

Report DDC-TR-70-4

AD 716 200

THE FUTURE OF INDEXING AND RETRIEVAL VOCABULARIES

PAUL H. KLINGBIEL

Directorate of Development

November 1970

DEFENSE DOCUMENTATION CENTER

Defense Supply Agency

Cameron Station

Alexandria, Virginia 22314

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
Springfield, Va. 22151

This document has been approved for public
release and sale; its distribution is unlimited.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)		2a. REPORT SECURITY CLASSIFICATION	
DEFENSE DOCUMENTATION CENTER		UNCLASSIFIED	
		2b. GROUP N/A	
3. REPORT TITLE THE FUTURE OF INDEXING AND RETRIEVAL VOCABULARIES			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates)			
5. AUTHOR(S) (First name, middle initial, last name) PAUL H. KLINGBIEL			
6. REPORT DATE NOVEMBER 1970		7a. TOTAL NO. OF PAGES 31	7b. NO. OF REFS 12
8a. CONTRACT OR GRANT NO.		9a. ORIGINATOR'S REPORT NUMBER(S) DDC-TR-70-4	
b. PROJECT NO.		9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	
c.			
d.			
10. DISTRIBUTION STATEMENT This document has been approved for public release and sale; its distribution is unlimited			
11. SUPPLEMENTARY NOTES		12. SPONSORING MILITARY ACTIVITY	
13. ABSTRACT The role of formal, controlled vocabularies for indexing and retrieval is contrasted with the use of natural language for these activities. The following credo is advanced for large central information processors as appropriate to the 1970's.			
a. Highly structured controlled vocabularies are obsolete for indexing and retrieval.			
b. The natural language of scientific prose is fully adequate for indexing and retrieval.			
c. Machine-aided indexing of natural language is within the state of the art.			
d. Natural language retrieval can be conducted on line if the request can be stated in a phrase or a sentence.			

DD FORM 1473
1 NOV 68

UNCLASSIFIED
Security Classification

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Controlled vocabularies Natural Language Machine-aided indexing Machine-aided searching						

UNCLASSIFIED

Security Classification



DEFENSE SUPPLY AGENCY
DEFENSE DOCUMENTATION CENTER
CAMERON STATION
ALEXANDRIA, VIRGINIA 22314

IN REPLY
REFER TO

PREFACE

Language has always been a serious problem in the storage and retrieval of report literature. The indexing of technical reports requires more depth and probably finer discrimination than the indexing of book or magazine literature. The traditional answer to the language problem for the large central processor has been the development and construction of specialized vocabularies.

In 1959 the predecessor organization of the Defense Documentation Center (DDC), the Armed Services Technical Information Agency (ASTIA), published a fourth edition Subject Heading List containing roughly 150,000 entries to control its collection of 200,000 reports. Concurrently, ASTIA moved into an automated operation. In retrospect, that fourth edition authority listing could have been used for indexing and retrieval in an automated mode since the computer was used simply as a large card file processor, and indexing by subject heading automatically provided either a one- or two-level Boolean search. However, at the time it appeared that a new kind of vocabulary was required and in May 1960 ASTIA published the first edition Thesaurus of ASTIA Descriptors. Other thesauri appeared in rapid order; the 1960's were truly the era of formalized, controlled vocabularies.

In contrast, the following credo for large central processors is proposed for the 1970's:

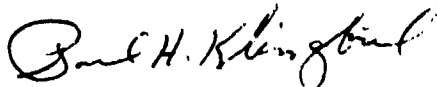
1. Highly structured controlled vocabularies are obsolete for indexing and retrieval.
2. The natural language of scientific prose is fully adequate for indexing and retrieval.
3. Machine-aided indexing of natural language text is within the state of the art.
4. Natural language retrieval can be conducted on line if the request can be stated in a phrase or a sentence.

Several factors have converged to force a reevaluation of the role of language in the transfer of information into and out of very large centralized stores. That reevaluation inescapably leads to the statement of the credo for the 1970's.

This document discusses the factors involved in language reevaluation at DDC and presents evidence which supports the statements of the credo.

Prepared By:

Approved By:


PAUL H. KLINGBIEL
Directorate of Development



HERMAN W. MILES
Director, Directorate of
Development

TABLE OF CONTENTS

	<u>PAGE</u>
INTRODUCTION.....	1
CONTROLLED VOCABULARIES	
Indexing Vocabularies.....	4
Retrieval Vocabularies.....	6
Natural Language.....	7
LANGUAGE DEVELOPMENT AT DDC	
Machine-Aided Indexing.....	10
The Natural Language Data Base.....	10
Machine-Aided Searching.....	11
PRINTED, FORMAL VOCABULARIES.....	11
REFERENCES.....	13
APPENDIX - LANGUAGE EVALUATION	
ASLIB-Cranfield I.....	14
Western Reserve University.....	15
ASLIB-Cranfield II.....	17
Medlars.....	18
Summary.....	21
References.....	22

INTRODUCTION

The following is an extract from the DDC Five-Year Development Program:

"Traditional terminological practices of central processors present barriers to direct information access by the information consumer in the areas of indexing, storage, and retrieval. Indexing tags are artificially constrained to conform with authority lists produced, maintained, and applied by trained specialists. Computer files to store the tags are structured by systems analysts and programmers concerned primarily with ease and economy of computer manipulation and maintenance. Direct retrieval by the consumer is impeded by both artificial indexing languages and practices and the inaccessibility of unknown file structures."

There are at least three factors which motivated that statement: (1) the size and growth rate of the DDC retrieval files; (2) the on-line environment, which presents special problems to users who are not professional bibliographers; and (3) the realization that machine-aided indexing and search formulation can be utilized on a routine basis in an operational environment.

Figure 1 illustrates the growth of the DDC inverted search file for the technical report literature. The file is reflected back to 1953 and shows only the subject postings. The caps on the last three bars show the postings attributable to subject identifiers and open-ended terms. That information is available only from 1968. The solid bars indicate descriptor postings. The lines emerging from the left of the graph indicate the percentage of the data base which is covered for each year by descriptors with postings of less than 5,000, 1,000, and 500 accessions. Percentages are based on the descriptor data base alone.

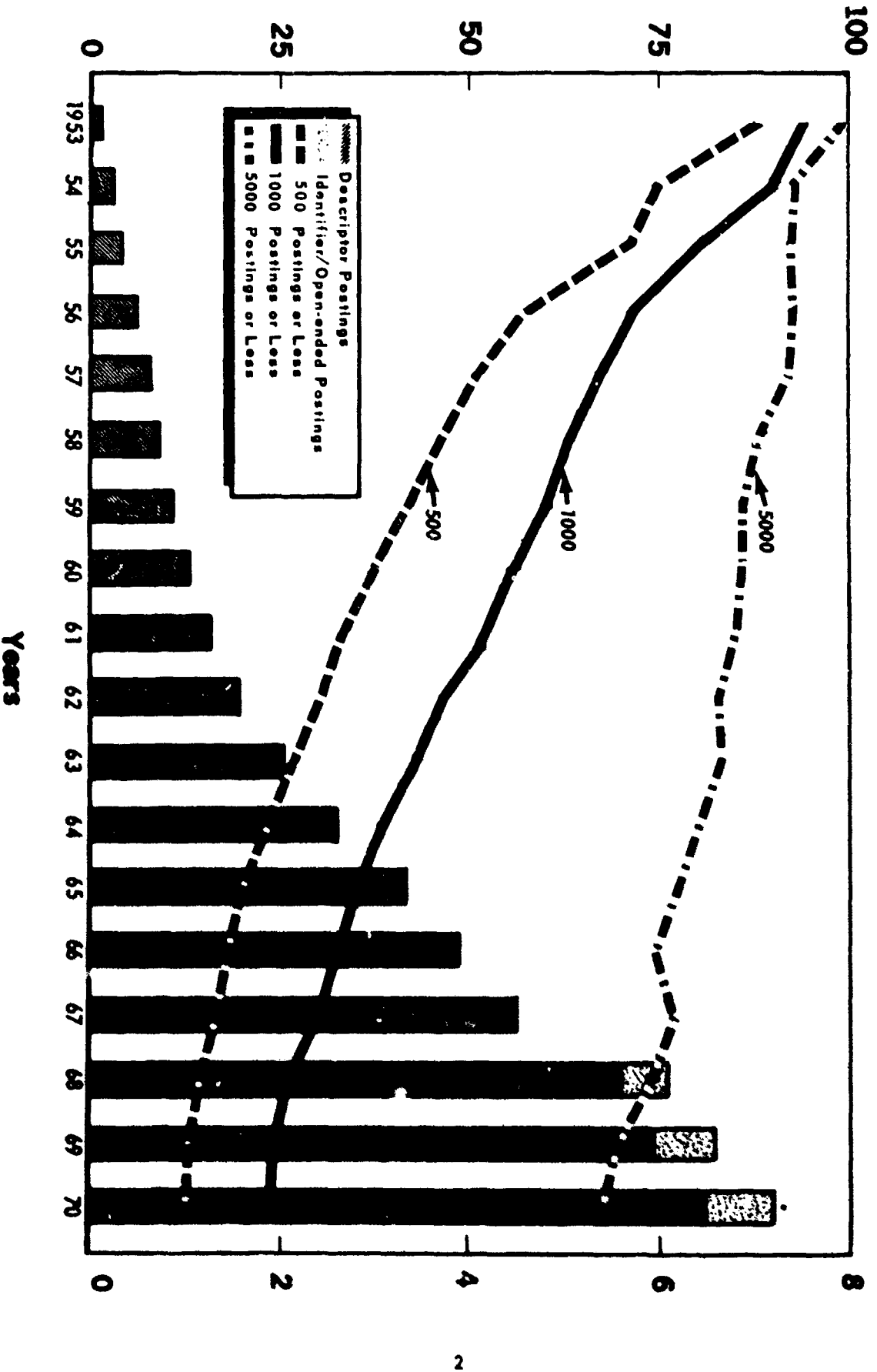
As this figure shows, the inverted search file has doubled about every four years and has passed the 7,000,000 mark. Since manually prepared vocabularies suffer a high rate of obsolescence - requiring updating every two or three years - the burden in time and money of converting files from one authorized language version to another has become absolutely prohibitive for the very large processors such as DDC.

Even if the time and cost of the computer process involved in the conversion were not prohibitive, the conversion is always partial. Very high postings cannot be converted except by manual reindexing. But no manager has the manpower to reindex a term with 50,000 or more postings, not even by attrition. DDC has two such terms. There are 56 terms with 10,000 or more postings. No conversion, as normally practiced, can modify these terms. What is a nonprofessional bibliographer to do when accessing the file on line only to discover that the terms he wants to use fall into this category?

PERCENTAGE OF
TOTAL POSTINGS

INVERTED SEARCH FILE GROWTH

TOTAL POSTINGS
MILLIONS



* As of 31 Oct 1970

On-line access cannot be limited to the use of information specialists (although the information specialist can best capitalize on the advantages of this new environment). All users require vocabulary displays as an adjunct to search statistics and as an aid in devising iterative search strategies. Internalizing present thesauri is not an adequate solution because, apart from the conversion difficulties mentioned above, (1) formal vocabularies are almost always constructed with a bias toward indexing requirements to the detriment of retrieval capability, and (2) formal vocabularies contain many devices intended to enhance precision and recall which are of questionable utility (see appendix, p 18).

Machine-aided indexing and machine-aided search formulation are both within the state of the art, with machine-aided indexing (MAI) the more advanced. MAI utilizes the natural language of whatever text is indexed while machine-aided searching (MAS) utilizes the natural language of the requester.

Since the central retrieval file must be highly controlled, an interface is needed between the raw indexing data and the actual posting point as well as between the natural language of the request and the search file. DDC is constructing such an interface under a project called the Natural Language Data Base (NLDB). This interface will not be biased towards either indexing or retrieval, and it is open ended for the user while maintaining tight control for the central processor.

Finally, managers are faced with the problem of the compatibility of on-line displays and printed, formal vocabularies. The need for on-line vocabulary displays need not be argued. But, what is the need for formal vocabularies? If needed, can they be derived as spin-offs from the various dictionaries internalized in an on-line mode? The appendix discusses very briefly both laboratory and operational tests devised to determine the value and utility of different forms of formal vocabularies. From these tests I conclude that:

1. Tests are extremely difficult to set up in terms of isolating and evaluating the many different factors involved.
2. The human use of language rather than the language itself is the major source of search failure as tested by recall and precision (relevance).
3. Only a 21-point spread exists between the best and worst vocabulary forms as measured by normalized recall, a single number devised from the Cranfield II test to compare vocabularies.
4. Within the 21-point spread, recall improves with vocabulary simplification. Conversely, the use of formalized, controlled vocabularies such as thesauri degrades performance when compared with the use of unstructured, less formalized vocabularies.

DDC is moving to implement the proposed four-point credo. The following chapters address themselves to each of its points in order. Indexing and

retrieval vocabularies will be discussed in order to provide some technical terminology and as a stage for a discussion of the DDC development activities in MAI, MAS, and the NLDB. Printed, formal vocabularies will then be discussed again, and a short, management-oriented survey of some major language tests is given in the appendix.

CONTROLLED VOCABULARIES

Highly structured controlled vocabularies are obsolete. Why? What constitutes control? Why was control thought to be necessary? What is a structured vocabulary, and who needs it? To answer these questions certain basic factors about indexing and retrieval vocabularies must be discussed.

Indexing Vocabularies

Cleverdon^{1/} lists three components which are required of all indexing vocabularies: a set of lead-in terms, a set of code terms, and a set of index terms. These designations are not mutually exclusive. They are, however, convenient topics about which to center a discussion of indexing vocabularies.

To use Cleverdon's definitions:

"A lead-in term represents a concept which is described by another term than itself. This may represent a synonym, e.g., speed use velocity, or may be a subordination of a specific term to a more general term, e.g., Hexagonal use shape.

"Code terms are those terms which are actually used in indexing, ... (e.g., Velocity, shape).

"Index terms are all Code terms, and additionally any combinations of code terms which make up or express new concepts. For instance, the Index term Peripheral speed (can be) expressed by the use of the two code terms Rotation and Velocity (and would appear in the vocabulary as Peripheral speed use Rotation and Velocity)."

The Thesaurus of DDC Descriptors contains 1,146 lead-in terms, 7,342 code terms, and 7,978 index terms. The Thesaurus of Engineering and Scientific Terms (TEST) contains 6,102 lead-in terms, 17,810 code terms, and 18,358 index terms (all of the code terms plus 548 additional combinations of code terms). However, code terms are lead-in terms which refer to themselves (a definition implicit in Cleverdon's discussion). On that basis, the total lead-in vocabulary for the Thesaurus of DDC Descriptors is 8,488 and that for TEST is 23,912.

With these statistics in mind, the following additional remarks by Cleverdon are of interest (underlining mine):

"While these three types of terms, i.e., lead-in terms, index terms and code terms, are normal ingredients of an index language, most index languages also make use of auxiliary devices or aids. In a completely simple system, lead-in terms would always be the index terms and the code terms, which is to say that terms would be used exactly as they appeared in the literature. As soon as the set of index terms is fewer in number than the set of lead-in terms, then a measure of control has been introduced. This normally takes the form of combining terms which are synonyms, and is only the first of many devices which are used in various ways to make up different index languages. There is nothing exclusive about such devices which restrict their use to any particular type of index language, precoordinate or postcoordinate, alphabetical or classified, any type of index language can potentially be given the same devices and thereby have the operational performance of any other index language."

If one takes the ratio of lead-in terms to code terms (the ratio will be 1.00 or larger), the greater the ratio, the more controlled the vocabulary. For TEST that ratio is 1.3; for the Thesaurus of DDC Descriptors it is 1.15; and for the total DDC indexing vocabulary the ratio is 1.0077. The latter figure indicates that DDC is presently operating under virtually an uncontrolled vocabulary condition.

The following remarks by Lancaster^{2/} are also appropriate in this context:

"The size and characteristics of the code vocabulary have very little influence on the performance of a retrieval system as measured by recall and precision,* although the number of code terms to be manipulated in the index will certainly affect searching times and costs. The index term vocabulary and the entry vocabulary...greatly influence the recall and precision performance of a retrieval system.

"The importance, to the retrieval function, of a rich entry vocabulary is generally overlooked. Ideally, an entry vocabulary should contain all words and phrases used in input documents to express items of subject matter that have been recognized in the conceptual analysis stage of indexing. The entry vocabulary will refer to the code terms used to express this subject matter. Words and phrases occurring in requests, and that can be translated into the system's code terms, should also appear in the entry vocabulary."

*Authors differ in terminology usage. The terms "precision" and "relevance" should be considered synonymous in this discussion.

To an indexer, lead-in terms are part of a control. He cannot index "speed." Whenever that concept is required to specify the subject content of a report, the indexer is constrained to use the code term "velocity." The lead-in term is a specifically designated mapping from a textual term to a term authorized as an index term.

Lead-in terms as useful devices to a bibliographer or user of a system will be discussed under retrieval vocabularies, and the relationship between indexing and retrieval vocabularies and natural language will be discussed under the topic of natural language.

Retrieval Vocabularies

The vocabulary used in stating a request in natural scientific prose is just as large as the vocabulary used by an author in producing a technical report or a work statement. But "equally large" does not imply identity. The prose of a technical report or work statement probably represents a more formal mode of writing than that of the written individual request. Differences in style imply differences in vocabulary.

In any event, a formalized retrieval vocabulary does not exist. The information user as well as the professional bibliographer must consider request words as entry terms to the formalized indexing vocabulary. If the entry term of a request matches exactly any of the three kinds of indexing terms (index terms, code terms, or entry terms) the obvious surface conditions for retrieval have been met.

Consider the case where an exact match between a request and the indexing vocabulary does not exist. Suppose the requester is interested in the topic of "ripple tanks." The term exists in neither TEST (1967) nor the Thesaurus of DDC Descriptors (1966). Probably the closest indexing term in the authorized DDC vocabulary is "Hydrodynamics." The broad term, "Hydrodynamics," has been used to index over 3,650 documents. To search all of them for those few concerned with "ripple tanks" requires more time than the professional bibliographer can afford. To send over 3,650 references to the requester is almost sure to invite the response, "low relevance." However, the DDC bibliographers actually have a vocabulary of 155,000 searchable terms. Moreover, this vocabulary has been listed and is available for reference. Within this wider vocabulary "ripple tanks" occurs as an open-ended term with a frequency of five.

Several factors should be mentioned. The term "ripple tanks" is available only in a listing classified "Secret," hence that listing is available only in-house. The term can be searched by anyone who has access to the terminals. However, since no listing is available in hard copy or on the scope, a user must utilize such terms blindly. Should such a user key in "ripple tank," he would draw a blank with no indication that the term is available in the plural form.

Given the initial search, a second search might be formulated as:

(Ocean waves or electromagnetic waves)

and

(Refraction or simulation or model tests)

based on the terms which are most pertinent as co-occurrences with "ripple tanks."

Considering the language needs of the bibliographer or searcher, one can describe a retrieval language as consisting of: lead-in terms, which represent concepts which are searched for by terms other than the requester's terms; code terms, which are requester terms identical to terms with postings; and search patterns, which are composed of code terms or combinations of code terms. (Search patterns are the analog of the indexer's use references and are limited to simple two-valued "and" and "or" statements. They are of necessity based on the known practices of the indexers using the indexing vocabulary.)

Such a formalized retrieval vocabulary does not exist, but it can be constructed. There are two necessary inputs: language extracted from bibliographic requests and the standard indexing vocabulary.

Summarizing, formal vocabularies are deficient in lead-in vocabulary and are almost always constructed with a bias toward indexing requirements. The lead-in problem cannot be solved with printed, formal vocabularies because of the security problem. Since the lead-in problem does not have a solution, one cannot simply say that highly structured controlled vocabularies as presently organized are deficient. Although that statement is true, a highly structured controlled vocabulary is inherently deficient. It is not possible to devise an adequate controlled vocabulary for general consumption. Consequently, highly structured controlled vocabularies are obsolete.

Natural Language

The natural language of scientific prose is fully adequate for indexing and retrieval. In one sense this is a truism. The objection to natural language in contrast to formalized indexing vocabularies is that natural language is somehow too diffuse. Further digging into what "diffuse" means reveals two basic problems: synonymy and lack of structure. Both seem to imply an uncontrolled situation incompatible with efficient storage and retrieval. Ultimately, that seems to imply that a search file utilizing natural language is not feasible.

The synonymy problem in indexing and retrieval refers to semantic identity in only a small number of cases. The issue is really that of quasi-synonymy of words and phrases which are lumped under one index or retrieval term for

convenience. Without attempting an exhaustive listing of types, there is the use of:

1. Singulars and plurals
2. Part-of-speech confounding (attenuated, attenuating, attenuation, attenuator)
3. Class membership (electric circuits use circuits)
4. Phrase substitution (electronic accounting machines use punched card equipment)

A codification such as a thesaurus is seen as a means of imposing order and structure on intractable natural language. Since a thesaurus is manually built (no matter what the computer support) for manual use, it is the end result of many intellectual decisions. Its proper use depends upon the user being aware of those decisions. Consequently, one cannot hope to use natural language via computer if, by that, one means to accept natural language, pass it by an internalized, standard thesaurus, and post on a file of terms defined by that standard thesaurus.

However, the idea of internalizing a standard thesaurus is simplistic. In addition to the deficiencies of formal vocabularies already noted, there are language features needed in an ADP but not in a manual environment, and vice versa. For instance, the usage of singular and plural forms is handled in a manual environment by a statement of rule and the presumption that if the indexer looks for a singular form and does not find it, he should look for the plural form. Obviously some other technique to handle singulars and plurals is needed in an ADP environment. Related terms in a formal thesaurus are meant as a guide for both the indexer and searcher. In an automated environment, related terms are useful only to the searcher. Consequently, they do not have to be tied to specific authorized terms and repeated for each search term. Hierarchy, too, is useful for retrieval, and a hierarchical search facility is required. However, it need not be explicitly tied to a given term. In short, the familiar, standardized thesaurus entry of index term, use reference, broader term, narrower term and related term is not what is internalized for automation.

What is internalized is a set of terms associated with postings, a variety of lead-in devices (words, grammars, and algorithms), search options including hierarchy and truncation, and vocabulary displays, created at the time they are needed and tailor-made to fit a particular need. The implementation of natural language indexing and searching under these circumstances is just the implementation of the natural language data base concept.

The question of file size is invariably raised in connection with natural language. The question is natural and the answer seems almost automatic. The files will be huge. This conclusion seems to be inevitable because of the richness of natural language.

Indeed the files will be large, but not unreasonably so. The use of natural language for indexing and retrieval will require a file larger than the standard inverted file used for formal vocabularies, but smaller than the file constructed for full-text systems.

Our experience to date in indexing over 500,000 words of text indicates fewer unique words than predicted by Kucera and Francis.^{3/} Because of the known frequency distribution of words - 40 percent of the unique words in a corpus appear only once - many terms need not be carried in a dictionary at all. That is, below a certain frequency of occurrence it is more economical to let certain words occur on an error listing for manual review than it is to carry them in a dictionary for suppression.

The statistics of word combinations or syntactic formats are of even more importance. What would, at first thought, appear to be an indefinitely large number of formats reduces to less than 100 distinct types. Several factors contribute to that surprising result. First, stretches of text longer than 5 words are generally not useful for either indexing or searching. They tend to be overly specific. The next factor is the number of parts of speech involved. Essentially these are adjectives and nouns. The conjunctions "and" and "or", the preposition "of", and the adjective "other" play a role but only in the sense that they signal combinations which can be transformed to word combinations that do not contain these words (there are a few exceptions). Even with two or three classes of adjectives and the same number of classes of nouns, the number of sensible combinations is severely limited in comparison with the mathematical combinatorial possibilities. Finally, the syntactic formats themselves seem to follow a log normal distribution, although this must be further investigated.

Consequently, the idea that unreasonably large files are required to utilize natural language is largely illusory. Standard formal vocabularies also contain about 100 different syntactic formats. One can conclude, therefore, that file size over and above that required for formal vocabularies is attributable to lead-in requirements, and since not all lead-in requirements involve the actual storage of lexical files, but can be handled by program, file size stays within manageable magnitudes. It is quite within today's state of the art to say that the natural language of scientific prose is fully adequate for indexing and retrieval.

LANGUAGE DEVELOPMENT AT DDC

Language development activities at DDC are designed to use the natural language of the text for indexing (MAI), the natural language of the user for searching (MAS), and an interface between natural language and the controlled posting points within a search file (NLDB).

Machine-Aided Indexing (MAI)

The machine-aided indexing effort is an attempt to automate and standardize the indexing effort to the degree that the state of the art will allow. The indexing programs read whatever text is available in machinable form. For technical reports this is limited to titles and abstracts, but could be extended to the Technical Report Condensation (TRC) document as well. For the work unit and similar systems (DD 1498 and DD 1634) the text consists of the title and the fields designated objective, approach, and progress.

The indexing programs require a recognition dictionary of individual words classified by part of speech. As a small dictionary, less than 10,000 words, the individual items are stored alphabetically. As the dictionary grows, other than alphabetical storage may prove more efficient. The dictionary can be printed out in a variety of formats, including an alphabetical sequence. However, the dictionary as such favors no prescribed form.

Text words are stored until they match dictionary words of indexing value. The cumulation is stopped by certain kinds of punctuation, or the occurrence of a nonindexable word. The cumulated words are then matched against a dictionary of less than 100 syntactic formats. Matches are passed on as candidate index terms, mismatches are printed out for human review. A report describing the logic of the system has been published. 4/

Over 500,000 words of text have been indexed. The system can become operational on an efficient basis as soon as a prototype natural language data base has been implemented. Machine-aided indexing of natural text is within the state of the art.

The Natural Language Data Base (NLDB)

This data base or file, which is being built in-house, is the real interface between both the indexers or user and the computer postings which control the DDC accession as to technical content or subject matter. When this file is of operational size, it will contain no less than 250,000 lead-in terms, 50,000 of which may be code terms. This is a control ratio of 5, which greatly exceeds the control ratio of current authority lists. Its vocabulary items will be taken from natural language text through the machine-aided indexing programs and from an in-house vocabulary study covering the bibliography requests to DDC for a 6-month interval.

The natural language data base will contain other elements besides lead-in and code terms. Possible additional features are suffix tables, which would be used in connection with stemming routines; grammars to recognize AN nomenclatures without the necessity for storing individual AN numbers as lead-in vocabulary; and grammars to recognize other standard data such as (possibly) personal authors, contracts, or alphanumeric combinations such as F-111.

A prototype is under construction which will interface with about 1400 documents from the DD 1634 data base. Programs have been written which will provide printouts of the vocabulary portions either as unstructured alphabetical lists or as more structured vocabularies as soon as the NLDB has been expanded to operational size in the sense that it spans a data base as large as the DD 1498 file (40,000 to 50,000 documents).

Machine-Aided Searching (MAS)

This system consists of a set of computer programs being written under contract by Bolt, Beranek, and Newman of Cambridge, Mass. The programs are being written in LISP 1.5, a higher level language adapted to text processing. These programs will utilize a large dictionary of single words classified by part of speech. It will differ from the dictionary used by the indexing programs in that words will be allowed multiple parts of speech, and words will be carried which are not useful for indexing - verbs, for instance. Whether or not this dictionary can be merged with the indexing recognition dictionary as a multi (computer) word record cannot be determined at the moment.

The dictionary will act as a recognition device for the syntactical analysis (parser) portion of the preprocessor. Each word in a bibliographic request must be looked up and its possible parts of speech noted so that a computer parse of the request can be attempted. Semantic information may be required as an added entry to some words as an aid in choosing among multiple parsings. The content and format of this dictionary will preclude its use to anyone except the professional staff needed to maintain, modify, or enlarge the dictionary.

A successful parse of the request will be converted to a Boolean statement utilizing the key words in the requester's language. Such key words will be matched against the natural language data base. Successful matches will result in search results displayed for the requester on the CRT. Unsuccessful matches will result in diagnostic statements to the requester with instructions as to how to proceed. These diagnostics may include suitable vocabulary displays. Interactive natural language retrieval for one-sentence requests can be conducted on line.

PRINTED, FORMAL VOCABULARIES

Formal vocabularies in the form of Subject Heading Lists have been a standard tool of librarians for a long time. When DDC automated its subject catalogue in 1958, the Subject Heading list was torn apart, reassembled with additional structure, and called a thesaurus. Both vocabulary formats were designed, built, and used by professionals. The formalism that makes the vocabularies attractive to professional users impedes their use by the casual user.

Individual users want vocabularies for quite different needs. A vocabulary structured for and by a specialist is not the best vocabulary for the manager. Consequently, no vocabulary (alpha list, thesaurus, Universal Decimal, etc.) can fit the needs of everyone.

The thrust of vocabulary development in DDC is to eliminate the need for a printed vocabulary by a user of the system. The machine-aided indexing system indexes from text in the natural language of the text. As an operational subsystem, the recognition dictionary will require maintenance. Primarily, such maintenance will involve the addition of new technical terms as they appear in text. The personnel who maintain the file could be on line, in which case no printed dictionary would be required. If the maintenance is not on line, dictionaries can be printed by part of speech, in alpha order, or in any other sequence desired.

Similarly, the dictionary associated with the natural language preprocessor (MAS) programs will require maintenance. The contractor is currently maintaining it on line. DDC can continue that practice or provide listings for the professionals who will work with that dictionary and its associated grammar. A significant difference between this and the indexing vocabulary is the source of new terms. The preprocessor dictionary acquires new terms on the basis of bibliographic requests rather than from text. This will be the first significant vocabulary built expressly to reflect the user's viewpoint.

The natural language data base represents the intersection of the above subsystems and reflects both the indexing vocabulary and the searching vocabulary. Its most visible characteristic will be the large number of lead-in terms in comparison with code (posting) terms. That ratio will not be less than 5 and potentially can be much greater. This ratio is a numerical indicator of the freedom from printed vocabularies provided the user while maintaining complete control of the actual posting points. This vocabulary can be printed out in any desired format for whatever need exists. However, except for the professional maintenance personnel, the need for such printouts may not exist.

The existence of a natural language data base creates a climate in which a printed vocabulary becomes less needed and less useful as the data base becomes more complete and comprehensive. But that is a natural and desirable spin-off from the language development efforts. The goals of automating the indexing and searching functions, together with the use of natural language, result in a situation where neither the contributor nor the user needs a formal vocabulary - their natural scientific prose is the vocabulary. The system is completely open to even the nonprofessional while the central processor maintains the carefully controlled posting points required for a successful information operation.

REFERENCES

1. Cleverdon, Mills, Keen, Factors Determining the Performance of Indexing Systems, Volume 1, Design. ASLIB Cranfield Research Project, Cranfield, England, 1966.
2. Lancaster, F. W., Information Retrieval Systems, John Wiley and Sons, Inc., 1968.
3. Kucera, Henry and Francis, W. Nelson, Computational Analysis of Present Day American English, Brown University Press, 1967.
4. Klingbiel, Paul H., Machine-Aided Indexing, June 1969, Defense Documentation Center, AD-696 200.

APPENDIX - LANGUAGE EVALUATION

ASLIB - Cranfield I

The idea of evaluating and measuring retrieval effectiveness became a popular issue with the publication of the results of the first ASLIB-Cranfield project in 1960.^{1/} Additional results were reported by Cleverdon^{2/} in 1962. The project was funded in part by the National Science Foundation (\$28,000), extended over a two-year period, employed three full-time indexers, an advisory committee, and 53 organizations which provided individuals to assist with supplementary indexing and question compilation. A collection of 18,000 documents in the subject area of aeronautics was indexed in five index languages: the Universal Decimal Classification, an alphabetical subject catalogue, a faceted classification scheme, and the Uniterm system.

The results of the Cranfield project were controversial (much more so in 1960 than they would be today). A good summary of the points at issue was given by Cleverdon^{3/} in 1965. At that time Cleverdon listed the following conclusions as resulting from the initial study.

1. No significant improvement in indexing is likely beyond an indexing time of four minutes.
2. Trained indexers are able to do consistently good indexing although they lack subject knowledge.
3. Indications are that information retrieval systems are operating normally at a recall ratio between 70 and 90 percent and in the range of 8-20 percent precision.
4. There is an optimum level of exhaustivity of indexing. To index beyond this limit will do little to improve recall ratio but will seriously weaken the precision ratio.
5. There is an inevitable inverse relationship between recall and precision.
6. Within the normal operating range of a system, a 1 percent improvement in relevance will result in a 3 percent drop in recall.
7. The most significant result of the main test program was that all four indexing methods were operating at about the same level of recall performance.
8. The most important factors to be measured in an information-retrieval (IR) system are recall and precision.
9. The physical form of the store has no effect on the efficiency of the system with regard to recall and precision.

10. The index language has a minor effect on the performance of an information-retrieval system. The main influence is the intellectual stage of concept indexing.

11. Given the same concept indexing, any two or more kinds of index languages will be potentially capable of similar performance in regard to recall and precision.

12. The more complex an index language (i.e., the more devices it incorporates), the greater the range of performance in regard to recall and precision.

13. Maximum recall is dependent on exhaustivity of indexing; maximum precision is dependent on the specificity of the index language.

Conclusions 10-13 are directly concerned with vocabulary considerations. Conclusion 12 was to be completely reversed on the basis of the later test (ASLIB-Cranfield II).

Western Reserve University

Cranfield I was a major test in a laboratory environment. One of the chief obstacles raised as to the validity of its conclusions was their extrapolation to an operational system. Consequently, in 1963 Aitchison and Cleverdon^{4/} reported on their collaboration with members of Western Reserve University (WRU) in a test of the operational Index to Metallurgical Literature.

This test involved a collection of 1300 documents each of which had been indexed by a modified version of the English Electric Facet Classification for Engineering and by a highly controlled vocabulary developed at WRU for the express purpose of indexing metallurgical literature. The final report by Aitchison and Cleverdon was reviewed by Rees.^{5/}

Aitchison and Cleverdon had concluded that the WRU test proved the validity of the ASLIB-Cranfield test method and the results could be summarized as follows:

	Recall Ratio	Relevance Ratio
WRU	75.8	17.7
Cranfield	69.5	33.7

Additionally, Aitchison and Cleverdon concluded that the WRU indexing was too exhaustive and that the complexity of the WRU indexing was not worth its cost. In contrast Rees states:

"An examination of their performance figures compels one to wonder why the WRU system, with its high power of discrimination provided by exhaustive and specific indexing, thesaural control

and syntactic relationships, did not perform substantially better than the faceted index. If control devices such as role indicators, punctuation, linking, and thesauri do not materially improve performance, then this has serious implications to the development of information retrieval systems, in that it has been assumed that these devices are desirable if not indispensable to the assurance of high relevance and recall.

"The stated purpose of the test was 'the evaluation of the operating efficiency of the (WRU) index, this involving evaluation of the code or index language, and of the intellectual processes of indexing and search programming.' There are several variables involved here, yet they were not separated in the experimental design, with the consequence that 'indexing' and 'index language' were confounded. For example, two of the variables are:

- a. The structure of the index language.
- b. The manner and effectiveness with which the index language is employed.

"As to the latter variable, it is evident that the consistency and effectiveness with which each index language is applied are of some significance since it is possible that the worst index language utilized by the best indexer may tend to approximate or even excel in performance the best index language utilized by the worst indexer....

"In the same manner that a lack of control was applied to the indexing, it is now apparent that all other variables were not held constant. Assuming that an information retrieval system is 'an integrated assembly of components that interact cooperatively to perform a predetermined function for a specific purpose,' it follows that tests of one component of the system, such as indexing, necessitate the neutralizing of the effect of all other component elements. Question analysis in particular was not held constant between the two indexing systems."

This last point, question analysis, was further investigated by WRU and the results were reported by Saracevic⁶ in 1968. The test was based on 600 documents randomly selected from the 1960 volume of the Tropical Disease Bulletin. Each document was indexed in five indexing languages: telegraphic abstract, key words selected by indexers, key words selected by computer, meta-language, and the Tropical Disease Bulletin index. Each index term, for each language, was treated in two

ways: as an independent English term with no specified relation with any other English term, and as a coded entry into a thesaurus where certain term relationships were defined. The performance effect of the thesaurus was then tested as a part of the question analysis procedure by enlarging the question terminology without the use of the thesaurus.

It is sufficient, I think, to state, in part, the conclusions rather than to cite the specific numerical data upon which the conclusions are based. These conclusions are:

"The handling of questions, methods of analysis and construction of logical statements for searching becomes increasingly important in the generation of IR systems using natural language with uncontrolled terminology, on-line systems, and/or stored dictionaries because there the 'secret' of the system performance lies overwhelmingly in the skillful handling of the questions.

"Enumeration of all related relevant terms by which an asked concept could be expressed is an elaborate, tedious, costly, time-consuming, but unmistakably necessary job. It seems that Thesauri, as constructed today, are not powerful tools for term analysis, when the expansion of terms depends exclusively on the related terms as found in a thesaurus."

ASLIB - Cranfield II

Cleverdon, Mills and Keen^{7/} state that:

"the original ASLIB-Cranfield investigation on the efficiency of indexing systems did not, by itself, produce firm answers to what is one of the basic problems in information retrieval, namely the decision as to which index language should be used...it had shown, by the analysis of search failures, that the decisions by the indexers in recognizing significant concepts in the documents were far more important than any variations in the structures of the various index languages."

ASLIB-Cranfield II was specifically designed to test index languages holding all other variables constant. A document collection of 1400 papers in the field of high-speed aerodynamics and aircraft structures was used as a base against 279 questions for which relevance judgements had been obtained from scientists in the field.

A complex indexing process is described from which index languages were created and classified on the basis of various devices presumed to affect recall and precision. Devices which increase recall are the confounding

of word forms, and hierarchical indexing. Devices which increase precision are links, rolls, weighting, and the coordination of terms. Test results in the form of a ranking are given for 33 different languages -- languages which differ from each other by having or not having one or more of the variety of recall and precision devices included in them. These results are shown in figure 1. Rankings for 47 languages are given in figure 2.

The Cranfield authors summarize the results of the charts as follows: 8/

"Within the environment of this test, it is shown that the best performance was obtained with the group of eight languages which used single terms. The group of fifteen indexing languages which were based on concepts gave the worst performance, while a group of six index languages based on the Thesaurus of Engineering Terms of the Engineers Joint Council were intermediary."

It was this test which reversed the idea that the greater the complexity of an indexing language, the greater its power in terms of recall and relevance. Not only are simple vocabularies effective, simple vocabularies are the most effective vocabularies.

Medlars

Lancaster^{9/} spent one year at the National Library of Medicine in an attempt to evaluate that system's operating characteristics. Lancaster was associated with the ASLIB-Cranfield II project and is probably the most knowledgeable practitioner of the evaluation of information retrieval systems in the United States. With a staff of about 20 people, Lancaster evaluated 302 searchers in terms of 797 recall failures and 3038 precision failures. He attributed those failures to the following principal system components: 10/

<u>Component</u>	<u>Recall Failure</u>	<u>Precision Failure</u>
Index Language	10.2%	36%
Indexing	37.4%	12.9%
Searching	35.0%	32.4%
Defective User/System Interaction	25.0%	16.6%
Other	1.4%	2.5%

<u>ORDER</u>	<u>NORMALIZED RECALL</u>	<u>INDEXING LANGUAGE</u>	
1	65.82	I-3	Single terms. Word forms
2	65.23	I-2	Single terms. Synonyms
3	65.00	I-1	Single terms. Natural Language
4	64.47	I-6	Single terms. Synonyms, word forms, quasi-synonyms
5	64.41	I-8	Single terms. Hierarchy second stage
6	64.05	I-7	Single terms. Hierarchy first stage
7	63.05	I-5	Single terms. Synonyms. Quasi-synonyms
7	63.05	II-11	Simple concepts. Hierarchical and alphabetical selection
9	62.88	II-10	Simple concepts. Alphabetical second stage selection
10	61.76	III-1	Controlled terms. Basic terms
10	61.76	III-2	Controlled terms. Narrower terms
12	61.17	I-9	Single terms. Hierarchy third stage
13	60.94	IV-3	Abstracts. Natural Language
14	60.82	IV-4	Abstracts. Word forms
15	60.11	III-3	Controlled terms. Broader terms
16	59.76	IV-2	Titles. Word forms
17	59.70	III-4	Controlled terms. Related terms
18	59.58	III-5	Controlled terms. Narrower and broader terms
19	59.17	III-6	Controlled terms. Narrower, broader and related terms
20	58.94	IV-1	Titles. Natural language
21	57.41	II-15	Simple concepts. Complete combination
22	57.11	II-9	Simple concepts. Alphabetical first stage selection
23	55.88	II-13	Simple concepts. Complete species and superordinate
24	55.76	II-8	Simple concepts. Hierarchical selection
25	55.41	II-12	Simple concepts. Complete species
26	55.05	II-5	Simple concepts. Selected species and superordinate
27	53.88	II-7	Simple concepts. Selected coordinate and collateral
28	53.52	II-3	Simple concepts. Selected species
29	52.47	II-14	Simple concepts. Complete collateral
30	52.05	II-4	Simple concepts. Superordinate
31	51.82	II-6	Simple concepts. Selected coordinate
32	47.41	II-2	Simple concepts. Synonyms
33	44.64	II-1	Simple concepts. Natural language

FIGURE 1 ORDER OF EFFECTIVENESS BASED ON NORMALIZED
RECALL FOR 33 CRANFIELD INDEX LANGUAGES
(AVERAGE OF NUMBERS)

<u>ORDER</u>	<u>NORMALIZED RECALL</u>	<u>INDEXING LANGUAGE</u>
1	65.82	I-3 Single terms. Word forms
2	65.23	I-2 Single terms. Synonyms
2	65.23	S-13 SMART Concon and indexing new qs.
4	65.13	S-9 SMART Abstract and indexing new qs.
5	65.00	I-1 Single terms. Natural language
6	64.94	S-11 SMART Indexing new qs. and f null
7	64.88	S-6 SMART Indexing new qs.
8	64.82	S-14 SMART Concon and indexing f null
9	64.47	I-6 Single terms. Synonyms, word forms, quasi-synonyms
10	64.41	I-8 Single terms. Hierarchy second stage
11	64.05	I-7 Single terms. Hierarchy first stage
12	63.64	S-8 SMART Abstracts and indexing f null
12	63.64	S-12 SMART Indexing new qs. and f null
14	63.05	I-5 Single terms. Synonyms. Quasi-synonyms
14	63.05	II-11 Simple concepts. Hierarchical and alphabetical selection
16	62.94	S-10 SMART Abstracts new qs. and indexing f null
17	62.88	II-10 Simple concepts. Alphabetical second stage selection
18	62.70	S-3 SMART Abstracts new qs.
19	62.41	S-5 SMART Indexing f null
20	61.76	III-1 Controlled terms
21	61.76	III-2 Controlled terms. Narrower terms
23	61.17	I-9 Single terms, Hierarchy third stage
24	61.06	S-2 SMART Abstracts f null
25	60.94	IV-3 Abstracts. Natural language
26	60.82	IV-4 Abstracts Word Forms
27	60.11	III-3 Controlled terms. Broader terms
28	59.76	IV-2 Titles. Word forms
29	59.70	III-4 Controlled terms. Related terms
30	59.58	III-5 Controlled terms. Narrower and broader terms
31	59.17	III-6 Controlled terms. Narrower, broader and related terms
32	58.94	IV-1 Titles. Natural Language
33	58.64	S-1 SMART Abstracts old qs.
34	58.58	S-4 SMART indexing old qs.
35	57.41	II-15 Simple concepts. Complete combination
36	57.11	II-9 Simple concepts. Alphabetical first stage selection
37	55.88	II-13 Simple concepts. Complete species and superordinate
38	55.76	II-8 Simple concepts. Hierarchical selection
39	55.41	II-12 Simple concepts. Complete species
40	55.05	II-5 Simple concepts. Selected species and superordinate
41	53.88	II-7 Simple concepts. Selected coordinate and collateral
42	53.32	II-3 Simple concepts. Selected species
43	52.47	II-14 Simple concepts. Complete collateral
44	52.05	II-4 Simple concepts. Superordinate
45	51.82	II-6 Simple concepts. Selected coordinate
46	47.41	II-2 Simple concepts. Synonyms
47	44.64	II-1 Simple concepts. Natural language

FIGURE 2 ORDER OF EFFECTIVENESS BASED ON NORMALIZED RECALL FOR 33 CRANFIELD AND 14 SMART INDEX LANGUAGES (AVERAGE OF NUMBERS)

Lancaster¹⁰/ comments as follows:

"Indexers and searches can perform only as well as the index language allows...A recall failure due to a lack of a specific term implies that the search topic, or some aspect of it, is not even covered in the system's entry vocabulary. A precision failure due to a lack of a specific term implies that the topic is not uniquely defined by the index term vocabulary."

System failure appears to be attributable more to human performance than to the tools provided the performer. This agrees with the experience at Western Reserve University.

Summary

Four major efforts to compare the effectiveness of vocabularies have been listed. Each of these has been a major effort involving large amounts of manpower, money, and time. If these are valid tests, there is no justification for repeating them. If these tests are held to be suspect in terms of design - which would cast doubt on the validity of the conclusions as stated by the testers - it is equally clear that no simple, economical, small-scale test will clearly demonstrate the fallacy of the previous tests.

A reasonable result which can be inferred from the large-scale investigations is that vocabulary format is not a major consideration in vocabulary effectiveness. Differences attributable to format are small enough so as to be insignificant in comparison with the skill of the practitioner.

REFERENCES

1. Cleverdon, C. W., Report on the first stage of an investigation into the comparative efficiency of indexing systems. Cranfield, 1960.
2. Cleverdon, C. W., Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Cranfield, 1962.
3. Cleverdon, C. W., The Cranfield Hypothesis. *Library Quarterly*, 35, 1965, pp. 121-124.
4. Aitchison, J., and Cleverdon, C. W., Report of a test on the index of metallurgical literature of Western Reserve University. Cranfield, 1963.
5. Rees, A. M., The ASLIB-Cranfield test of the Western Reserve University Indexing System for Metallurgical Literature: A review of the Final Report, *American Documentation*, Volume 16, No. 2, April 1965.
6. Saracevic, T., The Effect of Question Analysis and Searching Strategy on Performance of Retrieval Systems: Selected Results from an experimental study, Center for Documentation and Communication Research, Case Western Reserve University, May 1968.
7. Cleverdon, Mills, Keen, Factors Determining the Performance of Indexing Systems, Volume 1, Design. ASLIB Cranfield Research Project, Cranfield, England, 1966.
8. Cleverdon, C., Keen, M., Factors Determining the Performance of Indexing Systems, Volume 2. Test Results. ASLIB Cranfield Research Project, Cranfield, England, 1966.
9. Lancaster, F. W., Evaluation of the Medlars Demand Search Service, U. S. Department of Health, Education and Welfare, National Library of Medicine, 1968.
10. Lancaster, F. W., *Information Retrieval Systems*, John Wiley and Sons, Inc., 1968.

OTHER DEFENSE DOCUMENTATION CENTER DEVELOPMENT PUBLICATIONS

Blumberg, S. E., "An Interim Progress Report of Computer-Output-Microfilm Activities and Experiences at the Defense Documentation Center," Report DDC-TR-70-2, AD-708 600, July, 1970.

Gordon, R. F., "Microfiche Viewing Equipment," Report DDC-TR-70-1, AD-701 600, March, 1970.

Klingbiel, P. H., "Machine-Aided Indexing," Report DDC-TR-69-1, AD-696 200 June, 1969.

Miles, H. W., "Technical Information - Availability vs Selectivity vs Cost," Report DDC-TR-69-2, AD-713 200, September, 1969.

Skwarski, L. G., "Documentation Concerning the Installation of Terminal Equipment Associated with the Defense Documentation Center Remote On-Line Retrieval System Experiment," Report DDC-TR-70-3, AD-878 000L, October 1970.

Wicker, R. F., Neperud, R. M., Teplitz, A., "Microfiche Storage and Retrieval System Study: Final Report," SDC TM-WD-(L)-355/000/01, AD-710 000, August 1970. Prepared under Contract DANCI5-70-C-0188 by the System Development Corporation, Falls Church, Virginia.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D		
<i>(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)</i>		
1. ORIGINATING ACTIVITY (Corporate author)		2a. REPORT SECURITY CLASSIFICATION
DEFENSE DOCUMENTATION CENTER		UNCLASSIFIED
		2b. GROUP
		N/A
3. REPORT TITLE		
THE FUTURE OF INDEXING AND RETRIEVAL VOCABULARIES		
4. DESCRIPTIVE NOTES (Type of report and inclusive dates)		
5. AUTHOR(S) (First name, middle initial, last name)		
PAUL H. KLINGBIEL		
6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
NOVEMBER 1970	31	12
8a. CONTRACT OR GRANT NO.	8b. ORIGINATOR'S REPORT NUMBER(S)	
	DDC-TR-70-4	
b. PROJECT NO.		
c.	8d. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	
d.		
10. DISTRIBUTION STATEMENT		
This document has been approved for public release and sale; its distribution is unlimited.		
11. SUPPLEMENTARY NOTES		12. SPONSORING MILITARY ACTIVITY
13. ABSTRACT		
<p>>The role of formal, controlled vocabularies for indexing and retrieval is contrasted with the use of natural language for these activities. The following credo is advanced for large central information processors as appropriate to the 1970's.</p> <ul style="list-style-type: none">a. Highly structured controlled vocabularies are obsolete for indexing and retrieval.b. The natural language of scientific prose is fully adequate for indexing and retrieval.c. Machine-aided indexing of natural language is within the state of the art.d. Natural language retrieval can be conducted on line if the request can be stated in a phrase or a sentence.		

DD FORM 1473

UNCLASSIFIED

Security Classification

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Controlled vocabularies Natural Language Machine-aided indexing Machine-aided searching						

UNCLASSIFIED

Security Classification