DOCUMENT RESUME

ED 225 574 IR 050 055

AUTHOR Katzer, Jeffrey; And Others
TITLE A Study of the Impact of Representations in Information Retrieval Systems.
INSTITUTION Syracuse Univ., N.Y. School of Information Studies.
SPONS AGENCY National Science Foundation, Washington, D.C. Div. of Information Science and Technology.
PUB DATE Jul 82
CONTRACT IST-79-21468
NOTE 117p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Databases; *Information Retrieval; Models; *Online Systems; Records (Forms); *Reference Services; Relevance (Information Retrieval); Search Strategies; Statistical Analysis; *Subject Index Terms; Tables (Data); User Satisfaction (Information)
IDENTIFIERS *Free Text Searching

ABSTRACT

This report investigates seven document representations--configurations of controlled and free-text vocabulary--which can be used to search the INSPEC (Computer and Control Abstracts) and PsychInfo (Psychological Abstracts) databases. The performance of each representation is analyzed, as is overlap among the representations, i.e., the extent to which the same documents are retrieved when searching with different vocabulary configurations. The study's use of a DIALOG simulator known as DIATOM, the participation of 7 trained searching intermediaries, and the soliciting of search questions from 114 online users are described. Major differences between the two databases in terms of which representations perform most effectively, and consistently low overlaps among representations are reported. Results are also discussed in terms of the cumulative improvement on retrieval performance as representations are added sequentially. A probabilistic model of overlap is developed based on the assumption of random retrieval, and this model is fitted against the obtained asymmetric overlaps and the incremental improvements obtained by different overlaps. A total of 20 tables and 15 references are provided. Appendices comprise intermediary training materials, instructions to study participants regarding citation relevance judgements, directions to online users, and sample forms for searchers, as well as the study's Latin square and factorial design, analysis of variance summary results, and theoretical model proofs. (Author/ESR)

A STUDY OF THE IMPACT OF REPRESENTATIONS IN INFORMATION RETRIEVAL SYSTEMS

Final Report

July 1982

This material is based on research supported in part by the National Science Foundation, Division of Information Science and Technology, under Grant IST 79-21468. The opinions, findings and conclusions or recommendations expressed in this report are those of the authors and do not necessarily refelect the views of the National Science Foundation.

School of Information Studies
Syracuse University
Syracuse, New York 13210

This report was written by

Jeffrey Katzer, with the assistance of Judith A. Tessier, William Frakes and Padmini DasGupta

PROJECT STAFF

Principal Investigator	Jeffrey Katzer
Research Associate	Judith A. Tessier
Graduate Associate	William B. Frakes
Graduate Assistants	Padmini DasGupta Cheryl McAfee
Project Secretary	Margaret Montgomery
Consultants on Phase I	Terry Noreault Matthew Koll Robert Waldstein

ABSTRACT

Data bases of text materials such as English Language abstracts of documents are difficult to represent in an information system. Results of numerous investigations indicate that in many situations different document representations are, on the average, approximately equally effective. However, recent research findings indicate that different representations retrieve different subsets of documents (and relevant documents) from data bases.

This study investigated document representations in two different data bases and analyzed the overlap among the representations (extent to which the same documents were retrieved) as well as their performance. Using a technical data base, seven document representations were investigated. The study was repeated with a less technical data base using four representations.

Results indicate major differences between the two data bases in terms of which representations performed most effectively within each data base. The overlaps among the representations were consistently low. Differences were also found between search intermediaries and between the representations. Results were also discussed in terms of the incremental effectiveness of representations -- i.e. what is the cumulative improvement on retrieval performance as representations are added sequentially?

A probabilistic model of overlap was developed based on the assumption of random retrieval. The model was fitted against the obtained asymmetric overlaps and against the incremental improvements obtained by the different representations. In general, the model fit these data reasonably well.

ACKNOWLEDGEMENTS

This study required the help and support of many individuals and organizations in a variety of ways. I would like to take this opportunity to publicly acknowledge their assistance.

The Project Staff carried out the work with good cheer and quiet efficiency. Though they had their own responsibilities, they worked as a group and should be commended as a group: Padmini DasGupta, William Frakes, Cheryl McAfee, Margaret Montgomery and Judith Tessier. In addition, several individuals served as consultants to the Project: Matthew Koll, Terry Noreault and Robert Waldstein. Many others, not officially on the Project were also helpful -- especially Robert N. Oddy and Linda Smith. To all of these people: Thank You!

I also want to thank a few organizations for their assistance. Both INSPEC and PsychInfo were very helpful by making portions of their data bases available to the Project. Information Services and Research was responsible for obtaining professional intermediaries to carry out the searches in both Phases of the Project. Lastly, the School of Information Studies must be credited for providing an environment where research of this type is encouraged and supported.

Jeffrey Katzer
Principal Investigator

TABLE OF CONTENTS

TABLE OF CONTENTS, continued

TABLE OF TABLES

I. INTRODUCTION

This report presents the results of the Document Representation Overlap Study. The report contains the research background and objectives, the procedures used, the findings obtained, and a discussion of these findings. The study was designed to contribute to our knowledge of the effect of the representation of information items on information system performance.

Past studies have found that when using recall and precision as performance measures, the differences among various representations (such as free-text term, or descriptor phrase) have not been consistently evident. Studies to date have examined the precision and recall performance of two or more representations. The results of those studies are equivocal. For example, Cleverdon (1967), Keen, (1973), Salton (1968, pp. 316-349), and McGill (1979) report no sizeable differences among the representations they examined. On the other hand the results from the second Cranfield Project and from studies by Salton (1973), Sparck-Jones and Jackson (1970), Hersey, et al. (1971), and Sparck-Jones (1974) reported differences in average performance levels.

This study takes as its departure evidence that performance measures have masked real and systematic differences among the representations. Specifically, different representations result in the retrieval of different items.

One of the more recent studies supporting this assertion was conducted by Williams (1977). She computed the overlap among five different document representations in a random sample of 50 documents taken from Chemical Abstracts. No queries were obtained from users, rather representations were compared for matching terms. The results gave the degree of uniqueness or lack of overlap among representations. Title, for example is claimed to be an important representation for retrieval because an average of two title terms per document did not appear in other representations. Smith (1979) provided some indication of the overlap among seven document representations in a portion of the INSPEC data base. No users were employed; a random sample of 35 documents were selected and treated as queries. None of the average conditional probabilities (measures of asymmetrical overlap) exceeded .5, meaning that different document representations tended to retrieve different documents. A third study (McGill, 1979) compared documents retrieved using free-text and controlled terms in a portion of the ERIC data base. Users provided queries which were searched and relevance judgements obtained. Thirty-three of the queries were selected for a study of overlap. When each of the intermediaries searched both

document representations, the average overlap was only 14%. Other queries were searched by intermediaries using different representations. In this situation, the average overlap dropped to 5%. Both of these figures are surprisingly low indicating that users retrieve quite different sets of documents when the free and controlled representations are used.

These studies, as well as other investigations of the effectiveness of combined representations, have somewhat limited conclusions for three reasons: (1) usually only very few (usually two) representations were included, (2) often a single, very small data base was used, and (3) overlap was typically examined by itself, without any consideration of the effectiveness of the representations. The study reported here builds on the previous work, but examines both performance and overlap of up to seven representations in two different, moderately sized (12,000 document) data bases.

II. OBJECTIVES

The assessment of the various representations is concerned with a number of specific objectives:

(1) To determine if the information items retrieved by the differing representations are significantly and substantially different.

(2) To assess the effectiveness of representations or combinations of representations.

(3) To develop and test a theoretic model sufficient to explain any differences in information retrieval system operation based on changes in the representation of information items.

III. OVERVIEW

To achieve these objectives, it was necessary to submit search requests to alternative representations of a data base and to design the study so that measures of performance (of each representation) and overlap (among representations) could be obtained. The basic study was repeated a second time so that we could determine if the results were consistent when a different data base was employed.

The two phases of this investigation correspond to the two data bases employed. In general, both phases were similar: a data base was acquired and loaded into the DIATOM retrieval system. Real users provided written queries which were then given to trained intermediaries who were instructed to construct and submit high-recall searches to the system. The intermediaries were restricted to particular document representations for a given search, using a balanced design so that each intermediary used each document representation an equal number of times. The results of the searches entered for a given query were merged and given back to the user for relevance judgements.

Each phase of this study used a different data base. In addition, the two phases differed in two other important ways: (1) the analysis design differed, and as a result, (2) the number of document representations and intermediaries differed. In Phase I, seven representations were used. Each intermediary used each representation on one-seventh of the queries. Consequently, there was a possibility that intermediaries would be confounded with representations thereby hampering a clear interpretation of the results of overlap documents. This possibility was prevented in Phase II; each intermediary searched each query separately under all of the representations.

A summary of the characteristics of the two Phases of the study is presented in Table 1.

Table 1

Overview of Phase I and Phase II

	Phase I	Phase II
Duration	2/80 - 3/81	3/81 - 2/82
Data Base	INSPEC (Computer & Control Abstracts) 9/79 - 12/79	PsychInfo (Psychological Abstracts) 7/80 - 12/80
Number of Documents	~ 12,000	~ 12,000
Retrieval System	DIATOM	DIATOM
Number of Users	69	45
Number of Queries	84	52
Number of Intermediaries	7	4
Number of Representations	7	4
Type of Design	7x7 Latin Square replicated 12 times	4x4 factorial with repeated measures

IV. RETRIEVAL ENVIRONMENT

A. Data Bases

For Phase I, permission was granted by the Institution of Electrical Engineers to use the Computer and Control Abstracts portion (9/79 - 12/79) of the INSPEC data base. For Phase II, the PsychInfo Use Service granted permission to use a portion of the 1980 data base (July - December) whose printed counterpart is Psychological Abstracts. Each data base consisted of approximately 12,000 documents. The choice of these two data bases and the number of documents used insured that sufficient documents would be retrieved by each document representation.

Each document consisted of a series of bibliographic citation fields, the abstract, and some indexing information. The format of each document record as it was printed upon retrieval is given below.

INSPEC
- DNnumber (abstract numbers from INSPEC journals)
- Title
- Authors (separated by commas)
- Source Field: as follows
  - Publication: (volume and issue number) (part number) pagination data following this may be information in ( ). This is information on the cover-to-cover translation as follows: (publication; (volume and issue) pages, (date) (type of unconventional media) (availability) (Title of Conference) (location of conference) (sponsoring organization) (date) language).
- Abstract
- Indexing Information

PsychInfo
- DNnumber (abstract numbers from PsychAbs journals)
- Title
- Authors (separated by semi-colons)
- Source: as follows
  - Journal name
  - Publication date
  - Volume and issue number, pagination.
- Section Code: content classification assigned to sections of print PA
- Abstracts: Abstracts (75-175 words) used for articles directly relevant to psychology, annotations for less central items.
- Indexing Information: Descriptors
  - Identifiers

B. Retrieval System

DIATOM, an on-line retrieval system which was designed to simulate most of the features of Dialog, was used to conduct all the searches in this study. DIATOM was designed and programmed by Robert Waldstein (1981), a PhD student at the School of Information Studies.

The major differences between DIATOM and DIALOG are listed below.

1. DIATOM permitted the searchers to log on directly to a particular representation. All search statements were subsequently restricted to that representation only.

2. The system included a stemmer used for the stem representation in Phase I.

3. To restrict a search to a particular language, a Limit/ENG (for English) was used.

4. Adjacency (nW) could not be used with either truncation or stemming.

5. Adjacency at times ran very slow; the field operator (F) could be used instead.

C. Search Intermediaries

All of the intermediaries used in this study were professional librarians or information brokers with experience using computerized retrieval systems; all had some experience using DIALOG.

Before Phase I, the seven intermediaries took part in a day-long training session. Afterwards, each intermediary was required to become familiar with DIATOM and the INSPEC data base. Each intermediary submitted fourteen practice searches. A copy of the training materials provided the intermediaries is given in Appendix A.

Four of the search intermediaries employed in Phase I were used again in Phase II.* Each intermediary took part in a three hour training session and was required to submit two practice searches to the system.

D. Users and Queries

Users were solicited from Syracuse University and other institutions which were likely to have individuals with information needs related to the content of the two data bases. Our objective in accepting users was to come as close as possible to criteria used in operational search services so that queries and relevance judgements could be plausibly generalized.

Originally, the study design specified 98 users for Phase I and 60 for Phase II. Each user was to submit a single query. However, because of the difficulty in obtaining users, several users were permitted to submit more than one query. The number of users, their characteristics, and the number of queries for each Phase of the study are given in Tables 2 and 3.

E. Relevance Judgements

Relevance judgements were obtained from the users for all documents retrieved for the query.** A four point scale was used with "1" and "2" indicating relevant, "3" and "4" indicating non-relevant. The instructions which accompanied the search results are provided in Appendix B.

*One searcher left the project after completing 42 queries. The remaining queries were searched by a fifth intermediary who had the requisite experience and was trained for this study.

**After repeated attempts, four users in Phase I did not return their relevance judgements. In these few cases we identified other individuals in the specific topic area of the query who presumably could make relevance judgements. These surrogate users made the relevance judgements.

Table 2

Characteristics of Users in Phase I

Affiliation	No. of Users	Faculty	Students	Sci/Eng	Others	No. of Queries
Syracuse U.	35	26	8	0	1	41
General Electric	1	0	0	1	0	4
Univ. of Illinois	5	2	3	0	0	5
Univ. of Louisville	9	0	0	0	9	14
National Bureau of Standards	6	0	0	6	0	6
OCLC, Inc.	5	0	0	5	0	6
Environ. Protection Agency	6	0	0	6	0	6
OTISCA Industries	1	0	0	0	1	1
SUNY, College Environ. Sciences & Forestry	1	0	1	0	0	1
Total	69	28	12	18	11	84

Altogether, 69 individuals served as users in this study. 11 of these individuals submitted more than one query; 8 users submitted 2 queries, 2 users submitted 3 queries and 1 user submitted 4 queries.

Table 3

Characteristics of Users in Phase II

Affiliation	# of Users	Faculty	Students	Others	# of Queries
Syracuse University	39	11	28	0	44
Utica College	1	1	0	0	1
Madison Community Services	1	0	0	1	1
Social Service Dept OCC	1	1	0	0	3
BMW Cooperative Nursery	1	0	0	1	1
University of Illinois	1	0	0	1	1
SUNY Albany	1	0	0	1	1
Total	45	13	28	4	52

Altogether, 45 individuals served as users in this study. 6 of these individuals submitted more than 1 query, 5 users submitted 2 queries, and 1 user submitted 3 queries.

V. METHODOLOGY

A. Variables

The key experimental or independent variable was the representation used in searching the data base. Seven representations were used in Phase I, four were used in Phase II. The representations are described in Table 4.

The major dependent or criterion variables were performance measures (recall and precision), measures of overlap, and the total number of documents retrieved were also analyzed. These measures were operationalized as follows.

Recall: The recall ratios were formed by dividing the number of relevant documents retrieved by each representation by the total number of relevant documents retrieved by all of the representations.* Both "macro-" and "micro" recall ratios were used (Salton, 1968, p.299). Macro- (or "user") recall is computed by taking the average of the recalls calculated for each query. Micro- (or "system") recall totals the number of retrieved relevant documents across all queries and then divides that total by the sum across queries of all relevant documents.

Precision: The precision ratio was formed by dividing the number of relevant documents retrieved by each representation by the total number of documents retrieved by that representation. Both macro- and micro- versions of precision were computed.

Total-Retrieved: This measure is simply the number of documents retrieved by each representation; it is the denominator of the precision ratio. It was included because it is an indication of user effort required to read the output from the system.

*During Phase II another research investigation made use of a stemmed representation (similar to, but not identical with, the ST representation used in Phase I). Documents retrieved by this "fifth" representation were also judged for relevance by the user. The denominator of the recall ratios used in Phase II include relevant documents retrieved by the stemmed representation as well as the four major representations. No analysis of the stemmed representation for Phase II is included in this report. It should be noted, however, that the stemmed representation retrieved relevant documents not retrieved by the other four representations.

Table 4

Document Representation

Abbreviation	Description	Use
DD	Descriptor terms chosen by an indexer; a controlled vocabulary.	Phases I & II
AA	Free-text words from the abstract; trivial words excluded.	Phases I & II
TT	Free-text words from the title; trivial words excluded.	Phases I & II
II	Free-text phrases chosen by an indexer.	Phases I & II
DI	Indexer selected terms. A compound representation made up of DD and II.	Phase I
ST	A stemmed version (automatic suffix removal) of representation TA.	Phase I
TA	Free-text terms from the title and abstract. A compound representation made up of TT and AA.	Phase I

Asymmetric-Overlap: For two representations i and j, this measure is computed by dividing the number of documents retrieved by both representations by the number retrieved by one of the representations. If R_i and R_j are the sets of documents retrieved by representations i and j, then the asymmetrical-overlap measure can simple be given as

$$A_{ij} = \frac{n[R_i \cap R_j]}{n[R_i]}$$

where "n" is the counting operator. Seen this way, asymmetrical-overlap is the conditional probability of retrieval using representation j given that the data base is restricted to those retrieved by representation i.

Symmetric-Overlap: For two representations i and j, this measure is computed by dividing the number of documents retrieved in common by both representations by the total number of different documents retrieved by either. Or more formally, it is the number of retrieved documents in the intersection of the two representations divided by the number retrieved by the union of those representations.

$$S_{ij} = \frac{n[R_i \cap R_j]}{n[R_i \cup R_j]}$$

Union-Overlap: For two representations i and j, this measure is computed by dividing the number of documents retrieved by either of the representations by the number of documents retrieved by all r representations.

$$U_{ij} = \frac{n[R_i \cup R_j]}{n[R_i \cup R_j \cup \ldots \cup R_r]}$$

Thus, the union-overlap is more of a recall ratio for a combination of representations. It can be extended to combinations of more than two representations by expanding the numerator.

Different versions of these dependent variables were computed; they differed in terms of the stringency of the relevance criterion. In both Phases of this investigation, relevance was determined by the requestor. A four point continuum was used from 1 (definitely relevant) to 4 (definitely not relevant). Some analyses are based on a "strict" definition of relevance: only those judged "1" were included. Other analyses used a dichotomized relevance judgement and a broader definition of relevance was used: those documents judged with "1" or "2" were acceptable. Lastly, some analyses are based on all retrieved documents; relevance was not taken into account.

These alternative versions of the dependent variables are identified by an appended suffix. For example, Recall-1, Precision-1, Overlap-1, etc. are all based on the stricter definition of relevance; those measures with a suffix "2" are based on the broader definition.

B. Procedure

Queries obtained from users (see Appendix C for Directions to Users) were used as submitted; they were not screened for appropriateness to the data base or for on-line searching in Phase I; some screening was used in Phase II. Each intermediary was given a photocopy of the search request. In Phase I, each intermediary used a different representation to search each query, and across all the queries each intermediary used each representation an equal number of times. In Phase II, each intermediary searched each query four times using all four representations. In both phases, computer programs within the DIATOM system controlled the order that representations were used: according to the Latin Square Design in Phase I and randomly in Phase II (see Appendix E).

Search intermediaries used the DIATOM system to retrieve documents. Intermediaries were instructed to carry out "high-recall" searches. The directions given to each intermediary is provided in Appendix D.

After a query was completely searched (seven times in Phase I, sixteen times in Phase II), the retrieved document set was merged into a single listing and placed in reverse chronological order. This listing consisted of the citations and abstracts of the retrieved documents (if more than 200 documents were retrieved, a random sample of 200 was used). No clue was present which indicated either the intermediary or the representation used to retrieve the document.

Two copies of this listing were produced. Both copies were sent to the user with instructions (see Appendix B) to make relevance judgements on one copy which was to be returned to the project, the second copy was for the user.

C. Design and Analysis

The measures of macro-recall, macro-precision and total-retrieved were analyzed using standard analysis of variance (AOV) computations. The design and the analysis can control for extraneous variables and can identify separate effects for the representations, intermediaries, and other components of the study, including interaction effects if desired.

In Phase I, the overall design can be characterized as a 7x7 Latin Square replicated 12 times (hence 84 queries). The Latin Squares used in this study are given in Appendix E. The partitioning of the total variation can be determined from the various AOV Summary Tables given in Appendix F.

Approximately ten percent (66) of the precision results had to be excluded from the analysis because no documents were retrieved for a given query under a given representation. Fourteen queries had to be excluded from all Recall-1 analysis, and seven from the Recall-2 analysis, because in each situation no relevant documents were retrieved.

In Phase II, the overall design can be described as a factorial design containing sixteen cells (four searchers by four representations). Each of 57 queries was searched under all sixteen combinations. This design, in contrast with the Latin Square design used in Phase I, required that each intermediary use all representations when searching a query -- thereby enabling us to determine if representation effects interacted with intermediary effects.

VI. RESULTS

Our initial concern was to determine if the results from this study repeated the pattern noted earlier: relatively little difference in performance among the representations coupled with relatively little overlap. Table 5 presents these results. It is apparent that these results do repeat the pattern observed in other studies. Though some performance measures are significantly different, none of the differences exceed .18 -- which is clearly within the range of values reported in the literature. The overlaps range from a low of about 14% to a high of about 27%; these also correspond to the earlier results.

The remaining part of this section presents these findings in more detail. First the performance measures will be considered. Then the study of overlaps will be presented.

A. Analysis of Performance

The macro-performance measures of recall, precision, and total-retrieved are analyzed in terms of document representations. The design of the two studies also analyzes macro-performance in terms of search intermediary differences and (in Phase II) an interaction between searchers and representations. If interaction effects existed, any analysis or discussion of document representations would have to be tempered by their relationship with intermediary effects. Fortunately, that did not turn out to be necessary: the Phase II analyses (Appendix G) indicate an absence of searcher/representation interaction. Furthermore, the results show that searcher effects did not consistently appear: they were sizeable in Phase I and much smaller in Phase II (Appendix F and G.

Descriptive summary statistics for the macro-performance measures are presented in Tables 6 and 7. The macro-performance means were presented for statistically significant differences (see Appendix F and G for the AOV Summary Tables). A listing of the significant differences can be found in Table 7. It must be stated at the outset that there are some major differences in the results of the two Phases and consequently they need to be discussed separately.

Table 5

Overlaps Among "Best" and "Worst" Performing Representations*

		"Best" Performing Represent.	"Worst" Performing Represent.	Difference	Symmetric Overlap***
Phase I	Recall-1	.404	.229	.175**	.155
	Recall-2	.321	.200	.121**	.138
	Precision-1	.264	.173	.091	.172
	Precision-2	.422	.336	.086	.150
Phase II	Recall-1	.263	.179	.084**	.264
	Recall-2	.242	.153	.089**	.234
	Precision-1	.282	.219	.063	.273
	Precision-2	.539	.416	.123**	.256

*Macro-performance measures are taken from Table 6.

**Difference is statistically significant at .05 level.

***Symmetric overlap figures are taken from Tables 9 and 12 using the pairwise overlap between the "best" and "worst" performing representation.

Table 6

Macro-performance Means and Number of Queries

		DD	AA	TT	II	DI	ST	TA
Phase I	Recall-1	.229 (70)	.365 (70)	.273 (70)	.339 (70)	.330 (70)	.392 (70)	.404 (70)
	Recall-2	.200 (77)	.270 (77)	.205 (77)	.321 (77)	.284 (77)	.317 (77)	.290 (77)
	Precision-1	.173 (62)	.197 (77)	.264 (70)	.218 (79)	.221 (75)	.188 (81)	.224 (78)
	Precision-2	.336 (62)	.352 (77)	.422 (70)	.403 (79)	.361 (75)	.338 (81)	.352 (78)
	Total-Retr.	13.2 (84)	17.5 (84)	12.4 (84)	16.1 (84)	16.4 (84)	19.8 (84)	18.6 (84)
Phase II	Recall-1	.263 (176)	.256 (177)	.179 (177)	.205 (179)	--	--	--
	Recall-2	.242 (176)	.213 (177)	.153 (177)	.182 (179)	--	--	--
	Precision-1	.282 (176)	.219 (177)	.276 (177)	.255 (179)	--	--	--
	Precision-2	.532 (176)	.416 (177)	.539 (177)	.500 (179)	--	--	--
	Total-Retr.	18.6 (176)	17.9 (177)	10.3 (177)	12.6 (179)	--	--	--

Table 7

Significant Differences in Macro-performance Among Representations.

		Representation Poorer	Better	Average Difference*	Percent Improvement
Phase I	Recall-1	DD	TA	.175	76%
		DD	ST	.173	71%
		DD	AA	.136	59%
	Recall-2	DD	II	.121	60%
		DD	ST	.117	58%
		TT	II	.116	56%
		TT	ST	.112	55%
	Precision-1	--	--	--	--
	Precision-2	--	--	--	--
Phase II	Recall-1	TT	DD	.084	47%
		TT	AA	.077	43%
	Recall-2	TT	DD	.089	58%
		TT	AA	.060	39%
		II	DD	.060	33%
	Precision-1	--	--	--	--
	Precision-2	AA	TT	.123	30%
		AA	DD	.116	28%

*Differences are significant at .05 level using Tukey's HSD procedure. See Appendix F and G for details.

For Phase I results, representations differed significantly in (macro- Recall-1, Recall-2, and Total-Retrieved) scores. As indicated in Table 7, descriptors (DD) and titles (TT) performed rather poorly as representations on the recall measures, while identifiers (II) and title-abstracts (either TA or ST) performed much better.

Even though no pairs of representations differed significantly in either precision measure, it is useful to include some consideration of precision into these findings. Considering all five measures, the descriptor (DD) representation performs uniformly poorly on the recall and precision measures while title-abstract (TA) performs reasonably well on them -- though not as strongly as DD's negative performance. Interestingly, the free-text words assigned by indexers (II) perform moderately well over all five measures. Stemming (ST) which would tend to increase the total number retrieved performs quite well on the recall measures, but poorly on the precision measures. The title representation (TT) shows the opposite pattern -- high on the precision measures (and Tot-Ret.) and low for recall. The other representations fluctuate quite a bit over the five measures.

For Phase II the patterns of results are for the most part different. One important exception is titles (TT) which perform poorly here in terms of recall as in Phase I. The major difference between the two phases has to do with the relative performance of descriptors (DD) and free-index phrases (II). In Phase I, the index phrases perform much better than the descriptors, which in Phase II their results are somewhat reversed. And, somewhat surprisingly, this pattern occurs in terms of precision as well as recall. The precise cause of this reversal cannot be ascertained experimentally from the data collected in this study. Two possibilities should be considered: (1) the differences that exist between the two data bases especially in terms of specificity of terms, and (2) the differences that exist between the directions and training given the indexers at INSPEC and at PsycInfo.

Data base differences, however, are not likely to be the major cause of Phase II producing generally lower values in macro-recall and higher values in macro-precision than the comparable results in Phase I. Instead, these general trends in macro-performance between the two Phases are probably related to differences in the design of the two studies. In both Phase I and Phase II, the numerator of the macro-recalls was based on the results of one intermediary searching the data base once. The two phases differed, however, in the denominators; in Phase I it was based on seven intermediaries searching the query once, while in Phase II the denominator was based on 16 searches (four intermediaries each using all four representations.) Therefore, there was more opportunity to identify relevant documents for the recall denominator in Phase II, leading to a lower average

macro-recall. The macro-precision figures could easily have been affected by searching time. In Phase II each query had to be searched by an intermediary four times. Intermediaries may have reduced the search time so that the total time allotted to each query was comparable to the time spent in Phase I searches. To the extent that relevant documents are more likely to be retrieved early in the search process, the obtained higher levels of macro-precision found in Phase II can be attributed somewhat to decreased search times.

For both of these reasons, the differences between the two Phases in terms of macro-performance should not be attributed to the differences in the two data bases. The fact that the micro-performance results discussed below do not present a similar pattern between the two Phases strengthens this position.

The average micro-performance levels are reported in Table 8.* Micro-performance addresses the issue of how well the representations can do when multiple searchers pool their results. It is a more conservative approach; as indicators of system-level performance micro-measures are very helpful because they decrease the effect of single (perhaps atypical) searches or queries. In general, the results noted in the macro-performance data are also evident here. For Phase I, the index phrases (II) perform quite well overall, while the descriptors (DD) do poorly; the reverse is true for Phase II. For Phase II the micro-recall figures are higher than those of Phase I. This finding is much more intuitively reasonable than the macro-recall data suggest -- given the nature of the topics contained in the two data bases. This, plus the possible artifacts due to design (noted above) makes the micro-recall figures for Phase II better indicators of the recall obtained in that study.

*Because statistical inferential tests were not calculated on any of the micro-performance measures, it is not known if the observed differences are larger than what could be expected to occur by chance.

Table 8

Micro-performance Means

		DD	AA	TT	II	DI	ST	TA
Phase I	Recall-1	.237	.328	.285	.348	.309	.304	.369
	Recall-2	.216	.283	.229	.306	.268	.281	.294
	Precision-1	.173	.181	.221	.208	.182	.148	.192
	Precision-2	.335	.332	.378	.389	.336	.291	.324
Phase II	Recall-1	.520	.475	.322	.351	--	--	--
	Recall-2	.526	.440	.313	.350	--	--	--
	Precision-1	.133	.120	.141	.122	--	--	--
	Precision-2	.340	.283	.347	.309	--	--	--

B. Analysis of Overlaps

The simplest analysis of overlaps is pairwise, comparing each representation with every other representation. Tables 9-11 report the overlaps for Phase I data; Tables 12-14 for Phase II. Each table contains three overlap analyses: (1) most relevant documents, (2) all relevant documents, and (3) all documents retrieved. In these tables, a high value indicates greater overlap and therefore less of an independent contribution of the "second" representation.

In both Phases, the pairwise overlaps decrease as the number of documents under consideration increase. That is, the average overlap is highest when only most relevant documents are considered; it is lowest when all retrieved documents are included. A second general finding is that the overlap figures are lowest when overlap is defined symmetrically, they are the highest for the union overlap. This, of course, is a function of the definition of the three measures of overlap. And, there is a difference between the results of the two Phases. The average overlaps in Phase I are consistently lower than the corresponding averages for Phase II. At least part of this difference between the Phases is due to the different designs used. In Phase II, the design should have had a systematic effect of raising the overlaps -- first by excluding a searcher-representation interaction, and second by using the same intermediaries (with their individual understanding of the queries) to search each query on all four representations.

The major finding in these data is that the overlaps are quite small as indicated by the averages. For example, the highest symmetric overlap among the relevant documents is only about one-third -- .313 between ST and AA in Phase I, and .363 between AA and II in Phase II.

The low overlap between index-phrases and either titles or abstract terms can in part be attributed to the fact that indexers may have selected the II phrases from the body of the document, not from the title or abstract. But, in general, there is not any single or simple procedural explanation for these findings. Overlaps were even low between representations that should have retrieved very similar documents. This can be seen most clearly in the Phase I results by comparing the simple and the compound representations such as abstract (AA) and title-abstract (TA) or descriptor (DD) and descriptor-identifier (DI). One possible explanation for the small overlaps is searcher differences; which is the only possible explanation for low overlaps between simple and compound representations. But, as an explanation for the low overlaps among all representations, searcher differences are not likely to be the major cause even though the analysis of variance tables (see Appendix F and G) show that searcher effects occasionally account for significant portions of the variance. It is the data in the ranking study (McGill, 1979) that cast doubt on the contention that searchers

are the sole or major cause of the low amount of overlap. In the ranking study, overlaps between different representations searched by the same searcher only equalled 14% for retrieved documents. That figure certainly falls in the range of values reported here. Furthermore, the Phase II design required that each intermediary search each query under all representations; the overlap results were, at best, moderate.

In the symmetric measures (Tables 9 and 12) there is considerable consistency across representations -- especially when the inflating effect of the three compound representations in Phase I are excluded. In both Phases the maximum difference in overlaps does not exceed 0.10. Also, the free-index phrases (II) in both Phases show a tendency to share more relevant documents with title and abstract fields than with the descriptor field -- although the size of this overlap is still quite small.

The asymmetric measures indicate the proportion of documents that would have been retrieved "anyway" -- that is, by the other representation. For example, Table 13 reports an asymmetric overlap of .378 between DD and II for the most relevant documents. This can be interpreted as follows: of all the documents retrieved by the descriptor representation, approximately 38 percent of them can also be retrieved by the free-index phrases. Tables 10 and 13 provide both row and column average figures (the other tables are symmetrical and a single set of averages suffices). A useful interpretation of the difference between row and column averages for a single representation can be given in terms of the sequence the representations are used in searching. The averages of the columns of numbers (presented along the bottom of the table) can be interpreted in terms of being used "first" in the search process. Given a single representation (indicated by the column heading), the average at the bottom indicates the proportion of documents retrieved by this representation that could also be retrieved by other representations. The averages presented in the right column are understandable in terms of being used "last" in the search process. Given retrieved documents from other representations, the row average for a given representation indicates its effect if searching were resumed using it alone -- the lower the average, the more the new representation will contribute.

Given this distinction between using (or implementing) a representation "first" or "last", the asymmetric overlaps (in Tables 10 and 13) present a rather consistent picture -- especially for the most relevant documents. In Phase I, either descriptors or free-index phrases are slightly the best choice for "first" use; in Phase II it is clearly the descriptors. For "last" use, the data indicate titles in Phase I and descriptors again in Phase II. The distinction between first and last use of a representation will be important in the next section of this report.

Union overlaps presented in Tables 11 and 14 give an estimate of the combined effect of two representations; they are conceptually equivalent to the recall ratio for the two representations. Because the numerator of these pairwise union overlaps includes all distinct documents (in the appropriate version) retrieved by two representations, the union overlaps will have higher values than comparable figures for the symmetrical and asymmetrical overlaps. In principle, the diagonal elements in the union overlaps should be identical to micro-recall values presented in Table 8. And, that is true for Phase I data. However, as noted earlier in this report, Phase II micro-recalls were based on five representations -- (the fifth one was produced for another research investigation) while the overlaps in Table 14 are based on retrievals from four representations -- hence the discrepancy.

The union overlap results from Phase I shows that most pairs of representations achieve at least 50 percent recall levels, but not much higher. In contrast, the Phase II figures are higher. All pairs of representations (off-diagonals) provide over 50 percent recall and the combination of descriptors and abstracts gives over 80 percent of the most relevant documents and over 75 percent of all documents retrieved.

Union overlaps are one way to explore "marginal utility" or the "value added" of additional representations. Tables 11 and 14 provide only pairwise overlaps. The extension to more than two representations is necessary in order to get overall conclusions. The next section of this report takes this approach.

Table 9

Symmetric Pairwise Overlaps - Phase I

	AA	TT	TA	ST	II	DI	DD	AVG *
Version - Most Relevant								
AA	1.000	0.181	0.270	0.313	0.212	0.217	0.125	.220
TT	0.181	1.000	0.227	0.178	0.236	0.209	0.172	.200
TA	0.270	0.227	1.000	0.307	0.208	0.236	0.155	.234
ST	0.313	0.178	0.307	1.000	0.179	0.201	0.115	.215
II	0.212	0.236	0.208	0.179	1.000	0.314	0.173	.220
DI	0.217	0.209	0.236	0.201	0.314	1.000	0.270	.241
DD	0.125	0.172	0.155	0.115	0.173	0.270	1.000	.168
Version - All Relevant								
AA	1.000	0.141	0.215	0.235	0.167	0.186	0.112	.176
TT	0.141	1.000	0.154	0.133	0.173	0.172	0.150	.154
TA	0.215	0.154	1.000	0.245	0.167	0.173	0.114	.178
ST	0.235	0.133	0.245	1.000	0.138	0.137	0.081	.161
II	0.167	0.173	0.167	0.138	1.000	0.242	0.138	.171
DI	0.186	0.172	0.173	0.137	0.242	1.000	0.258	.195
DD	0.112	0.150	0.114	0.081	0.138	0.258	1.000	.142
Version - All Documents								
AA	1.000	0.064	0.148	0.138	0.112	0.103	0.046	.102
TT	0.064	1.000	0.072	0.057	0.086	0.080	0.068	.071
TA	0.148	0.072	1.000	0.156	0.096	0.092	0.052	.103
ST	0.138	0.057	0.156	1.000	0.077	0.063	0.033	.087
II	0.112	0.086	0.096	0.077	1.000	0.131	0.063	.094
DI	0.103	0.080	0.092	0.063	0.131	1.000	0.120	.098
DD	0.046	0.068	0.052	0.033	0.063	0.120	1.000	.064

* Averages were computed with the diagonal element omitted.

Table 10

Asymmetric Pairwise Overlaps** - Phase I

	AA	TT	TA	ST	II	DI	DD	AVG.*
Version - Most Relevant								
AA	1.000	0.329	0.401	0.496	0.340	0.368	0.266	0.367
TT	0.286	1.000	0.328	0.293	0.348	0.332	0.323	0.318
TA	0.451	0.424	1.000	0.520	0.355	0.420	0.344	0.419
ST	0.459	0.312	0.428	1.000	0.284	0.332	0.234	0.341
II	0.361	0.424	0.334	0.325	1.000	0.508	0.365	0.386
DI	0.346	0.359	0.351	0.337	0.450	1.000	0.490	0.389
DD	0.192	0.268	0.221	0.183	0.248	0.376	1.000	0.248
AVG	0.349	0.353	0.344	0.359	0.338	0.389	0.337	
Version - All relevant								
AA	1.000	0.276	0.348	0.381	0.275	0.323	0.233	0.306
TT	0.223	1.000	0.237	0.212	0.258	0.274	0.268	0.245
TA	0.361	0.304	1.000	0.402	0.281	0.310	0.241	0.316
ST	0.379	0.261	0.385	1.000	0.233	0.247	0.172	0.279
II	0.297	0.344	0.292	0.254	1.000	0.418	0.292	0.316
DI	0.305	0.319	0.283	0.235	0.366	1.000	0.458	0.328
DD	0.178	0.253	0.178	0.132	0.207	0.370	1.000	0.220
AVG	0.291	0.293	0.287	0.269	0.270	0.324	0.277	
Version - All Documents								
AA	1.000	0.145	0.250	0.229	0.210	0.193	0.103	0.188
TT	0.103	1.000	0.113	0.088	0.140	0.131	0.123	0.116
TA	0.265	0.169	1.000	0.262	0.188	0.180	0.119	0.197
ST	0.259	0.141	0.279	1.000	0.159	0.131	0.080	0.175
II	0.193	0.182	0.163	0.129	1.000	0.230	0.131	0.171
DI	0.180	0.172	0.158	0.108	0.233	1.000	0.240	0.182
DD	0.078	0.131	0.085	0.053	0.108	0.194	1.000	0.108
AVG	0.180	0.157	0.175	0.145	0.173	0.177	0.133	

* Averages were computed with the diagonal element omitted.

** The representations in the columns form the denominator of the overlap measure.

Table 11

Union Pairwise Overlaps - Phase I

	AA	TT	TA	ST	II	DI	DD	AVG. *
Version - Most Relevant								
AA	0.328	0.520	0.549	0.481	0.558	0.523	0.502	0.522
TT	0.520	0.285	0.533	0.500	0.512	0.491	0.446	0.500
TA	0.549	0.533	0.369	0.515	0.594	0.548	0.525	0.544
ST	0.481	0.500	0.515	0.304	0.553	0.510	0.485	0.507
II	0.558	0.512	0.594	0.553	0.348	0.500	0.499	0.536
DI	0.523	0.491	0.548	0.510	0.500	0.309	0.430	0.500
DD	0.502	0.446	0.525	0.485	0.499	0.430	0.237	0.481
Version - All Relevant								
AA	0.283	0.449	0.475	0.457	0.505	0.465	0.449	0.467
TT	0.449	0.229	0.453	0.451	0.456	0.424	0.388	0.437
TA	0.475	0.453	0.294	0.462	0.514	0.479	0.458	0.474
ST	0.457	0.451	0.462	0.281	0.516	0.483	0.461	0.472
II	0.505	0.456	0.514	0.516	0.306	0.462	0.459	0.485
DI	0.465	0.424	0.479	0.483	0.462	0.268	0.385	0.450
DD	0.449	0.388	0.458	0.461	0.459	0.385	0.216	0.433
Version - All Documents								
AA	0.220	0.353	0.395	0.412	0.380	0.386	0.369	0.382
TT	0.353	0.156	0.363	0.384	0.331	0.335	0.302	0.345
TA	0.395	0.363	0.234	0.418	0.398	0.402	0.380	0.393
ST	0.412	0.384	0.418	0.249	0.420	0.428	0.402	0.411
II	0.380	0.331	0.398	0.420	0.203	0.361	0.347	0.373
DI	0.386	0.335	0.402	0.428	0.361	0.206	0.332	0.374
DD	0.369	0.302	0.380	0.402	0.347	0.332	0.166	0.355

* Averages were computed with the diagonal element omitted.

Table 12

Symmetric Pairwise Overlaps -- Phase II

	II	DD	AA	TT	AVG *
Version - Most Relevant					
II	1.000	0.289	0.363	0.351	0.334
DD	0.289	1.000	0.273	0.264	0.275
AA	0.363	0.273	1.000	0.277	0.304
TT	0.351	0.264	0.277	1.000	0.297
Version - All Relevant					
II	1.000	0.269	0.319	0.328	0.305
DD	0.269	1.000	0.233	0.234	0.245
AA	0.319	0.233	1.000	0.256	0.269
TT	0.328	0.234	0.256	1.000	0.273
Version - All Documents					
II	1.000	0.199	0.182	0.215	0.199
DD	0.199	1.000	0.150	0.159	0.169
AA	0.182	0.150	1.000	0.127	0.153
TT	0.215	0.159	0.127	1.000	0.167

*Averages were computed with the diagonal element omitted.

Table 13

Asymmetric Pairwise Overlaps**-- Phase II

	II	DD	AA	TT	AVG *
Version - Most Relevant					
II	1.000	0.378	0.469	0.551	0.466
DD	0.552	1.000	0.452	0.551	0.518
AA	0.616	0.407	1.000	0.536	0.520
TT	0.491	0.336	0.364	1.000	0.397
AVG*	0.553	0.374	0.428	0.546	
Version - All Relevant					
II	1.000	0.357	0.437	0.523	0.439
DD	0.524	1.000	0.413	0.500	0.479
AA	0.54	0.348	1.000	0.485	0.458
TT	0.468	0.305	0.351	1.000	0.375
AVG*	0.511	0.337	0.401	0.503	
Version - All Documents					
II	1.000	0.289	0.264	0.394	0.316
DD	0.39	1.000	0.256	0.364	0.337
AA	0.371	0.267	1.000	0.307	0.315
TT	0.321	0.220	0.178	1.000	0.240
AVG*	0.361	0.259	0.233	0.355	

* Averages were computed with the diagonal element omitted.

** The representations in the columns form the denominator of the overlap measure.

Table 14

Union Pairwise Overlaps -- Phase II

	II	DD	AA	TT	AVG *
Version - Most Relevant					
II	0.377	0.719	0.640	0.528	0.629
DD	0.719	0.550	0.821	0.701	0.747
AA	0.64	0.821	0.495	0.651	0.704
TT	0.528	0.701	0.651	0.336	0.627
Version - All Relevant					
II	0.368	0.715	0.624	0.525	0.621
DD	0.715	0.539	0.806	0.704	0.742
AA	0.624	0.806	0.454	0.624	0.685
TT	0.525	0.704	0.624	0.329	0.618
Version - All Documents					
II	0.314	0.616	0.640	0.469	0.575
DD	0.616	0.424	0.753	0.587	0.652
AA	0.640	0.753	0.442	0.619	0.671
TT	0.469	0.587	0.619	0.256	0.558

* Averages were computed with the diagonal element omitted.

VII. DISCUSSION

What are the factors which explain these findings? Are the results simply due to chance variations or are there some systematic components that can be identified? This section of the report responds to these questions. First, differences in data bases and indexer instructions will be reviewed. Then different overlap models of the data will be presented and explored from several viewpoints.

A. Data Bases and Indexing

As noted earlier, there are two related factors that might have contributed to the differences in performance of descriptors (DD) and free-index phrases (II) in the two data bases. They are the differences in the indexing procedures used and the avowed purpose of the representations in the data bases. Indexing procedures are not so much a function of the written indexing rules (though such rules exist, for example INSPEC, 1970) but are more a matter of what the indexers actually do.

At INSPEC, indexers read the title and abstract while at PsychAbs, the indexers focus on the abstract only. Both groups of indexers then identify the main concepts of the document. At INSPEC the concepts are taken in the form of the actual phrases used in the document. To this list of phrases the INSPEC indexers add any concepts implicit in the document not already representated by the selected phrases. The phrases plus the implicit concepts form the II representation. The descriptor terms (DD) at INSPEC are then generated from a thesaurus; the goal being to select terms that represent the concepts noted in the title and abstract.

At PsychInfo the indexers reverse this process. First they use the thesaurus to select descriptor terms that best represent the concepts found in the document abstract. The free-index phrases are then generated from the abstract to provide supplementary information. For documents reporting experimental research the supplementary information (in the form of II phrases) further describes the details of the study -- information about the variables used and the subject population. For nonexperimental or theoretical articles, the free-index phrases are more general descriptions of the documents.

Thus, to some extent there is a relationship between the II phrases used in INSPEC and the descriptors used in PsychAbs. Both are generated from the document and more importantly, both attempt to capture the main concepts of the document. In comparison, descriptors assigned by INSPEC indexers may not

exhaustively capture all of the concepts in the document because the procedure used misses implicit concepts and also because the descriptors used at INSPEC were developed for a manual system and as a result are not as exhaustive as they could be. The identifier phrases in PsychAbs are not meant to exhaustively represent all of the concepts in the document. For these reasons, we could expect the descriptors in PsychAbs and the II representation in INSPEC to perform quite well in comparison with the other representations used in these data bases in their ability to retrieve relevant documents.

Precision is a function of specificity. The II phrases used by INSPEC are for the most part composed of the author's own words and are therefore as specific as free-index terms. And, as noted earlier, the II phrases in Psychabs may be much more general. In PsychAbs, however, it is the descriptor field that is designed to be specific as well as exhaustive (APA, 1976).

From this analysis it seems possible that the (relative) superior performance of II in INSPEC and DD in PsychAbs in terms of both recall and precision may be a function of their similarity of purpose and the method by which they are produced: both are generated from the concepts found in the document and both aim at exhaustivity while maximizing the specificity of the terms selected.

B. Descriptive Models of Overlap

Overlaps between pairs of representations were discussed earlier. The question of concern here focuses on the relationship among all of the representations: what is the optimum combination of representations, or more precisely, the optimum ordering of representations. That is, if a retrieval environment were limited to a single representation, which one would it be? If a second could be added, which of the remaining representations contributes the most over and above the effect of the first representation? A third representation could be added over and above the first two, and so on.

The most sensible measure to use in answering this question is based on the union overlap.* Tables 15 and 16 present the results of this analysis. Table 15 uses all seven representations for the Phase I data and analyzes both the highly relevant as well as the total relevant measures across queries.

*Union overlaps are recall estimates and the discussion in this section is based on these recalls only -- precision is not considered.

Table 15

Representations Ordered by Incremental Improvement

Phase I

	Order	1st	2nd	3rd	4th	5th	6th	7th
Most Relevant	Representation	TA	II	AA	DD	TT	ST	DI
	Cum. No. Docs.	299	444	574	656	722	768	810
	Cum. Percentage	.369	.548	.709	.810	.891	.948	1.000
All Relevant	Representation	II	ST	DI	TA	TT	AA	DD
	Cum. No. Docs.	527	889	1118	1318	1466	1602	1723
	Cum. Percentage	.306	.516	.649	.765	.850	.930	1.00

Table 16

Representations Ordered by Incremental Improvement

Phases I* and II*

		Order	1st	2nd	3rd	4th
Most Relevant	Phase I	Representation	II	AA	TT	DD
		Cum. No. Docs.	282	452	554	634
		Cum. Percentage	.445	.713	.874	1.000
	Phase II	Representation	DD	AA	TT	II
		Cum. No. Docs.	339	506	573	616
		Cum. Percentage	.550	.821	.930	1.000
All Relevant	Phase I	Representation	II	AA	DD	TT
		Cum. No. Docs.	527	870	1093	1275
		Cum. Percentage	.413	.682	.857	1.000
	Phase II	Representation	DD	AA	TT	II
		Cum. No. Docs.	871	1302	1489	1615
		Cum. Percentage	.539	.806	.922	1.000

*Compound Representations Omitted

Since three representations (TA, DI, ST) are composed of other representations, the analysis was repeated in Table 16 omitting these "compound" representations. Table 16 also includes the comparable results from Phase II.

Tables 15 and 16 present different models -- different orderings of representations. Such models, if consistent, would allow a searcher to know which combinations of fields would be most likely to retrieve relevant documents. Such models would also point to obvious economies in the design and operation of retrieval systems. Unfortunately, these data suggest that the models are not totally consistent. There are differences within data bases which depend upon the definition of relevance used (most relevant versus all relevant), there is also the presence of the compound representations in the Phase I study which hampers our ability to see a pattern in the other fields, and most dramatically, there are differences in the orderings between Phase I and Phase II -- differences which could be a function of the data bases themselves (e.g. specificity of terms), or a function of how they were constructed (e.g. instructions given to indexers) or an interaction between these two.

There are also some interesting similarities evident in Table 16. Though the models (orderings) differ between Phases, they are very similar within Phases. For Phase II the order doesn't change as a function of relevance stringency, and the change for Phase I is both small and less important (involving the third and four representations). There are also similarities in the growth rates within each Phase -- as evident in the cumulative percentages.

What appears to be highly consistent is the cumulative increase in the percentage of relevant documents accounted for as each additional representation is included. This similarity may simply be due to the fact that the models are based on highly interrelated data -- within each phase data are subsets of one another. When the cumulative percentages are plotted against the order, the resulting curves appear to be hyperbolic in form. The next section of this report presents one theoretical interpretation for this finding.

The overlap among document representations can also be viewed from the perspective of a representation's "unique" contribution. For a given representation, what documents does it contribute to the relevant retrieved that were not retrieved under any other representation? The question is equivalent to the observed improvements in the models when the representation is the last entered into the model. Tables 17 and 18 report the effect of each representation, assuming the representation entered the model first or last. These are the maximum and minmum incremental improvements for each representation.

Table 17

Maximum and Minimum Contribution of Seven Representations

Phase I

	Repr.	Maximum Contribution* No. Docs.	Percent**	Minimum Contribution* No. Docs.	Percent**
Most Relevant	AA	266	.328	49	.060
	DD	192	.237	44	.054
	DI	250	.309	42	.052
	II	282	.348	74	.091
	ST	246	.304	44	.054
	TA	299	.369	53	.065
	TT	231	.285	52	.064
					.440
All Relevant	AA	488	.283	137	.080
	DD	373	.216	127	.074
	DI	462	.268	120	.070
	II	527	.306	196	.114
	ST	485	.281	149	.086
	TA	506	.294	134	.078
	TT	395	.229	133	.077
					.579

*Maximum contribution is the effect of that representation alone -- either it is the sole representation in the data base or it is used (entered) first, before the others are used. Maximum contribution is therefore equivalent to macro-recall (see Table 8). Minimum contribution is the "unique" effect of that representation after all documents retrieved by the other six representations have been removed; thus it can be considered to have entered the search process last.

**Percentages are based on all documents retrieved in each category: 810 for the most relevant and 1723 for all relevant.

Table 18

Maximum and Minimum Contributions of Four Representations

Phase I and Phase II

		Repr.	Maximum Contribution* No.Docs.	Percent**	Minimum Contribution* No.Docs.	Percent**
Most Relevant	I	AA	266	.328	125	.154
		DD	192	.237	85	.105
		II	282	.348	114	.141
		TT	231	.285	88	.109
						.509
	II	AA	310	.475	112	.172
		DD	339	.520	158	.242
		II	229	.351	42	.064
		TT	210	.322	50	.077
						.555
All Relevant	I	AA	488	.283	269	.156
		DD	373	.216	197	.114
		II	527	.306	271	.157
		TT	395	.229	182	.106
						.533
	II	AA	728	.440	286	.173
		DD	870	.526	429	.259
		II	579	.350	120	.072
		TT	518	.313	131	.079
						.583

*Maximum contribution is the effect of that representation alone--either it is the sole representation in the data base or it was used (entered) first, before the others are used. Maximum contribution is therefore equivalent to micro-recall. (see Table 8).

Minimum contribution is the "unique" effect of that representation after all documents retrieved by the other three representations have been removed; thus, it can be considered to have entered the search process last.

**Percentages are based on all documents retrieved by all representations in each category. For Phase I that number is 810 for most relevant and 1723 for all relevant. For Phase II the numbers are 652 for most relevant and 1653 for all relevant.

The "unique" effect of each representation is reported as the minimum contribution.

The lack of overlap among representations is again evident in the unique percentages. Given a data base with four representations, the fourth representation can contribute a sizeable number of additional relevant documents -- approximately 25 percent for the DD representation in Phase II, and approximately 15 percent for the II representation in Phase I. Even when the number of document representations is increased to seven (see Table 17), there is an approximate 10 percent contribution of relevant documents by the seventh representation (II in the INSPEC data base).

One final indicator of the lack of overlap among document representations is the sum of the unique contributions (Tables 17 and 18). Considering Phase I and Phase II, these totals range from 44 percent to about 58 percent. Thus, the amount of overlapping documents range from 42 percent to a high of 56 percent.

The incremental contributions reported in these Tables can also be used to provide some measure of the effect of human intervention in preparing documents for inclusion in a retrieval system. Taylor (in press) writes of the "value-added" process in document preparation. Document indexing is believed to add value to the document because it makes the document more readily accessible. Among the four basic representations used in the two studies reported here, II and DD require intellectual intervention. Between these two representations, DD can be thought of as making more use of intellectual contribution because it is based on the human produced thesaurus. As viewed from this perspective, the strong showing of both DD and II in terms of maximum and minimum contributions provides support for intellectual-based representations. Though the actual figures given in Tables 17 and 18 are useful in this regard, they are essentially recalls and a better quantification of value-added would combine these with measures of precision (e.g. van Rijsbergen, 1979, p. 167).

C. Theoretical Model of Overlaps

Can the obtained overlap results presented earlier in this report be understood or interpreted in terms of some theoretical model? Of the several possible approaches which could be developed one of the most basic is a probabilistic model based on the assumption that relevant retrievals are independent in the different representations - a plausible assumption given the low levels of recall obtained. It is assumed that each representation retrieves an independent random sample of the

relevant documents. Given this conservative assumption, what overlaps would be predicted for the different observations and how well do these predictions agree with the obtained results?

Such a derivation of a model is presented the first part of Appendix H. That model is then used to predict asymmetrical overlaps. Given the independence assumption, asymmetrical overlaps being conditional probabilities simplify to the micro-recall value of the second representation (see Appendix H, part 2 for a more formal proof).

The predicted values are presented in Table 19. The patterns in the two Phases are similar. The model fits the data remarkably well, given the single, simple assumption on which it was based. The greatest deviations from the model are identified by very large or very small values in the (obser/pre) data: (1) there are substantially lower than expected overlaps between AA and DD, and (2) substantially higher than expected overlaps between TT and II. In Phase II there is also a higher than predicted overlap between free-text abstract terms and identifier terms; this finding did not also occur in Phase I.

The obtained low overlap between AA and DD is not surprising, reflecting the contrast between controlled and "free" vocabulary. In fact, these two representations are at opposite ends of the continuum from least to most controlled: AA, TT, II, DD. The high overlaps between titles and index phrases may indicate that titles are well chosen by authors. That is, they contain many of the same key words as an indexer would select. The high overlaps between AA and II in Phase II could be a function of indexer practice at PsychAbs -- indexers may not go beyond the abstract to find identifier phrases. Or in the INSPEC data base (where the overlap is lower), perhaps the indexers find that they need to frequently go beyond the abstract to choose the key II phrases.

This same model can also be used to predict the incremental effects on recall through use of additional representations (as in Tables 15 and 16). Given four representations, the predicted recall using the model can be determined for a single representation, for two representations, etc., as shown below.

Table 19

Predicted* and Obtained Asymmetrical Overlaps

		II	DD	AA	TT	AVG
Phase I						
	Predicted		.348	.348	.348	.348
II	Observed		.365	.361	.424	.383
	Obser/pre		(1.05)	(1.04)	(1.22)	(1.10)
	Predicted	.237		.237	.237	.237
DD	Observed	.248		.192	.268	.236
	Obser/pre	(1.05)		(0.81)	(1.13)	(1.00)
	Predicted	.328	.328		.328	.328
AA	Observed	.340	.266		.329	.312
	Obser/pre	(1.04)	(0.81)		(1.00)	(0.95)
	Predicted	.285	.285	.285		.285
TT	Observed	.348	.323	.286		.319
	Obser/pre	(1.22)	(1.13)	(1.00)		(1.12)
	Predicted	.283	.320	.290	.304	.300
AVG	Observed	.312	.318	.280	.340	.312
	Obser/pre	(1.10)	(0.99)	(0.97)	(1.12)	(1.04)
Phase II						
	Predicted		.351	.351	.351	.351
II	Observed		.378	.469	.551	.466
	Obser/pre		(1.08)	(1.34)	(1.57)	(1.33)
	Predicted	.520		.520	.520	.520
DD	Observed	.552		.452	.551	.518
	Obser/pre	(1.06)		(0.87)	(1.06)	(1.00)
	Predicted	.475	.475		.475	.475
AA	Observed	.616	.407		.536	.520
	Obser/pre	(1.30)	(0.86)		(1.13)	(1.09)
	Predicted	.322	.322	.322		.322
TT	Observed	.491	.336	.364		.397
	Obser/pre	(1.52)	(1.04)	(1.13)		(1.23)
	Predicted	.439	.383	.398	.449	.417
AVG	Observed	.553	.374	.428	.546	.475
	Obser/pre	(1.26)	(0.98)	(1.08)	(1.22)	(1.14)

*Based on the model, predicted values are micro-recalls.

Representation(s)	Predicted Micro-Recall*
Any single representation	$1 - (1-r_1)$
Any two representations	$1 - (1-r_1)(1-r_2)$
Any three representations	$1 - (1-r_1)(1-r_2)(1-r_3)$
All four representations	$1 - (1-r_1)(1-r_2)(1-r_3)(1-r_4)$

*See Appendix H, part 1.

To get the maximal increments as each representation is added, we simply need to order the four representations by their micro-recall values from Table 8. The results of applying the model to the Phase I data are presented in Table 20.

So, at least for the data in Phase I, the model predicts quite well. Predictions are not made for the Phase II data because the obtained relative recall is not an accurate enough estimate of actual recall -- there are not a sufficient number of relevant documents known to be in the data base beyond those retrieved by the four representations.

The overall conclusion is that overlaps are much as might be expected if the representations were selecting relevant documents from the data base at random. The problem of finding truly complementary representations is largely unsolved, but the contrast between abstract words (AA) and descriptors (DD) is a small step in the right direction. If these results generalized to other data bases, then one interpretation is that systems should have both controlled and "free" document representation vocabularies.

Table 20

Predicted and Obtained Incremental Improvements in Recall - Phase I

	Order	Repr.	Micro-recall	Combined Representations	Predicted Recall	Observed Recall
Most Relevant	1st	II	.348	I	.348	.349
	2nd	AA	.328	I, A	.562	.558
	3rd	TT	.285	I, A, T	.687	.684
	4th	DD	.237	I, A, T, D	.761	.783
All Relevant	1st	II	.306	I	.306	.306
	2nd	AA	.283	I, A	.502	.505
	3rd	TT	.229	I, A, T	.616	.634
	4th	DD	.216	I, A, T, D	.699	.740

NOTES: (1) Micro-recall values are taken from Table 8.

(2) Predicted recall computed from formulas in text of report.

(3) Observed recall are computed from number of relevant documents retrieved (Table 16) divided by either 810 or 1723 (Table 15). Observed recalls are relative recalls based on seven representations. These figures will, therefore, overestimate actual recall.

REFERENCES

American Psychological Association. Psychological Abstracts Information Services Users' Reference Manual. APA, 1976.

Cleverdon, C.W. The Cranfield Tests on Index Language Devices. ASLIB Proceedings, 19, No. 6, June 1967. Pp. 173-194.

Hersey, D.F.; Foster, W.R.; Stalder, E.W., and Carlson, W.T. Free text work retrieval and scientist indexing; Performance profiles and costs. Journal of Documentation, 1971, 27, 167-183.

INSPEC. Free-Indexing Specification. The Institution of Electrical Engineers. London, England, December 9, 1970.

Keen, E. Michael. The Aberystwyth Index Language Test. The Journal of Documentation, Volume 29, No. 1, March 1973, pp. 1-35.

McGill, Michael J. An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems. Final Report for Grant NSF-IST-78-10454 to the National Science Foundation, 1979.

Salton, Gerard. A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). Journal of the American Society for Information Science. 1973, 23, 75-84.

Salton, Gerard. The evaluation of computer-based retrieval systems. In Automatic Information Organization and Retrieval. New York: McGraw Hill. 1968.

Smith, Linda C. Selected Artificial Intelligence Techniques in Information Retrieval Systems. Unpublished doctoral dissertation. Syracuse University School of Information Studies, 1979.

Sparck-Jones, K. Automatic Indexing. Journal of Documentation. 1974, 30, 393-432.

Sparck-Jones, K., and Jackson, D.M. The use of automatically-obtained keyword classification for information retrieval, Information Storage and Retrieval, 1970, 5, 175-201.

Taylor, Robert S. Value-added Processes in the Information Life Cycle. Journal of the American Society for Information Science. In Press.

van Rijsbergen, K. Information Retrieval (2nd ed.). Butterworths, 1979.

Waldstein, Robert. DIATOM: A DIALOG Simulator. On-Line, 1981, 5, 68-72.

Williams, M.E. Analysis of terminology in various CAS data files as access points for retrieval. Journal of Chemical Information and Computer Sciences. 1977, 17, 16-20.

APPENDICES

APPENDIX A

Phase I

Appendix A-1

PROJECT DESCRIPTION

This project will examine the relation between the relevance of retrieved citations and the fields that were searched to obtain them. Retrieval from seven different document representations will be studied. A representation consists of one or two designated search fields.

The data base for the study is Computer and Control Abstracts (a subfile of INSPEC). The system you will use is a local simulator of DIALOG, mounted on the S.U. computer. Almost all DIALOG features are available for you to use, but some restrictions will be made to achieve the study objectives.

The objectives of the study require you to conduct high recall searches, but with a limit of no more than 50 citations per query.

In all, you will be asked to search 98 queries. Over the course of the study, you will use all seven representations, but for each query only one representation will be assigned.

For each query, you will be asked to search from a request form; the statement of the query was prepared by a real user who will receive the output. The request form will also prescribe the representation you are to use. The unique password assigned to the request will automatically "lock" the search so that you can only search on the designated parts of the citations.

After you have completed each search (including the essential print command), return the search request form and a copy of your interaction with the system to Brian McLaughlin.

(5/2/80)

Phase I

DATA BASE

Appendix A-2

Computers and Control Abstracts is that portion of the INSPEC Data Base dealing with all areas of computing and information science. The specific data base that will be searched in this study consists of four months (Sept. - Dec. 1979) of Computer and Control Abstracts.

The citations you will retreive will be organized as follows:

DNnumber (abstract numbers from INSPEC journals)
Title
Authors (separated by commas)
Source field: as follows
Publication: (volume and issue number)(part number) pagination data
Following this may be information in [ ]. This is information on the cover-to-cover translation as follows: [publication; (volume and issue) pages date] (type of unconventional media)(availability) (Title of conference), (location of conference); (sponsoring organization) (date) language
Abstract
Indexing information

NOT all the citations will contain each of these items of information.

Phase I

DIALOG - SIMULATOR DIFFERENCES

The DIALOG simulator you will be using to conduct the searches is almost identical to "regular" DIALOG. In general, searching should be performed in the same way as any DIALOG search.

The restrictions, cautions and limitations are noted below.

1. Each new query you search must be started with the full BEGIN.
2. To restrict a search to a particular language, use a Limit /ENG (for English), or whatever language you wish.
3. Adjacency (nW) cannot be used with either truncation or stemming.
4. Adjacency may run very slow; the field operator (F) can be used instead.

(5/2/80)

Phase I

THE REPRESENTATIONS

Appendix A-3

You will be using seven different representations during the study. A representation names the one or two fields of the citation to which your search must be restricted. You will search on only one representation for any given query. The representation you are supposed to search on will be designated on the request form we give to you. A unique password will be given with each request and this password will automatically lock the search onto the assigned representation.

The seven representations and the fields they will search are as follows:

TT - will search terms in title only.

AA - will search terms in abstract only.

DD - will search descriptor terms only. A thesaurus will be provided to you for use with this controlled vocabulary representation. (The thesaurus may only be used on this project).

II - will search identifier terms only.

TA - will search terms in title and abstract only.

ST - will search stemmed terms in title and abstract only. The computer will automatically take the logical root of any entered term. Truncation cannot be used with this representation.

DI - will search terms in descriptor and identifier fields. The thesaurus will be provided for use with this controlled vocabulary representation.

One representation with which you may be unfamiliar is stemming (ST), which will be used with title and abstract words only. A stemmed term is a word that has been shortened by the computer to its logical root. This is similar to truncation in that the stem LIBRAR would retrieve LIBRARY, LIBRARIES, LIBRARIAN, etc. For truncation however, the root is determined by the searcher. For example, if you entered LIBRARY under the ST representation, the computer would automatically be reduced to its logical root and LIBRARY, LIBRARIES, LIBRARIAN, LIBRARIANS, etc. would all be retrieved.

Truncation is not to be used with the stemming representation. In fact, the simulator will reject any attempts to use truncation in this representation.

(5/2/80)

Query # 003 - Practice Search

Phase I

NAME: _____ DATE: _____

SCHOOL ADDRESS: _____ PHONE: _____

HOME ADDRESS: _____ PHONE: _____

We would like a description of your topic of interest. This statement should be clear enough so that any person who also knows about this topic would, on the basis of this statement alone, be able to pick out citations of interest for you.

Please write your description here;

I am interested in information about voice recognition systems and the used of speech recognition in man-machine systems. I am particularly interested in the use of interactive terminals and continuous speech recognition. I do not want citations that deal only with computer pattern recognition. The information must also include voice recognition.

Given your purposes in requesting this search, how many citations do you want? _____

About how many citations on your topic do you expect to receive from this computer search? _____

YOU MAY FOLD THIS REQUEST FORM IN THIRDS. STAPLE SECURELY, AND DROP IN CAMPUS MAIL.

4/4/80

Query # 004 - Practice Search

Phase I

NAME: _____ DATE: _____

SCHOOL ADDRESS: _____ PHONE: _____

HOME ADDRESS: _____ PHONE: _____

We would like a description of your topic of interest. This statement should be clear enough so that any person who also knows about this topic would, on the basis of this statement alone, be able to pick out citations of interest for you.

Please write your description here;

My topic of interest involves national and international policy issues as they relate to computers and information. I would like information about how the political structure affects the communications market and how different policies affect database usage, applications, and cost. Although I am especially interested in policies with regard to management information systems and EDP management, I would like as many citations as possible concerning the broader area of policy issues.

Given your purposes in requesting this search, how many citations do you want? _____

About how many citations on your topic do you expect to receive from this computer search? _____

YOU MAY FOLD THIS REQUEST FORM IN THIRDS. STAPLE SECURELY, AND DROP IN CAMPUS MAIL.

4/4/80

Phase II

SEARCHER INFORMATION

PROJECT DESCRIPTION:

This project will examine the relation between the relevance of retrieved citations and the fields or representations that were searched to obtain them. The database for the study is a portion of Psychological Abstracts. Searchers will be asked to search each query four times - once under each of the four representations.

REPRESENTATIONS:

A representation names the field of the citation to which a search must be restricted. The four representations to be used for each query by each searcher are:

1) TT - search terms in TITLE only.
2) AA - search terms in ABSTRACT only.
3) DD - search terms in DESCRIPTORS only.
4) II - search terms in IDENTIFIERS only.

DATA BASE DESCRIPTION:

The database consists of journal articles written in English from Psychological Abstracts (PA) published during six months (July-December 1980). This file contains both clinical and research aspects of psychology and includes subjects such as cognitive processes, educational psychology, psychometrics and statistics, and guidance and counseling.

PA citations printed on-line exhibit the following categories of information, when available:

Document Number
Title
Author
Source
Section Code
Abstract
Descriptors
Identifiers

SYSTEM FEATURES:

You will be using DIATOM, a system mounted on the S.U. computer which is a local simulator of DIALOG, and almost identical to it. Some of the major features you will probably make use of are.

-- Select or Select Steps.
-- Boolean operators with a Select or Combine statement.
-- Full text operators, (W), (NW), (F), (C).
-- Truncation with any operator (boolean or full text).

SEARCHER INFORMATION, Page 2, Phase II

Refer to the DIATOM-DIALOG simulator handout which lists all the possible commands. Use only those which are in both systems.

Although a stemmer and some other "automatic" features are available on DIATOM, do not use them as DIALOG does not have them.

SEARCH PROCEDURE:

Each searcher will search on 40 to 60 queries. Four searches will have to be conducted by each searcher for each query, one for each of the four representations. The four representations must be searched in a pre-specified order.

Your job as searcher is to prepare and conduct high recall searches.

For each search you will be given a request form. The query will be prepared by a real user who will receive the output. You will receive no information regarding the user's request other than what is designated on the request form. This form will also have the order of the representations to be used designated on it.

You are to pick up the search requests on Mondays and Thursdays, and return the completed searches by the Monday or Thursday that follows. You will have 2-3 days to complete each search.

You may perform the search on any terminal that is or can be connected to S.U., that is convenient to you, as long as a hard copy can be printed.

Here it is important to note that each search on a query should be started with a BEGIN command (which together with the query number and searcher password) locks the search to a particular representation. The next BEGIN command for the same query locks it to a different representation according to a pre-assigned order of representations. This way the order of representations to be used cannot be changed. You will be given a thesaurus for controlled vocabulary searching.

When you have completed a search, use a PRINT command with Format 1, to get the document numbers of the retrieved set. If no documents have been retrieved, type in NOTHING FOUND and print out any one document with FORMAT 1.

Return 1) the search request packet, filling in the needed information and

2) a copy of your interaction with the system to Brian McLaughlin.

Phase II

DIATOM - a DIALOG simulator

DIATOM (Dialog Implementation - Augmented To Overcome Magic) was implemented at Syracuse University by Robert Waldstein as both a teaching device and a research tool. It incorporates most of the features of DIALOG and has a few additional features. The comparison in the following description is accurate as of May 1980.

Command Summary

BEGINn, Bn, !n
To start a search in file n. Erases work done to that point; restarts set numbers at 1.
Examples: BEGINn; B1; !1

BEGIN
Equivalent to BEGINn but includes a routine for labeling the search.
Examples: BEGIN

BEGIN BYPASS, BB, !B
This command is the equivalent of BEGIN1.
Examples: BEGIN BYPASS; !B; BB

EXPAND, E
To display a part of an index. May be used with words, prefix codes, or online thesaurus.
Examples: EXPAND ART; ELIBRARY; EAU=Waldstein, R?; E R1
Simulator difference: Only one expand list exists at a time. I.e. you can't have both an E and R list at the same time.

EXPAND (word)
To display subject related terms from a thesaurus.
Examples: EXPAND (ENERGY); E (READING)

SELECT, S
To request postings to be retrieved from the index. May be used with words, prefix codes, or EXPAND numbers.
EXAMPLES: S MIRAGE; SAU=BOB; SE1,E4-E7; SR2,R4-R6,R9
SELECT can also be used with boolean operators. In that case it Selects a full boolean set description; with AND, OR, NOT, and parentheses operators. Note that boolean hierarchy is used in the following order: (), NOT, AND, OR. Set numbers may be used as an item, e.g. S DOG AND S1; S DOG AND #1.
Examples: SELECT DOGS AND CATS; S DOG/DE,AB OR E3
S (AU=BOB OR JO=JASIS) NOT E1-E5
Simulator difference: DIALOG always creates the sets in the order given. E.g.

```
S DOG AND LIBRAR?  NOT R2,R5
          150 DOG
         2053 LIBRAR?
           12 R2,R5
    1     35 DOG AND LIBRAR?  NOT R2,R5
```

The simulator may create the sets in a different order for internal optimization reasons.

Phase II

SELECT STEPS, SS, S STEPS
Equivalent of SELECT with boolean operators except that each term results in a numbered set. For example:
SS DOG AND LIBRAR? NOT R2,R5
1 150 DOG
2 2053 LIBRAR?
3 12 R2,R5
4 35 DOG AND LIBRAR? NOT R2,R5

SELECT [word]
SELECTs the thesaural entries for this word. It selects all entries except RTs (related terms) and BTs (broader terms).
Examples: SELECT [ENERGY]; S [READING]
Simulator difference: DIALOG has no comparable capability.

COMBINE, C
Used with boolean operators AND, OR, NOT to relate sets. May only be used with set numbers.
Examples: COMBINE 1 AND 2; C6-8/OR; C 4 AND (5 OR 6); C7-4

TYPE, T
To type record(s) online at a terminal. Used with either set numbers or DIALOG accession numbers: set/format/range. Formats 1-8 are used.
Examples: TYPE 10; T12/2/1-6; TDN1023

DISPLAY, D
Displays a record online. Same as TYPE.
Examples: DISPLAY 10/3/2-4,7; D DN312

PRINT
To request offline prints. Used with either set numbers or DIALOG accession numbers.
Examples: PRINT 7/5/1-49
Simulator difference: A print creates a set on disk named by the password used at LOGON. It is of the form
<1st 6 chars of password>.<last 2 chars>.
To get an offline print once the simulator is left then use monitor PRINT command.

PRINT -
To cancel the previous print request. Must be used before LOGOFF, BEGIN, .FILE, or END commands.
Examples: PRINT -

END
Gives time elapsed and cost estimates since last BEGIN or END or file change. Does not interfere with search strategy. Starts new costing.
Examples: END

.COST
Gives the elapsed time and cost estimate since last BEGIN. Does not interfere with search strategy.
Examples: .COST

DISPLAY SETS, DS
To display all sets made since previous BEGIN. Used for a recap of search strategy used.
Examples: DISPLAY SETS; DS

DISPLAY SETS n-m,x,y-z
Used to display a certain set of the created sets.
Examples: DISPLAY SETS 15-18,26; DS 3
Simulator difference: This capacity is a little broader than DIALOG.

EXPLAIN, ?
To request online explanations of command and file features.
Examples: EXPLAIN FILE1; ?NEWS; ?NEGDIC

.FILEn
To change to another file. Use not recommended on either DIALOG or the simulator.
Examples: .FILE 1

FEEDBACK, F
This enables the user to do feedback on a known relevant document. Feedback can be done on four fields: title, abstract, descriptors, and identifiers. For the title and abstract the terms fedback on are those separated by spaces while for the descriptors and identifiers the terms separated by semicolons are those fedback on. For this reason it will not work to combine free and controlled representation feedback. Note feedback can also be done on major fields (e.g. DE*). The default field is the title. There are 3 different types of feedback available:
FEEDBACK 1 - This type of FEEDBACK ORs all the terms of the desired field(s). Note that this is the default.
FEEDBACK 2 - This type of FEEDBACK ANDs all the terms of the desired field(s). Note that usually this will give no documents.
FEEDBACK 3 - This FEEDBACK uses the ERIC thesaurus. Note that it is therefore meaningful only on the descriptor field.
Examples : FEEDBACK2 DN1234/TI; F DN5/ID*; F3 DN2543/DE; F DN3456
Simulator difference: No equivalent feature in DIALOG.

NATURAL, N
Does a search on the words of a natural language request. Takes the words of the command string and ORs their stems together.
Examples: NATURAL THE USE OF INFORMATION RETRIEVAL SYSTEMS
Simulator differences: No equivalent feature in DIALOG

NATURAL RANK, NR
Does a search as in NATURAL but unstemmed and ranks the retrieved documents by inverse document frequencies. Important note: the sets created by this command can not be combined with other sets! Note that format 12 gives the rank weights of the retrieved documents.
Example: NR THE USE OF INFORMATION RETRIEVAL SYSTEMS
Simulator differences: no equivalent features in DIALOG.

Phase II

LIMIT, L
To restrict SELECTed set to specified requirements. Capability varies by file.
Examples: LIMIT 5/MAJ; L2/MIN; L 8/MAJ,MIN; L3/TI,AB
Simulator difference: DIALOG does not permit LIMITing by field, DIATOM does. In general, DIALOG has more LIMITs per file then DIATOM. Check file documentation for details.

LIMIT ALL, LALL
Used before SELECTing sets to limit all subsequent SELECTing to specified requirements. Capability varies by file.
Examples: LIMIT ALL/MAJ; LALL/STEM; LALL/DE,ID*,TI
Simulator difference: The simulator can't limit by accession number. However, DIALOG can't limit by stem or by field.

LIMIT ALL/ALL
To cancel a LIMIT ALL command
Examples: LIMIT ALL/ALL; LALL/ALL

PAGE, P
To request another screen (or page) of display after an EXPAND
Examples: PAGE; P

LOGOFF
To signoff and disconnect from DIALOG or the simulator. Automatically gives cost estimate of connect time.
Examples: LOGOFF
Simulator comments: The pause that sometimes occurs during logoff is caused by two processes: all TMP files created by the user are deleted and all PRINT commands are executed.

Search Save Commands: END/SAVE, .EXECUTE, .RELEASE, .RECALL
Simulator difference: None of these are implemented on the simulator. Note however that they all give appropriate messages when their use is attempted.

Search features

Truncation - ? (question mark)
There are four capabilities in truncation:
1) Unlimited number of characters after the stem.
SELECT EMPLOY?
2) Specified maximum number of characters after the stem.
SELECT HORSE? ?
SELECT THEAT?? ?
3) Embedded variable character
SELECT WOM?N
SELECT ADVERTI?E
4) Combination of the above.
SELECT WORKM?N?

Stemming - @ (ampersand) There are two capacities in stemming.
1) SELECT all words with same stem.
SELECT LIBRARIAN@
2) In combination with internal truncation.

Phase II

SELECT WOM?N@

Simulator difference: No comparable feature in DIALOG.

Basic index field indicators

Suffix symbols; used to specify searching on field(s) which make up the basic index. Fields vary per database.

../AB	Abstract
../DE, .../DE*	Descriptors
../DF, .../DF*	Full descriptors (single word)
../ID, .../ID*	Identifiers
../IF, .../IF*	Full identifiers (single word)
../TI	Title

* indicate MAJOR

1) SELECTing single terms:
   SELECT BUDGETS/TI
2) Specifying more than one field:
   SELECT TENSION/TI,DE,ID
3) With full text operators:
   SELECT POP(W)TOP(F)CANS/TI,AB

Additional indexes

Always used with two-letter prefix code. Prefixes vary per database.

AU=	Author
JO=	Journal

Full text operators

Used only with SELECT command.

(W) To request a word immediately adjacent to another in the given sequence.

Example: S SOLAR(W)ENERGY

(nW) To request a word within n words of another in the given order.

Example: S SOLAR(3W)ENERGY

(F) To request a word in the same field as another; in any order in any field.

Example: S SOLAR(F)ENERGY

(C) To request a word in the same citation as another; in any order. Note that this is the same as AND.

Example: S SOLAR(C)ENERGY

Simulator difference: The simulator does not recognize (L) or (S).

Simulator comment: Adjacency searching (W) is very slow. E.g S INFORMATION(W)RETRIEVAL may take around 3 minutes.

Full text operators used with truncation or stemming

A recent addition to DIALOG is the ability to use full field features in conjunction with stemmed or truncation features.

Examples: S LIBRA???? ?(F)AUTOMAT???? ?; S WOM?N(F)SOCIETY@

Simulator difference: The simulator cannot use internal truncation when adjacency is used. E.g. S WOM?N(W)HISTORY will not work. Note that simulator will give an unimplemented DIALOG

Phase II

feature message.

Range searching
Simulator difference: The simulator does not recognize range searching requests.

Using Boolean terms
Apostrophies (') may be used to select a term with a boolean operator.
Example: S 'ARMY AND NAVY'
Simulator difference: The simulator works slightly more generally than DIALOG. The difference will not be apparent in normal use. However, DIALOG improperly handles
S CAN'T AND WON'T
while the simulator handles it correctly.

Command entry and output features

Stacking
Use a semicolon (;) to seperate a series of commands to be executed with one carriage return.
Example: S E1-E3;S AU=BOB;L 2/MAJ; C 1 AND 3

BREAK
Use the break key to stop output and stop execution of present command
Example: T 1/5/1-400 [BREAK]
Simulator difference: Unfortunately this doesn't work till the DEC clears its output buffer of approximately 150 characters. <cntl O> will stop output immediately. Note that <cntlO> does not stop execution and it is important to hit [break] as well.

Backspace and erase
Use <cntl H> or <backspace key> or <delete> to erase last characters typed before carriage return.

Erasing a line
Use <escape> key followed by the <return> key. The system will ignore the line and give another prompt.

Width control at logon
When giving your 8 character password a terminal width may be specified. This can range from 30 to 115. Just follow the password with "Wnnn" where nnn is the desired width.

Output Control

Format Options
The following options are available and may be used with the TYPE, DISPLAY, or PRINT commands.
Format 1 - DIALOG accession number
Format 2 - Full Record except abstract
Format 3 - Bibliographic citation
Format 4 - Abstract and title
Format 5 - Full record

Phase II

Format 6 - Title and accession number
Format 7 - Bibliographic citation and abstract
Format 8 - Title and indexing

TYPE set #/format #/range
If no range is given defaults to the first citation. If no format is given defaults to 5.
DISPLAY set #/Format #/range
Same as for TYPE
PRINT set #/Format #/range
Same as for TYPE

Files

Presently there are six files in the system.

ERIC - File 1
This file consists of 8,573 citations from the ERIC database. It contains all the RIE and CIJE documents for four clearinghouses: IR, EA, TM, and TE from 1980. Note this was a transition year for the ERIC thesaurus.
Suffixes: AB,TI,DE,DE*,DF,DF*,ID,ID*,IF,IF*
Prefixes: JO=, AU=, CH=, DT=
Limits: MIN, MAJ, ED, EJ

CIJE - File 2
This file consists of 10,885 citations from the ERIC database. These are all from current index to journals in education (EJ numbers) from four clearinghouses: IR, EA, TM, and TE from 1974-1978.
Suffixes: AB,TI,DE,DE*,DF,DF*,ID,ID*,IF,IF*
Prefixes: JO=, AU=
Limits: MIN, MAJ
DN numbers are used in place of ED or EJ numbers.

INSPEC - file #3
This file consists of 12,864 documents which is the last 4 months of the 1979 Computer and Control file.
Suffixes: AB,TI,DE,DF,ID,IF
Note that the ID fields are the free text terms assigned by INSPEC indexers.
Prefixes: JO=,AU=
Limits: FRN, ENG
DN numbers are used for internal access.

OSP - file #4 This file consists of the research being conducted presently at Syracuse University. It is produced by the Office of Sponsered Programs under Bill Wilson. It is (presumably) being continually updated.
Suffixes: TI, AB, DE, DF
Prefixes: Sponser's Name (SN=), Project Director (PD=), Department Name (DN=)

LRAP - File #5
This file contains bibliographic citations for books, reports, dissertations, and other items of importance to the Local Revenue

Phase II

Administration Project. Funded by U.S. Agency for International Development through Syracuse University Maxwell School, the project is directed by D. Glynn Cochraine.
Suffixes: TI, AB, ID, IF, DE, DF, GE, GR (Geographical region)
Prefixes: Author (AU=), Affiliation (AF=), Source (SO=), Date of Publication (PD=), Document Type (DT=), Contract Number (CN=), Historical Period (HP=), CAll number (CA=)
Limits: ENG, FRN, MAP, BIB (Bibliography), TAB (Table)

PSYABS - File 6
This file consists of 11,662 citations from the Psychological Abstracts database. It consists of all documents from issue 64 with a DT (document type) of journal.
Suffixes: TI,AB,DE,DF,IF
Prefixes: AU=, JO=, SH=

Simulator file limitations

Thesaurus
There are no RT entries in the main inverted file. However, descriptors are listed with a ? in the related term column during an EXPAND. These items can have a thesaural expansion done by doing an E E9 (in the case where E9 has a ? in the RT column). Also no posting information is included in the thesaurus EXPANDs.

Other simulator features for the head honcho

EXPLAIN files
When any file is created under the main PPN (e.g [3434,12]) or the PPN from which the simulator is being executed with a DIA extension it is accessable from the simulator using an EXPLAIN command. E.g. if a file is created called BOB.DIA then ?BOB will type out this file on-line. If a file called LOGON.DIA is created it is printed whenever anyone logs on.

Required passwords
When a file called passwd.DIA is created in the account from which the simulator is being executed then only the passwords in that file can use the simulator. A form of an entry in this file is:

<8 letter password><space><file number><space><LALL Command>

The file number and the LALL command are both optional. An example entry is

WALDSTEI 1 /STEM

will cause a person using password WALDSTEI to logon into file 1 with a LIMIT ALL to STEM.

ERIC file size on the DEC 10
The size needed for storage of the ERIC file in blocks (128 DEC10 words) is as follows:

ERIC.DAT	- document file	12720 blocks
ERIC.INV	- main inverted file	5369 blocks
ERIC.JO	- journal inverted file	79 blocks
ERIC.CH	- Clearing house file	19 blocks

Phase II

ERIC.DT	- Document type file	59 blocks
ERIC.AU	- author inverted file	429 blocks
ERIC.BIG	- inverted file of terms with >1100 postings	691 blocks
ERIC.THE	- ERIC thesaurus	2338 blocks

CIJE file size on the DEC 10

The size needed for storage of the CIJE file in blocks (128 DEC10 words) is as follows:

CIJE.DAT	- document file	8467 blocks
CIJE.INV	- main inverted file	3749 blocks
CIJE.JO	- journal inverted file	119 blocks
CIJE.AU	- author inverted file	539 blocks
CIJE.BIG	- inverted file of terms with >1100 postings	326 blocks
CIJE.THE	- CIJE thesaurus	2066 blocks

An indeterminate amount of space can be used by the EXPLAIN commands as described above.

5/5/23
DN11137 13063
TITLE: 3-Methoxy-4-hydroxyphenylglycol excretion in acutely schizophrenic patients during a controlled clinical trial of the isomers of flupenthixol.
AUTHOR: Joseph, M. H.; Baker, H. F.; Johnstone, Eve C.; Crow, T. J.
SOURCE: Psychopharmacology. 1979 Vol 64(1) 35-40
SECTION CODE: 3340; 2520
ABSTRACT: Urinary 3-methoxy-4-hydroxyphenylglycol (MHPG) excretion in 45 acute schizophrenics was studied before and during a trial of the isomers of flupenthixol and placebo. Pretrial MHPG excretion was not related to severity of illness before the trial or to other pretrial clinical variables. In male Ss, higher pretrial MHPG excretion was associated with a better outcome 1 yr posttrial. However, in females, no relationship between MHPG excretion and outcome was established. During the trial there was a reduction in MHPG excretion in Ss treated with beta-flupenthixol but no decrease in the group treated with alpha-flupenthixol or chlorpromazine. In Ss on placebo, there was a reduction in MHPG excretion in those who did well clinically but not in those who did poorly. Thus low MHPG excretion may be a predictor of poor outcome in schizophrenia, but MHPG excretion also changes as a function of clinical state and neuroleptic drug administration. (35 ref)
DESCRIPTORS: URINATION; NOREPINEPHRINE; METABOLITES; ACUTE SCHIZOPHRENIA; NEUROLEPTIC DRUGS; HUMAN SEX DIFFERENCES; DRUG THERAPY; NEUROCHEMISTRY; PREDICTION
IDENTIFIERS: isomers of flupentixol, urinary excretion of 3-methoxy-4-hydroxyphenylglycol & relationship of metabolite levels to clinical variables & prediction of drug response, male vs female acute schizophrenics

5/5/30
DN11111 13029
TITLE: Treatment of severe dog phobia in childhood by flooding: A case report.
AUTHOR: Sreenivasan, Uma; Manocha, S. N.; Jain, V. K.
SOURCE: Journal of Child Psychology & Psychiatry & Allied Disciplines. 1979 Jul Vol 20(3) 255-260
SECTION CODE: 3330
ABSTRACT: An 11-yr-old girl with a 5-yr history of severe phobia of dogs was treated with flooding after desensitization failed. 19 mo after flooding the S was free of the phobia and symptoms of a tension state. (10 ref)
DESCRIPTORS: IMPLOSIVE THERAPY; PHOBIAS; SCHOOL AGE CHILDREN; CASE REPORT; HUMAN FEMALES
IDENTIFIERS: flooding treatment, treatment of dog phobia, 11 yr old female

ABSTRACT: Utilized consumer-descriptive and behavioral-descriptive data to examine the factors that influence overall magazine readership levels within a sample of US men and women (2,819 women and 3,186 men). Over 70% of the total variance in readership could be predicted with a combination of demographic, psychographic, media-usage, TV-program-choice, and magazine-choice variables. Psychographic dimensions were more important predictors for women than men, and TV program factors were more important for men than women. These patterns may develop from the (generalized) differences in the uses of media for each sex, or from sexually based differences in how individuals perceive the gratifications available from the different media. Further research would be necessary to confirm the suspicion, the author notes, but congruity of TV and magazine preference patterns could be expected more frequently where psychographically related functions of the media (for ""other-directedness'') were weaker. Men may perceive TV and magazines as similar media (for relaxation, perhaps), whereas women's use of these print and broadcast media differs and therefore their selection patterns differ. It is also noted that the pattern of demographic and psychographic predictors confirms previous findings on the positive relationship between higher socioeconomic characteristics and higher magazine readership. (48 ref)

ABSTRACT: 35 patients (mean age 34.7 yrs) with premenstrual syndrome recorded their symptoms daily using the Moos Menstrual Distress Questionnaire. These were analyzed by a least mean square method of fitting sine waves. After recording an untreated cycle, Ss were given progesterone (200 mg) and placebo in a double-blind crossover manner; 75% of the Ss were then given progesterone (400 mg) and placebo in a similar manner. Treated cycles were rated by both daily menstrual distress questionnaires and retrospective self-assessment. Both rating methods showed there was no significant difference between progesterone and placebo in reducing symptoms of premenstrual syndrome, and in the majority of cases placebo was more effective, although never significantly so. (13 ref)

ABSTRACT: In a replication of a study by H. Garland and K. H. Price (see PA, Vol 61:1020), 143 male and 83 female advanced university business students read descriptions of the success or failure of a fictional female manager in the 1st yr of her job, completed the Women as Managers Scale (WAMS), and rated 4 possible causes for the manager's success or failure (ability, effort, luck, or nature of job). Garland and Price's finding that WAMS scores were not affected by success or failure descriptions was replicated for both male and female Ss, and additional data show that males and females tended to attribute success and failure to similar factors. (10 ref)

TITLE: Psychophysiological investigations of post lunch state in male and female subjects.
AUTHOR: Christie, Margaret J.; McBrearty, Eileen M.
DESCRIPTORS: FOOD INTAKE; HUMAN BIOLOGICAL RHYTHMS; METABOLISM; EMOTIONAL STATES; PSYCHOPHYSIOLOGY; PERFORMANCE; PARASYMPATHETIC NERVOUS SYSTEM; HUMAN SEX DIFFERENCES; BODY TEMPERATURE
IDENTIFIERS: lunch, diurnal variation in blood glucose & autonomic factors & body temperature & mood & performance efficiency, male vs female Ss, implications for parasympathetic involvement in deactivated mood

TITLE: A developmental attributional analysis of sex role stereotypes for sport performance.
AUTHOR: Bird, Anne M.; Williams, Jean M.
DESCRIPTORS: SCHOOL AGE CHILDREN; ADOLESCENTS; AGE DIFFERENCES; STEREOTYPED ATTITUDES; SEX ROLE ATTITUDES; SPORTS; ATTRIBUTION
IDENTIFIERS: age & sex of athlete & outcome & sport, attributions of ability vs luck to sports performances & sex role stereotypes, male & female 7-9 vs 10-12 vs 13-15 vs 16-18 yr old Ss

TITLE: Human social attitudes affected by androstenol.
AUTHOR: Kirk-Smith, Michael; Booth, D. A.; Carroll, D.; Davies, P.
DESCRIPTORS: HUMAN SEX DIFFERENCES; SOCIAL PERCEPTION; EMOTIONAL RESPONSES; EMOTIONAL STATES; ANDROGENS; DRUG EFFECTS
IDENTIFIERS: androstenol, mood & personality ratings of people in photographs, male vs female Ss

TITLE: Adults' conceptions of children's cognitive abilities.
AUTHOR: Miller, Scott A.; White, Nancy; Delgado, Maria
DESCRIPTORS: COGNITIVE ABILITY; COGNITIVE DEVELOPMENT; HUMAN SEX DIFFERENCES; PARENTS; PIAGETIAN TASKS; ADULTS; DEVELOPMENTAL DIFFERENCES; SOCIAL PERCEPTION
IDENTIFIERS: various Piagetian cognitive ability tasks & type of question asked of adults, adult conceptions of children's abilities, male vs female & parent vs nonparent Ss

TITLE: Performance-self-esteem and dominance behavior in mixed-sex dyads.
AUTHOR: Stake, Jayne E.; Stake, Michael N.
DESCRIPTORS: HUMAN SEX DIFFERENCES; SELF ESTEEM; PERFORMANCE; SEX ROLES; DOMINANCE/; GROUP DECISION MAKING; DYADS; INTERPERSONAL INFLUENCES
IDENTIFIERS: decision making dominance in mixed sex dyads & performance self esteem, male & female Ss

TITLE: Crowding, contagion, and laughter.
AUTHOR: Freedman, Jonathan L.; Perlick, Deborah
DESCRIPTORS: CROWDING; LAUGHTER; INTERPERSONAL INFLUENCES; GROUP DYNAMICS
IDENTIFIERS: low vs high density crowding conditions & confederate laughing vs not laughing during humorous tapes, amount of laughter by Ss, female college students

TITLE: Severity of psychiatric disorder and the 30-item General Health Questionnaire.
AUTHOR: Finlay-Jones, Robert A.; Murphy, Elaine
DESCRIPTORS: TEST VALIDITY; QUESTIONNAIRES; MENTAL HEALTH; MENTAL DISORDERS/; PSYCHODIAGNOSIS
IDENTIFIERS: validity of 30-item General Health Questionnaire, diagnosis of severity of psychiatric disorder, 18-40 yr old female general practice patients vs 18-65 yr old Ss with recent severe physical symptoms

TITLE: Consequences for targets of aggression as a function of aggressor and instigator roles: Three experiments.
AUTHOR: Gaebelein, Jacquelyn W.; Mander, Anthony
DESCRIPTORS: ROLES; AGGRESSIVE BEHAVIOR; ROLE PERCEPTION; ROLE EXPECTATIONS
IDENTIFIERS: aggressor vs instigator role of Ss, intensity of aggression toward opponent, female college students

TITLE: Aggression against a remorseful wrongdoer: The effects of self-blame and concern for the victim.
AUTHOR: Harrell, W. Andrew
DESCRIPTORS: GUILT; THEFT; CRIMINALS; AGGRESSIVE BEHAVIOR
IDENTIFIERS: remorseful vs nonremorseful thief, aggressive behavior towards thief, female Ss

TITLE: Interpersonal gaze and helping behavior.
AUTHOR: Valentine, Mary E.; Ehrlichman, Howard
DESCRIPTORS: EYE CONTACT; HUMAN SEX DIFFERENCES; ALTRUISM; ASSISTANCE (SOCIAL BEHAVIOR)
IDENTIFIERS: eye contact, helping behavior, male vs female Ss

TITLE: Importance of imagery in maintenance of feedback-assisted relaxation over extinction trials.
AUTHOR: LeBoeuf, Alan; Wilson, Clare
DESCRIPTORS: IMAGERY; BIOFEEDBACK TRAINING; RELAXATION THERAPY; EXTINCTION (LEARNING)
IDENTIFIERS: use of imagery vs passive concentration during frontalis EMG feedback training, maintenance of relaxation during extinction trials, female Ss

TITLE: Subjective estimates of body tilt and the rod-and-frame test.
AUTHOR: Sigman, Eric; Goodenough, Donald R.; Flannagan, Michael
DESCRIPTORS: ROD AND FRAME TEST; ILLUSIONS (PERCEPTION); ESTIMATION; VISUAL PERCEPTION
IDENTIFIERS: magnitude estimation procedure, illusory self tilt effect in rod

Phase II

SEARCH QUERY SHEET

Searcher Staff

Query Number 201 (Practice)

Date Search Collected 7/8

Order of Representations ADIT

Date Search to be returned 7/31

DIALOG Password STORMBON

Date Returned to Brian McLaughlin

Date Returned to NSF

Some Important Points:

1. Each new search must be started by the full BEGIN command.
2. Be sure to print the documents retrieved before typing the next BEGIN command.
3. If no documents are retrieved, type NOTHING FOUND and print using Format 1, any one document.
4. You do not need to LOGOFF after each search before starting the next search.

TO LOGON AND LOGOFF:

The step-by-step sequence for connecting with the computer, for conducting a DIALOG search, and for disconnecting from the computer, is given below

1. If you are using a dial-up terminal, the phone number is 423-1313.
2. Turn power on and hit carriage return.
3. Type: LOG 3434,14
4. Type: NSF
5. DO DIALOG

   The computer will ask for your dialog password. It is given at the top of this page.
6. Type: BEGIN

   The computer will ask for the query number and will lock the search to a particular representation code.

Phase II

SEARCH QUERY SHEET - Page 2

7. Carry out the search for this query.

Remember, we want a high recall search. Refer to the DIATOM-DIALOG Simulator handout for a description of possible commands.

Before starting a new search, use the PRINT command, the format should be 1, to have a set of the retrieved documents printed. If no documents have been retrieved, type in NOTHING FOUND and print out any 1 document with FORMAT 1.

8. If you want to conduct another search (for the same query) begin at Step 6.

If you are completely done searching for now, go to Step 9.

9. Type: LOGOFF

10. Type: K/F

11. Return all the materials to Brian McLaughlin.

HELP AND ASSISTANCE:

1.	Brian McLaughlin 210 Hubbell Avenue Syracuse, New York	476-7359 423-2091	(Home) (Work)
2.	NSF Retrieval Project 113 Euclid Avenue Syracuse, New York	423-4549	(Room 304) or (Room 306)

Phase II

SEARCH QUERY SHEET

Searcher Storm Query Number 202 (Practice)

Date Search Collected 7/8 Order of Representations ITAD

Date Search to be returned 7/31 DIALOG Password STORMBON

Date Returned to Brian McLaughlin Date Returned to NSF

- -

Some Important Points:

1. Each new search must be started by the full BEGIN command.
2. Be sure to print the documents retrieved before typing the next BEGIN command.
3. If no documents are retrieved, type NOTHING FOUND and print using Format 1, any one document.
4. You do not need to LOGOFF after each search before starting the next search.

TO LOGON AND LOGOFF:

The step-by-step sequence for connecting with the computer, for conducting a DIALOG search, and for disconnecting from the computer, is given below.

1. If you are using a dial-up terminal, the phone number is 423-1313.
2. Turn power on and hit carriage return.
3. Type: LOG 3434,14
4. Type: NSF
5. DO DIALOG

   The computer will ask for your dialog password. It is given at the top of this page.
6. Type: BEGIN

   The computer will ask for the query number and will lock the search to a particular representation code.

Phase II

SEARCH QUERY SHEET - Page 2

7. Carry out the search for this query.

   Remember, we want a high recall search. Refer to the DIATOM-DIALOG Simulator handout for a description of possible commands.

   Before starting a new search, use the PRINT command, the format should be 1, to have a set of the retrieved documents printed. If no documents have been retrieved, type in NOTHING FOUND and print out any 1 document with FORMAT 1.

8. If you want to conduct another search (for the same query) begin at Step 6.

   If you are completely done searching for now, go to Step 9.

9. Type: LOGOFF

10. Type: K/F

11. Return all the materials to Brian McLaughlin.

HELP AND ASSISTANCE:

1. Brian McLaughlin
   210 Hubbell Avenue
   Syracuse, New York
   476-7359 (Home)
   423-2091 (Work)

2. NSF Retrieval Project
   113 Euclid Avenue
   Syracuse, New York
   423-4549 (Room 304) or (Room 306)

APPENDIX B

Phase I

NSF INFORMATION RETRIEVAL PROJECT

INSTRUCTIONS TO PARTICIPANTS

Attached you will find a copy of your interest statement and two copies of a list of references. List (a) is to be used as part of the study and should be returned after you make your judgements of relevance. Copy (b) is yours to keep.

Each citation is organized into seven parts:

DN - Document identification number
TI - Title
AU - Author
SO - Source of the citation (i.e. journal title)
AB - Abstract
DT - Date
DE - Descriptors of the citation

Please read each citation and abstract to form an idea of what that particular document (book, article, report) is about. Compare this to your interest statement, and for each citation listed, decided how closely that citation is related to your topic. Based on the information in front of you, is the citation relevant to your topic, or not relevant to what you had in mind.

Use the following scale for your judgement:

1 - Definitely relevant to your topic.

2 - Probably relevant to your topic.

3 - Probably not relevant to your topic.

4 - Definitely not relevant to your topic.

Please rate each citation by placing the number corresponding to your judgement in the box immediately following each citation. After you have checked all the citations to see whether or not they are relevant to your interest statement, please return the copy with the judgements to us in the pre-addressed envelope through campus mail. If you are not on campus, these envelopes should be used to return the completed forms to us through the regular mail service. Thank you for your cooperation.

If you have any questions, please contact us at:

School of Information Studies
Syracuse University
113 Euclid Avenue
Syracuse, New York 13210
423-4549

6/16/80

Phase II

NSF INFORMATION RETRIEVAL PROJECT

INSTRUCTIONS TO PARTICIPANTS (A)

Attached you will find a copy of your interest statement and two copies of a list of references. Copy (A) is to be used as part of the study and should be returned after you make your judgements of relevance. Copy (B) is yours to keep.

Each citation is organized into eight parts:

Document identification number
Title
Author
Source of the citation
Section Code
Abstract
Descriptors of the citation
Identifiers

Please read each citation and abstract to form an idea of what that particular document is about. Compare this to your interest statement, and for each citation listed, decide how closely that citation is related to your topic. Based on the information in front of you, is the citation relevant to your topic, or not relevant to what you had in mind.

Use the following scale for your judgement:

1 - Definitely relevant to your topic.

2 - Probably relevant to your topic.

3 - Probably not relevant to your topic.

4 - Definitely not relevant to your topic.

Please rate each citation by placing the number corresponding to your judgement in the box immediately following each citation. After you have checked all citations to see whether or not they are relevant to your interest statement, please return the copy with the judgements to us in the pre-addressed envelope through campus mail. If you are not on campus, these envelopes should be used to return the completed forms to us through the regular mail service. Thank you for your cooperation.

If you have any questions, please contact us at

School of Information Studies
Syracuse University
113 Euclid Avenue
Syracuse, New York 13210
423-4549

APPENDIX C

SYRACUSE UNIVERSITY
SCHOOL OF INFORMATION STUDIES

Phase I

113 EUCLID AVENUE SYRACUSE, NEW YORK 13210 PHONE (315) 423-2911

NSF INFORMATION RETRIEVAL PROJECT

We are working on a project which will help us understand how the pertinence of information retrieved by computer is related to the method by which it is searched.

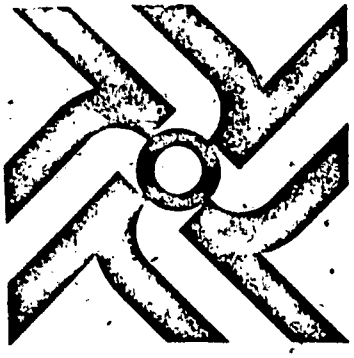
For this project, we need information requests which will be searched in Computer and Computer Control Abstracts (from October 1979 to January 1980). If you need information in the area of computers and information science, we will conduct a search for you free of charge. All you have to do is submit a search request to us and give us information on how we did after the search.

For the search request we would like you to describe a topic of interest to you; one you are working on or are familiar with, in the computer field. Several days later you will receive a list of citations that have been retrieved by computer. You will be asked at that time to indicate which of these are pertinent to your interest. One copy of the computer output will be returned to us and the other copy will be for your own use.

We would very much appreciate your cooperation and participation in this project. If you are willing to participate, please read the attached pages and write your search request in the space provided.

If you do not need a search, please pass this form to a student.

7/24/80



SYRACUSE UNIVERSITY

SCHOOL OF INFORMATION STUDIES

113 EUCLID AVENUE SYRACUSE, NEW YORK 13210 PHONE (315) 423-2911

Phase I

NSF INFORMATION RETRIEVAL PROJECT

As a participant in this project we would like you to submit a search request (on the attached form) about some aspect of computers and information science.

We will take your request and search the current issues of COMPUTER AND COMPUTER CONTROL ABSTRACTS. The results of this search will be a list of citations to books and journal articles.

We will then give you this list of citations and ask that you let us know which of these are most pertinent to your search request.

* * * * * * * * * * * * *

The enclosed form is for you to describe your topic of interest. If you are planning a talk or doing a paper, you probably have a topic in mind; if you don't have a topic you are working on, consider one with which you are familiar. Using this form, write down your information requirements as if you were talking to a colleague who understands the field as well as you do. Don't worry about trying to say it in "computerese"; we have trained people to make sure that your search is conducted professionally.

* * * * * * * * * * * * *

Thank you for your cooperation. If you have any questions, please feel free to contact us.

NSF Information Retrieval Project
School of Information Studies
113 Euclid Avenue
Syracuse, New York 13210
(315) 423-4522

4/4/80

Phase I

Appendix C-3

NAME: _____ DATE. _____

SCHOOL ADDRESS: _____ PHONE: _____

HOME ADDRESS: _____ PHONE: _____

We would like a description of your topic of interest. This statement should be clear enough so that any person who also knows about this topic would, on the basis of this statement alone, be able to pick out citations of interest for you.

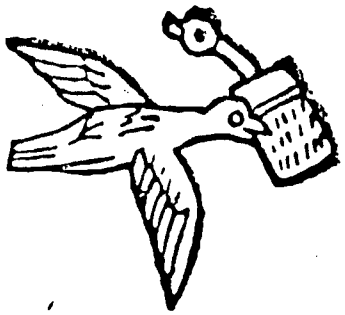
Please write your description here;

Given your purposes in requesting this search, how many citations do you want? _____

About how many citations on your topic do you expect to receive from this computer search? _____

YOU MAY FOLD THIS REQUEST FORM IN THIRDS. STAPLE SECURELY, AND DROP IN CAMPUS MAIL.

4/4/80



SYRACUSE UNIVERSITY

SCHOOL OF INFORMATION STUDIES

113 EUCLID AVENUE SYRACUSE, NEW YORK 13210 PHONE (315) 423-2911

Phase II

NSF INFORMATION RETRIEVAL PROJECT

We are working on a project which will help us understand how the pertinence of information retrieved by computer is related to the method by which it is searched.

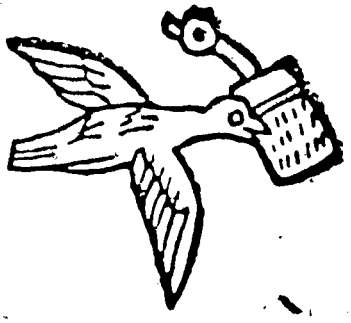
For this project, we need information requests which will be searched in Psychological Abstracts (from July to December 1980). If you need information in the area of psychology or related fields included in Psychological Abstracts, we will conduct a search for you free of charge. All you have to do is submit a search request to us and give us information on how we did after the search.

For the search request, we would like you to describe a topic of interest to you; one you are working on or are familiar with, in the psychology field. Several days later, you will receive a list of citations that have been retrieved by the computer. You will be asked at that time to indicate which of these is pertinent to your interest. One copy of the computer output will be returned to us, and the other copy will be for your own use.

We would very much appreciate your cooperation and participation in this project. Please read the attached pages and write your search request in the space provided, if you are willing to participate.

If you do not need a search, please pass this form to a student or fellow colleague.

JULY 1981



SYRACUSE UNIVERSITY

SCHOOL OF INFORMATION STUDIES

113 EUCLID AVENUE SYRACUSE, NEW YORK 13210 PHONE (315) 423-2911

Phase II

NSF INFORMATION RETRIEVAL PROJECT

As a participant in this project, we would like you to submit a search request (on the attached form) about some aspect of psychology or a related field.

We will take your request and search in Psychological Abstracts (July 1980 - December 1980). The results of this search will be a list of citations to journal articles.

We will then give you this list of citations and ask that you let us know which of these are most pertinent to your search request.

* * * * * * * * * * * * * * * * * *

The attached form is for you to describe your topic of interest. If you are planning a talk or doing a paper, you probably have a topic in mind; if you do not have a topic you are working on, consider one with which you are familiar. Using this form, write down your information requirements as if you were talking to a colleague who understands the field as well as you do.

* * * * * * * * * * * * * * * * * *

Thank you for your cooperation. If you have any questions, please feel free to contact us.

NSF Information Retrieval Project
School of Information Studies
113 Euclid Avenue
Syracuse, New York 13210
(315) 423-4549

JULY 1981

Phase II

NAME: _____ DATE: _____

SCHOOL ADDRESS: _____ PHONE: _____

HOME ADDRESS: _____ PHONE: _____

We would like a description of your topic of interest. This statement should be clear enough so that any person who also knows about this topic would, on the basis of this statement alone, be able to pick out citations of interest for you.

Please write your description here:

Given your purposes in requesting this search, how many citations do you want? _____

About how many citations on your topic do you expect to receive from this computer search? _____

YOU MAY FOLD THIS REQUEST FORM IN THIRDS. STAPLE SECURELY, AND DROP IN CAMPUS MAIL.

APPENDIX D

Phase I

SEARCH QUERY COVER SHEET

Searcher: _____ Search Query Number _____

Date to Searcher: _____ Representation Code this Query: ___

Date to be Returned: _____ DIALOG Password _____

Some Important Notes:

1. Each new query to be searched must be started by the full BEGIN command.
2. You do not need to LOGOFF after each query before starting the next query. You do need to PRINT the documents retrieved before typing the BEGIN command for the new query.
3. Truncation cannot be used with the stemming representation (ST); it can be used with other representations.
4. Though you can use adjacency, you should know that it may run very slowly. Instead, you may choose to use the field operator (F). This implementation of DIALOG will not allow the use of adjacency with truncation, or adjacency with stemming.

To LOGON and LOGOFF

The step-by-step sequence for connecting with the computer, for conducting a DIALOG search, and for disconnecting from the computer is given below.

Everything you type at the terminal must be sent to the computer with a carriage return.

The computer responses to some of these commands are not given here.

---

1. If you are using a dial-up terminal, the phone number is 423-1313. Remember, it must be a hard-copy terminal.
2. Turn power on and hit carriage return.
3. Type: LOG 3434,14
4. Type: NSF
5. Type: DO DIALOG

   The computer will ask for your dialog password. It is given at the top of this page.

Date Returned to Brian McLaughlin: _____ Date Returned to NSF: _____

(5/2/80)

Phase I

SEARCH QUERY COVER SHEET -- Page 2

6. Type: BEGIN

   The computer will ask for the query number and the representation code. Both can be found at the top of Page I.

7. Carry out the search for this query.

   Remember, we want a high recall search with a maximum of 50 documents retrieved.

   Before starting a new query you need to have the set of retrieved documents printed. Use the PRINT command; the format should always be 1.

8. If you want to search another query, look at the COVER SHEET for that query and begin at Step 6.

   If you are completely done searching for now, go to Step 9.

9. Type: LOGOFF

10. Type: K/F

11. Turn power off, collect your materials and submit them to Brian McLaughlin.

Submitting Searches

Brian McLaughlin will distribute and collect all searches. When a search is completed, you need to submit this COVER SHEET and a copy of your interaction. Queries should be searched and returned within 48 hours after receiving them.

Help and Assistance

1. Brian McLaughlin — 476-7359 (Home)
   210 Hubbell Avenue — 423-2091 (Work)
   Syracuse, New York

2. NSF Retrieval Project — 423-4522
   113 Euclid Avenue
   Syracuse, New York

(5/2/80)

Phase II

Appendix D-3

SEARCH QUERY SHEET

Page 1

Searcher_____ Query Number_____

Date Search Collected_____ Order of Representations_____

Date Search to be returned_____ DIALOG Password_____

Date Returned to Brian McLaughlin_____ Date Returned to NSF_____

- -

Some Important Points:

1. Each new search must be started by the full BEGIN command.
2. Be sure to print the documents retrieved before typing the next BEGIN command.
3. If no documents are retrieved, type NOTHING FOUND and print using Format 1, any one document.
4. You do not need to LOGOFF after each search before starting the next search.

TO LOGON AND LOGOFF:

The step-by-step sequence for connecting with the computer, for conducting a DIALOG search, and for disconnecting from the computer, is given below

1. If you are using a dial-up terminal, the phone number is 423-1313.
2. Turn power on and hit carriage return.
3. Type: LOG 3434,14
4. Type: NSF
5. DO DIALOG

   The computer will ask for your dialog password. It is given at the top of this page.
6. Type: BEGIN

   The computer will ask for the query number and will lock the search to a particular representation code.

Appendix D-4

Phase II

SEARCH QUERY SHEET - Page 2

7. Carry out the search for this query.

   Remember, we want a high recall search. Refer to the DIATOM-DIALOG Simulator handout for a description of possible commands.

   Before starting a new search, use the PRINT command, the format should be 1, to have a set of the retrieved documents printed. If no documents have been retrieved, type in NOTHING FOUND and print out any 1 document with FORMAT 1.

8. If you want to conduct another search (for the same query) begin at Step 6.

   If you are completely done searching for now, go to Step 9.

9. Type: LOGOFF

10. Type: K/F

11. Return all the materials to Brian McLaughlin.

HELP AND ASSISTANCE:

1.	Brian McLaughlin 210 Hubbell Avenue Syracuse, New York	476-7359 423-2091	(Home) (Work)
2.	NSF Retrieval Project 113 Euclid Avenue Syracuse, New York	423-4549	(Room 304) or (Room 306)

APPENDIX E

Phase I

14 LS T

SQUARE 1

	101	102	103	104	105	106	107
EDWA	DD	AA	TA	DI	ST	TT	II
VAUG	ST	II	AA	DD	TT	TA	DI
MINO	DI	TA	TT	II	DD	ST	AA
SETT	TA	DD	DI	TT	AA	II	ST
LAUB	AA	ST	DD	TA	II	DI	TT
MCLA	II	TT	ST	AA	DI	DD	TA
ABBO	TT	DI	II	ST	TA	AA	DD

SQUARE 2

	108	109	110	111	112	113	114
EDWA	II	DD	ST	DI	AA	TA	TT
VAUG	AA	DI	DD	II	TA	TT	ST
MINO	DI	ST	TT	DD	II	AA	TA
SETT	DD	TT	TA	ST	DI	II	AA
LAUB	TT	AA	II	TA	ST	DD	DI
MCLA	ST	TA	AA	TT	DD	DI	II
ABBO	TA	II	DI	AA	TT	ST	DD

SQUARE 3

	115	116	117	118	119	120	121
EDWA	DD	ST	DI	AA	TT	II	TA
VAUG	AA	II	TA	ST	DI	TT	DD
MINO	ST	TT	DD	II	TA	DI	AA
SETT	TT	TA	ST	DI	AA	DD	II
LAUB	TA	AA	TT	DD	II	ST	DI
MCLA	II	DI	AA	TT	DD	TA	ST
ABBO	DI	DD	II	TA	ST	AA	TT

SQUARE 4

	122	123	124	125	126	127	128
EDWA	TA	ST	II	TT	DI	AA	DD
VAUG	DD	II	TT	DI	TA	ST	AA
MINO	DI	AA	ST	II	TT	DD	TA
SETT	AA	TT	DI	TA	DD	II	ST
LAUB	II	TA	DD	AA	ST	DI	TT
MCLA	TT	DD	AA	ST	II	TA	DI
ABBO	ST	DI	TA	DD	AA	TT	II

Phase I

SQUARE 5

	129	130	131	132	133	134	135
EDWA	DI	II	TA	DD	AA	TT	ST
VAUG	TT	ST	DI	TA	DD	II	AA
MINO	II	AA	TT	DI	TA	ST	DD
SETT	ST	DD	II	TT	DI	AA	TA
LAUB	TA	TT	DD	AA	ST	DI	II
MCLA	DD	DI	AA	ST	II	TA	TT
ABBO	AA	TA	ST	II	TT	DD	DI

SQUARE 6

	136	137	138	139	140	141	142
EDWA	TT	TA	ST	DI	II	AA	DD
VAUG	ST	TT	DD	II	AA	TA	DI
MINO	AA	II	TA	ST	DD	DI	TT
SETT	TA	AA	TT	DD	DI	II	ST
LAUB	DI	DD	II	TA	TT	ST	AA
MCLA	DD	ST	DI	AA	TA	TT	II
ABBO	II	DI	AA	TT	ST	DD	TA

SQUARE 7

	143	144	145	146	147	148	149
EDWA	TA	TT	ST	II	DI	AA	DD
VAUG	DD	DI	II	TT	TA	ST	AA
MINO	DI	II	AA	ST	TT	DD	TA
SETT	AA	TA	TT	DI	DD	II	ST
LAUB	II	AA	TA	DD	ST	DI	TT
MCLA	ST	DD	DI	TA	AA	TT	II
ABBO	TT	ST	DD	AA	II	TA	DI

SQUARE 8

	150	151	152	153	154	155	156
EDWA	II	TT	DD	AA	TA	DI	ST
VAUG	DD	AA	TT	DI	II	ST	TA
MINO	TA	DD	II	TT	ST	AA	DI
SETT	ST	II	TA	DD	DI	TT	AA
LAUB	DI	TA	ST	II	AA	DD	TT
MCLA	AA	ST	DI	TA	TT	II	DD
ABBO	TT	DI	AA	ST	DD	TA	II

Phase I

SQUARE 9

	157	158	159	160	161	162	163
EDWA	AA	ST	II	DI	TA	TT	DD
VAUG	TT	DI	TA	AA	ST	DD	II
MINO	ST	II	TT	TA	DD	DI	AA
SETT	II	TT	DI	DD	AA	TA	ST
LAUB	DD	AA	ST	TT	DI	II	TA
MCLA	DI	TA	DD	ST	II	AA	TT
ABBO	TA	DD	AA	II	TT	ST	DI

SQUARE 10

	164	165	166	167	168	169	170
EDWA	AA	TT	DI	ST	DD	II	TA
VAUG	DI	AA	ST	TA	TT	DD	II
MINO	TT	DD	AA	DI	II	TA	ST
SETT	ST	DI	TA	II	AA	TT	DD
LAUB	DD	II	TT	AA	TA	ST	DI
MCLA	TA	ST	II	DD	DI	AA	TT
ABBO	II	TA	DD	TT	ST	DI	AA

SQUARE 11

	171	172	173	174	175	176	177
EDWA	TT	ST	DI	II	AA	TA	DD
VAUG	ST	DD	II	AA	TA	TT	DI
MINO	II	AA	TT	ST	DD	DI	TA
SETT	AA	TA	ST	DD	DI	II	TT
LAUB	DD	DI	AA	TA	TT	ST	II
MCLA	TA	TT	DD	DI	II	AA	ST
ABBO	DI	II	TA	TT	ST	DD	AA

SQUARE 12

	178	179	180	181	182	183	184
EDWA	AA	TT	TA	DI	DD	II	ST
VAUG	DI	AA	II	ST	TT	DD	TA
MINO	TT	DD	ST	AA	II	TA	DI
SETT	DD	II	DI	TT	TA	ST	AA
LAUB	II	TA	AA	DD	ST	DI	TT
MCLA	ST	DI	DD	TA	AA	TT	II
ABBO	TA	ST	TT	II	DI	AA	DD

Phase I

SQUARE 13

	185	186	187	188	189	190	191
EDWA	TA	II	TT	AA	ST	DI	DD
VAUG	DD	TT	DI	ST	II	TA	AA
MINO	AA	DI	TA	II	TT	DD	ST
SETT	ST	TA	DD	TT	DI	AA	II
LAUB	II	DD	AA	DI	TA	ST	TT
MCLA	DI	ST	II	DD	AA	TT	TA
ABBO	TT	AA	ST	TA	DD	II	DI

SQUARE 14

	192	193	194	195	196	197	198
EDWA	TT	DD	AA	DI	ST	TA	II
VAUG	DD	II	TT	AA	DI	ST	TA
MINO	DI	AA	ST	TA	II	DD	TT
SETT	II	TA	DD	TT	AA	DI	ST
LAUB	AA	TT	DI	ST	TA	II	DD
MCLA	ST	DI	TA	II	DD	TT	AA
ABBO	TA	ST	II	DD	TT	AA	DI

Phase II

Random Query Order

201 LAUB D I T A
201 MCLA D I T A
201 MINO A T D I
201 STOR A D I T

202 LAUB A I T D
202 MCLA D A I T
202 MINO D A I T
202 STOR I T A D

203 LAUB A I T D
203 MCLA T D I A
203 MINO D I A T
203 STOR T D I A

204 LAUB I D T A
204 MCLA T D A I
204 MINO A T I D
204 STOR D I A T

205 LAUB D I A T
205 MCLA I A D T
205 MINO A D T I
205 STOR D T I A

206 LAUB D T I A
206 MCLA I T D A
206 MINO A T D I
206 STOR A T I D

207 LAUB D A T I
207 MCLA I D T A
207 MINO I A D T
207 STOR T D A I

208 LAUB D A T I
208 MCLA D A T I
208 MINO D A I T
208 STOR T A I D

209 LAUB D A T I
209 MCLA D I T A
209 MINO A T D I
209 STOR T D A I

210 LAUB D I T A
210 MCLA I A D T
210 MINO D T I A
210 STOR T I A D

211 LAUB T I A D
211 MCLA A D T I
211 MINO T A D I
211 STOR I D T A

212 LAUB I D T A
212 MCLA T D I A
212 MINO D A I T
212 STOR A I D T

213 LAUB A I D T
213 MCLA I T A D
213 MINO D A T I
213 STOR A D T I

214 LAUB D A T I
214 MCLA A D T I
214 MINO T A D I
214 STOR T A I D

215 LAUB A I T D
215 MCLA A I D T
215 MINO A I D T
215 STOR I T A D

216 LAUB I D T A
216 MCLA A D T I
216 MINO A D I T
216 STOR D A T I

217 LAUB A T D I
217 MCLA A D I T
217 MINO I T A D
217 STOR A D I T

218 LAUB D I A T
218 MCLA T I D A
218 MINO T A D I
218 STOR A I T D

219 LAUB I T A D
219 MCLA I D A T
219 MINO D I T A
219 STOR D T A I

220 LAUB D I T A
220 MCLA D I A T
220 MINO A I T D
220 STOR T I A D

221 LAUB T D A I
221 MCLA D A I T
221 MINO I T A D
221 STOR A I D T

222 LAUB I A D T
222 MCLA T D A I
222 MINO I D A T
222 STOR T A D I

223 LAUB D T I A
223 MCLA I T A D
223 MINO A T D I
223 STOR A D I T

224 LAUB I T D A
224 MCLA T D I A
224 MINO D I A T
224 STOR I A D T

225 LAUB A T I D
225 MCLA A D T I
225 MINO I T A D
225 STOR A T D I

226 LAUB D A T I
226 MCLA A D I T
226 MINO T I A D
226 STOR D I T A

227 LAUB T A I D
227 MCLA D I T A
227 MINO A I T D
227 STOR T D I A

228 LAUB T D A I
228 MCLA I D A T
228 MINO D I A T
228 STOR A T I D

229 LAUB D I T A
229 MCLA D A I T
229 MINO A T D I
229 STOR I A D T

230 LAUB A D I T
230 MCLA A I T D
230 MINO I A D T
230 STOR A T I D

231 LAUB T I D A
231 MCLA D T A I
231 MINO T I D A
231 STOR T I A D

232 LAUB T A D I
232 MCLA I D T A
232 MINO I T A D
232 STOR I T A D

233 LAUB A D T I
233 MCLA A T D I
233 MINO D A T I
233 STOR T D I A

234 LAUB I T D A
234 MCLA T I D A
234 MINO A D T I
234 STOR T I D A

235 LAUB A D I T
235 MCLA D T I A
235 MINO D A I T
235 STOR A I T D

236 LAUB D T A I
236 MCLA I T D A
236 MINO T D A I
236 STOR A T I D

237 LAUB T D A I
237 MCLA A T D I
237 MINO T D I A
237 STOR A T I D

238 LAUB I A D T
238 MCLA I D T A
238 MINO T D I A
238 STOR I D T A

239 LAUB D A I T
239 MCLA T A I D
239 MINO I D T A
239 STOR I D T A

240 LAUB D T I A
240 MCLA A T I D
240 MINO A I T D
240 STOR T D A I

241 LAUB D T A I
241 MCLA D I T A
241 MINO T D A I
241 STOR A T I D

242 LAUB T I A D
242 MCLA D I T A
242 MINO T D A I
242 STOR A D T I

243 LAUB D A T I
243 MCLA D A I T
243 MINO A T I D
243 STOR A D I T

244 LAUB T A I D
244 MCLA D T A I
244 MINO A T I D
244 STOR T A D I

245 LAUB I T D A
245 MCLA T A I D
245 MINO T I D A
245 STOR T I A D

246 LAUB D T I A
246 MCLA I T A D
246 MINO A D T I
246 STOR T A I D

247 LAUB A T I D
247 MCLA T I A D
247 MINO D I A T
247 STOR A T D I

248 LAUB D I A T
248 MCLA I A T D
248 MINO T I A D
248 STOR D A T I

249 LAUB T D A I
249 MCLA A I T D
249 MINO A T D I
249 STOR A T D I

250 LAUB T D A I
250 MCLA T I D A
250 MINO D I T A
250 STOR I T A D

251 LAUB I T A D
251 MCLA D I A T
251 MINO I A T D
251 STOR D T I A

252 LAUB T A D I
252 MCLA D T A I
252 MINO I A D T
252 STOR I D A T

253 LAUB A T D I
253 MCLA A T I D
253 MINO A I T D
253 STOR D A I T

254 LAUB I A T D
254 MCLA I T A D
254 MINO T A D I
254 STOR T A I D

255 LAUB D I T A
255 MCLA A T I D
255 MINO A I D T
255 STOR I T D A

256 LAUB D I A T
256 MCLA T I D A
256 MINO T I A D
256 STOR D A T I

257 LAUB D A T I
257 MCLA A T I D
257 MINO I D A T
257 STOR D T A I

APPENDIX F

Appendix F

Phase I

AOV SUMMARY TABLE: Recall-1

Source	Sum of Squares	df	Mean Square	F
Between Squares	2.624	11	.239	
Queries in Squares	10.415	58	.180	
Searchers	4.072	6	.679	
Squares X Searcher	7.940	66	.120	
Representations	1.415	6	.236	3.324*
Square X Representation	6.021	66	.091	1.282**
Residual (by subtraction)	19.714	276	.071	
Total	52.201	489		

*Region of rejection begins at 2.14 (α =.05) or 2.89 (α =.01)

**Region of rejection begins at 1.12 (α =.25). Since obtained value falls within the region of rejection, the square X representation source of variation is not pooled into the residual.

NOTE 1: Tukey's HSD region of rejection = 4.17
standard error - .0318

NOTE 2: Missing values in the data (14 queries retrieved no highly relevant documents) required a least squares solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed.

Appendix F

Phase I

AOV SUMMARY TABLE: Recall-2

Source	Sum of Squares	df	Mean Square	F
Squares	.963	11	.088	
Queries in Squares	5.678	65	.087	
Searchers	4.088	6	.681	
Squares X Searchers	4.842	66	.073	
Representations	1.032	6	.172	3.44*
Pooled Error (by subtraction)	19.038	384	.050	
Total	35.641	538		

*Region of rejection begins at 2.14 (α =.05) or 2.89 (α =.01)

NOTE 1: Tukey's HSD region of rejection = 4.17
standard error = .0255

NOTE 2: Missing values in the data (7 queries retrieved no relevant documents at all) required a least squares solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed.

Appendix F

Phase I

AOV SUMMARY TABLE: Precision-1

Sources	Sum of Squares	df	Mean Square	F
Squares	3.536	11	.321	
Queries in Squares*	15.066	72	.209	
Searchers	0.528	6	.088	
Squares X Searchers	3.740	66	.057	
Representations	0.219	6	.0365	.829 (n.s.)
Pooled Error (by subtraction)	15.829	360	.044	
Total	38.918	521		

*Missing values in the data (66 cases with documents retrieved) required a least squares solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed which results in more than one value for the Queries in Squares sum of squares. The value given above is the smaller of the two values, which led to a slightly larger value for the Error sum of squares. The approach is conservative in the sense that if the effect of representations were to be significant, it would also be significant if the other value for the Queries in Squares sum of squares were used.

Appendix F

Phase I

AOV SUMMARY TABLE: Precision-2

Sources	Sum of Squares	df	Mean Square	F
Squares	5.489	11	.499	
Queries in Squares*	19.886	72	.276	
Searchers	0.691	6	.115	
Squares X Searchers	5.348	66	.081	
Representation	0.364	6	.0607	1.05 (n.s.)
Pooled Error (by subtraction)	20.788	360	.0577	
Total	52.566	521		

*Missing values in the data (66 cases with no documents retrieved) required a least squares solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed which resulted in more than one value for the Queries in Squares sum of squares. The value given above is the smaller of the two values, which led to a slightly larger value for the Error sum of squares. The approach is conservative in the sense that if the effect of representations were to be significant, it would also be significant if the other value for the Queries in Squares sum of squares were used.

Appendix F

Phase I

AOV SUMMARY TABLE: Total-Retrieved

Sources	Sums of Squares	df	Mean Square	
Between Squares	10688.347	11	971.668	
Queries in Squares	40273.878	72	559.359	
Searchers	19316.177	6	3219.363	
Squares X Searchers	13719.415	66	270.870	
Representations	3654.511	6	609.085	4.24*
Residual	61236.183	426	143.747	
Total	148888.511	587		

*Region of rejection begins at 2.14 (α =.05) or 2.89 (α =.01)

NOTE: Tukey's HSD region of rejection = 4.17
standard error - 1.308

APPENDIX G

Appendix G

Phase II

AOV SUMMARY TABLE: Recall-1

Source	Sum of Squares	df	Mean Square	F
Searcher	0.652	3	0.217	3.91**
Representation	0.868	3	0.289	5.20**
Searcher X Representation	0.101	9	0.011	0.20
Within Cell	38.535	693	0.056	

*attached to an F statistic indicates that the probability of obtaining that value by chance alone is less than 5%.

**attached to an F statistic indicates that the probability of obtaining that value by chance alone is less than 1%.

NOTE 1: Analysis of variance of the Phase II data was preceded by a multivariate test of all five dependent variables. Any observation that was "missing" on one or more of these variables was automatically eliminated for all five of the variables. Consequently, the degrees of freedom for the Analysis of Variance Summary tables are based on the remaining observations. The Tables of Means (Table 6 and 8), however, are based on the number of observations remaining after the missing values were eliminated from that variable only.

Appendix G

Phase II

AOV SUMMARY TABLE: Recall-2

Source	Sum of Squares	df	Mean Square	F
Searcher	0.628	3	0.209	6.92**
Representation	0.778	3	0.259	8.57**
Searcher X Representation	0.153	9	0.017	0.56
Within Cell	20.952	693	0.030	

*attached to an F statistic indicates that the probability of obtaining that value by chance alone is less than 5%.

**attached to an F statistic indicates that the probability of obtaining that value by chance alone is less than 1%.

NOTE 1: Analysis of variance of the Phase II data was preceded by a multivariate test of all five dependent variables. Any observation that was "missing" on one or more of these variables was automatically eliminated for all five of the variables. Consequently, the degrees of freedom for the Analysis of Variance Summary tables are based on the remaining observations. The Tables of Means (Tables 6 and 8), however, are based on the number of observations remaining after the missing values were eliminated from that variable only.

NOTE 2: Using Tukey's HSD procedure for the PsychAbs data base results, the region of rejection (α =.05) begins at 3.63. The standard error and the minimal difference that would be significant between any two representation means are 0.013 and 0.047.

Appendix G

Phase II

AOV SUMMARY TABLE: Precision-1

Source	Sum of Squares	df	Mean Square	F
Searcher	0.216	3	0.072	0.86
Representation	0.417	3	0.139	1.66
Searcher X Representation	0.198	9	0.022	0.26
Within Cell	58.128	693	0.084	

*attached to an F statistic indicates that the probability of obtaining that value by chance alone is less than 5%.

**attached to an F statistic indicates that the probability of obtaining that value by chance alone is less than 1%.

NOTE 1: Analysis of variance of the Phase II data was preceded by a multivariate test of all five dependent variables. Any observation that was "missing" on one or more of these variables was automatically eliminated for all five of the variables. Consequently, the degrees of freedom for the Analysis of Variance Summary tables are based on the remaining observations. The Tables of Means (Table 6 and 8), however, are based on the number of observations remaining after the missing values were eliminated from that variable only.

Appendix G

Phase II

AOV SUMMARY TABLE: Precision-2

Source	Sum of Squares	df	Mean Square	F
Searcher	0.337	3	0.112	1.19
Representation	1.670	3	0.557	5.91**
Searcher X Representation	0.289	9	0.032	0.34
Within Cell	65.250	693	0.094	

*attached to an F statistic indicates that the probability of obtaining that value by chance alone is less than 5%.

**attached to an F statistic indicates that the probability of obtaining that value by chance alone is less than 1%.

NOTE 1: Analysis of variance of the Phase II data was preceded by a multivariate test of all five dependent variables. Any observation that was "missing" on one or more of these variables was automatically eliminated for all five of the variables. Consequently, the degrees of freedom for the Analysis of Variance Summary tables are based on the remaining observations. The Tables of Means (Tables 6 and 8), however, are based on the number of observations remaining after the missing values were eliminated from that variable only.

NOTE2: Using Tukey's HSD procedure for the PsychAbs data base results, the region of rejection ($\alpha = .05$) begins at 3.63. The standard error and the minimal difference that would be significant between any two representation means are 0.023 and 0.084.

Appendix G

Phase II

AOV SUMMARY TABLE: Total-Retrieved

Source	Sum of Squares	df	Mean Square	F
Searcher	6379.012	3	2126.337	9.54**
Representation	8673.786	3	2891.262	12.98**
Searcher X Representation	4463.481	9	495.942	2.23*
Within Cell	154393.334	693	222.790	

*attached to an F statistic indicates that the probability of obtaining that value by chance alone is less than 5%.

**attached to an F statistic indicates that the probability of obtaining that value by chance alone is less than 1%.

NOTE 1: Analysis of variance of the Phase II data was preceded by a multivariate test of all five dependent variables. Any observation that was "missing" on one or more of these variables was automatically eliminated for all five of the variables. Consequently, the degrees of freedom for the Analysis of Variance Summary tables are based on the remaining observations. The Tables of Means (Tables 6 and 8), however, are based on the number of observations remaining after the missing values were eliminated from that variable only.

NOTE 2: Using Tukey's HSD procedure for the PsychAbs data base results, the region of rejection (=.05) begins at 3.63. The standard error and the minimal difference that would be significant between any two representation means are 0.023 and 0.084.

APPENDIX H

Appendix H

1. Proof that $r_{123...n}$ is a product of the r_i's.

Let d be a relevant retrieved document, R_i is the i^{th} representation and r_i is the recall achieved by that representation. Then,

$$r_{123...n} = \text{Prob(d is retrieved by at least one of the } R_i\text{)}$$

$$= 1 - \text{Prob(d is not retrieved by any of the } R_i\text{)}$$

$$= 1 - \prod_{i=1}^{n} \text{Prob(d is not retrieved by } R_i\text{)}^*$$

$$= 1 - \prod_{i=1}^{n} (1 - \text{Prob(d is retrieved by } R_i\text{))}$$

$$= 1 - \prod_{i=1}^{n} (1 - r_i)$$

*NOTE: This step depends upon the independence assumption. See section VII-C of this report.

Appendix H

2. Proof that asymmetric overlap equals r_2 under the independence assumption.

For R_1 and R_2,

$$A_{12} = \frac{n[R_1 \cap R_2]}{n[R_1]} = \frac{n[R_1] + n[R_2] - n[R_1 \cup R_2]}{n[R_1]}$$

$$= \frac{r_1 + r_2 - r_{12}}{r_1}$$

$$= \frac{r_1 + r_2 - 1 + (1 - r_1)(1 - r_2)}{r_1}$$

$$= r_2$$

NOTE: r_{12} is recall obtained by relevant documents retrieved by either R_1 or R_2.