction with the text scan. The Free Text Inverted File search narrows the search to documents which contain the words to be scanned.

Example: Free text search for black cats.

Inverted File Search

@STR@

black

and

cats

end

Text Scan Search

@SRTAB@

black cats

end

8. Synonyms should be considered as part of the search strategy:

Example: MARIJUANA, MARIHUANA, POT, GRASS, WEED, MARY JANE

APPENDIX F:

FREE TEXT VERSUS UNCONTOLLED VOCABULARY

A PERFORMANCE COMPARISON

Thomas F. Lahr

Information Science Intern Program

Six Month Paper

24 December 1980

TABLE OF CONTENTS

BACKGROUND

The objective of this study is to evaluate the retrieval effectiveness of the Technical Reports (TR) database of the Defense RDT&E On-Line System (DROLS) when a free text generated index file is used instead of indexer assigned uncontrolled vocabulary.

The documents in the Technical Reports database are assigned various posting terms from the thesaurus or controlled vocabulary (DTIC Retrieval and Indexing Terminology-DRIT). In addition, indexers have the option of assigning terms not found in the controlled vocabulary which are known as identifiers or open-ended terms.

These terms have historically been assigned along with the controlled vocabulary, to pick up topics where a main idea or concept of a report is not covered in the thesaurus. An identifier was assigned to describe a very specific item, usually an alpha-numeric, which would represent a project, code name, equipment model number, etc. Examples of identifiers are: F 104 Fighter, AN/SPS-39, and Plumbob Project. Open-ended terms have been assigned to describe new technology or concepts, acronyms, author suggested terms, etc. Previously a distinction was made in the database as to whether a term was an identifier or an open-ended term, but currently they are both labelled as identifiers in the Techical Reports file.

The free text file in the TR database contains single words taken from the titles and abstracts and are directly searchable. The free text inverted file consists of:

(1) Alphabetic, alphanumeric or numeric strings of characters up to 60 characters in length.

(2) All special characters (commas, periods, slash marks, colons, etc.) are converted to blanks which serve as term delimiters.

(3) A term which is present on the stop word list is discarded (see Attachment A).

As an example, the following Technical Report abstract will provide the listed free text terms.

> A SELECTIVE DETECTION SCHEME FOR ATOMS IN THE METASTABLE 2S STATE OF HYDROGEN THAT PROVIDES THE HIGH SPATIAL RESOLUTION (0.1 CM) NECESSARY FOR TIME-OF-FLIGHT ATOMIC BEAM STUDIES IS DESCRIBED. THE SCHEME UTILIZES THE LYMAN PHOTON EMITTED WHEN THE METASTABLE IS DE-EXCITED IN AN ELECTRIC FIELD VIA THE STARK EFFECT. DETAILS OF CONSTRUCTION AND OPERATION ARE DISCUSSED.

SELECTIVE	FLIGHT	STARK
DETECTION	ATOMIC	DETAILS
SCHEME	BEAM	CONSTRUCTION
ATOMS	STUDIES	OPERATION
METASTABLE	DESCRIBED	DISCUSSED
2S	SCHEME	
HYDROGEN	UTILIZES	
PROVIDES	LYMAN	
HIGH	PHOTON	
SPATIAL	EMITTED	
RESOLUTION	METASTABLE	
0	DE	
1	EXCITED	
CM	ELECTRIC	
NECESSARY	FIELD	
TIME	VIA	

This report will assess the retrieval effectiveness of specific terms searched in both the free text system and as identifiers/open-ended terms using the records for 150,000 entries in the Technical Reports data base.

METHODOLOGY

A collection of terms was put together which are felt to be representative of typical terms that a DROLS user might come up with during a search, not found in the DTIC Retrieval and Indexing Terminology. A total of 212 terms were chosen to be searched. Of these, 100 to 125 were chosen from the "Combined Frequency Count" which is a multivolume, alphabetical listing of DRIT terms and identifiers, along with their frequency of occurence in the DROLS databases. These specific words were used in order to assure a number of search terms with known hits as identifiers in the Technical Reports file. In contrast to this, approximately 75 to 100 words or word phrases were chosen without reference to the "Combined Frequency Count". Most of them relate in some way to the subject content of the TR database. Also included is a sampling of subject areas not normally connected to the Department of Defense, but which may be representative of certain needs of DROLS users, and of which research may have been performed by DoD.

The test was done on each individual term (a term may be one or more words, or alphanumerics, not found in the DRIT) not on specific searches, strategies, or combination of terms.

After the approximately 200 search terms were chosen, they were individually searched in the Technical Reports database using the terms as indexer assigned keywords. Since the free text file was only loaded for a certain set of AD (Accessioned Document) number ranges (AD900000-AD924000, ADA000001-ADA075000, ADB000001-ADB045000), the searches were

limited to those ranges only. It was decided that terms having up to 15 hits would be included in the relevancy check. Occurences greater than 15 were included in the overall totals, but not in the relevancy count. Bibliographies were ordered for the search terms with up to 15 hits. All bibliographies were then checked for relevancy to the term searched. If there were any questions as to the relevancy of a specific item, a copy of the document itself was reviewed and rated. All items in each bibliography were designated as relevant, marginally relevant, or not relevant.

Searches of the same terms were done on the free text file. Terms containing more than one word were searched using the Boolean operator "AND" (in the keyword system they had been searched on one level as a single multiword index term). A term such as AGENT ORANGE would be searched in the following manner:

As an identifier-

```
@STR@
AGENT ORANGE
END
```

In the free text file-

```
@STR@
AGENT
AND
ORANGE
END
```

In the free text test, if a search term resulted in more than 15 hits, a qualifying search was done, performing a text scan on the hits.

This utilized the ability to string search (search based on the physical relationship of the words in the term). Text scan was done only if the term had more than one word or alphanumeric grouping in it. Single words were not qualified by string searching. Bibliographies were then ordered for those terms having 15 hits or less.

Again, all the bibliographic references for each term were checked for relevancy to the term searched, and each item was designated as relevant, marginally relevant, or not relevant. In any instances where the relevancy was in doubt, a copy of the document itself was looked at and checked. Relevancy statistics are presented only for terms having 15 or less hits in both the identifier and free text systems.

RESULTS

The 212 terms searched produced a total of 334 hits as identifiers, and 5998 hits in the free text system (this was reduced to 3930 hits after string searching of multiple word terms). Of these, 52 terms (24.53%) had no hits in either system and 38 terms (17.92%) had greater than 15 hits in both systems and therefore not checked for relevancy.

Of the 212 terms, slightly greater than 50% (122 terms) provided a number of hits (0-15) which were then checked for relevancy Twelve terms produced hits as identifiers, with no hits in the free text system. Forty eight terms had hits in the free text, with none as identifiers. Sixty two terms resulted in hits in both systems. In totaling these up for the relevancy count, the 122 terms searched as identifiers resulted in 187 hits, and in the free text system 596 hits. The 187 hits from the identifier searches consist of 103 that were determined to be relevant (55.08%), 73 that were marginally relevant (39.04%), and 11 hits not relevant (5.88%). In the 596 free text hits, 313 were found to be relevant (52.53%), 217 marginally relevant (36.41%), and 66 hits not relevant (11.07%).

DISCUSSION

The conclusion that one can draw from the results of the test is that the use of the free test searching produces approximately three times as many hits as using identifiers/open-ended terminology, with only a slight (2.5%) decrease in the relevancy of the items retrieved. There are instances of items not found in the controlled vocabulary where the use of free text is beneficial, such as variant spellings and word forms, alphanumerics, chemical terminology, foreign names, proper names, etc. The free text searching technique becomes an additional means for the search analyst to augment the search performance of the system. It allows the searcher to get at specifics that the controlled vocabulary does not directly address. Naturally, there are terms that one would not normally use in a free text system, where use of the controlled vocabulary and a defined search strategy would be necessary to narrow down the results. In this study, some of the searches that were not checked for relevancy, because the results numbered in the hundreds, would have to be further defined using search alternatives and perhaps some of them would not necessarily be searched using free text. Free text searching provides a viable alternative to the use of uncontrolled vocabulary and in any retrieval system can prove to be a valuable tool.

ATTACHMENT A

TEST STOP WORDS IN THE TECHNICAL REPORT DATA BASE

A
ABOUT
ACCORDINGLY
AFFECT
AFFECTED
AFFECTION
AFFECTS
AFTER
AGAIN
AGAINST
ALL
ALMOST
ALREADY
ALSO
ALTHOUGH
ALWAYS
AMONG
AN
AND
ANOTHER
ANY
ANYONE
APPARENTLY
ARE
ARISE
AS
ASIDE
AT
AUTHOR
AWAY
BE
BECAME
BECAUSE
BECOME
BECOMES
BEEN
BEFORE
BEING
BETWEEN
BOTH
BRIEFLY
BUT
BY
C
CAME
CAN
CANNOT HOWEVER
CERTAIN
CERTAINLY
CFRD
COPYRIGHT
COULD
CRD
DID
DIFFERENT
DO
DOES
DONE
DUE
DURING
EACH
EFFECT
EFFECTS
EITHER
ELSE
ENOUGH
ESPECIALLY
ETC
EVER
EVERY
FOLLOW
FOLLOWING
FOR
FOUND
FROM
FURTHER
FY
GAVE
GETS
GIVE
GIVEN
GIVING
GONE
GOT
HAD
HARDLY
HAS
HAVE
HAVING
HERE
HOW
HOWEVER
IF
IMMEDIATELY
IMPORTANCE
IMPORTANT
IN
INTO
IS
IT
ITS
ITSELF
JUST
KEEP
KEPT
KG
KM
KNOWLEDGE
LARGELY
LIKE
MADE
MAINLY
MAKE
MANY
MAY
MG
MIGHT
ML
MORE
MOST
MOSTLY
MUCH
MUG
MUST
NEARLY
NEARLY
NECESSARILY
NEITHER
NEXT
NO
NONE
NOR
NORMALLY
NOS
NOT

TEST STOP WORDS (cont.)

NOTED
NOW
OBTAIN
OBTAINED
OF
OFTEN
ON
ONLY
OR
OTHER
OUGHT
OUR
OUT
OVERALL
OWING
PARTICULARLY
PAST
PERHAPS
PLEASE
POORLY
POSSIBLE
POSSIBLY
POTENTIALLY
PREDOMINANTLY
PRESENT
PREVIOUSLY
PRIMARILY
PROBABLY
PROMPT
PROMPTLY
QUICKLY
QUITE
RATHER
READILY
REALLY
RECENTLY
REGARDING
REGARDLESS
RELATIVELY
RESPECTIVELY
RESULTED
RESULTING
RESULTS
S
SAID
SAME
SEEM
SEEN
SEVERAL
SFRD
SHALL
SHOULD
SHOW
SHOWED
SHOWN
SHOWS
SIGNIFICANTLY
SIMIALR
SIMILARLY
SINCE
SLIGHTLY
SO
SOME
SOMETIME
SOMEWHAT
SOON
SPECIFICALLY
SRD
STATE
STATES
STRONGLY
SUBSTANTIALLY
SUCCESSFULLY
SUCH
SUFFICIENTLY
THAN
THAT
THE
THEIR
THEIRS
THEM
THEN
THERE
THEREFORE
THESE
THEY
THIS
THOSE
THOUGH
THROUGH
THROUGHOUT
TO
TOO
TOWARD
U
UNDER
UNLESS
UNTIL
UP
UPON
USE
USED
USEFULLY
USEFULNESS
USING
USUALLY
VARIOUS
VERY
WAS
WERE
WHAT
WHEN
WHERE
WHETHER
WHICH
WHILE
WHO
WHOSE
WHY
WIDELY
WILL
WITH
WITHIN
WITHOUT
WOULD
YES
YET

ATTACHMENT B-TOTALS, STATISTICS AND RELEVANCY

TOTALS

	HITS	PERCENTAGE
212 Terms Searched	6332	
Identifiers/Open-ended	334	5.27%
Free Text	5998	94.73%
After string search of 30 of the 212 terms (in free text only)	4264	
Identifiers/Open-ended	334	7.83%
Free Text	3930	92.17%

STATISTICS

	HITS/TERM	PERCENTAGE
52 Terms		24.53%
Identifiers/Open-ended	0	
Free Text	0	
38 Terms		17.92%
Identifiers/Open-ended	GT 15	
Free Text	GT 15	
122 Terms Checked for Relevancy		57.55%
Identifiers/Open-ended	0-15	
Free Text	0-15	
(a) 12 Terms		5.66%
Identifiers/Open-ended	1-15	
Free Text	0	
(b) 48 Terms		22.64%
Identifiers/Open-ended	0	
Free Text	1-15	
(c) 62 Terms		29.25%
Identifiers/Open-ended	1-15	
Free Text	1-15	

RELEVANCEY

	HITS/TERM	HITS TOTAL	PERCENTAGE
122 TERMS		783	
Identifiers/Open-ended	0-15	187	23.88%
103 Hits relative			55.08%
73 Hits marginally relative			39.04%
11 Hits not relative			5.88%
Free Text	0-15	596	76.12%
313 Hits relative			52.53%
217 Hits marginally relative			39.04%
66 Hits not relative			11.07%

ATTACHMENT C-TERMS

TERM	IDENTIFIERS/OPEN-ENDED HITS	REL	MARG	NOT	FREE TEXT HITS	STRING	REL	MARG	NOT
155 MM HOWITZERS	0				14		5	6	3
ACETAMETAPHYN	0				0				
AGENT ORANGE	0				3			1	2
AIRCRAFT PICKETING	0				0				
AN/GSQ-120	3	2	1		3		3		
AN/SPS-39	3		2	1	2			2	
AN/TSW	1	1			12		12		
APERTURE CARDS	0				3		1		2
ARAMID	3	1	2		15		11	4	
ARCADENE PROPELLANT	0				0				
ARMY CHAPLAINS	1	1			4		3		1
ARSENALS	5				17				
ARTIFICAL FOG	1	1			8		5	1	2
ARTIFICAL INSEMINATION	0				0				
ATMOSPHERIC ABSORPTION	4				179	43			
ATOMIC ENERGY COMMISSION	0				18	17			
ASIAN INFLUENZA	0				1		1		
BAGGAGE	3	1	1	1	10		2	8	
BAKED HAM	1				1			1	
BAKERY PRODUCTS	0				0				
BASEMENT SHELTERS	0				13		12	1	
BATTLEFIELD ILLUMINATION	2	2			12		7	2	3
BENZENE HEXACHLORIDE	0				1		1		
BETA MODELS	0				21	0			
BEVERLY	0				0				
BICYCLES	3	1		2	2		1	1	
BIONIC SONAR	0				3		3		
BLACKOUT	0				33				
BLACK HOLES	3	3			3		2		1
BOMB TESTS	0				188	9	4	5	
BURG MAXIMUM ENTROPY METHOD	0				4		3	1	
C3I	0				2		2		
CARRIER PIGEONS	0				0				
CHAIN GUN	0				6		2		4
CHEMICAL DEMULSIFICATION	1	1			1		1		
CHIRT	0				0				
CLONES	3	2	1		11		2	9	
CODE READING	0				7		3	4	
CODEINE	3	1	2		3		1	2	
CRUDE EMULSIONS	0				0				
D2 PLASMA	0				1			1	
DRAGON	0				32				
EC135C	0				0				

TERMS

TERM	IDENTIFIERS/OPEN-ENDED				FREE TEXT				
	HITS	REL	MARG	NOT	HITS	STRING	REL	MARG	NOT
ELECTROMETALLURGY	4	3	1		7		4	3	
FALCON MISSILES	0				0				
FISH IMPINGEMENT	0				0				
FLAKE CUTTING	0				0				
FLEET SATELLITE COMMUNICATIONS	0				0				
FLT SAT COM	0				0				
FLYING SQUIRRELS	0				3		1	2	
FOAM IN PLACE	2				19				
FREEBASE	0				0				
FRIENDSHIP	1	1			8		1	3	4
FROGMEN	1	1			2		1	1	
FUNGUS-EATER GAMES	0				0				
GAS GENERATORS	4				92	42			
GAVIONS	0				0				
GEODSS	0				0				
GROYNES	0				0				
H-21C AIRCRAFT	0				0				
HABITAT	0				130				
HARPOON	1				48				
HEART FAILURE	0				15		7	2	6
HELMET ANTENNAS	1		1		0				
HELMET IMAGING AND POINTING SYSTEM	0				2		2		
HIPS	0				3		1	2	
HOSPITAL SHIPS	1	1			2		2		
HOT GOGGLE	0				0				
HOT SPOT SIGNATURES	1		1		1		1		
HYDRA	0				6			5	1
HYPODERMIC SYRINGES	1	1			1		1		
IADT	0				0				
ICE BORING	0				0				
ICEBERG PROJECT	0				1		1		
IFV	0				10		9	1	
IGLOO	0				21				
INFANTRY FIGHTING VEHICLE	1	1			11		7	3	1
INFLATABLE WINGS	0				0				
INFORMATION HIDING MODULES	0				0				
INTEGRATED AUTOMATIC DETECTION AND TRACKING	0				5		3	1	1
ION CONCENTRATION	0				128	28			
J-52-W-16A ENGINES	0				0				
JND GLASS LASER	0				0				
JP-1 FUEL	0				40	0			
KAUAI	0				8		2	6	
KH2PO4	0				3		2	1	

TERMS

TERM	IDENTIFIERS/OPEN-ENDED				FREE TEXT				
	HITS	REL	MARG	NOT	HITS	STRING	REL	MARG	NOT
KNAPSACKS	1	1			1			1	
KRAKATOA	1		1		0				
LABORATORY RODENTS	0				10		3	6	1
LAND BASED OPERATIONS	1		1		18	1	1		
LASER GUNS	0				10		5	1	4
LASER SCATTERING	4				253	19			
LATRINES	4	4			3		2	1	
LAUNCHING CHUTES	1	1			1		1		
LAUNDRY WASTE WATER	0				10		3	7	
LEAD SULFIDE	0				19	11	6	5	
LEADERSHIP STYLE	6				28	19			
LIGHTNING BUGS	0				0				
LNG TERMINALS	0				2		1	1	
MARK 48 TORPEDO	0				2		2		
MARLINESPIKE SEAMANSHIP	0				0				
MAXIMUM LIKELIHOOD METHOD	1				86	23			
MAYDAY	0				1		1		
METAL CARBONITRIDES	0				0				
MIDSHIPMEN	8	6	2		14		4	6	4
MINE CLIMATOLOGY	0				0				
MINIATURE BALLBEARING	0				0				
MIXED GAS BREATHING APPARATUS	1	1			10		9	1	
MOLASSES	1	1			2		1	1	
MOMCOP	0				0				
MONGOLIAN GERBILS	2	1	1		4			4	
MONOGAMY	0				0				
MOONWATCH	0				0				
MOONQUAKES	0				0				
MOPED	1	1			1		1		
MOVING MINES	0				1				1
MUSK OX	0				0				
MX	0				155				
NATIONAL ENVIRONMENTAL POLICY ACT	4		4		8		6	2	
NATIONAL GOALS	11	3	7	1	59	9	4	4	1
NATIONAL GOVERNMENT	8				109	0			
NATIONAL HEALTH INSURANCE	2	2			5		5		
NATIONAL INTEREST	2				72	23			
NERVE GASES	2	2			0				
NEWT	0				1			1	
NUCLEAR CIVIL PREPAREDNESS	1		1		15		13	2	
NUCLEAR POWER	1				275	79			
NUDE MICE	1	1			8		2	6	
NUDET	0				0				
OIL EMBARGO	1	1			21	14	8	2	4

TERMS

T E R M	IDENTIFIERS/OPEN-ENDED				FREE TEXT				
	HITS	REL	MARG	NOT	HITS	STRING	REL	MARG	NOT
OPTICAL BOMBS	0				5		3	2	
PEANUT BUTTER	0					0			
PHASE TRANSITION MATERIALS	0				25	0			
PHOTOVOLATIC SYSTEMS	0				0				
PILING	0				18				
PLANOGRAPH	0				0				
PICTOMAP	0				0				
PRECAST CONCRETE	1		1		13		4	8	1
PRESSURE POINTS	0				93	4	1		3
PINK WATER	6	5	1		17	15	11	4	
RACISM	2	2			8		7	1	
RADAR ECHO	0				54	15	7	8	
RADIATIONS EFFECTS (HUMANS)	2	2			0				
RAFTS	3		2	1	14		10	3	1
RAINFORESTS	3	1	2		1		1		
RARE GAS COMPOUNDS	1		1		2			1	1
RED FUMING NITRIC ACID	0				4		1	3	
RED WATER	6	4	2		59	12	3	9	
RIFLEMANS ASSUALT WEAPON	0				0				
RING CUSP	0				2			2	
RIPCORD HANDLES	0				0				
RIPRAP	17				54				
ROSS ICE SHELF	2	2			3		2	1	
SACCHARIN	2	2			2		2		
SAILBOATS	0				1			1	
SALIMANDERS	0				0				
SALT DOMES	3	1	1	1	0				
SALT MARSHES	22				9				
SAND DUNES	7				16				
SATELLITE DETECTION	4	3	1		51	3	1	2	
SATURN	7		7		13		5	8	
SCIENCE	3				716				
SCHOOL DROPOUTS	0				0				
SCUBA DIVING EQUIPMENT	0				6		5	1	
SCUBA EQUIPMENT	0				8		6	1	1
SEA COWS	1	1			1			1	
SHIP ACCIDENTS	3	2	1		6		6		
SKYCRANE	0				4		2	2	
SOIL LIQUIFICATION	0				0				
SPS-39 RADAR	0				1			1	
SQUIB	0				27				
SSN688	0				2		1	1	
STINGER	0				23				
SUPERTANKERS	3	2	1		0				
SUPERRESOLUTION	0				2		2		

APPENDIX G: FULL TEXT SEARCH SYSTEM TEST

A test of the full text search system was performed on a selected portion of the technical reports data base by retrieval analysts. Our conclusion based on this limited experience was that the full text search system may possibly be implemented as a supplement to our present system, but it cannot adequately replace current methods of indexing and retrieval by itself.

No technical problems in implementation were found with the full text system during the experiment period. The statistics pages for the test did not list the number of hits for each full text term searched. This should be programmed when the full data base is implemented. In our experience, awareness of stop words was important, as they tended to occur in sample searches that were made and care must be taken to avoid them.

Indexing will continue to be important with implementation of full text searching as a supplement to the current system. Document abstracts will need to be examined when submitted for completeness and accuracy. At the indexing point, the entire document and not jhust the abstract needs to be examined. Trained subject area specialists in indexing should consider variations in phrasing and inferred concepts that are not explicit and amend the abstracts if necessary, in addition to appropriate indexing, including weighting of terms. This indexing process also should include open-ended or identifier terms. In comparison with prior indexed terms, reliance on the full text search to retrieve open-ended and identifier terms would produce an output inferior in both quality and quantity. DTIC indexing and input personnel should have an opportunity to present their comments on the full text experiment before the system is implemented.

As for retrieval, full text searching should be a supplement to the present retrieval system. This capability must be integrated into retrieval as presently practiced. Such an integration must take into consideration that currently search strategies frequently contain several one word phrases. Some multiple word posting terms also have several single word posting terms as part of their hierarchy. Thus, to avoid problems with one word phrases, a method is needed to distinguish a word to be full text searched from a single word posting term. The most straight forward approach would be an inverted file search by means of a role code when a full text search of a word is desired. One word phrases occuring at any level of the search strategy would not be full text searched unless preceded by a role code of two numeric (or alphabetic) digits. Mixed role code full text and subject term searches could then be made as desired and further refined if necessary, by means of the already implemented on line title and abstract searching capability.

A full text search system would not enable personnel untrained in scientific and technical subject areas to retrieve acceptable products. But implemented as a supplemental tool, overall retrieval should improve for trained analysts.

RECOMMENDATIONS:

The proposed full text search system may be implemented as a supplement to the present system, taking into account the following points:

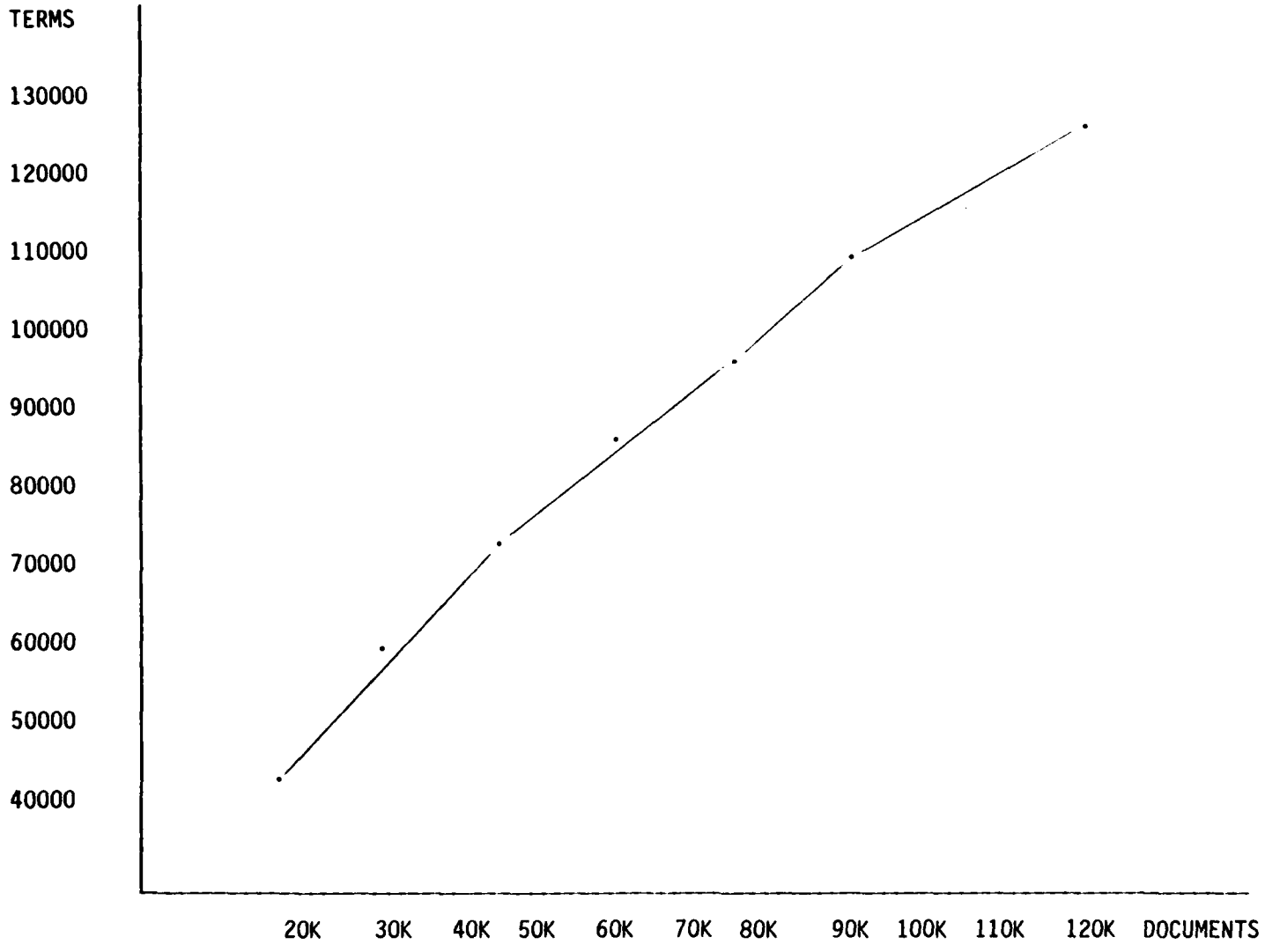
Indexing, including indexing of identifiers and open-ended terms cannot satisfactorily be replaced by full text searching.

Input and indexing personnel should have an opportunity to comment on the proposed implementation.

The full text system must be compatible with the present retrieval system. Retrieval of full text single word phrases could be by means of a role code or similar method to distinguish the words from other retrieval terms.

Implementation could be in stages. The full text system could be first put in operation on in-house terminals for further testing before implementation system wide.

APPENDIX H: TERM GROWTH IN THE TECHNICAL REPORTS DATA BASE




DATE
ILME