

UNCLASSIFIED

AD 414713

DEFENSE DOCUMENTATION CENTER

FOR

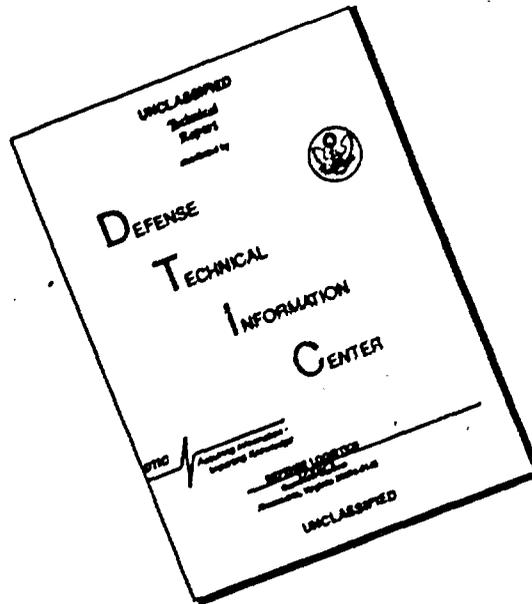
SCIENTIFIC AND TECHNICAL INFORMATION

CAMERON STATION, ALEXANDRIA, VIRGINIA



UNCLASSIFIED

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

AD No. 414713

DDC FILE COPY

414713

*3.60

1

NAVAL COMMAND SYSTEMS SUPPORT ACTIVITY



State

DOCUMENT STORAGE AND RETRIEVAL TECHNIQUES

APPENDIX

INDEXING TECHNIQUES, DESCRIPTION
AND BACKGROUND

APPENDIX TO

NAVCOSSACT REPORT NO. 0024

(PRC D-634A)

Revised

13 June 1963

1245

This Document Consists of 29 Pages.
COPY No. 15 of 29 copies

705 900

NAVAL COMMAND SYSTEMS
SUPPORT ACTIVITY

(13) (6) DOCUMENT STORAGE AND RETRIEVAL TECHNIQUES
APPENDIX

INDEXING TECHNIQUES, DESCRIPTION
AND BACKGROUND,

(21) APPENDIX TO
NAVCOSACT REPORT NO. 0024
1-41 (PRC D-634A)

Revised

(11) 13 June 1963, (12)

Prepared by

Donald V. Black

(13) PLANNING RESEARCH CORPORATION
Under Contract Number Nonr-3665(00)

For

Commander in Chief Pacific

file

FOREWORD

This appendix was produced as a part of PRC D-634 and is intended as background material to that document. In addition, a technique of automatic indexing is described, and that description does not necessarily relate to the content of PRC D-634 as background, but rather as an extension thereof.

TABLE OF CONTENTS

	<u>Page</u>
FOREWORD	iii
I. BACKGROUND.	1
II. NEWER TECHNIQUES.	7
A. The Uniterm and Coordinate Indexing.	7
B. The Descriptor	11
III. PROBLEMS.	13
A. Viewpoint	13
B. Generics.	13
C. Semantics	13
D. Syntactics	13
E. Miscellaneous	14
IV. AUTOMATIC INDEXING AND OTHER CONSIDERATIONS .	15
V. FURTHER EXPERIMENTS.	23
VI. ECONOMICS	27
VII. REFERENCES.	29

LIST OF EXHIBITS

	<u>Page</u>
1. Universe and Array of Coordinate Classes	5
2. Retrieval Results Averaged Over 50 Questions in Nuclear Physics	24

I. BACKGROUND

Up until the beginning of the 1950's there were really only two methods in use in the United States, and in general throughout the rest of the world (with one minor exception), of analyzing the subject content of materials.

One was a system of alphabetic subject cataloging¹ as exemplified, primarily, by the system in use at the Library of Congress and elsewhere in large research libraries, and the system of cataloging with a hierarchical classification system, best exemplified by the Dewey Decimal Classification and its offshoot, the Universal Decimal Classification. Both of these methods of analyzing subject content of printed materials stem from the work of two Massachusetts librarians, Charles A. Cutter and Melvil Dewey. Both systems were first published in 1876. Despite the more general familiarity of a large majority of people in the United States with Dewey's Decimal Classification (because of its widespread use in public libraries), on the basis of current subject cataloging practices in large libraries throughout the world one can hardly question the claim that Cutter's work has had the greater influence on the development of subject cataloging or indexing techniques. While both alphabetic subject cataloging, as exemplified by Cutter's Rules for a Printed Dictionary Catalog (3), and decimal classification, as set forth in Dewey's A Classification and Subject Index for Cataloging and Arranging the Books and Pamphlets of a Library (4), are used for the same purpose--namely, that of arranging materials on the basis of their subject content and presenting a printed representation thereof--there is a completely different philosophy behind the two. Alphabetic subject cataloging attempts to specify, precisely, the topic. In a hierarchical classification system, on the other

¹ Throughout this document the terms indexing and cataloging are used as if they were synonymous, though in fact there are technical differences.

hand, specific entry (i. e., specifying the topic as in alphabetic subject cataloging) may be used, but the goal is to assign the material, from the work at hand, to a class of objects and to fit it into a position with respect to the hierarchical arrangement of that class. For example, under specific entry rules a book about cats would be entered under that term: "Cats." In a hierarchical classification system, however, the work would be entered under Zoology, or Mammals, or Domestic Animals, all of which are classes containing the class Cats. It is difficult in actual practice, of course, to avoid making class entries in alphabetic subject cataloging systems. It seems to be a human tendency to classify items; that is, to attempt to put concepts into logical classes bearing generic relationships one to another. There are other similarities between the two systems. For example, classification systems must always have an index to the generic schedule, which is, typically, alphabetically arranged. This is necessary, because not everyone who uses a given system is aware of the hierarchical structure of all subjects which may be contained in the printed index for the system. There are four phases in alphabetic subject cataloging, which we shall list as follows:

1. Deciding which subject(s) the item (i. e., book, document) is to be entered under.
2. Choosing a name for each subject decided upon in 1.
3. For each name chosen in 2 that consists of more than one word, determining which word is to serve as the entry word.¹
4. Making references, for example, from some subject name synonymous with a name chosen in 2.

The argument for preferring specific entry, then, is a major portion of the idea of alphabetic subject cataloging: any specific subject can be subsumed under a variety of subject classes, depending in part upon the aspect in which it is used; but the user's access to a cataloged item

¹The entry word in an alphabetically arranged subject index or catalog is the first word of a multiword phrase, and is the first word which is considered when placing the phrase in its proper alphabetical sequence.

should not have to depend on whether he can guess correctly the subject class to which the specific subject, in which he is interested, has been assigned. The alphabetic subject catalog has as its purpose, then, to show in one view all the sides of each object. The classed catalog shows together the same side of many objects.¹

There are many problems of a linguistic nature concerned with alphabetic subject cataloging. These have to do with the types of compound subject names, etc., which are so prevalent in the English language. For example, five types of compound subject names can be distinguished:

1. A noun preceded by an adjective (e. g., ferrous metals).
2. A noun preceded by an attributive noun (e. g., death penalty).
3. A noun connected with another by a preposition (e. g., penalty of death).
4. A noun connected with another by "and" (e. g., fun and games).
5. A phrase or sentence (e. g., physics as a profession).

Moreover, in many cases, it is possible to combine compound subject names into one word, for example, "Moral philosophy" to "Ethics," "Social science" to "Sociology." Such reduction is not possible, of course, in all cases, as, for example, "Agricultural chemistry." The current edition of the subject heading list, used by the Library of Congress, is a volume fully as large as Webster's Unabridged Dictionary and it is being updated monthly. One can appreciate, then, the magnitude of the task of devising specific entries for the whole universe of knowledge. It is probable that this great difficulty is what has led to the virtual absence of this type of subject analysis in any operating intelligence library.

¹For example, consider the topic aluminum. In an alphabetic-specific catalog, i. e., an alphabetically arranged subject catalog containing specific entries, the word "aluminum" would be used to indicate the subject aluminum. The term would cover all aspects of aluminum: production, properties, and use. On the other hand, a classed catalog would carry information on aluminum under metals or metallurgy, which would have various subclasses considering, in turn, all metals from various viewpoints, e. g., production, properties, and use; but all metals would be included under each viewpoint, not just aluminum.

This type of entry was used for a long time by the Armed Services Technical Information Agency, but recently ASTIA has changed to a coordinate indexing system. Other major agencies making use of this type of entry, outside of large general research libraries, are the Atomic Energy Commission (in their Nuclear Science Abstracts) and the National Library of Medicine (in its publication Index Medicus).

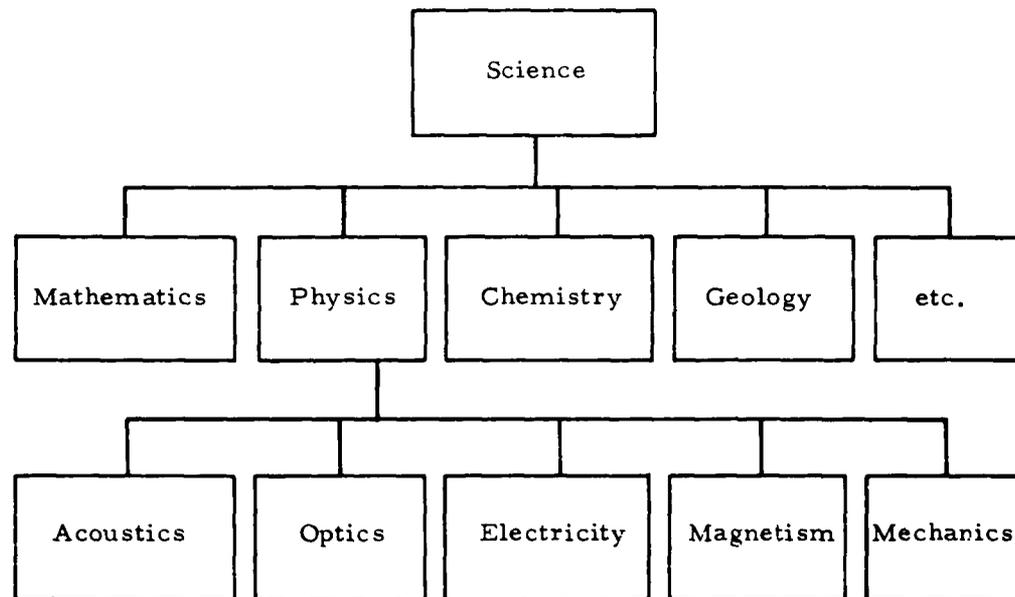
The Intelligence Subject Code, without doubt, stems from Dewey's Decimal Classification. Obviously, the ISC has been created especially for the field of intelligence. Nevertheless, it is obviously based on the principles exemplified in Dewey's work. The basic feature of decimal classification is that it is an enumerative classification; that is, it attempts to enumerate all major classes, minor classes, and subdivisions thereof. Thus, all that is necessary to enter a work in such a classification system is to locate the proper class, or subclass, at the most specific point within the classification schedules. For example, a work on politics is entered under "Political science," which is a subdivision of a main class "Sociology." The primary purpose of a classification, of course, is to arrange items into useful groupings, rather than to individuate the items. Other principles of classification which are important are as follows (see Exhibit 1):

1. One and the same characteristic must be used consistently to derive an array of coordinate classes from a universe.
2. The coordinate classes of a universe should be mutually exclusive.
3. An array of classes must contain an independent and exclusive place for every one of the classes that can be derived from its immediate universe.
4. The classes in an array must totally exhaust the universe from which the array is derived.

A brief example from the Dewey Decimal Classification will show the problems in following these principles of classification.

370	Education
372	Elementary education
373	Secondary education
375	Curriculum
375.514	Mathematical curriculum

EXHIBIT 1 - UNIVERSE AND ARRAY OF COORDINATE CLASSES



Mathematics, physics, chemistry, geology, etc., are an array of coordinate classes formed from the universe "science" above. Acoustics, optics, electricity, magnetism, and mechanics form an array of coordinate classes derived from their universe above, "physics." The example is not necessarily complete, as there are other subclasses of science and physics, which are on the same level of subordination as those given.

As Olney has pointed out (10), the D. C. numbers show that Elementary education, Secondary education, and Curriculum are being treated as an array (i. e., series of coordinate classes) at the first level of subordination from Education. But Dewey here has violated the principle of consistency, because two distinct characteristics have been used to derive these classes from the universe, i. e., Education: the level or class of persons being taught (elementary, secondary) and teaching techniques or goals (curriculum). As often happens, the violation of the principle of consistency here produces a violation of the principle of exclusiveness. As a subject class, Secondary education includes some topics subordinate to Curriculum, and vice versa. There is no clear preference between entering the subclass "Mathematical curriculum for secondary schools" under Secondary education or under Mathematical curriculum. A choice probably must be made,¹ but this kind of arbitrary choice on the part of the classifier leaves the user of the classified catalog with no clue as to where the topic has been placed (unless he is able to find a reference to it in the alphabetic index to the classification), and it tends to undermine his faith in the system.

¹Faceted classification attempts to overcome this difficulty by providing a shifting viewpoint from which to construct an array of coordinate classes. Ranganathan's "Colon Classification"(6) is the most widely known of the faceted classifications, and Ranganathan is the father of this type of classification. Such classifications have received very little use, and there is considerable doubt as to their validity.

II. NEWER TECHNIQUES

A. The Uniterm and Coordinate Indexing¹

About 1950 it was observed, by Mortimer Taube, that in subdividing certain alphabetical subject headings the Library of Congress had, seemingly, departed from certain principles of subject heading theory in such headings as: Liver-Disease. In this usage Disease seems to be a subdivision of Liver, but it is not a form subdivision nor a logical subdivision. Here two separate concepts have been brought together in a form of coordination which appeared to be a subdivision similar to customary subdivision practices in traditional alphabetic library cataloging. In this particular instance, the subdivision might have worked in the other direction: Disease-Liver (and in a general library would undoubtedly have been so designated). Here again Liver is certainly not a subdivision, in any logical sense, of Disease. Why could not other concepts be coordinated also in a similar manner? And so Uniterm, or coordinate indexing, was born. The reader is referred to the many writings of Taube and his associates for a complete description of the Uniterm system of coordinate indexing. Taube was not the first to suggest a coordination of common concepts in this manner, however. Earlier, Batten (2) had done the same thing in a small system for his personal use. In this connection, there are also references to a French system during the early part of the 20th century, but certainly Taube was the one who, almost single-handedly, gave impetus to the Uniterm system and promoted it actively, so that it has become widely used at the present time. The word Uniterm, of course, refers to Taube's original proposal that subject concepts could be broken down into single words, i.e., unit-terms. This idea soon proved unworkable in many fields, since the concepts and terminology used in many subject disciplines frequently involve phrases,

¹For the most comprehensive and recent work on coordinate indexing, see reference 5.

and it is found expedient to use the complete phrase rather than to break it into its constituent parts. Basic philosophies of the Uniterm system seem to be that subject analysis can be performed by less skilled personnel because the words of the author are taken and entered into the system, without hunting for so-called subject concepts or attempting to fit the subject content into a hierarchical scheme of classification. To do the latter involves the necessity for considerable subject knowledge on the part of the indexer. The effort, then, is shifted to the output end (i. e., in retrieval) rather than being expended at the input end. That this is successful in practice, in certain instances, cannot be denied. That it is unsuccessful in certain instances, likewise, cannot be denied.

Perhaps not too many have been aware of the reasons for the success of coordinate indexing. It is not necessarily that the system itself is superior, or that its practitioners are more knowledgeable or diligent or fortunate, but that it uses terminology that is perhaps more familiar to the users of the indexed information. This is by virtue of the fact that the system is using terminology derived from documents which is, presumably, a vocabulary more familiar to the users of these documents. Traditional library systems, in cataloging an item, more frequently take the viewpoint of the author of the item being entered into the system, and change his terminology into the library's terminology, whereas the systems using coordinate indexing frequently use the author's language as it stands, without translation into the arbitrary fixed language of a subject heading system or the language of a classification scheme. The users of the system thus find the language of the Uniterm subject indexing to be more nearly their own language.

It has been pointed out that effective coordinate vocabularies can be developed by any one of three methods, or perhaps combinations thereof: (1) development from scratch, as might occur in a new library or system; (2) development from a previous subject heading list; and (3) conversion of a previously existing classification system. With

the latter two techniques, the terminology is still essentially that of the library or the classification system, rather than that of the documents or that of the user. This has important ramifications in considerations of indexing theory. Some concern has been indicated, by more than one writer, for the question of control of coordinate vocabularies during use. Evidently, some systems have operated with essentially no control. That is, the indexer chooses terms from the documents without regard to previous practices in the system. On the other hand, some systems have a very rigid control, which may, or may not, change the language from that of the documents to that of the library. Arguments have been made that there should be minimum constraints on the language, since the less information is transformed into a rigid structure in the input, the more it will lend itself to a dynamic interpretation in the output phase. It may be that this is a specious argument.

Mere co-occurrence of several terms in a document is a very weak form of coupling. All syntactic relationships are lost, especially where single-word terms are used heavily. Let us illustrate some of the problems. Suppose we have a document (number 1000) on "Glass coatings for steel pipe." If single-word terms are used we have:

Glass
Coating
Steel
Pipe

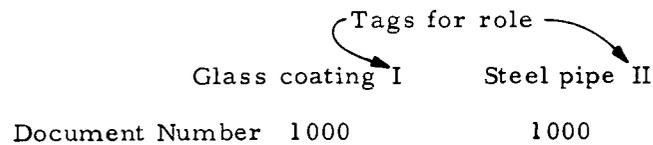
From the four terms we could derive the following combinations:

Glass coating	Glass pipe	Pipe coating
Steel pipe	Steel coating	Pipe glass

Four of the six are clearly erroneous.

While the simplest solution would be to precoordinate the terms, that is, use Glass coating, and Steel pipe as coordinate terms, documentalists are not always as logical as they should be. Instead, systems of roles and links have been developed to solve the problem. It must be confessed that role indicators are necessary (if mere co-occurrence within a document is sufficient to effect retrieval), since even though

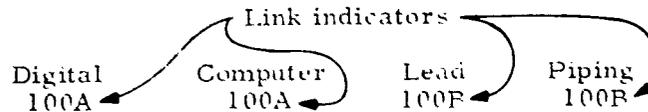
Glass coating eliminates the possibility of pairing Glass and Pipe, there is still no indication that the document concerns "glass as a coating for steel pipe" and not "glass coating" and "steel pipe," as well as a whole host of other items considered as separate products. Thus some tagging system is devised to indicate that Glass is the coating medium referred to in our example document. Such a tag indicates that terms so tagged are members of an ordered pair, and not merely separate terms.



Tags

- I = A material being used in relation to some other material as a covering, wrapping, coating, etc.
- II = A material which is the recipient of some action, process, or material, e.g., forming, extruding, cutting, coating, packaging, etc.

Links become necessary in two circumstances: (1) if the number of single-word terms extracted from a document is very large and contains verbs as well as nouns and adjectives; and (2) if a document considers two separate topics within one set of covers, e.g., digital computers and lead piping. There are enough references in the literature to the first instance, so that we shall not consider the topic further. The second instance may be illustrated as follows:



This would indicate that Digital and Computer go together in one section of document number 100, and that Lead and Piping are associated in another section of document number 100.

B. The Descriptor

Developed at about the same time as Uniterm was Calvin Mooers' "descriptor." (1) Just how descriptors differ from conventional subject headings or Uniterms has not been especially clear, evidently, to many persons. It seems however, that the basic difference between both descriptors and subject headings and descriptors and Uniterms is that descriptors are much more oriented to the viewpoint of a specialized group of users than are either subject headings or Uniterms. The various "Zatocoding"¹ systems, which Mooers has set up, and which used descriptors, were designed in consultation with a relatively small body of users in every instance. The users themselves were the ones who chose the terminology which was to be applied. Descriptors, however, are manipulated in very much the same way as Uniterms except that, in the Zatocoding system, descriptors on cards are manipulated mechanically on a special sorting block. The descriptor system has two main portions. One is an alphabetically arranged listing of descriptors and terms which are not descriptors, with complete scope notes for each descriptor, and cross-references from the terms which are not descriptors to those which are. The meaning of a descriptor is usually much broader than the meaning of a subject heading in traditional library practice. Specificity is achieved by the use of several descriptors in concert. In the scope notes for each descriptor, terms which are frequently used in conjunction with any given descriptor are listed, along with a rather detailed description of each topic in which a given descriptor may be found. The meaning of any descriptor is assigned in such a way that the descriptor will be of maximum use within a specified collection of material. This would mean, then, that only individuals oriented to the specific system will be able to use it with immediate effectiveness. This restriction is quite justified in a specialized collection. The second part of a descriptor system is a large sheet of paper containing the descriptive schedule. In assigning

¹The proprietary system of the Zator Corporation, which uses the "descriptor" technique (see reference 1).

descriptors to documents, an analyst must go through the entire schedule of descriptors. This would seem to be a laborious procedure, yet a very clever scheme of grouping is utilized, with a series of questions at the beginning of each group. The questions are made up to reflect the fields of interest of the organization which is using the system. Thus, it is necessary only to look at the question at the top of each group. If the answer is "yes," the analyst will pick one or more descriptors from that group. If the answer is "no," he continues to the next question. It must be emphasized that the grouping of descriptors is not a classification system. There are no generics, or specific terms. An important point to be seen here is that the specialized living vocabulary of a specific group of users is, by far, more oriented to successful retrieval of information than can be any system using an index vocabulary which is derived from written material, without regard to the specialized viewpoint of those who will be using the materials or those for whose use the system is designed. It is important to recognize the fact that many systems are designed for use by the operators of the system rather than the true ultimate users of the information extracted from the system. The operator is the one who manipulates the system in whatever way it is designed to be manipulated; but the ultimate consumer is not the operator, necessarily.

While peripheral (and perhaps mysterious) reference has been made to Faceted Classification, we have not discussed it. It has had no impact whatsoever on the indexing of intelligence materials, and, in the opinion of the writer, it will have none. It may be, perhaps, fairly well suited to the classification of scientific subjects (for example, physics or chemistry), but it does not lend itself well to nonhomogeneous collections of subject matter; that is, collections of materials of which the subject is not homogeneous. The topic of faceted classification is deeply enmeshed in philosophical discussions of ordered universe, logic, etc., and it does not seem to be germane to our discussion here.

III. PROBLEMS

There are a number of technical problems in the subject analysis of information contained in printed material. The following, perhaps, can serve as a summary of the problems.

A. Viewpoint

Human beings are unable inherently to place any given concept (an idea or thought image of one specific subject or thing) in only one logical class. For example, is a bow a weapon or a piece of sporting equipment? The problem becomes even worse when combinations of concepts are involved.

B. Generics

This problem involves the family tree of concepts, because each concept implies narrower concepts. Literature searching based upon broad concepts should effectively retrieve the information based upon narrower but related concepts. For example, the retrieval of all information pertaining to "publication" should also (and automatically) result in obtaining all information on books, magazines, newsletters, etc. Before the generic problem can be solved, the solution to the viewpoint problem must be in hand, because each class in which a concept is placed has its own family tree which must be considered.

C. Semantics

This involves the relationships between symbols (words) and the concepts themselves; i. e., the relationships between words and their meanings. Not only are we concerned here with definitions, with words, but also with synonyms, near synonyms, homonyms, homographs, etc. Examples of semantically confusing words include: base, color, lead, finish, tank.

D. Syntactics

This problem is concerned with syntax or word order. If syntax is disregarded we may erroneously equate "cooling (of) water" to "cooling (by means of) water."

E. Miscellaneous

1. We tend to overlook the fact that, when we use an alphabetically derived subject heading system, we are in essence classifying by alphabet. Is this less logical than classification based on some supposed natural order or synthetic (e. g., faceted) classification scheme? We have two tools for the English language itself which bear correspondence to the alphabetical subject index and the classified index for collections of materials. The dictionary is an alphabetical index to the language, and the thesaurus, such as Roget's, represents a classification system for certain parts of the language. Is either one more logical than the other? No one questions the usefulness of both. Yet, one of the arguments of the proponents of classification systems is that classification is more logical than is alphabetical indexing.

2. Finally, when one is faced with an existing collection of indexed materials, how does one assess the effectiveness of any retrieval system? Suppose that one receives 20 documents as a result of a query to the system. Suppose further that all 20 documents are quite pertinent to the topic of interest. Is there any way to assess the amount of pertinent information still unretrieved from the file? Or is there any way of learning whether the retrieved information is more pertinent than the unretrieved information? The answer is "No!" The use of any retrieval system is, then, an act of faith in the quality of indexing.

IV. AUTOMATIC INDEXING AND OTHER CONSIDERATIONS

A series of experiments, beginning in about 1959 and continuing up to the present, has gradually built up a technique of utilizing a computer for wholly automatic indexing and retrieval of printed materials. Such a system would, of course, be especially attractive wherever the materials are already available in machine-readable form, e. g. , incoming messages on teletype. Were a computer system available, no human indexers would need to handle the material from the teletype tape; it would be directly input into a computer. Whereas the idea is attractive, there are, of course, a number of characteristics of teletype messages which tend to create problems for automatic data processing. However, in terms of practicality, it seems possible to overcome these. Despite the fact that for regular printed materials no such automatic input is available, a computerized process, such as will be described, would seem to be especially attractive, insofar as the personnel needs of the system are minimal.

Existing intelligence library systems with any claim to efficient operation, as well as accurate subject analysis of incoming materials, are characterized by large numbers of fairly well-trained and experienced personnel who perform indexing and do other administrative or clerical tasks within the system. In many cases the quality of the personnel operating some given system is barely marginal, because of the fairly low-level job classification which has been established for these positions. Good indexers are probably born and not made; that is, there are relatively few people who really have the necessary qualities to perform consistently high-level indexing day after day. Careful training of less adept personnel may produce adequate results. On the other hand, one would feel that only the best personnel obtainable should really be utilized in a system of importance, such as intelligence library systems would seem to be. The fully automatic computer system, however, would need relatively low-level clerical personnel to perform input or output functions which are, in essence, merely the transformation of the printed or written materials into machine readable form; that is, keypunching or typing on

a tape-producing typewriter, of which there are a number of examples (Flexowriter, teletype, etc.). There is considerable evidence that a representation of the printed materials (which representation must be stored in the computer memory to serve as an index to the materials) can be created from such things as tables of contents, summary paragraphs, titles, individual indexes contained within a printed item, etc. Studies have shown that, despite protestations to the contrary, most human indexing, even of a supposed high level, actually is more machine-like in nature than the indexers would care to admit. (9) Justification for any indexing technique must ultimately be based on successful retrieval. Success can only be evaluated in terms of a closed system; that is, a system wherein sufficient knowledge is available of the entire contents of the materials, so that an evaluation can be made of various techniques as to their retrieval effectiveness. The various systems described herein cannot really be weighed except on the basis of a test comparing one against the other. This has not been done in any place. Thus, there are two alternatives: (1) to decide on a system on the basis of its attributes as set forth on paper, weighing the needs of the system for high-level personnel, etc., or (2) actually to perform a series of experiments utilizing various systems, so that it can be shown that one system is clearly superior to another in its retrieval effectiveness, and to settle on that system, despite its other attributes such as personnel needs, equipment costs, etc.

In all systems utilizing human beings as indexers and retrievers, the success with which information can be retrieved depends critically on the ingenuity exercised in formulating a search instruction, as well as the care and effort expended in analyzing the materials as they were entered into the system.

It is well known that human beings do not perform optimally at all times. The level of human performance is quite variable from one period to another. There have been enough experiments to indicate that there is no consistency, or very little, between one indexing performance by a given individual and another indexing performance, at a later date, by the same individual. The same inconsistency has been discovered among different individuals all indexing the same documents.

Thus there is neither inter-indexer consistency nor intra-indexer consistency in any system that depends on human performance. Human performance must be considered as an integral part of that system. Then, no matter how effective or sophisticated a system as a whole may be, its overall performance is no better than the performance of any one link within the system. As long as humans are links in such systems, their performance will have considerable effect on the total system performance. While there are undoubtedly both indexing geniuses and retrieval geniuses, such persons are in short supply and, even so, cannot be said to operate at a constant level of efficiency at all times. A fully automatic system, on the other hand, depends solely on the ingenuity and resourcefulness expended upon the development of the initial system, as well as the continued proper operation of the machinery which is a part of the system. Mechanical and electronic reliability can almost be said to be specified or specifiable beforehand, whereas human reliability cannot be predicted. Reducing the argument, then, to a few considerations: provided that retrieval effectiveness is nearly equal, the system which depends less on the human element would clearly seem to be the more desirable from a reliability and efficiency standpoint, if not necessarily from an economic standpoint. It is perhaps a truism that input is even more important than output, for if there is failure at the output end one can always try again. If there is failure at the input end, however, the material is most likely irretrievably lost. In a human system, performance of the overall system is a direct function of input quality. The quality of input is, in turn, a direct function of four elements concerning personnel:

1. Availability of qualified indexers.
2. The adequacy of the system by which they will index.
3. The quality of their training.
4. Continuing coordination and auditing of their work.

In a completely automatic system only one of the above four items is a factor; that is, the continuing coordination and auditing of the system performance. The other three factors are taken care of, once and for all, at the beginning, during the establishment of the system.

There is no need to outline in great detail here the precise course and results of various experiments which have been made on fully automatic indexing and retrieval. However, certain landmarks in the course of experimentation will be mentioned, inasmuch as they tend to show the relative merit of the concept of automatic indexing.

Early experiments (7) pitted human indexing and retrieval against machine retrieval using material in nuclear physics as an experimental library. In the early experiments retrieval (both manually and by means of the computer) was performed by designing search instructions which would be carried out by various groups of retrievers. Much latitude existed for human ingenuity in transforming the original retrieval question into a search instruction. At the conclusion of the project, search instructions formulated by seven different individuals, all of whom held the Ph. D. degree in nuclear physics, were studied to determine their intellectual content. It was found that many of these search instructions embodied elementary oversights, and few exhibited any significant level of insight or ingenuity. Furthermore, it was determined that the failure to achieve relevant retrieval had little to do with the absence of ingenuity in formulating search requests. It was concluded that the process of translating the original natural language question into a search instruction probably could have been done better by a machine, since the machine could be made to follow a more systematic and consistent procedure.

Let us now state some premises regarding automatic full-text processing for indexing and retrieval.

1. The frequency of occurrence of a natural language term within a document is not necessarily related to its relevance, or lack thereof, for storage or retrieval.
2. For a particular type of collection a thesaurus can be compiled in which groups of nearly equivalent terms can be constructed such that the equivalence of the terms within a group is largely independent of context.
3. These groups can be weighted independently of context to reflect varying degrees of importance for information retrieval.

4. Syntactic relationships can be usefully approximated by proximity; that is, spans of one word to several sentences.
5. Couplings¹ of two or three terms can be weighted to improve automatic processing by a significant amount.

In the various experiments (cited above) carried out in retrieval, it was discovered that the frequency of occurrence of terms in the relevant documents was no different than the frequency of occurrence of terms in irrelevant documents. Thus, premise 1 above seems to be borne out. Premise 2 above depends upon the processing of approximately 1 million running words of text.² Weights are assigned to each of the terms within a thesaurus in consultation with subject specialists. It is predicted that for most systems a vocabulary will be developed of some 20 to 25 thousand terms. This vocabulary will comprise the thesaurus. After the initial thesaurus is created, a series of tests must be performed in consultation with subject specialists. Each test serves to add polish to the thesaurus, in the weightings of the vocabulary within the thesaurus, so that ultimately it becomes a rather precise instrument for automatic retrieval and processing. The premise stated in 4 above was developed when it was discovered that a four-sentence span of proximity was sufficient to eliminate some 60 percent of irrelevant retrieval due to lack of syntactical specification. That is, specifying co-occurrence of words

¹The coupling of terms is a process whereby two or more words are customarily used together in a phrase: e. g., the terms "nuclear," "power," and "submarine" coupled together form the concept "nuclear powered submarine." The three terms might appear separately in a given document without carrying the concept mentioned above, but when they are coupled together they very clearly carry one, and only one, connotation.

²In the experiments which have been performed, only several thousand running words of text have been analyzed, yet, judged by the criterion of retrieval effectiveness, it was possible to develop groups of words which were relatively synonymous even with this small sample of actual vocabulary. With the large sample mentioned above the effectiveness of the system is greatly enhanced.

within a limited span of up to four sentences prevented irrelevant material from being retrieved along with relevant material, and did not decrease relevant retrieval significantly. This has an important further advantage: paragraphs or even smaller units could be retrieved, thus reducing the amount of material that must be reviewed by the user of the material. The reduction is some 90 percent,¹ and this allows, then, a higher tolerance in the final selection process of a higher ratio of irrelevant to relevant retrieval. Premise 5 above seemed to be borne out, in the experiments, when it was discovered that word couplings were responsible, in some cases, for relevant retrieval and helped to prevent a significant amount of irrelevant material. This is readily apparent if we consider that there are many subject concepts in English which cannot be expressed adequately by one word.²

Any retrieval system which makes use of various techniques to make specifications narrower in order to eliminate irrelevant material is subject to the error of overspecification. This error will cause relevant material, to a greater or lesser extent, to be missed. The problem then is not to find a solution which will preclude irrelevant retrieval as well as specifying all relevant materials, for such a solution exists only as a fantasy of overzealous documentalists. Rather, we must strive to produce a system which will minimize irrelevant³ retrieval and maximize relevant retrieval to the greatest extent practical.

¹The figure 90 percent is derived from experience in previous experiments, wherein the amount of relevant material was scanned and a subjective judgment was formed that the relevant material was actually about 10 percent of the total verbiage retrieved. That is, about 10 percent of each document contained the relevant material; 90 percent of the document was of no relevance but the document as a whole was relevant. Even if this figure is off as much as 20 percent the reduction is still significant.

²See examples of this phenomenon in the discussions concerning the Dewey Decimal Classification.

³It is, perhaps, appropriate to mention that there is a technical difference between relevance and pertinence which we have not observed in this paper. For our purposes here the distinction does not seem important.

At this point, let us summarize the outlines of the system. Each word of an item (message, document) is examined by the computer. The computer looks up in the thesaurus every word and contiguous word pair of each sentence.¹ The thesaurus contains a vocabulary of words which are weighted for retrieval importance. In addition the thesaurus contains cross-references,² and all words are grouped into synonym classes. Each item is assigned a "relevance score" which depends upon the number of meaningful terms present and their weights, plus proximity and pairing factors.³ Retrieval is performed by processing a natural language question through the regular input routine. The meaningful terms are weighted, as are proximity and pairing factors. The result⁴ is the

¹In this sentence contiguous word pairs would be, for example, "every word," "contiguous word," "word pair," etc.

²"Cross-reference" in this context means that each word in a group of words would be given the same code number in the computer memory; that is, the terms "politics" and "political science" could both have the same code number, even though they might be located in different positions within the computer memory. The synonymity of terms is decided a priori by humans when initially creating the thesaurus.

³As explained above, proximity is important in reducing irrelevant retrieval. Thus, if two given terms occurred within a span of three sentences they would receive a bonus weight, whereas if they occurred at a span greater than three sentences a penalty would be assessed against the weight which the terms themselves carry. The same process is followed with pairing factors. Terms may have weights as separate entities in the thesaurus, but if they occur together the weight given is not merely the sum of the two individual weights, but may be quite a bit greater.

⁴Relevance scores are important in retrieval, inasmuch as, while keywords occurring in a question may be matched exactly by the keywords contained in a given document, the relevance scores may be different. This is due to proximity and pairing factors which occurred in the document, and in the question to a lesser extent. The divergence of scores indicates that the document does not quite contain the concepts contained in the retrieval question. This may be because the terms which occurred in the document may be scattered throughout the document, and really have little relation one to another, whereas the terms in the question do have relation one to another. Such a document would not be an especially good response to that particular question.

PRC D-634A

22

"relevance score" of the question, which is matched against the "relevance scores" of each item contained in the file. The output consists of a list of items sequenced according to relevance score. It is possible to tag specific paragraphs, or even sentences, which produced the particular relevance score of a given item. This is of potentially great benefit to users, since it could reduce by a large factor the amount of material which must be examined.

V. FURTHER EXPERIMENTS

Using the material gathered together for the experiment in retrieval from nuclear physics material, a new experiment (8) was designed based on the aforesated premises of fully automatic indexing. The computer examined each article in the experimental collection, word by word and phrase by phrase. The score of a given article was calculated by summing the weights of the words in the article which coincided with the words in the search instruction, adding a premium score to take into account occurrence of phrases, and subtracting a penalty factor if the co-occurring terms were separated by more than four sentences. We wish to reiterate that the search instructions were computer-produced. The output consisted of a list of article numbers sequenced according to relevance score. Along with each article number appeared a list of those words in the search instruction which were found within that article. A second list of article numbers was also printed out which indicated for each question the true relevance scores based on human judgment. A comparison with the relevance based on human judgment permits one to determine the percentage of relevant information retrieved for some given acceptable quantity of irrelevant information. This can be stated in another way, by saying that the amount of irrelevant information that must be retrieved, in order to achieve some specified relevance percentage, can be determined in comparison with human performance. Exhibit 2 summarizes the respective performances of human and machine.

In conclusion we should like to suggest an experiment which could be performed to test the relative merits of various systems of indexing. We should like to suggest that this experiment make use of selective dissemination.¹ Dissemination of materials is equivalent to retrieval

¹Dissemination is the process of routing documents, as they are received, to certain individuals. This is a common enough occurrence, whereby certain topics are listed on a routing form along with names of recipients. Documents on specified topics are then routed to the individuals indicated on the routing list. The terms "standing query" and "spot query" have also been used to distinguish dissemination from specific retrieval.

EXHIBIT 2 - RETRIEVAL RESULTS AVERAGED OVER 50 QUESTIONS
IN NUCLEAR PHYSICS

	<u>Man</u>	<u>Machine</u>	<u>Man⁽¹⁾</u>	<u>Machine</u>
Percent Relevant Retrieval	20	20	55	55
Number of Irrelevant Documents	1.6	0.9	12.8	7
Percent of Source ⁽²⁾ Documents Retrieved	88	100		
Number of Irrelevant Documents	12.8	5		

Notes: (1) Highest retrieval achieved by the humans who were involved.

(2) Source documents were those which had elicited the questions (that is, material in the source documents had suggested each question).

of the same materials, with the sole difference being one of time. That is, instead of presenting a retrieval question at the moment in time when a user needs material on a particular topic, the user specifies beforehand a particular topic of interest in precisely the same manner he would use in presenting a retrieval question. Thus, selection of materials for dissemination is done to meet a continuous need, whereas specific retrieval is done to meet a discrete need. The description of the proposed experiment follows.

It is suggested that a teletype data base, consisting of approximately 1,500 messages, be processed by a computer program constructed according to the premises of automatic indexing given above. The data base of messages should be tested in three batches. The first batch of messages would consist of approximately 100 to 200 messages. It would be used to check out the program and to remove obvious errors from the dissemination dictionary (i. e., thesaurus). After studying the first batch, any changes would be tested on a second batch of approximately half of the remaining messages. Any changes produced as a result of the second test would be embodied into a new program, which would then be tested on the remaining messages. By consistently adopting a procedure of testing the program and thesaurus only on new batches of messages not previously studied or examined, the effect of ad hoc rules that correct specific errors, but do not work generally, can thus be eliminated.¹

Results of the automatic dissemination could be checked against human performance. There will be two major categories of machine error, i. e., over-dissemination and tags² missed. Based on previous experiments we would predict that any single instance (in a particular

¹ It is not possible to tell, precisely, when a rule is "ad hoc" until a separate, new batch of data has been processed using the new rule. If errors of the same general type as that which elicited the new rule remain uncorrected in the new batch of data, then the rule was, in fact, "ad hoc," and not a good general rule.

² Tags are simply the thesaurus terms which result in specific dissemination. These tags would be printed at the beginning of a message with their relevance weights summed to indicate the total relevance.

message) of either over-dissemination or under-dissemination would be correctable by means of an ad hoc rule, but without testing such a rule on a large number of messages no true assessment can be made of its effectiveness. Thus we should qualify the term "correctable" with "probably." It will not, however, be necessary to judge the results on an intuitive basis if sufficient care is exercised in the experimental procedure to be followed.

VI. ECONOMICS

The question of whether people should index and disseminate or whether machines should take over this job is somewhat a matter of economics. Since it is a routine, high-volume task, machines are probably more consistent and reliable, particularly if the process we have described is refined and improved to the extent which we believe is feasible. Rough estimates can be made of the cost on the following basis: The speed of the computer program is approximately one message per second and machines of the type programmed¹ can be rented commercially for about 20 cents per second. They can probably be operated for less, but, even so, 20 cents per message is probably right within a factor of two or so. Estimates of the time people spend in the process of indexing and dissemination, particularly if they are to try to do as thorough a job as can be done by machine, would considerably exceed this amount. To determine the true comparative cost figures would require a study of the particular situation being considered for development of a system.

We believe, however, that the matching of the text of a "document" to a question, phrased in natural language, is more accurate with the aid of an automatic thesaurus than is the human process of assigning appropriate subject index terms to a "document" and to a retrieval question.

¹IRM 7090.

VII. REFERENCES

1. Casey, R. S. (ed.). Punched Cards; Their Application to Science and Industry. Second edition. New York: Reinhold, 1958.
2. Collison, R. L. Indexes and Indexing. New York: John deGraff, 1959.
3. Cutter, C. A. "Rules for a Dictionary Catalog," Special Report on Public Libraries, Part II. United States Bureau of Education, fourth edition. Washington: Government Printing Office, 1904.
4. Dewey, M. A Classification and Subject Index for Cataloging and Arranging the Books and Pamphlets of a Library. Amherst, Massachusetts: Dewey, 1876.
5. Jaster, J. J., and others. State of the Art of Coordinate Indexing. Preliminary edition. Washington: Documentation, Inc., 1962.
6. Ranganathan, S. R. Colon Classification. Sixth edition. London: Asia Publishing House, 1960.
7. Swanson, D. R. "Searching Natural Language Texts by Computer," Science, 132:1099, 1960.
8. Swanson, D. R. "Interrogating a Computer in Natural Language." Presented at the Congress of the International Federation of Information Processing Societies, Munich, 1962.
9. Montgomery, C., and D. R. Swanson. "Machine-Like Indexing by People," American Documentation, 13:359, October, 1962.
10. Olney, J. C. Library Cataloging and Classification. Report No. TM-1192. Santa Monica: System Development Corporation, 1963.