DOCUMENTATION

INCORPORATED

WASHINGTON, D. C.

THE STATE OF THE ART OF COORDINATE INDEXING

Prepared for

Office of Science Information Service
National Science Foundation

under

Contract No. NSF-C-147

by

Josephine J. Jaster, Barbara R. Murray
and Mortimer Taube

DOCUMENTATION INCORPORATED
Washington 14, D. C.

February 1962

(Preliminary Edition)

1. INTRODUCTION

1.1 SCOPE OF STUDY

1.2 ARRANGEMENT OF REPORT

1.3 ACKNOWLEDGMENTS

1. INTRODUCTION

1.1 Scope of Study

Many state-of-the-art studies have been criticised because of their narrow definition of the subject covered. In an attempt to avoid this, coordinate indexing in this study has been defined to include all systems in which the logical operation of intersection, union and negation are brought into play in the manipulation of index terms. This therefore covers nearly every mechanized system. It includes both term-on-item and item-on-term systems. It excludes only those systems which are precoordinated, i.e, nonmanipulative. Among the many names given to coordinate indexing are correlative indexing, multiple aspect indexing, concept coordination, and "Enriched Co-ordinate Indexing."

Just as the definition of coordinate indexing was made as broad as possible, so also were subsidiary definitions. For example, whether systems use Uniterms, keywords, descriptors, unit concepts, or structerms, they are included in this study. A distinction is made between pure and constrained systems, but not so as to eliminate the latter from the study.

The same broadness in definition is allowed in the study on the various constraints. For example, in the chapter on roles and links, although there is an attempt to distinguish among the various kinds of

roles and links, none is omitted because it does not fit a particular definition.

In any state-of-the-art study, there is a danger of not having given full credit and not having done justice to all existing work, to say nothing of the inevitability of accidental exclusion. Doubtlessly this study will suffer from all of these defects. It is particularly regretted that the study is limited to the United States.

1.2 Arrangement of Report

There are unavoidable overlaps among the various sections of this report. To maintain completeness within each section, a high degree of redundancy has purposely been introduced and material from other sections is either summarized, repeated, or referenced.

Each section has its own reference-footnotes. However, at the end of the report is an annotated bibliography of the more significant papers that were published between 1947 and 1960.

1.3 Acknowledgments

Documentation Incorporated is pleased to report that it received support, encouragement, and assistance in all respects.

Except for certain proprietary information which was justifiably withheld, other data were generously made available. The Documentation Incorporated study group was made welcome at all facilities visited.

The field of documentation has too often been characterized by close-mouthed secrecy. It is hoped that the cooperation we have

received indicates a future trend and that the report justifies its support.

Although all who cooperated cannot be recognized, special thanks are due to the following individuals for their assistance:

Harold Wooster	(Air Force Office of Scientific Research)
J. Heston Heald	(ASTIA)
Paul Klinefelter	"
Paul Klingbiel	"
Loffie Lee	(CEIR)
Harry Marx	"
C. Dake Gull	(G.E., Bethesda)
Earl Stroup	"
B. K. Dennis	(G.E., Evendale)
James Hubbell	"
Jean MacCauley	"
Saul & Mary Herner	(Herner & Company)
Lynn Bartlett	(Information for Industry Inc.)
J. C. Costello	(Jonker Business Machines, Inc.)
Gloria Smith	"
Joseph Hoffer	(National Conference on Social Welfare)
R. A. Fairthorne	(Western Reserve University)
Allen Kent	"
Jessica Melton	"
Paul Olejar	(U. S. Army Chemical Research & Development Laboratories)
Edward Sullivan	(U. S. Atomic Energy Commission
Julius Frome	(U. S. Patent Office)

Thanks are also due to W.T. Brandhorst of Documentation Incorporated for his assistance in compiling the bibliography, and to Eugene Wall for his contributions to Chapters 4 and 5.

Mrs. Helen Brownson of the National Science Foundation very graciously made available to us certain unpublished material which will be a part of Nonconventional Technical Information Systems in Current

Use, No. 3, which is in preparation. This directed our attention to certain companies whose work in coordinate indexing might otherwise have remained unknown to us.

2. THE LITERATURE OF COORDINATE INDEXING

2. THE LITERATURE OF COORDINATE INDEXING

The materials listed in this survey are selected from a much larger literature which has developed in this field in the past fifteen years. The reader will note certain omissions and will have a right to inquire concerning the basis of selection.

The primary basis for selection was whether or not the document contributed to the state of the art of coordinate indexing.

For example, this is not a study of "machine literature searching". There are many papers on the use of machines which make no contribution to the logic of machine search. Because coordinate indexes are manipulative and are predicated on logical operations, such papers are not included in the bibliography (although many have been read during the course of this study).

Since 1947 (the earliest paper in the bibliography) the number of workers in the field and the number of papers have grown enormously. From an odd and esoteric concern of a few people, the field of I. R. has grown until it has even been noticed by Fortune, which says:

> "A number of I. R. machines have already been devised and built. There are also in use systems that have adapted for I. R. purposes conventional punched-card machines and computers. I.B.M. estimates that industry is now spending about $2 million a year on systems of the latter sort. By 1965, according to I.B.M., the expenditure on I. R. will

jump to over $100 million a year, and thereafter double every three years." *

Any selection process involves a value judgment and the reader, if he wishes, can insist that such evaluations are subjective, private, and lack objective validity. There are, however, certain internal criteria which a good bibliography can use, especially in a field as new and active as this one. As the field has grown, it has been notorious for the controversies it has engendered and for the large number of self-proclaimed "experts" it contains. One basis of objectively evaluating the literature in a field and the experts who produce it is by ascertaining how many of the experts read the works of the other experts. In other words, if bibliography is significant, it must be assumed that in general, a man who writes about any subject without any knowledge of or reference to other writings on the subject is not likely to be making a significant contribution to it. There are always exceptions to such a conclusion. Any field has its small percentage of meteoric personalities who may enter the field glowing like a comet. Out of their own internal genius or knowledge they may provide new insights which are important and effective. But since knowledge in general, and science in particular, is cumulative,

---

* Fortune, September 1960, pp. 162-192.

in most cases it is reasonably safe to disregard the "contribution" which is made either with complete unawareness or complete unconcern for the literature in the field.

In another context, Etienne Gilson described certain innovators as being "in that blessed state of ignorance which makes it easy for a clever man to be original." It is the cream of the jest that in this field, basically concerned with the value of information and the means of retrieving it so that science in general can be an orderly, cumulative process, there is a high percentage of papers which appear without footnotes and without reference to the works of others. The elimination of many such papers from this bibliography has a scientific and objective basis.

3. COORDINATE INDEXING AND CLASSIFICATION

THEORY -- THE DEVELOPMENT OF COORDINATE INDEXING

3. COORDINATE INDEXING AND CLASSIFICATION THEORY -- THE DEVELOPMENT OF COORDINATE INDEXING

3.1 General

In this historical account of coordinate indexing, as in accounts of similar developments, the most difficult task is arriving at some statement of "how it all began." Most intellectual movements do not begin ab ovo but usually have roots which can be traced as far back as anyone has energy to trace them. But the account must begin somewhere, and it will be helpful in determining this place if there is first set forth a definition of coordinate indexing which will distinguish it from other forms, namely, alphabetical indexing and systematic classification, which go back almost to the beginning of intellectual history. Even if coordinate indexing is considered to be only a variant of these two older forms, it remains necessary to trace the beginnings of this variant, if only to determine how it has developed from these beginnings. Hence, the first task is to define coordinate indexing in a manner which distinguishes it quite sharply from alphabetical indexing, on one hand, and systematic hierarchical classification on the other.

3.2 Class Manipulation

One of the essential characteristics of coordinate indexing is that it is manipulative, as opposed to fixed. For example, in a system of alphabetical subject headings or in standard alphabetical indexing, each heading is fixed and complete. Any item in such systems may be

indexed by a plurality of headings; but it is not necessary to "match", "coordinate", or "intersect" headings in a search of the system. (The logic of such intersection or matching will be discussed below.)

Quite early in the development of coordinate indexing it was pointed out that subject headings, like "liver - radiation injuries" and "gamma rays - pathological effects",[1] were each intended to be complete and to provide a fixed and complete mode of access to the item indexed. A system using such headings is essentially based on the assumption that the headings can be as complete as required and that it is never necessary to intersect headings or to use two or more headings to retrieve any indexed item. Similarly, with reference to classification systems, with or without notations, any subclass is defined by all the classes of which it is a subclass; and classification systems do not require, in any instance, that a search be made by looking for identical members of different subclasses. On the other hand, it is the essence of coordinate indexing that a search be made in a system of coordinate indexing by performing operations on the classes defined by the headings in the system. It is because of these operations that the word "manipulative" to characterize systems of coordinate indexing is apt and descriptive. (The credit for this term should be given to Charles Bernier, formerly of Chemical Abstracts and now of ASTIA.)[2]

It is now fairly generally accepted that many different types of

machines can be used to perform the manipulations required in systems of coordinate indexing. On the other hand, in terms of the question of origins, does the use of machinery in coordinate indexing imply that coordinate indexing is traceable back to the first recorded descriptions of the use of punched cards or edge-notched cards for indexing collections of data or information? It may be thought that the answer to this question does not matter too much. One could say that coordinate indexing began when manipulative indexing began, that is, when any device was used, such as a sorting machine or a set of needles, to select items from a file. However, this conclusion would hide the fact that the meaning of "manipulative" is wider than the meaning of "coordinate indexing". What is manipulated in a system of coordinate indexing are classes in order to deliver certain intersections (or other logical functions) of classes. It is possible to manipulate a file for quite other reasons. For example, one might use a sorting machine to organize a file into an alphabetical array; or one might use a set of needles to select from a file all items indexed under any particular class designation. In other words, although coordinate indexing implies the use of some type of mechanism for manipulating classes, the use of a manipulative mechanism does not imply that a system using such a mechanism is a system of coordinate indexing.

3.3 Class Intersection

In many of the early descriptions of edge-notched card systems and punched card systems the advantages claimed for such devices were:

(1) They could eliminate the requirement of maintaining the files in any organized array.

(2) They could eliminate the necessity of multiple cards for items indexed under multiple headings.

If systems are used in connection with machines for these purposes and these purposes alone, they cannot be considered systems of coordinate indexing. For example, Mooers' careful delineation of descriptors and his development of superimposed coding for multiple indexing in restricted coding fields, while important contributions to the art of information retrieval, do not by themselves make Zatocoding into a system of coordinate indexing. Of course, if a Zatocoding system is designed and set up for the purpose of making searches on the logical functions of descriptors, then such a system would be a system of coordinate indexing.

One of the earliest papers to recognize that mechanical devices could be used to intersect classes was a paper presented by Dr. J. E. Holmstrom at the Berne Conference of the International Federation for Documentation in 1947 and reprinted as Paper No. 33 in The Royal Society Scientific Conference.[3] In this paper Holmstrom specifically

distinguishes mechanical indexing from classification and alphabetical indexing in the following terms:

> "Punched card systems have been allotted a separate place in our diagram as they differ fundamentally from any of the above. Requiring no regular filing sequence at all, nor any cross references or multiple entries, they offer, in the following respect, an outstanding advantage over any other system: the possibility of sifting out from an accumulated mass of records, at any time, any desired conjunctions of information..."[4]

Here we see the recognition that punched card equipment can eliminate the necessity for ordered files and the necessity for multiple cards; but there is also the recognition that such devices can be used for the "conjunction" of headings. Holmstrom, in this paper, also recognizes that search by conjunction of headings was also recommended by Batten in a 1947 paper, "A Punched Card System of Indexing to Meet Special Requirements."[5]

Although it is clear that Holmstrom recognized that machines could be used for a single search on a number of headings, he did not emphasize this point because he believed that the relationships among terms were more important than the terms themselves. Thus, he indicated that punched cards could be used for handling material classified according to UDC numbers. Such numbers would preserve the relationships and the relationship of classes to one another and still make it possible to index an item under a plurality of different class headings on one card.

3.4 Boolean Functions

At the Third Plenary Session of the Royal Society Conference, Dr. Holmstrom, in his general remarks, referred to the Berne paper and reiterated the possibility of searching for an item by the two terms, "man" and "dog", but he goes on to say that searching for "man" and "dog" is not as effective as searching for "dog bites man" and he emphasizes that

> "... it is precisely these relations between the concepts, not the concepts themselves, which we want to be able to select from the total mass of literature. Unless we have some means of doing that there is perhaps some danger of falling into the misconception of Mark Twain's over-sanguine young journalist (I think the story is Mark Twain's) who undertook to produce at short notice an article on Chinese philosophy which happened to be urgently required, and when asked how he proposed to do it said the task was quite simple: he would look up 'China' in the encyclopaedia and he would look up 'philosophy' in the encyclopaedia, and then he would have merely to combine his material."[6]

(The story is, in fact, Dickens'.)

It is fair to say that by thus emphasizing relations which may or may not be important in the indexing problem, Holmstrom failed to recognize the logic underlying the pure form of coordinate indexing. Calvin Mooers has called to our attention the fact that the early Hollerith patents of 1884 and 1889 and the Taylor patent of 1950 exhibited a recognition of the application of the logic of classes to search and selection by multiple terms, although they did not use the

logic of classes explicitly in describing the operation of their devices. The earliest explicit recognition of the logic of search which has been found is a paper by Fairthorne, "The Mathematics of Classification."[7]

The work of H.P. Luhn with indexing systems and the Luhn Scanner also began in the late '40's. In a paper prepared in 1948 he concerned himself with retrieval of material by any combination of terms coded on a card and also with the development of free field coding for such terms. In addition to this early work, he made a major contribution to coordinate indexing, namely,"KWIC" indexing, which is discussed in the conclusion of this section.

The work of Documentation Incorporated in this area began in 1952, but this work was in part based on an earlier paper, "The Coordinate Indexing of Scientific Fields," read by Mortimer Taube before the Symposium on Mechanical Aids to Chemical Documentation of the Division of Chemical Literature of the American Chemical Society, September 4, 1951, and a technical report prepared by Irma Wachtel in 1951 for the Atomic Energy Commission (TID-469). The paper on "The Coordinate Indexing of Scientific Fields" was stimulated, in turn, by Gilbert King's paper, "Application of Card Methods to Scientific Computation," especially the following passage:

> "The capabilities of the machines could be learned by examples, but it would be much more satisfactory if the algebra of the machines were worked out. The scientist in deciding how to solve a problem and the detailed programming of number (i.e., cards, through the machines) is not ... interested in the circuitry, and he would only be unnecessarily limited by examples. He needs facts, namely, what can be fed into the machine, what can happen to this information, and what can come out. The machine carries out operations of symbolic

> logic, and for proper coding and programming the scientist should know the basic algebra of the black boxes into which he will put cards. At present, an installation of punched-card machines requires a supervisor who is thoroughly acquainted with what the machines can do and equally at home with the science of the problem. It would be much more efficient if every scientist on the staff knew what the machines can do. This can be accomplished now only by those who learn by experience; this is a long process, and one which requires continual practice to insure efficiency. On the other hand, if the algebra or symbolic logic were formulated, it could be rapidly learned and readily applied."[8]

In the 1951 paper, "The Coordinate Indexing of Scientific Fields," the utility of machines which can manipulate classes of items in order to determine Boolean functions of such classes is made to rest upon a hypothetical statement. This hypothesis was stated as follows:

> "If any given scientific field can be analyzed into simple terms or ideas so that all or most of the concepts of the field can be expressed in these terms or in truth functions* of them (i.e., logical combinations) then mechanical aids can make important contributions to organizing and searching for information in such a field... In order to see more clearly why this is true, we can turn now to a more formal presentation of the nature of coordinate indexing for machines... In a field, Q, with terms a, b, c, ..., etc., coordinate indexing allows only the relations
>
> a
>
> not a
>
> a and b
>
> a or b." [9]

---

* For further development of these ideas, see the distinction between logic and mechanics of retrieval in "Studies in Coordinate Indexing," Vol. III.

> "(In terms of machine searching, if we search for 'a or b' we will get everything that is 'a and not b', everything that is both 'a and b'. The search will reject only those things which are 'not a and not b'.)"[10]

The initial Uniterm experiment, conducted by Documentation Incorporated for ASTIA, was an attempt to determine whether or not a specific body of literature could be indexed by assigning to each document in the collection a subset of a set of terms with the restriction that a search for any document could only be made by a logical operation on the terms of the subset. The hypothesis expressed in the 1951 paper was repeated in a series of lectures given at the University of Chicago in the Spring of 1952.*

Most of the early criticism of the pure system of coordinate indexing was based on the failure to recognize its experimental or hypothetical nature. It was always recognized that the elimination of syntax and the restriction of relations between terms to pure Boolean functions might lead to some degree of noise in the retrieval process. The initial experiments were designed to find out how much noise.

In considering the subsequent history of coordinate indexing, with its introduction of various types of constraints, such as pre-

---

* Since the 1951 paper was unpublished, Dr. Shera, with the knowledge and approval of the author, presented the hypothesis in a paper given at a symposium at Columbia University in 1953. It is interesting to note that in a paper by Jack Morris, published in American Documentation in 1954, which in general is most critical of the Uniterm System, the same hypothesis is stated with complete approval.[11]

coordinations, roles, links, categorization of terms, etc., it is important to realize that in its initial presentation, the theory was presented in the pure form, that is, with maximum freedom among the terms. It was felt that a system with this degree of freedom was necessary to provide compatibility between the differently structured CADO system and the TID system, which were to be combined in the then recently established ASTIA. (Cf. Section 5.5.) The use of Boolean algebra to describe the relationships among the terms in a pure system of coordinate indexing is not accidental, nor even pedantic, as has been supposed by some critics.[12] It was used to emphasize the possibility of constructing an indexing system which would dispense with syntactical relationships of any kind among terms and which would restrict itself to the class operations allowed in a Boolean algebra. For example, even the noncommutative relationship of ordered pairs which can be handled in an extended theory of classes was ruled out of pure coordinate indexing on the grounds that such relationships are outside the domain of a restricted Boolean algebra. Since the theory was presented in a pure form, it followed that all subsequent innovations could not be in the direction of making the system purer, i.e., freer, but in adding constraints of one kind of another. (See Paragraph 3.7, on Luhn's keyword-in-context indexing.)

Before going on to consider the history of attempts to reintroduce various types of constraints into coordinate indexing, it is important to point out the relationship between certain conclusions in information theory and coordinate indexing. It must be stated that although these relationships were not immediately evident during the early days

of the work in coordinate indexing, they were noted subsequently and used to throw light on the advantages of pure coordinate indexing.

3.5 Information Theory

Calvin Mooers and others had noted that certain conclusions of information theory were relevant to coding problems, especially problems of superimposed coding. Formulas can be used to determine the optimum depth of superimposed indexing in limited coding fields on cards which are analogous to formulas used in calculating the amount of information which can be transmitted through limited channels. The relationship between information theory and coordinate indexing noted by Documentation Incorporated was not this relationship between channel capacity and coding, but between questions of maximizing information and the free form of coordinate indexing. In other words, the hypothesis that an indexing system could be based upon Boolean relations of classes is not ad hoc. Rather, it arises from the fact that such a system is in principle the simplest and most economic, providing only that it does not lead to too much noise. An example can illustrate what is meant here: Suppose a coordinate indexing system to consist of a vocabulary of 2,000 terms, and let it be supposed that in using such a system for indexing and searching, it is always discovered that the amount of noise delivered by any search is negligible. It would obviously be redundant to set up in such a system additional terms which consisted of pre-coordinations of terms selected from the basic

vocabulary. It must be noted, however, that as in information theory, if any constraint is definitely predictable, such a constraint can be used in the design of a coordinate indexing system and will lead to a more efficient system without increasing costs. For example, if in any system there were 5,000 documents on guided missiles and no documents on unguided missiles and no documents on guidance of anything other than missiles, it would certainly be silly to set up a term for "guided" and a term for "missiles". This conclusion has been expressed symbolically in a discussion of what terms should be free and what terms not free in a pure system of coordinate indexing.[13]

Aside from this relationship to the theoretical conclusions from information theory and conclusions derived from the simple elegance of Boolean algebra, undoubtedly one of the motivating factors behind the hypothesis was the recognition that in ordinary English the order of words is more often redundant than not. It was recognized that in ordinary discourse, dispensing with word order and prepositions would certainly lead to inelegances and difficulties in the transfer of meaning. On the other hand, it was supposed that for retrieving information from a file, order and prepositional or syntactical relations could be dispensed with at considerable saving without appreciably increasing the noise in the system.

3.6 Introduction of Constraints

Various subsequent innovations in the system of pure coordinate

indexing in effect represent various types of constraints. Some of these constraints were the subject of experimentation by Documentation Incorporated. Similar and additional constraints have been suggested by other organizations and have been described in the literature which has appeared on this subject. It is this literature which is set forth in the appended annotated bibliography. The constraints have been variously named and described:

(1) Categorization

The categorization of terms in a vocabulary involves the division of a vocabulary under a number of generic headings. These generic headings can be thought of as supplying a grid or a set of questions put to any document in order to insure complete and uniform description of each document. Documentation Incorporated, as a result of its research, concluded that such categories were only useful in limited and carefully defined fields of information or in data processing, as distinct from information handling, systems. In its own experience it has successfully used categorization in three projects: The Nuclide Index, Project ECHO for the Air Force Office of Scientific Research, and the Cancer Chemotherapy data processing system for the National Institutes of Health.

Other descriptions of categorization are also included in the bibliography.

(2) Pre-Coordination

The particular distribution of documents in any collection might indicate that certain terms initially treated as separate terms should be pre-coordinated in order to cut down the time requirements for searching. This device of pre-coordination has been alluded to above in discussing the use of the term "guided missiles", rather than the terms "guidance" and "missiles". This technique can be followed as far as experience warrants. A pre-coordination essentially creates a more specific term than the terms which make it up. Being more specific, it has the effect of adding another term to the vocabulary and cutting down, by so doing, the average number of postings on each term in the vocabulary. At the opposite end of the spectrum, experience may disclose a number of highly specific terms with a very low incidence of posting. Periodic examination of such terms may indicate several possibilities:

(A) The terms may represent synonyms for other terms in the system and can be eliminated in favor of cross-references. Duplicate posting on synonyms may also be employed.

(B) The terms may be important terms which just happen to be relevant to only a few documents

in the system. If so they must be retained, even though they represent an increase in the vocabulary size.

(C) Finally, lightly posted terms may represent oblique synonyms which create a problem for the systems designer. He must determine whether they are useful enough to be retained or whether they can be eliminated in favor of combining their postings with the postings on other, more heavily used terms in the system.

It should be apparent that in a system of coordinate indexing the danger points are the two extremes, namely, terms with a great many postings, and terms with too few postings. It can be assumed that in the middle of the spectrum, the terms which have been used frequently by the indexer will also be those terms used frequently by the searcher, since in both cases they represent terms used frequently in the literature. The systems designer must, then, continually attend to the extremes as his system grows.

(3) The Association of Ideas and Thesauri

Quite early in its work Documentation Incorporated realized that a free vocabulary might present certain problems to the searcher unfamiliar with the vocabulary. This problem can be handled, in part, in any system which employs categories because each category can be displayed to the searcher. If categories are not employed, the searcher faces the problem of knowing what terms to ask for. For example, a searcher

approaching the system with any particular term in mind may not know what terms to coordinate with it, that is, what terms in general have been used with any given term in analyzing items in a collection. As an example of how this problem might be solved, a device known as "EDIAC" (Electronic Display of Indexing Association and Content) was created. This device consisted of an electrified panel having a vocabulary and a set of numbers. If any word in the vocabulary was selected as the beginning of a search, there would also be displayed on the panel the logical sum of all other words used in indexing any document indexed by the word first entered into the device. This display would enable the searcher to select an additional word to coordinate with the first word. He could do this with assurance that the system contained at least one document indexed by both terms.

This display of the vocabulary and its connections has some relationship to the process of redirecting a search as a result of browsing in a manual catalog. An initial search in a manual catalog might disclose not only a number of items but also give information to the searcher concerning what other terms he should use in further steps

in his search. The display of vocabulary on the EDIAC does exactly this. At least two thesauri which have been compiled are intended to provide information about the relationships of terms in an indexing system. These are the ASTIA and A.I.Ch.E. thesauri. Closely related to thesauri are semantic or generic codes. WRU uses the former and the U.S. Patent Office has studied the latter. These, as well as the thesauri mentioned above, are described in greater detail in Paragraph 5.2 of this report.

(4) The Use of Roles and Links

Quite early in the development of free forms of coordinate indexing which utilized only the logical functions of intersection, union, and negation as connections among terms, it was realized that the search of such systems would deliver a certain amount of noise or false answers. Consider, for example, a document containing information on "lead as a coating" and another document on "coatings for lead". If the system has terms for only "lead" and coating", the system will deliver noise. It is to avoid problems of this kind that there have been advanced in the past years the devices known as role indicators and links

as a means of supplying syntactical correctives in addition to logical correctives in an indexing system. Roles and links are described in greater detail in Section 6 of this report.

3.7 Conclusions

The development of a discipline sometimes resembles an ascending spiral, rather than a straight line, and in this development of constraints in coordinate indexing, there has been in some sense a return to the beginning.

H.P. Luhn has suggested a method of extracting keywords from a document and displaying them in an index in a context of other terms.[14] The context, in this case, constitutes a constraint, that is, a method of narrowing the significance of the keyword and making it more specific. At the same time, it must be recognized that the selection of keywords by machine matching, as opposed to the use of Uniterms, descriptors, etc., by an indexer, may be considered an advance in the direction of freedom. The original form of coordinate indexing certainly envisioned that the terms are selected not by a pure matching process or on the basis of any statistical counts, but by trained indexers who use their own background and understanding of the meaning of the document as determinants of the terms selected. To substitute for this rational process a process of machine matching

based upon statistical counts or other considerations relevant to information theory which can be programmed into a machine is to go beyond the original form of coordinate indexing in the direction of "digitalizing" information. In other words, the original form of coordinate indexing depends upon formal considerations only for the matching of terms and relies on intuitive considerations for their selection or creation. Keyword-in-context indexing relies on mechanical means for selecting the terms themselves from titles. (It should be noted here that the KWIC technique is presently more useful as a quick and convenient announcement technique for relatively small numbers of accessioned documents than as an accumulative index useful as a tool for detailed information retrieval. However, as authors recognize the need for and construct more informative titles, the KWIC technique might have broader capabilities.) Nevertheless, by showing how the context of terms can fulfill the same role as other more artificial syntactical constraints, Luhn moves coordinate indexing back in the direction of freedom -- the direction in which information theory indicates it should go.

Unfortunately, it is now fashionable in the field of information retrieval to deny the relevance of information theory because its application is limited to engineering and operating problems in the I.R. field. The bulk of present research in the information retrieval field seems to be concentrated in the domains of linguistics and semantics, even though no less an authority than Quine has warned

that the search for the meaning of meaning carries an investigator beyond positive science into the arcane field of metaphysics*

---

*Quine, Willard Van Orman, From a Logical Point of View, Cambridge, Harvard University Press, 1953. "Lexicography is concerned, or seems to be concerned with identification of meanings, and the investigation of semantic change is concerned with change of meaning. Pending a satisfactory explanation of the notion of meaning, linguists in semantic fields are in the situation of not knowing what they are talking about."

REFERENCES

1. Taube, M. and Associates. Studies in Coordinate Indexing. Documentation Incorporated, Vol. I, 1953, p. 34.

2. Bernier, Charles L. Correlative Indexes. A set of five papers published in American Documentation:
   Vol. 7, No. 4, Oct. 1956, pp. 283-288.
   Vol. 8, No. 1, Jan. 1957, pp. 47- 50.
   Vol. 3, No. 3, July 1957, pp. 211-220.
   Vol. 8, No. 8, Oct. 1957, pp. 306-313.
   Vol. 9, No. 1, Jan. 1958, pp. 32- 41.

3. The Royal Society Scientific Information Conference, 21 June - 2 July 1948, Report and Papers Submitted. The Royal Society, London, 1948, pp. 501-516.

4. Ibid., p. 511.

5. Ibid., p. 516.

6. Ibid., p. 86.

7. Fairthorne, Robert Arthur. The Mathematics of Classification. Proc. Brit. Soc. for International Bibliography, 9, 14 October 1947, pp. 35-42.

8. King, Gilbert, "Application of Card Methods to Scientific Computation," Punched Cards, edited by Robert Casey and James Perry, New York, Reinhold Publishing Corp., 1951, pp. 407-422.

9. Taube, M. and Thompson, A. F., "The Coordinate Indexing of Scientific Fields." Unpublished paper read before the Symposium on Mechanical Aids to Chemical Documentation of the Division of Chemical Literature of the American Chemical Society, September 4, 1951.

10. Ibid.

11. Morris, Jack S. The Duality Concept in Subject Analysis. American Documentation, Vol. V, No. 3, August 1954, pp. 117-146.

12. Bar-Hillel, Y. A Logician's Reaction to Recent Theorizing on Information Search Systems. American Documentation, Vol. 8, No. 2,

April 1957, pp. 103-113. See also Calvin N. Mooers' comments on this paper in the same issue, pp. 114-116.

13. Taube, M. and Associates. Studies in Coordinate Indexing. Documentation Incorporated, Vol. II, Chapter VIII, 1954.

14. Luhn, H. P. Keyword-in-Context Index for Technical Literature. IBM Advanced Systems Development Division, Yorktown Heights, N. Y., 1959. (ASDD Report RC-127.)

4. VOCABULARY GENERATION

4. VOCABULARY GENERATION

4.1 General

The vocabularies of retrieval terms employed by various operators of coordinate indexing systems have evolved in several ways. When the initial development of these vocabularies is considered, however, the *modi operandi* are found to fall into only a few broad classes, and the seeming variations in developmental techniques within the few broad classes are seen to be merely matters of degree. These variations in degree have resulted from different environmental and capability factors with which the system developers have had to contend. For example, adequate funding and development time permit much more vocabulary refinement and "editing" to be undertaken than do inadequate funding and timing. Similarly, personnel experienced in coordinate techniques will (given the same time and money) develop more refined vocabularies than will inexperienced personnel. Thus, in many (probably most) instances of vocabulary development, the degree of refinement initially achieved depended only indirectly upon a knowledgeable desire for adequate refinement (i.e., upon rationally chosen goals) but more often upon comparatively unrelated environmental factors. This statement, of course, is almost impossible to prove; system operators are understandably loath to admit fully the effects of the exigencies of their situations during

the embryonic stages of system development.

As noted above, coordinate vocabularies have been developed via only a few basic routes: (1) empirically, (2) from treatment of existing subject heading lists, or (3) from treatment of existing classification schemes. These three methods, and some variations, are described below.

4.2 Empirical Development of Vocabularies

The empirical development of vocabularies is typically based upon the technique of "free indexing". By "free indexing" is meant the assignment as index terms (for a given document) of words or phrases chosen by the indexer (based on judgment of relative importance) from the set of words or phrases conceived by the indexer to be appropriate indexing terms even though they may not have been employed in the document by the author, i.e., words or phrases expressing what the indexer considers to be the true meaning of the ideas expressed in the report.

The set of terms resulting directly from "free indexing" a collection of documents can be considered as the most rudimentary form of a coordinate index vocabulary; such a set of terms has usually been considered as preliminary in that a certain amount of refinement effort has nearly always been expended upon the "raw" vocabulary. For example, attempts (varying in their orderliness

and effectiveness) have usually been made to detect synonyms and to create "see" references or their equivalent.

The initial development of the du Pont Engineering Department vocabulary is typical (perhaps slightly "better" than typical) of initial vocabulary development efforts. This project was well documented shortly after the work had been completed.[1]

The report on the du Pont development implies several important environmental considerations. First, funding was adequate for the mounting of a major, concerted effort. Second, the time permitted for completion of the work was limited. Third, the personnel involved in the effort were largely inexperienced in coordinate techniques, although they did engage the advice of a firm in the documentation field. Briefly, the development proceeded as described below.

A complete subcollection of 2,100 reports was chosen for test because of both breadth of technological coverage and importance of the reports -- the particular subcollection having had codes assigned to its members indicating such relative importance. Indexing was carried out mostly from the report summaries (two to five pages in length) but in a few instances from abstracts only. The report states that when a phrase was chosen as an indexing term, each word of the phrase was also used as an entry (e.g., "shielded arc welding", "arc welding", "welding") although examination of tracing

sheets indicates that all indexers may not have followed this rule consistently.

Thus far, the du Pont development was typical of the empirical method. The next step, however, employed data processing equipment not available to many system developers. Accordingly, many system developers were resigned at this point to posting on Uniterm cards manually from the original tracings, to post each entry on each tracing on the appropriate Uniterm card, and then to examine the result for synonyms. At du Pont, however, the tracings were key-punched, one card being made for each entry on the tracing. The 28,000 resulting cards were then sorted alphabetically and numerically by machine. They were listed alphabetically by index term and by report number within index term. It was found that 9,400 " unique " index terms had been employed by the indexers.

The listing was then examined to permit combination of terms which should have been identical except for misspellings and inconsistent assignment of singular and plural forms, etc. After this, each term in the listing was transferred manually to a 5" X 8" Uniterm card; the vocabulary was thus reduced to about 4,000 terms.

The index was then edited. "See" and "see also" references were created. When the word components of phrases could be coordinated conveniently and apparently without "noise", the multiword

terms (i.e., phrases) were deleted and their postings transferred to the appropriate single-word term cards. Certain of the original terms were removed altogether (e.g, terms with very broad meanings). Hierarchical (or generic) posting from one term to a broader, more inclusive term was not consciously employed.

The final result was a vocabulary of 2,667 terms and "see" references. The number of "see" references or terms alone was not reported.

The effort represented, insofar as can be determined, a somewhat better than typical initial effort and one which is helpful in illustrating the initial generation, via empirical means, of a vocabulary for use in further coordinate indexing efforts.

4.3 Development of Vocabularies from Subject Headings

A number of coordinate vocabularies have been developed from pre-existing subject heading lists. Ignoring the reasons for instituting coordinate indexing systems at all, at least two factors appear operative: (1) a relatively large document collection exists which has already been subject cataloged, but which would have to be re-indexed at considerable cost should the empirical method be employed, and/or (2) the system developers have a familiarity with and confidence in the subject heading list, as well as a natural desire to preserve its advantages.

Whatever the reasons for choice of the "converted subject heading approach", however, the method seems to have much to commend it from a vocabulary-building viewpoint. The terminology is likely to reflect not only the technological scope of the collection to a fair degree but also a certain amount of standardization. Semantically imprecise terms (e.g., "columns") are likely already to have been detected and treated in some fashion, either implicitly or explicitly. A careful, thoughtful conversion process can retain these advantages as well as avoid the re-indexing of any pre-existing collection (although it is probable that the converted index will be less effective for the pre-existing collection than for new accessions).

One of the earliest studies in converting subject headings to Uniterms is described in a paper by Thomas and Gull.[2] In the experiment, the subject headings in the List of Subject Headings of the Technical Information Division (TID) of the Library of Congress and the Subject Heading List of the Document Service Center (DSC) of the U.S. Armed Services Technical Information Agency were converted to Uniterms. (Since their merger in March 1953, these activities have been under the general direction of ASTIA.)

The DSC Subject Heading List contained about 10,380 headings listed alphabetically by the inverted method in contrast to the TID list of more universal headings. This difference was the most serious

impediment to a merger of DSC and TID catalogs by subject headings since such a merger would have required changing one set of headings to conform to the other. The TID List of Subject Headings contained approximately 25,000 subject headings and 24,000 cross references in one alphabet. Subject headings contained one or more words and were followed by single or multiple subdivisions. The cross references provided for various relationships between words and phrases.

When the conversion of the DSC and TID lists had been completed, the terms and cross references comprised the following files:

1. One alphabet of 3607 unit terms converted from the TID list; 2287 terms were common to both lists and 1320 were derived from the TID list only.
2. One alphabet of 3275 unit terms converted from the DSC list only.
3. Approximately 720 synonymous references from the TID list.
4. Approximately 100 references from the TID list considered unnecessary.
5. Approximately 800 references from the DSC list, not separated as 3 and 4 above.

The five files listed above were completely edited before their merger into one alphabet of Uniterm headings. The final product of conversion was:

1. One alphabet of 6582 Uniterm headings, of which 1320 were derived from the TID list, 2975 were derived from the DSC list, and 2287 were common to both.

2. One alphabet of 549 synonymous cross references (single words, abbreviations, and notations).

3. One alphabet of 2106 cross references considered unnecessary (showing specificity, generalization, and synonymity in phrases).

Another vocabulary development utilizing the "converted subject heading approach" is that of ASTIA's Project Mars.[3] The principal headings were divorced from their subdivisions and, in one move, the list was reduced from 70,000 to about 8300 main headings. The 850 subdivisions were reduced to about 700 by the removal of words such as "application" that were no longer useful. These first steps resulted in a tentative vocabulary of some 9000 terms.

The tentative vocabulary was then edited in a fashion similar to that described above in the du Pont example. Synonyms were detected and appropriate "use" (i.e., "see") references were created. Also, certain only slightly used terms were included in other closely related terms via the "use" reference route (e.g., "miticides use acaricides"). These actions reduced the number of terms to less than 7000.

In addition, ASTIA provided "also see" (i.e., "see also") references among terms. Finally, terms were assigned to <u>groups</u> which were somewhat artificially designed according to the organizational arrangement of ASTIA; the group name was also displayed

parenthetically with the term itself in the resulting authoritative publication.[4] The cross references noted above were also displayed.

Following the initial vocabulary development and as a foundation for the institution of a mechanized retrieval system at ASTIA, the subject heading index was simultaneously converted both to the new coordinate vocabulary and to a machine-readable medium.

Despite apparent shortcomings in time permitted and funding, the ASTIA conversion represents a landmark. As contrasted with the experimental conversion of the ASTIA vocabularies by Documentation Incorporated, this conversion was actually used. The thesaurus was known to be incomplete at the time of publication, was intended only for internal use at ASTIA, and has been undergoing continuous review by the ASTIA staff. (Section 5.2 contains a brief review of the ASTIA Thesaurus revision effort.)

It should be noted that, according to ASTIA, the Uniterms developed during the contract with Documentation Incorporated were employed as aids in defining the descriptors, although in some cases they became descriptors.[5]

4.4 Development of Vocabularies from Classification Systems

Although in the original experimentation by Documentation Incorporated a certain number of classes were converted in order to demonstrate the method,[6,7] the conversion of classified indexes to coordinate indexes appears to have occurred only infrequently. This

is probably because the operators of hierarchically classified systems may have had little to gain by converting their indexes. The coordinate index which would result would require more effort in retrieval (i.e., manipulation during retrieval) than would the original tool and (unless the subclassification structure of the original tool were much more detailed than is usually the case) any resulting coordinate index (barring re-indexing of the collection) would be inadequate because the individual documents would originally have been "forced" into classes and the fine detail of information retrieval thereby lost at the outset. Another reason for the paucity of such conversions may be that very few progressive retrieval systems based principally on hierarchical classification actually exist.

The picture is somewhat different, however, with respect to faceted classification systems; few of these exist (at least in the U.S.A.). Faceted classification systems are likely to have captured at the outset much of the fine detail of information contained in accessioned documents, but at the expense of convenience in notation and file arrangement. Hence, they can be converted advantageously to coordinate systems because the manipulative characteristics of the latter can be employed to overcome the disadvantages of long complex notations and complex rules for file arrangement.

One conversion from a faceted classification to a coordinate index has been described by Wadington.[8] Wadington noted that the desire for extremely detailed indexing resulted in 1949 in the rejection of both hierarchical classifications and subject headings as candidates for retrieval tools in the reported system. Rather, a faceted classification derived from Ranganathan's colon system was tried and adopted. He reports that "although the component parts of the (classification) code were simple, the finished codes for information were long and complex. Manual filing by complete code was ruled out because it was too complicated and subject to too many errors."

"Extensive use of cross referencing was next considered. A card would be made out for each part of the code. While this would simplify the filing, it would greatly increase the number of cards to be processed." (This is assuming about eight cross-index cards per document.)

" Punched-card techniques were then considered, assuming one punched card per document; it was found that sorting time would run as high as 16 hours to make sure that all the information on an average question was retrieved."

Finally a coordinate indexing system was chosen and installed in 1954. Wadington reports that excessive noise resulted. This

difficulty was remedied by converting the previously developed faceted classification to a coordinate system. Apparently each facet or code component was converted to a subject term and the appropriate document numbers posted thereon.

Wadington noted that generic terms were consciously (and easily) generated from the original classification and that "development of these generic relationships would have been virtually impossible or at best exceedingly difficult" otherwise.

4.5 Conclusions

It can be concluded that effective coordinate vocabularies can be developed by any of the three above techniques. The empirical technique is based first upon "free indexing", and the results are modified (according to need) to provide adequate control; thus, the cost of vocabulary development is normally spread over a greater or lesser period of time, perhaps at the expense of achieving truly satisfactory retrieval of the first documents entering the system. The conversion of a subject heading list, on the other hand, can be made at moderate initial cost with the assurance that retrieval effectiveness will be no worse than that provided by the original tool; in this instance at least part of the cost of development of the subject heading list is recovered. The same is true of conversion of classifications, particularly faceted classifications, with the added advantage of

being able easily to detect generic relationships among terms.

References

1. Wall, Eugene. Use of Concept Coordination in the du Pont Engineering Department. Conference on Multiple Aspect Searching for Information Retrieval, Washington, D.C., 1957, Armed Services Technical Information Agency, pp. 12-25.

2. Thomas, Richard B. and C.D. Gull. The Choice of Terms for a Uniterm Coordinate Index of Scientific and Technical Reports. Studies in Coordinate Indexing, Vol. 1, pp. 47-55.

3. Heald, J. Heston. In Automation of ASTIA, December 1959, AD 227 000.

4. The Thesaurus of ASTIA Descriptors. First Edition, May 1960.

5. see 3.

6. Gull, C.D. Alphabetic Subject Indexes and Uniterm Coordinate Indexes, An Experimental Comparison. Studies in Coordinate Indexing, Vol. 1, pp. 56-64.

7. Wachtel, Irma. Classification and Categorization in Information Systems. Studies in Coordinate Indexing, Vol. 1, pp. 65-72.

8. Wadington, J.P. Modification of a Multiple Aspect System for Company Use. Conference on Multiple Aspect Searching for Information Retrieval, Washington, D.C., 1957, Armed Services Technical Information Agency, pp. 36-43.

5. CONTROL, MODIFICATION AND GROWTH OF VOCABULARIES

5.1 General

5.2 Thesauri, Authority Lists, and Generic Codes

5.3 Editing during Indexing, Posting, and/or Retrieval

5.4 Modifying the Vocabulary

5.5 Compatibility

5. CONTROL, MODIFICATION AND GROWTH OF VOCABULARIES

5.1 GENERAL

The effort needed for controlling vocabularies (once they have been developed) and for modifying and adding to them, would seem to be equivalent for all systems, except (as noted above) that conversion of a classification appears to provide somewhat better control of generic relationships from the outset.

Control of a coordinate vocabulary, once it has been developed, can be said to consist of two facets: (1) the proper use of the established vocabulary throughout the system, and (2) the modification of the vocabulary to conform to changing characteristics of inquiries and of accessioned documents. Both of these forms of control are discussed below.

An increasing number of operators of coordinate systems are coming to rely upon some form of authority for controlling the use of an established vocabulary. Typically, those systems which achieved the greater degree of refinement of their initial vocabularies are also those which exercise the greater degree of vocabulary control during use of their vocabularies. This is understandable because "vocabulary refinement" and "vocabulary control" are closely related and those environmental factors which affect the exercise of one of these two

functions will also affect the exercise of the other.

Thus the control during use of coordinate vocabularies varies from essentially "no control" (i.e., indexers continue to choose terms from the literature without regard to those previously chosen -- and little if any control of synonyms or even word forms is exercised during posting of the tracings to the index) to rather complete control, i.e., where an authority is employed either by indexers or by editors to keep the terminology used consistent with the established vocabulary. (The consensus seems to reflect the knowledge that the best indexing is performed by competent personnel, having both a subject specialty and some familiarity with the preparation of indexes. When such a combination of talents exists in the same person, the need for extensive editing becomes less imperative.)

Such control can, of course, vary in its degree of rigidity. It may, at a minimum, consist of instructing indexers to employ the word forms previously accepted and established by a simple authority list of terms, "see" references and "see also" references (e.g., "sulphuric acid" see "sulfuric acid" or "distilling" see "distillation"). New terminology, when encountered by an indexer, can be "flagged" for review at least to insure that synonymous relationships with the established vocabulary are not involved. In addition, new "see also"

references may be established both to and from any accepted new terms. And, in practice, these actions constitute the limit of exercisable control using a simple authority list.

Whether or not structure, i.e., vocabulary control, should be introduced into a system at the input level has not been determined. An argument in favor of at least minimum constraints at the input is represented in the following statement by H. P. Luhn.

> "Excessive editing obviously increases the likelihood of bias due to current interest, experiences, and points of view. In consequence the usefulness of a system will be reduced as emphasis and interest change. It would therefore appear that the less information is classified and contracted at the input, the more it will lend itself to dynamic interpretation at the output phase."[1]

5.2 THESAURI, AUTHORITY LISTS, AND GENERIC CODES

An increasing degree of interest and activity in using thesauri (or their equivalents) as authorities is apparent. The index to Current Research and Development in Scientific Documentation, No. 9, lists eight references to organizations working with or developing thesauri. There may be others, since there is no reference in the index to the work on thesauri now going on at ASTIA.

There are other aids to indexers that are employed internally in the working systems we have encountered. Most of these are rather

informal and are not available for distribution.

The existence of a thesaurus provides an opportunity for the operators of a system to exercise control and consistent use of the vocabulary at any or all phases of system operation: indexing, posting, or inquiry processing. Synonyms are indicated in a suitable thesaurus, as are "up and down" generic relationships and other relationships of unspecified, indeterminate or variable type (i.e., "see also's"). Thus a thesaurus may serve as the means of bringing to the attention of the indexer those vocabulary terms which might be employed in making a search for the document at hand and/or to the attention of the searcher those vocabulary terms which might have been employed by the indexer in describing documents pertinent to the question at hand. The result is greatly improved retrieval effectiveness.

The WRU Semantic Code and the U. S. Patent Office Generic Codes fall into the class of thesauri because they have incorporated, in some way, all of the three characteristics mentioned above. They differ from more conventional thesauri in that they are designed to be brought into play during mechanized posting or search rather than by a searcher during his formulation of a query or by a human editor.

The creation of a thesaurus for information retrieval purposes is, however, an enormously expensive task -- although this cost may for a given system be reduced markedly by referring to thesauri created

earlier by others with technological interests sufficiently similar to that of the system in question such that much of the earlier work may be utilized intact.

There is currently a great deal of time and energy being expended in the general area of thesaurus preparation. Although there is far from universal agreement on the necessity or desirability for such thesauri, many installations would probably utilize such an authority if its coverage were extensive enough.

Agreement is also lacking on the point of maximal utility of the thesaurus: Is its prime function to assist the indexer or the searcher?

Described below are the ASTIA and A.I.Ch.E. thesauri. Other thesauri exist or are being developed, but these represent the best efforts to date. Also described below are Western Reserve's Semantic Code and the Patent Office's Generic Code.*

5.2.1 THESAURUS OF ASTIA DESCRIPTORS

The Thesaurus of ASTIA Descriptors (First Edition), ASTIA, 1960, entangled its technical vocabulary (which is extensive, although oriented largely to military terminology) in multiword terms too much to permit it to be truly useful to others working with different collections. Further, its "also see" references are limited, and it does not exhibit generic relationships. However, it is understood that these shortcomings will be at least alleviated in the second edition now being

*For information on expense, see "Cost of Generic Coding," by Mortimer Taube, in Studies in Coordinate Indexing, Vol. III, pp. 34-57.

prepared. The preparation of the first edition was described by J. Heston Heald in Section III of the previously cited ASTIA report AD-227000. This description serves not only to report on one method of thesaurus development but also on how to build a thesaurus upon an earlier subject heading list.

5.2.2 CHEMICAL ENGINEERING THESAURUS

The Chemical Engineering Thesaurus, 1961, by the American Institute of Chemical Engineers, covers an extensive technical vocabulary (albeit somewhat more attentive to engineering terminology) and provides extensive synonymous, generic, and "related term" (i.e., "see also") cross-referencing. Because it was designed to be used in conjunction with a system having syntactical controls, however, it lacks definition of terminology sufficient to permit its use by others without some effort. For example, terms standing for materials (e.g., "water") do not have indicated their roles played in the system -- e.g., "water" used for "cooling" or "water" being "cooled"; the editors of this work intended that the indexers add an appropriate role indicator to such terms.

Nevertheless, the Chemical Engineering Thesaurus seems to be, to date, the nearest approach to a true and generally useful information retrieval thesaurus. It was first developed as an internal work by du Pont and the development has been described in detail by B. E. Holm and L. E. Rasmussen.[2] This description is accompanied by a valuable review of other work with (or related to) technical thesauri and by an exceptionally complete list of references to the work of others.

5.2.3 WRU SEMANTIC CODE*

The terms used by an indexer are converted to semantic codes automatically if the terms are already in the system, but if the terms are new, editors create new codes.

The purpose of the semantic code is to automatically bring into play all applicable aspects of a term. That is, it attempts to cover both higher and lower generic levels of terms as well as to encompass semantically related concepts.

An example is the word "diamond" coded as CERB≠CWRS≠PYPR♂1028, which may be interpreted as "a crystalline form composed of carbon and characterized by hardness." (The codes have been arbitrarily limited to four factors.) Any one of the factors may be searched. That is, the items indexed by "diamond" would be retrieved whether the question called for that specifically or for "things composed of carbon","hard things", or"crystalline forms".

Actually, CERB≠CWRS≠PYPR♂1028 is not the complete code for diamond, since under the principles of the semantic code this would be true not only of diamonds but of any other, say, hard crystals of carbon. To specify that diamonds and only diamonds are wanted, a further element -- the numerical suffix -- must be given. This is a

*American Documentation will publish "A Note on the Evaluation of the WRU Semantic Code as an Example of Generic Coding," by Mortimer Taube, Documentation Incorporated, in the April 1962 issue.

four-digit figure, the first numeral of which is that of the number of factors in the term. The other three are those peculiar to the particular concept. In this instance the numerical suffix might be 3001. Note that one of the factors, PYPR, is followed by the numerical <u>infix</u> 1028. This 1028 is the identifying numerical suffix for that particular physical property (P-PR), "hardness". Its use as a numerical infix here shows that the particular physical property characterizing (Y) diamonds is hardness. Only numerical suffixes beginning with 1 can be used for infixes. Since 2's, 3's, and 4's indicate that the code for the concept has more than one factor, their use as infixes would make it impossible to particularize specific concepts within the generic framework of a code.

This is further explained by the definitions given below.

<u>Semantic factor</u>. By this term is meant the separate units of a code, expressed by three consonants. In RAML≠RWHT≠TOMS$^{+}_{O}$1002≠3679, the semantic factors are R-ML, R-HT, and T-MS. Each semantic factor represents one of a number of highly generic concepts. Together they form, as it were, the building blocks of the code. It should be noted that within a code composed of more than one semantic factor, the separate semantic factors are arranged alphabetically ignoring the infixes.

<u>Infix</u>. By this term is meant certain symbols used with the semantic factors in a code. In RAML≠RWHT≠TQMS$^{+}_{O}$1002≠3679, the infixes are A, W, Q, and 1002.

<u>Alphabetic infix</u>. By this term is meant the infixes represented by alphabetic symbols. They show the analytic relationships of the semantic factors in which they appear to the concept represented by the code.

Numerical infixes. By this term is meant the infixes represented by numerals following the symbol $\overset{+}{0}$. They show, where used, a degree of particularization in the semantic factor to which they are affixed. Actually every semantic factor may be thought of as possessing a numerical infix; however, only in certain instances are they explicit, that is,they actually appear in the code. In the majority of instances, they are implicit, that is, they represent a numerical infix "1001" which is not actually printed out.

Numerical suffix. By this term is meant the particularizing number assigned each individual code to distinguish it from all other codes which, though they represent different concepts, contain the same semantic factors.

The semantic code, then, attempts to eliminate the manual or machine "see also" or "see" references. Furthermore, it attempts to include generic levels and general characteristics.

The code, of course, varies according to the system requirements. For example, in some systems the following code for diamond may be more valuable: CERB≠CWRS≠GUMM≠MANR; this may be interpreted as a "mineral in crystalline form composed of carbon and used as a gem".

For comparison, the A.I.Ch.E. thesaurus entry for "diamond" is shown below.

DIAMOND

PO Carbon

RT Abrasives

RT Crystal

PO = Post (also) on, and RT = Related term

It is obvious that it is not necessary to include all possible meanings, relations, or generic levels. For example, diamond in the sense of a baseball diamond would certainly be superfluous in a metallurgical system. However, whether or not WRU's four factors are sufficient and whether or not the factors chosen best suit the users' needs are both open to debate.

5.2.4 U. S. PATENT OFFICE GENERIC CODING

Many organizations, notably the Patent Office, have proposed that a coordinate indexing system include certain generic relationships among terms in the system, such generic relationships to be provided by codes which exhibit such relationships. From one point of view, a total code can be considered as a term, and a search can be made for the members of the classes defined as logical functions of such codes. From another point of view, a code can be considered as a class number, so that different parts of the number can provide for generic searches on different levels. For example, in one system proposed by the Patent Office, the established generic relationships of chemical names are replaced by a system of generic codes, as follows:

Heterocyclic Compounds	1313
Para-N-benzene Sulfoxy	1313-1512
Azoles	1313-2512

Thiazoles	1313-2512-1423
Oxazoles	1313-2512-1523

It is possible, of course, to provide for generic searches by indexing and posting on several levels without employing numerical codes. On the other hand, the numerical code is presumed to provide automatic generic indexing whenever the specific term is indexed.

5.3 EDITING, DURING INDEXING, POSTING, AND/OR RETRIEVAL

Vocabulary control (as distinct from vocabulary modification) can thus be exercised with the aid of some device such as an authority list or preferably a thesaurus; the types and degree of possible control have been described above. This section of the report will deal with where in the system the control can be exercised and note the reasons why different systems are reported to exercise control at different points. The control can be imposed by the indexer, by an editor, or by the searcher; thesauri or authority lists might be used in the process.

A. Vocabulary Control During Indexing

Control of the vocabulary during indexing places the burden of use of the authority (list or thesaurus) upon the indexer. In short, the Indexer is expected to analyze the document at hand, to develop a tentative list of retrieval terms therefor, and then to refer to the authority for the following purposes:

(1) To determine if each tentatively-chosen term is <u>in</u> the vocabulary; if so, to proceed to step (2) below; if not, to take the appropriate one of the three actions listed immediately below:

-- choose an accepted synonym (or sufficiently near-synonym) for the tentatively-chosen term

-- choose an accepted spelling (or word form) for the tentatively-chosen term

-- decide whether the tentatively-chosen term justifies being noted for consideration as an addition to the vocabulary; if not, delete the term from the tracing for the document.

(2) To examine the specified generic relationships with other terms in the vocabulary and:

-- to add to the tracing generically-higher terms if the chosen term is of sufficient importance to the information involved to justify such action, <u>and</u>

-- to add to the tracing <u>appropriate</u> generically-lower terms if such specificity is appropriate, considering the specificity of the information involved.

(3) To replace the tentatively-chosen term with two or more accepted terms.

(4) To add to the tracing appropriate additional terms chosen from the list of terms "seen also" from the chosen term.

Vocabulary control at this point is obviously costly in indexer time, yet it insures the best possible indexing. It does, however, tend to develop such "deep" indexing that the use of syntactical controls (e.g., role indicators and links) may be required to prevent excessive noise during retrieval; the use of syntactical controls adds further to the cost of indexing. In spite of this, if the ratio of accessions to inquiries is low, and if the need for speed and ease of answering inquiries is great, vocabulary control during indexing may well prove to be most economical of time and money.

B. Vocabulary Control During Indexing and Posting

This method places a lesser burden upon the indexer. In short, the indexer might only exercise functions (1), (3) and (4) as described in the above paragraph. Function (2), the addition of generic relationships, then is automatically performed (either clerically or by machine) at the time the individual postings are made to the index. However, only generically-higher relationships can be added automatically (the proper addition of generically-lower relationships being impossible to perform at this step) and such generically-higher relationships must

be added blindly, irrespective of the importance to a given document of the term upon which they are based. The advantage of this method is that it does save a certain amount of costly indexer effort.

C. Vocabulary Control During Posting Only

This method relies upon competent nonprofessional personnel or upon sophisticated machine programs (or upon both) to carry out the four functions described in paragraph (A) above. The function of "normalizing" tentatively-chosen terms into accepted terms may be quite well performed at this point. The function of the addition of generic relationships is performed just as in paragraph (B) above. However, the function of replacing the tentatively-chosen terms with two or more accepted terms and of the function of adding terms implied by the information in a given document is difficult, if not impossible, to perform at this point.

The advantage of this method is that it saves a large amount of costly indexer effort.

D. Vocabulary Control During Indexing and Inquiry

This method requires that the indexer perform only function (1), "normalizing" tentatively-chosen terms into accepted terms, and relies upon inquiry-processing operations to compensate for the non-performance at an earlier time of functions (2), (3) and (4). That is, rather than phrasing inquiries as simple logical intersections of terms

(e.g., all information on A and B and C, etc.), the phrasing will take the form of intersections of term unions (e.g., all information on A and/or Z and/or Y, etc. intersected with -- and -- all information on B and/or X and/or W, etc.). The terms chosen to constitute these unions are based upon the initial terms of the inquiry and are developed from the generic relationships (particularly lower generic relationships) and "see also" relationships exhibited in the authority for the initial terms of the inquiry.

This method is practicable only when suitable retrieval machines are available; without such machines the inquiry becomes excessively laborious and complex. This fact explains why operators of mechanized systems usually report that their inquiries involve unions of terms[3] whereas numerous operators of manual retrieval systems have reported that their inquiries are usually expressed merely as intersections of terms.[4]

The advantages of this method are that the indexer effort is reduced markedly and that choice of additional terms during retrieval is based upon a "real-world" situation (as defined by one inquiry) versus choice of additional terms during indexing having to be based upon hypothetical situations (as implied by all possible inquiries). Furthermore, the actual bulk of the index is decreased. On the other hand, the time and cost of processing inquiries is increased.

E. Vocabulary Control During Posting and Retrieval

This method is essentially the same as that described in the preceding paragraph except that function (1) -- "normalization" -- is performed by competent nonprofessional personnel rather than by the indexer. It reduces somewhat the indexing work load, but all other advantages and disadvantages remain as described above.

Given the possession of suitable retrieval machines, and assuming the usual (high) ratio of accessions to inquiries, it would seem that this method of vocabulary control offers the most advantageous choice for most retrieval systems.

An editor or anyone else concerned in the control and improvement of a vocabulary must rely on "feedback". If the user or searcher reports effectiveness of terms or coordinations of terms, the vocabulary can be controlled accordingly.

5.4 MODIFYING THE VOCABULARY

It is generally recognized that system vocabularies cannot remain static but must change to accommodate the changing technology reported in new accessions to collections. Basically, of course, there exist two forms of modification: addition of new terms and deletion of terms found to be of little value. However, the means of achieving these two forms of modification vary.

Under any circumstances, a route must be provided for adding new

terms to the vocabulary. For example, one can well imagine the difficulties and unnecessary work which would be created had the now-heavily-used acronym "radar" been excluded initially from the ASTIA vocabulary with the insistence that all documents dealing with radar be posted instead on "radio", "detection" and "ranging".

One (and a principal) means of addition to a vocabulary was noted above: i.e., the "flagging" of newly-encountered terminology during indexing (or posting) for consideration for addition to the vocabulary. Ideally, such new terminology should not be accorded status as retrieval terms until at least a moderate degree of usefulness has been proven; hence, accurate records of proposed new terminology and of its frequency of use must be kept. Further, if the proposed new terms are held aside in a subsidiary index until their usefulness has been proven, it will be easy to transfer them (together with their postings of document numbers) to the main index when their utility has been proven or to merely delete them (with their postings) from the subsidiary index when their utility has been conclusively denied. ASTIA follows these techniques to a large degree by maintaining its "Identifier Listing".[5]

A second means of adding to vocabularies is that of pre-coordination of existing terms. When it is found that two terms may be combined to create a third term, the two original terms are retained.

Deletion of vocabulary terms may be merely that, the action being based upon low usage during indexing and/or retrieval. More usefully, however, terms should be deleted by transferring their postings to either a generically-higher term (or terms) or to a sufficiently near-synonym. Such deletions should have the term replaced by appropriate "see" references.

Whatever the form taken by vocabulary modification, the authority for the vocabulary must reflect the change. New terms must have exhibited their generic and other related references -- and the pre-existing terms must similarly be referred appropriately to the new terms. When terms are deleted, all references to them in the authority (except an appropriate "seen from" reference) must also be deleted.

Little has been reported in the literature on formal modification of vocabularies. Perhaps this is because relatively few vocabularies are yet truly well controlled. ASTIA is now preparing the second edition of its Thesaurus, and this operation is being well documented in a continuing series of publications.[6,7,8]

ASTIA has stated the following.

> "A major objective in revising the Thesaurus of ASTIA Descriptors is to provide an improved ASTIA indexing authority in a form most useful (1) to assist analysts in making consistent and sufficiently complete assignment of descriptors to accessioned technical information and (2) to assist bibliographers in making a corresponding consistent use of

> the descriptors during the formulation of inquiries for mechanized retrieval.
>
> "A second major objective in revising the Thesaurus is to create a device which will be as useful as possible to reference personnel in organizations other than ASTIA. In this connection, ASTIA is anxious during revision of the Thesaurus to have the cooperation and active participation of all individuals and organizations who can assist in making the Thesaurus more useful both to themselves and to ASTIA."[9]

An important factor in ASTIA's request for cooperation was to inject both user orientation and user feedback into the effort.

Although the initial publication has been continually reviewed since it was issued, revisions were kept to a minimum in order to insure that the statistics obtained during its use were meaningful. (These statistics are currently being employed to determine the need for coalescing closely related descriptors or reducing the number of broad descriptors.)

The current revision program, initiated formally in mid-1961 and continuing to date, represents a large cooperative effort, and one of considerable magnitude. The new edition of the Thesaurus is scheduled for distribution in the Autumn of 1962.

5.5 Compatibility

No report would be complete without some reference to the many efforts involved in trying to achieve compatibility among various information systems. Several major endeavors in this area are, accordingly, briefly described.

An Ad Hoc Interagency Study Group on Language Compatibility in Mechanized Storage and Retrieval Systems was organized in October 1961. Looking forward to the day when many agencies will have mechanized information systems, this group is directing its attention to total systems compatibility, rather than compatibility of thesauri.

In a 1950 paper,[10] Mortimer Taube pointed out that the long-time effort to secure uniformity of subject cataloging and subject control on an international and national level had broken down. Even with reference to our three national libraries, the Library of Congress, the National Medical Library, and the Department of Agriculture Library, different authority lists, different methods of indexing, and different methods of classification are employed, reflecting differences in the collection and types of service. It was suggested in that paper that the way towards compatibility was in the direction of free indexing and unstructured vocabularies. The initial work on coordinate indexing was essentially a practical demonstration of this view. The authority lists used by the Air Force and Navy sections of ASTIA were structured

differently and, hence, incompatible. One of the specific tasks carried out by Documentation Incorporated and reported to ASTIA was the construction of a single, minimally structured list from the two existing lists.

In the current revision of the ASTIA Thesaurus, compatibility of the vocabulary was given prime consideration:

> "Thus the vocabulary should be as compatible as possible with other similarly used vocabularies -- and the Thesaurus, as the principal means for achieving such compatibility, should make it possible for other organizations to 'translate' their vocabularies to or from that of ASTIA -- and for ASTIA to do the same with other vocabularies. In this respect, the assistance of organizations other than ASTIA will prove invaluable." [11]

It has been pointed out, however, that the complexity of structure as reflected in a thesaurus does not contribute to compatibility of indexing systems. Undoubtedly on the basis of his own experience with thesaurus construction, Paul Klingbiel of ASTIA has said, for example:

> "But the documentalist is not at all concerned with the possible meanings a term may have. He is concerned only with that variety of meanings which occur in his collection. Moreover, he is charged with the responsibility of storing and retrieving that segment of man's knowledge represented by his library."[12] (italics Klingbiel's)

That is not to say that a thesaurus or any similar highly structured vocabulary may not be of great value as a description of the vocabulary structure of a special collection and, hence, as an important

tool for the searcher, but it does say that there can be no universal thesaurus which has prescriptive significance for indexers working in different collections in different information centers.

It is understood that the Engineers Joint Council is becoming interested in a plan of abstracting and indexing journal articles at the time of publication to render them more readily and promptly suitable for effective storage and retrieval by the various systems employed at point of use. This system has been instituted by the American Institute of Chemical Engineers, and their three journals now carry index terms and abstracts for the separate articles. The program has been well accepted by the A.I.Ch.E. membership and has been adopted as a permanent part of their literature service. The Chemical Engineering Thesaurus was a major tool in the implementation of this program.[13,14]

References

1. Luhn, H.P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, Vol. 1, No. 4, October 1957, p. 310.

2. Holm, B.E. and L.E. Rasmussen. Development of a Technical Thesaurus. American Documentation, Vol. 12, No. 3, July 1961.

3. Whaley, Fred. A Deep Index for Internal Technical Reports. ASTIA Conference on Multiple Aspect Searching for Information Retrieval, Washington, D.C., 1957.

4. Herner, Saul and Mary Herner. Determining Requirements for Atomic Energy Information from Reference Questions. International Conference on Scientific Information, Washington, D.C., 1958, pp. 181-193.

5. Automation of ASTIA - 1960. AD 247 000.

6. Gillum, T.L., P.H. Klingbiel, C.N. Mooers, and E. Wall. Philosophy of and Guidelines for Revision of the ASTIA Thesaurus. ASTIA, 1 Nov. 1961.

7. Klingbiel, Paul H. and Eugene Wall. Revision of the Thesaurus of ASTIA Descriptors. ASTIA, Progress Report No. 1, 17 Nov. 1961.

8. Ibid. Progress Report No. 2, 31 Jan. 1962.

9. see 6.

10. Taube, Mortimer. Functional Approach to Bibliographic Organization: A Critique and a Proposal. Presented before the Fifteenth Annual Conference of the University of Chicago Graduate Library School, July 24-29, 1950.

11. see 6.

12. Klingbiel, Paul H. Language Oriented Retrieval Systems. ASTIA, February 1962. AD 271 600.

13. Morse, Rollin. Information Retrieval. Chemical Engineering Progress, Vol. 57, 1961, pp. 55-58.

14. Holm, B.E. Information Retrieval - A Solution. Chemical Engineering Progress, Vol. 57, 1961, pp. 72-78.

6. ROLES AND LINKS

6.1 Description and Definition

6.2 Operating Experience with Roles and Links

6. ROLES AND LINKS

6.1 Description and Definition

Although they are called by several names and implemented in many forms (which will be described in the second part of this section), roles and links can be defined quite simply:

> Roles narrow the definition of the terms by designating the role of a word in context; they are generally, but not necessarily, symbols appended to terms or term numbers.
>
> Links are grouping devices; they are generally, but not necessarily, symbols appended to item numbers.

Links show that terms are related. Roles in themselves do not show how terms are related, because the role affixed to a term only describes the use of that term in context; it is only upon coordination of terms which have roles that a relationship among terms is established.

One of the most perplexing problems encountered in the study of roles and links is in nomenclature. Roles are variously called role indicators, modifiers, and modulants; occasionally, scope notes are called roles or role indicators. Terms with roles attached have been called "structerms".

Links are variously called interfixes, link letters, punctuation, and association.

Scope notes are erroneously called roles, because although scope notes do "narrow the definition of terms" (as specified in the

definition of roles) they do not "designate the role of a word in context". Scope notes are used generally only to differentiate homographs.

The nomenclature confusion has arisen for one main reason: when one modifies a concept even ever so slightly, he indicates this fact by giving it a new name. Sometimes it is only a modification in scope; sometimes it is a modification in form.

For example Western Reserve University's indexers indicate grouping of terms into sentences or phrases by dots. They call this "punctuation" rather than links. (Their links have taken this form because they use a term-on-item system rather than an inverted system. For further details, see Section 9.4.) Because it is possible to symbolize a link either by a letter or a number, one group at du Pont has called their links "link letters".

There is at least one person who uses links and does not call them links -- nor any of the other terms mentioned above. In a paper presented at the Fourth Institute of Information Storage and Retrieval, at The American University, William B. Kehl described work in storage and retrieval of legal material at the University of Pittsburgh.[1] In this system, as in WRU's, the index is a term-on-item file; the text of the document, including punctuation, is entered intact on computer tapes. Prof. Kehl said that the computer program was such that it could accept

the instruction "retrieve this document if and only if this word is within five words (or some other number of words) of that word". An example he gave included the terms "unable" and "pay" or "payment". This reduces (but does not completely eliminate the possibility of retrieving documents which in fact have these two terms but not the expression "unable to pay". The computer can also accept the instruction, "retrieve this document if and only if this word and that word are in the same sentence". These instructions are grouping devices, hence, links.

Prof. Kehl's system also uses "roles", although here again, he did not so call them. The computer can accept the instruction, "retrieve this document if and only if this word precedes that word". This will increase (though not guarantee) the chances of retrieving, e.g., 'water for cooling" rather than "cooling of water". The example he gave was "district school" versus "school district".

Even descriptive catalogers use links when they analyze a document in two or more separate parts, numbering the report 2051a, 2051b, etc.

In the final analysis, it is possible to say that, except for their mode of implementation, roles, role indicators, modifiers, and

modulants are synonomous; and that links, interfixes, link letters, punctuation, and association are synonomous.

Mortimer Taube of Documentation Incorporated has recently published a significant paper on roles and links.[2] In this paper he shows that the use of links as syntactical devices may lead to loss of information.

6.2 Operating Experience with Roles and Links

Very little operating experience with roles and links has been reported. Certain opinions have been offered, informally, as to the cost and/or efficiency of these devices, but they have not been completely substantiated.

It is generally admitted that the addition of roles or links will increase input costs; by how much, however, is not known. Estimates range from 5 percent to 50 percent.

Whether or not roles and links truly improve system efficiency has not been proven. Even those who insist that the system is more effective do not indicate whether or not the percent of increase in efficiency is equal to or in fair proportion to the increase in cost.

In one informal disclosure to the Documentation Incorporated study group it was said that the proportion of false drops retrieved without roles and links to the false drops retrieved with these devices in the system was 4.5 to 1. It was admitted, however, that the test was neither exhaustive nor conclusive; no cost comparison was made.

According to available information and material submitted to NSF for Nonconventional Technical Information Systems in Current Use, No. 3, which is in preparation, the following organizations are using or about to use roles, but not necessarily links.

U.S. Patent Office, Research and Development Division
Central Intelligence Agency
Western Reserve University, Center for Documentation and Communication Research
National Lead Company, Titanium Division
Armour and Company, Patent Law Department
Philip Morris, Inc., Research Center
Linde Company, Research Laboratory
Eastman Kodak Company, Research Laboratories
E. I. du Pont de Nemours & Company:
Engineering Information Center
Polychemicals Information Center
Monsanto Chemical Company
Jonker Business Machines, Inc.
General Electric Co., Defense Systems Department

WRU,[3] du Pont,[4,5,6,7] and the Patent Office[8] have published on their application of roles and links. These papers describe implementation more than operating experience; that is they do not give

Information on effectiveness, cost, etc.

It should be noted that roles must be tailored to the system. For example, the A.I.Ch.E. roles would be of no use in legal material; in fact, several chemical systems use different roles. Examples of the A.I.Ch.E. roles are given below.

A. Input to a chemical reaction, physical production, operation, electrical, or mathematical system.
C. Waste, contaminant, impurity.
D. Agent
H. Active concept, subject of study.

Roles in a legal system, on the other hand, may be as follows:

A. Plaintiff
B. Defendant
C. Witness
D. Statute cited -- applicable
E. Statute cited -- not applicable.

Although roles and links generally are of the form described in the preceding section, i.e., they are letters or numbers attached to terms or item numbers, they do appear in other forms.

An example of a different configuration is shown below. This is selected from a Western Reserve University coding sheet. In a non-inverted system like WRU's, links especially must take a different form.

1.	..KEJ,	2.	ROD
3.	.KUJ,	4.	ALLOY
5.	.KUJ,	6.	AL
7.	..KAM,	8.	ANNEALING

The double dots indicate the grouping of the terms "rod", "alloy", and "aluminum". The comma indicates that KEJ is one unit of information and that rod is another which is associated with it. The single dot indicates that KUJ and alloy (which are associated with each other by the comma) are associated with the other terms between the double dots.

KEJ is a role indicator. It shows that the word "rod" is the name of a material which is acted on by a process. When the other role indicators are translated, the information in the sample is as follows: a rod is shown as being composed of an alloy whose major constituent is aluminum and this aluminum alloy rod is being subjected to the process of annealing.

Now that several groups have incorporated roles and/or links in their systems, it is hoped that thorough studies on effectiveness and cost will be made. Prima facie, it seems that in systems of deep indexing and limited vocabulary, such devices may be useful and worth the cost.

References

1. Kehl, William B. The Communication between the Computer and the User in Information Searching. Paper presented at the Fourth Institute on Information Storage and Retrieval, The American University, February 12-16, 1962.

2. Taube, Mortimer. Notes on the Use of Roles and Links in Coordinate Indexing. American Documentation, Vol. 12, No. 2, April 1961.

3. Kent, Allen, Jessica Melton and Alan M. Rees. Test Program for Evaluating Procedures for the Exploitation of Literature of Interest to Metallurgists -- Part III, Analysis and Quality Control. Western University, Center for Documentation and Communication Research, Cleveland 6, Ohio.

4. Costello, J.C., Jr. Storage and Retrieval of Chemical Research and Patent Information by Links and Roles in du Pont. American Documentation, Vol. 12, No. 2, April 1961.

5. Holm, B.E. Information Retrieval -- A Solution. Chemical Engineering Progress, Vol. 57, No. 6, June 1961.

6. Montague, Barbara. Patent Indexing by Concept Coordination Using Links and Roles. Paper presented at ADI Annual Convention, Boston, Mass., Nov. 5-8, 1961.

7. Morse, Rollin. Information Retrieval. Chemical Engineering Progress, Vol. 57, No. 5, May 1961.

8. Andrews, Don D. Interrelated Logic Accumulating Scanner (ILAS). Patent Office Research and Development Reports, No. 6, June 25, 1957. (For other Patent Office Reports, see Annotated Bibliography, Section 12, of this report.)

7. EVALUATING COORDINATE INDEXING SYSTEMS

7. EVALUATING COORDINATE INDEXING SYSTEMS

There is a marked absence of rigorous tests and controlled experiments in the theory and practice of coordinate indexing today. Various attempts have been made to evaluate systems and applications of theories of coordinate indexing; these have at best met with limited success and at worst have been prejudiced and invalid.

This section of the report is an attempt to discover the reasons for the inconclusiveness of evaluative studies, to indicate the problems involved, and to report and suggest methods of analysis.

By "retrieval system" or "information system" here is meant any means of providing access to a body of information. Thus, even a book's index is a system.

7.1 Factors in Evaluation

The evaluation problem has several levels. In the first place, coordinate indexing systems may be compared among themselves or with other indexing systems, such as classification systems. Secondly, the system may be evaluated within itself in order to determine its effectiveness. This is an opinion of the users, who differ among themselves. It is the opinion of the designer, who might say that his system is the best available given the present state of the art. It is the opinion of the indexers, who might say they are indexing as well as possible. It is the opinion of the administrator, who might say that

It is the best system for the price.

Suppose, however, that the ineffectiveness of the system is admitted. (Even if it is not, there are very few people who would say there is no room for improvement in their systems.)

The faults in an ineffective retrieval system may lie in any, all, or any combination of the following:

(1) Acquisition

(2) Indexing
  (a) point of view
  (b) depth of indexing

(3) Index
  (a) term arrangement
  (b) format
  (c) size

(4) Constraints
  (a) categorization
  (b) pre-coordination
  (c) "see" and "see also"
  (d) role indicators
  (e) links
  (f) scope notes
  (g) codes

(5) Search procedure
  (a) logical operations
  (b) equipment capabilities

(6) User education
  (a) knowledge of system capabilities
  (b) question formulation

Because this report is on the state of the art of coordinate indexing and not on the larger subject, information storage and

retrieval, this discussion of evaluation will be restricted to studies of indexing and of indexes with and without constraints.

It is understood, of course, that indexes and indexing are dependent variables in a total system, and that a complete evaluation must take the total system into account.[1,2.]

It should be noted that the problem of evaluation is threefold: (1) determining whether or not a system is effective in itself or as compared to another, (2) determining where deficiencies lie, and (3) determining the causes. Furthermore, the evaluation itself must be conducted in a formal, logical, and valid way.

7.2 The Difference between Evaluating Indexing and Evaluating the Index

The difference between indexing and the index must be made clear. This is important because in any system, the indexing may be perfectly good, but the index so arranged that it is unwieldly or overcomplicated. The reverse situation can also exist: the indexing might really be poor, and no matter how perfectly arranged the index, the user will not retrieve information.

Saying that the index is poor, then, implies something about the arrangement of the index (including term arrangement, constraints, format, and size).

Saying that indexing is poor implies something about the intelligence of the indexer, that is, the indexer's experience in

indexing, his knowledge of the subject (or at least of the terminology of the subject), and his capacity for being educated to the point of view of the system.

Coordinate indexing in its free form has a great advantage over other kinds of indexing in that the indexer is not required to "fit" the terms of the document into a greater system, as he must in a structured system. Nor is he required to formulate the concepts of the document into concise phrases, as he must for an alphabetic index such as a book index.

Nevertheless, several demands are made on the indexer's intelligence. He must have an appreciation for the nature and use of the index; he must know the subject matter well enough to be able to choose the key words which best describe the document, and he must also know how that document or the terms which describe it fit into the total system.

Furthermore, with the addition of constraints, such as roles and links or generic posting, the indexer must enlarge his effort in order to recognize and name functions or relationships among terms.

In any attempt to evaluate indexing, these intellectual operations of the indexer must somehow be taken into consideration. It is for this reason that there has been little or no success in evolving criteria by which indexing quality can be measured. There have been

studies made of indexing, however, and certain useful information has been derived from the work. These studies are described in greater detail below.

7.3 Indexing Studies

Nearly everyone who has ever been involved in indexing has conducted the following very simple experiment: several people (of related or different backgrounds) are given the same document; no limit is placed on time or on the number of terms that can be used.

It is obvious that there are many variables in this experiment and this partially accounts for the fact that results differ from experiment to experiment and that the results from any one experiment are difficult to analyze.

It may be of interest here to note a few findings of such an informal experiment in Uniterm Indexing in Documentation Incorporated's Man-Machine Information Center.

(1) Number of terms indexed varied from 10 - 45.

(2) Time varied from 5 - 25 minutes.

(3) The number of terms used was not proportional to time expended.

(4) The number of terms in common was about six. (This was an average figure; in comparison of some sets of two results only one or two terms were found in common.)

(5) Point of view was significant.

(a) Indexers with different technical backgrounds indexed differently. (For example, a report on pressure effects on aircraft pilots at high altitudes would be indexed differently by a psychologist and by a physicist. Both, however, would probably choose the major terms: pressure, high altitudes, pilot, aircraft.)

(b) Because the indexers, in this system, also retrieved and used the information in the store for providing answers to M-MIC clients, their indexing was influenced by their knowledge of system use. (For example, one indexer who was aware of a particular client's wish for all information on thin films was inclined to index any related information on the subject, even though the general value was small.)

(6) Under different circumstances, the same indexer might index the same document differently. That is, even from one day to the next, although a certain number of terms (about seven) might be the same, the choice of less important terms varied.

Especially from results (5) and (6) the importance of depth of indexing becomes obvious. If the system were restricted to four or five terms the differences in choice of terms would be negligible. This is generally not found practical, however, because system users require detailed information.

For example, the report mentioned above on effects of pressure on aircraft pilots at high altitudes might be pertinent to a search for information on pressure tests made with a particular kind of device. If that device had not been indexed, the report would not be retrieved by

correlation of the terms in the question.

There are two possible "fail-safe" procedures: (1) to formulate questions in the most general terms possible, or (2) to index 50 or 60 terms per document. Both of these methods result in retrieval of excessive non-pertinent information. (Information that is "nonpertinent" is not necessarily a "false drop". This will be discussed further in a later section.)

The M-MIC personnel frequently found that overindexing contributed to the retrieval of excess material. For example, in a search for information on digital differential analyzers, many reports were retrieved which mentioned such devices only in passing.

Because of these informal findings, M-MIC indexers were instructed to use only 10 - 12 terms per document where possible. This was not a formal deduction from a scientifically correct experiment, but its application was found to produce less extraneous information. It was not determined whether or not information was being lost, but it was believed that when the indexers were more restricted in number of terms which could be used, they were more likely to weigh the relative merit of that term in the context of the document.

In reviewing the experiment outlined above, several failings are seen: (1) the experiment was not formalized, (2) it was not a "controlled" experiment, (3) results were not carefully analyzed, and (4) the

(intuitive) implementation of the findings was not subsequently subjected to rigorous test.

Study of the reports on coordinate indexing systems has disclosed no fundamentally better experiments in indexing, nor any more formal deductions from experiments. Many groups have reported that "it was found" that ten terms best indexed an article, or that chemists indexed chemical literature better than others did, or that to terms that an author chose it was necessary to add others.

It is significant that the converse of each of these findings is reported with equal confidence.

Studies in indexing or in evaluating indexing are almost entirely intuitive. Perhaps, because the indexer's intellectual operations must be taken into consideration, deductive studies are not possible.

Described below are four possible exceptions to one or both of these statements. One is an attempt to formulate a generalized theory of indexing; the second and third are attempts to make deductions from actual indexing experience; the fourth is the work in automatic indexing.

Jonker's Descriptive Continuum

Frederick Jonker's "The Descriptive Continuum -- A 'Generalized' Theory of Indexing"[3] describes a study based on the premise that "no true understanding of existing systems and problems seems possible, unless all systems can be seen in the light of more general common concepts, linking

all these systems together into a single 'closed' system."

The generalized theory of indexing postulated in this article looks upon all indexing systems as a continuum, the descriptive continuum. The main parameter of this continuum is the average length of the "entries" or "headings" used. At one end of the continuum or "spectrum" is keyword indexing; subject heading indexing is somewhere in the middle, while hierarchic classifications are at the other extreme. The average length of the headings or descriptive terms used determines the position in the continuum.

Throughout the continuum, all other parameters behave as functions of the average term length. Some of these parameters are:

-- Potential depth of indexing
-- Permutability of indexing criteria
-- Degree of hierarchial definition of indexing
-- Potential need for a coordinating mechanism
-- Retrieval noise
-- Size of the access apparatus
-- False coordinations
-- Capacity for handling semantic indeterminacy.

The theory indicates that once the main parameter, average term length, is determined, all other properties of the indexing system are fixed. For every information collection there is an "optimum" position in the continuum, according to which the collection should be organized. This optimum position is determined by the diffuseness of the information in that particular field.

MacMillan and Welt Indexing Study

"A Study of Indexing Procedures in a Limited Area of the Medical Sciences,"[4] by Judith T. MacMillan and Isaac D. Welt, reports an attempt in making deductions from actual operating experience.

The Cardiovascular Literature Project from which this study emerged was an indexing-abstracting project, which over a period of six years encompassed 31,000 articles.

The paper points up the point-of-view problem discussed earlier. The indexing differences among chemists, physiologists, etc., were found "more obvious than the similarities." Figures are given on numbers of doubly-indexed papers which were indexed the same or synonymously, which were duplicated but incompletely, etc.

Although the findings of this report are more often inferences than deductions, this work does represent an attempt to make deductions from actual indexing experience.

Documentation Incorporated -- Rome Air Development Center Study

A study is now in progress at Documentation Incorporated, supported by the Rome Air Development Center. This is a large-scale, statistical investigation to determine indexing consistency among indexers, as well as any one indexer's consistency, and to measure the effects of certain learning or teaching aids and tools on indexer efficiency.

The experiment is being performed on chemical patents indexed by

experienced indexers, and should be completed by November 1962.

Automatic Indexing

It is curious that some of the basic studies of indexing have been made in connection with automatic indexing.

These are, however, only studies or experiments. Even these investigators have not found means of truly testing or evaluating the effectiveness of their indexing because in nearly all experiments with machine indexing, the "control" or the basis of comparison is human indexing. The machine indexing is considered "good" if it compares favorably with human indexing. This is paradoxical, because no one knows when human indexing is "good".

Maron[5] has said that if one is willing to collect enough statistical data relating words and categories, and if one is prepared to consider more and more of the relationships that exist between individual words, word combinations, word type, etc., and categories, one can index by machine with increasing accuracy. This statement refers to indexing by categories.

Similar statements were made by Cherenin[6] in an early paper in discussing the problem of an information language.

Luhn's experiments in automatic indexing[7] are directed to what may be called a simpler problem: indexing by Uniterms. Here, as above, statistical data are important. Luhn proposes to index by counting

frequency of occurrence of "notion" words in the document. Notion words are opposed to "connective tissue" -- the and's, of's, for's, etc., in a sentence.

Before any of these systems can be proved effective it must be shown that the point of view of the system can be taken into account, that is, that terms can be selected on the basis of the design and use of the system.

It should be pointed out that mechanization is not a solution to indexing problems. That automatic indexing will in certain cases substitute for human indexing is possible; that it will eliminate the need for human indexing (or solve any of the associated difficulties) is unlikely.

In conclusion, it can be said that the indexing studies described in preceding sections have been scientifically unsatisfactory; i.e., they have been intuitive, informal, and without rigid controls. This does not make them invalid nor useless, but serves to underline the fact that indexing is a thought process and as such cannot be subject to deductive analysis. The product of indexing, i.e., the indexer's tracing sheet, may lend itself to deductive or comparative analysis, but the study of indexing per se is necessarily intuitive.

7.4 Index and Index Constraint Studies

7.4.1 General Studies

Although indexing studies have been few, there is no lack of index and index constraint studies. They range in scope from comparisons of coordinate indexing and classification systems to measuring the value of the addition of particular constraints (e.g., roles) to an index.

The constraints, as mentioned before, are

(a) categorization
(b) pre-coordination
(c) "see" and "see also"
(d) roles
(e) links
(f) scope notes
(g) codes

It should be noted that although many of the tests have been conducted with at least an attempt to be scientifically valid, results (just as in indexing studies) are not always universally valid, and perhaps with reason. Trying to apply someone else's experience with roles, for example, is rather like expecting the medicine in another man's cabinet to cure your own ills.

At least two major studies of comparison of coordinate indexing with classification or other systems should be noted here. These are Cleverdon's Aslib Cranfield Research Project[8] and Schüller's report of experience with UDC and Uniterms[9]. Several other studies have, of course, been made. Many of these are reported in ASTIA's report on the Conference on Multiple Aspect Searching[10].

The Cleverdon paper reports results of five tests of retrieval efficiency and gives an "efficiency percentage" for UDC, alphabetical, faceted, and Uniterm indexes.

Alphabetical indexes had the highest percentage in Test 1, retrieval efficiency based on 300 searches by project staff. UDC

scored highest in Test 2, searches by engineering staff.

In Test 3, the indexes were rated by indexing time. With 16 minutes allowed, alphabetical indexes had the highest efficiency percentage; with 12 minutes, UDC; with 8 minutes, alphabetical; with 4 minutes, Uniterm; with 2 minutes, UDC.

Test 4 compared results according to three indexers. For two of the indexers, alphabetical indexes had the highest retrieval efficiency; for the third, UDC.

Test 5 compared results according to subject. Uniterm indexes had but a slightly higher percentage of retrieval for aerodynamic papers (with alphabetical indexes close behind); for "other subjects" the alphabetical indexes had the highest percentage.

Before one draws too many conclusions from these findings, he should be reminded that the tests were on documents dealing with aeronautics and allied subjects, and the results might be entirely different in other areas.

The Schüller study is on a much smaller scale than Cleverdon's. One test result (on 100 queries) shows the following.

	Time (Min)	Irrelevant Documents	Relevant Documents
UDC	4042	4908	698
Uniterms	3981	528	739

with 482 found by both systems.

The Schüller study recommends, based on these and other tests, the use of the Uniterm system for a collection of technical reports.

This is not the conclusion one might reach from the Cleverdon study, and serves to point out what has been constantly repeated: that test results, even when correct, are not universally applicable.

Studies on coordinate indexing systems with or without constraints are too numerous to detail. Reports on such studies are referenced throughout the bibliography. (Work on roles and links is detailed to some degree in Section 6 of this report.)

In the following three sections, however, are discussed three particular methods of evaluating or analyzing indexes.

7.4.2 False Drops as Test of Index Effectiveness

It would seem possible to devise means of testing system effectiveness by analyzing false drops. The question immediately arises: What is a false drop?

As Mr. B. K. Dennis suggested (in a private communication), it is very difficult to define, let alone analyze, a false drop. Whether or not the material answers the query is subject to the opinion of (1) the man who indexes, (2) the searcher who formulates the question, (3) the technical evaluator who screens retrieved material before forwarding it to the user, and (4) the user (one individual may perform several or all of these functions). A system, he pointed out, may correctly retrieve nonpertinent information.

It may, for example, retrieve an item correct in all respects except, say, temperature range or geographical location. Perhaps the user did not specify these criteria in his question, or perhaps such terms were never entered into the index.

Another possible "correct" nonpertinent retrieval was suggested previously. A document may indeed mention digital differential analyzers and yet not provide the user the information he wants.

Still another kind of "correct" false drop is encountered in certain systems using generic levels. Documents discussing cocker spaniels may be retrieved when the system is queried for information on fox terriers, if both were posted on "dogs", and the user included that generic term to insure that no items on fox terriers were missed.

Although analysis of false drops seems a possible method of evaluating system performance, there have been no reports of success in this method. Nor has anyone disclosed a means of analyzing the reasons for nonpertinency of material received.

The only profitable use made of false drops, or what are more properly called false coordinations, is in the checking of search procedures or equipment.

7.4.3 Use Statistics

Studies in frequency of term usage both in index and search, and studies of the "association factor", are under way. Find-

ings from such investigations are potentially useful not only in evaluating but in improving a system.

At ASTIA, three computer-prepared aids are under study. These are printouts showing (1) Descriptor Frequency Listing by Document Assignment and by Bibliography Usage. An example of such a printout is shown below.

DESCRIPTOR	ASSIGNMENT	REQUEST
Jet Planes	2216	37
Jet Seaplanes	22	9
Tungsten Wire	23	0
Water Entry	123	2

(This and following examples are taken from a memo to all ASTIA divisions dated 30th October 1961, and signed by J. Heston Heald, Chief, Document Processing Division. The examples are "sample formats of proposed reference tools".)

(2) Low Frequency Descriptor Manual File.

Descriptor	AD Numbers
Alpha Chambers	204 929
Demerol	209 426
First Aid Kits	219 127 222 912
Hail Damage	219 427 220 921

Thus, where a descriptor is used infrequently, the term and reference will be maintained in a manual file.

(3) List of Context Descriptor Sets

Cameras (138)

Aerial Cameras (11)

Airborne (11)

Amplitude Modulation (15)

Combustion (1)

Design (48)

Detection (43)

Instrumentation (25)

Photography (14)

Power Supplies (16)

Training (3)

Transducers (1)

Underwater Photo. (1)

Upper Atmosphere (2)

Vibration (2)

This printout shows all terms which have been indexed with the term "camera" and how often. ASTIA has not yet an economically practical (i.e., fast) way to get this count.

From list (1), the terms which are never, or infrequently used, can easily be found and eliminated or modified.

From list (2) infrequently used terms which are retained in the system are noted and removed to manual files so as to eliminate extra search time on the computer.

(The reverse of (2) could be useful in a system for document inventory control. That is, a printout showing number of times a document was retrieved in a search would indicate the document's relevancy to a system.)

The usefulness of a tool such as list (3) has long been recognized. Documentation Incorporated's "EDIAC" (Electronic Display of Indexing Association and Content) was a device which displayed the logical sum of all other words used in indexing any document indexed by the first word entered in the device. This, as the ASTIA printout, would enable a searcher to select an additional word to coordinate with the first word. In the EDIAC method the searcher would know that at least one document in the system was indexed by both terms; in the ASTIA method, he would also know how many documents are indexed by both words

A further refinement of this method has been proposed by Stiles.[11] The first step in the procedure is to develop a list of terms arranged according to their degree of association with a given term. ASTIA's method does not give this, but only <u>frequency</u> of association. In the paper referenced, the formula used for the association factor (to measure <u>degree</u>) and further details of the procedure are given. The

method is still under trial, and no conclusive results have been achieved; the economic feasibility of such an approach has not yet been shown. The goal, however, of such a method is indeed a worthy one: to find documents related to a request even though they have not been indexed by the exact terms of the request, and to present the documents in the order of their relevance to the request.

Mary Elizabeth Stevens[12] is also working on a means of leading the searcher to an answer even if he does not ask the question properly. In her "I.Q." or "Selective Recall" system a computer program is evolved to permit multiple levels of generic, specific, and associative access to items in an index store.

Claire Schultz's work in use statistics should also be noted. In a study of the indexing terms of the Merck Sharp & Dohme Research Laboratories indexing system (reported by Schultz and Shepherd),[13] the frequency of use of the indexing terms, both singly and in combination with one another, was determined. In another study, Mrs. Schultz and her associates conducted a comparative study of the dictionaries of the Merck Sharp and Dohme punched card system and the ASTIA computer system.[14] Computer routines were used in both studies. The summary of the results of the latter study is quoted below to illustrate how such findings might be used in system or index evaluation.

"(1) In both systems the shape of the curve for the occurrence of single descriptors in the sample is one-half of a U. The same is true for the occurrence of pairs of descriptors.

(2) One system generates more double, triple and quadruple combinations of descriptors than the other. This is because that system has a higher number of descriptors per document. Descriptors are ordinarily searched for in triple and quadruple combinations. Whether or not the system providing the higher number of descriptor combinations has the more useful file remains to be scored.

(3) A curve has been drawn for each system, measuring the use of single descriptors in terms of the average number of uses of descriptors in that system. The two curves nearly coincide. It is hypothesized that the shape of this curve (nearly a straight line for cumulated frequency of use of descriptors) is an intrinsic characteristic of dictionary use, to which dictionary use in all systems can be related. More systems will have to be analyzed before an acceptable standard can be chosen for interpreting the meaning of deviations in the curves of individual systems."

7.4.4 User Studies

User studies differ from the use statistics studies described above in that the former refer to actually studying the user rather than the system.

Most of the user studies are determinations of the requirements of the users. Mortimer Taube prepared "An Evaluation of 'Use Studies' of Scientific Information"[15] in 1958 in which he analyzes such work. At that time, he says, the consensus was that the studies

were not of much value in the design of a system. Some objected to the studies on technical grounds and questioned the methods of sampling, interviewing, etc. Dr. Taube analyzes the reasons for the "generally accepted failure" of use studies by establishing a distinction between consumer services and professional services. He concluded that the organization and dissemination of scientific information is a professional activity, and that such responses cannot supply directions for the design of more effective scientific information and reference systems.

Consumer acceptance of information systems is without a doubt important. This is, however, primarily a packaging problem and a user education problem. A consumer will tend to buy an attractive box and will not want something he finds difficult to use.

User studies based on interviews with scientists in order to determine where they seek information fit in the consumer acceptance area, but use studies based on analysis of the kind of reference questions a library receives more nearly apply to the design and evaluation problem. Saul and Mary Herner (Herner and Company) have conducted both kinds of studies. Especially the latter, they believe, can be used in the design or "tailoring" of classification systems. Quoted below are two paragraphs from a report by the Herners; in this statement is outlined the kind of useful knowledge they believe can be gained from user studies.[16]

"Unfortunately, for the time being, we have to content ourselves with by-products. The first was an analysis of the subject content and logical structure of the reference questions we had collected in our course of the study (2)*. The two most striking findings of this by-product study were that over a fifth of the questions, all of which had been delegated to nuclear energy libraries, were completely non-technical and the bulk of the technical questions involved two concepts, these concepts being almost always related as logical products. Of course, it can be argued that the questions we collected were addressed to manual retrieval systems that did not lend themselves to multiple-subject correlations. Whether requestors would address more complex questions to correlative systems is a question worthy of study on a rigorous basis and susceptible to such study.

"Our second by-product -- the finding of small coincidence between nuclear energy reference questions and nuclear energy reports -- is also worthy of further consideration. If, as we suspect, reports are almost exclusively tools of current awareness having little retrospective value, the tremendous and costly efforts that are being made to organize reports for searches would seem to be futile in the extreme. On the other hand, efforts to use subject analysis of reports as a means of pin-pointing current disseminations to interested individuals or groups would seem to be extremely promising. Both would make useful subjects of study."

7.5 Conclusions

If a distinction is made between the internal and external factors in information systems, it is seen that both indexing and user studies fall in the "external". Index studies, however, are "internal."

---

* Reference here is to an earlier study, "An Experiment in the Use of Reference Questions in the Design of a Classification System." [17]

That indexing studies are external criteria of system design is not generally recognized. As mentioned in Paragraph 7.2 above, the indexer's experience and knowledge, both of the system and of the subject, play an important role in indexing. Just as consumer acceptance or user requirements are subjective, so is indexing.

The internal and external factors, of course, are not mutually exclusive. For example, if it can be determined that a cross-reference structure increases the consumer acceptance of a system, the design of the system could be affected.

The key to an understanding of the interplay of internal and external criteria is found in the distinction between individual short-term and massive long-term consumer response.[18] The designers of information systems must exhibit a professional competence which implies the ability to utilize internal criteria as a measure of consumer acceptability; and they can only be required to question their criteria if massive and long-term consumer dissatisfaction proves the criteria and the systems based upon them to be inadequate.

References

1. Bourne, C.P., G.D. Peterson, B. Lefkowitz, and D. Ford. Requirements, Criteria, and Measures of Performance of Information Storage and Retrieval Systems. Stanford Research Institute, SRI Project No. 3741, December 1961.

2. Taube, Mortimer. Evaluation of Information Systems for Report Utilization. Studies in Coordinate Indexing, Vol. I, Documentation Incorporated, 1953, pp. 96-110.

3. Jonker, Frederick. The Descriptive Continuum -- A "Generalized" Theory of Indexing. AFOSR TN 57-287, prepared by Documentation Incorporated, under Contract AF 49(638)91, June 1957.

4. MacMillan, Judith T. and Isaac D. Welt. A study of Indexing Procedures in a Limited Area of the Medical Sciences. Paper presented before the annual meeting of the American Documentation Institute, Berkeley, California, October 25, 1960.

5. Maron, M.E. Automatic Indexing, An Experimental Inquiry. The RAND Corporation, Santa Monica, California, January 1961.

6. Cherenin, V.P. Certain Problems of Documentation and Mechanization of Information Search. Moscow, 1955, pp. 3-37, 74-76.

7. Luhn, H. P. The Automatic Derivation of Information Retrieval Encodements from Machine-Readable Texts. IBM Advanced Systems Development Division, Yorktown Heights, N.Y., 1959, 9 p. (ASDD Report L #438.)

8. Cleverdon, C.W. The Aslib Cranfield Research Project on the Comparative Efficiency of Indexing Systems. 35th Annual Conference, Brighton, September 1960.

9. Schüller, J. A. Experience with Indexing and Retrieving by UDC and Uniterms. 35th Annual Conference, Brighton, September 1960.

10. Armed Services Technical Information Agency. Conference on Multiple Aspect Searching for Information Retrieval. Washington, D. C., Feb. 12-13, 1957.

11. Stiles, H. Edmund. The Association Factor in Information Retrieval. Journal of the Association for Computing Machinery. Vol.8, No. 2,

April 1961.

12. Stevens, Mary E. The Domain of Information Selection and Retrieval Research. NBS Report 6683, Feb. 15, 1960. (unpublished)

13. Schultz, Claire K. and Clayton A. Shepherd. A Computer Analysis of the Merck Sharp and Dohme Laboratories Indexing System. American Documentation, Vol. 12, No. 2, April 1961, pp. 83-92.

14. Schultz, Claire K., Phyllis D. Schwartz, and Leon Steinberg. A Comparison of Dictionary Use within Two Information Retrieval Systems. American Documentation, Vol. 12, No. 4, October 1961, pp. 247-253.

15. Taube, Mortimer. An Evaluation of "Use Studies" of Scientific Information. Report No. AFOSR-TU-58-1050, Contract AF 49(638)91. AD 206987.

16. Herner, Saul and Mary Herner. Determining Requirements for Atomic Energy Information from Reference Questions. Proceedings of the International Conference on Scientific Information, Washington, D.C. National Academy of Sciences, National Research Council, 1959.

17. Herner, Saul and Mary Herner. An Experiment in the Use of Reference Questions in the Design of a Classification System. Report to National Science Foundation, January 4, 1962.

18. See 2.

8. MECHANIZING COORDINATE INDEX SYSTEMS

8. MECHANIZING COORDINATE INDEXING SYSTEMS

Because of the intrinsic simplicity of a coordinate indexing system, there are many methods and techniques of mechanization. In order to fully appreciate this simplicity, a brief review of the principles of coordinate indexing and their implementation is presented below.

8.1 Description and Definition

In the original Uniterm system of coordinate indexing,[1] (1) the index term was a class name (airplane, computer, control panel, lighting, etc.), (2) the posting in the system was item-on-term and not term-on-item, and (3) the user found relevant material by matching or comparing item numbers.

This is still a correct description of a coordinate indexing system, but certain variations and modifications have appeared. Now, for example, there are also coordinate indexing systems which compare term numbers in term-on-item systems. Furthermore, although all coordinate indexing systems use class names,* certain hierarchical classification schemes have been brought into play. This is discussed in Sections 4 and 5 of this report.

---

* "Class names" covers all the following: Uniterms, descriptors, keywords, unit terms, selectors, locators, etc.

A coordinate index is distinguished by the fact that the user himself (perhaps with mechanical assistance) performs the coordinations. He does this with terms and items coded and arranged in such a way that they can be readily susceptible to logical operations.

8.2 Logical Operations

A typical logical operation of coordination is finding the logical intersection: this class and that. However, all the Boolean functions may be represented by machine operations in the system, as required by the search techniques.

The possible logical operations are listed below:

1. This class and that (intersection)
2. " " or " (union)
3. " " but not that (negation)
4. Combination of this class and that plus either or both of two others (combination of logical intersections and/or logical unions)
5. Other complex combinations of (1), (2), (3).

8.3 Coding and Arrangement

Coding and arrangement vary from system to system. Certain generalizations may be made, however, about all coordinate indexing systems.

Either the term or the item or both may be designated by a number code. In an inverted (item-on-term) system, where terms are compared

for item matches, it is obviously expedient to have the item in the briefest possible form. If a six-digit number is used, 999,999 items can be designated by means of only the numbers 0 through 9.

The same holds true in term-on-item systems where it is expedient to designate terms by, say, a six-digit number.

Some systems use numbers for both the term and the item. Others leave either the item or term, whichever is not "read" for matching purposes, in alphanumeric form.

In manual systems, the item and/or term numbers are usually Arabic numerals. However, as in Batten or peek-a-boo and edge-notched cards, these numbers may then be designated by hole positions.

In mechanized systems, the Arabic numbers become binary numbers and are coded by various means. Binary numbers are used because they require only two codes, one for zeros and another for ones.

(1) In punched card or tape systems, the binary numbers are coded by "hole" and "no hole".

(2) In microfilm systems, the binary numbers are coded by opaque and transparent markings.

(3) In magnetic tape or magnetic card systems, the binary numbers are coded by magnetized and nonmagnetized areas.

In a mechanized system, the search question must also be coded. Not only must the terms be specified, but also the logical operations.

The logical operation may also be coded by binary numbers, e.g., 01 = union, 11 = intersection, 10 = negation. There are, of course, many other instructions and codes that are used, e.g., in programming a computer, but we will not concern ourselves with these.

The method of carrying out mechanized searches varies with the arrangement of the store: whether it is a term-on-item or an item-on-term system and whether codes are randomly or sequentially ordered. This will be discussed in further detail in Paragraph 8.5.

Because, as we have seen, coordinate indexing systems rely only on logical operations with numerical codes for terms and items, they are ideally suited to mechanization. When roles, links, or semantic codes are introduced into a system, certain modifications must be made, but the procedures remain generally the same.

There are many means presently available for mechanizing or partially mechanizing an information storage and retrieval system based on coordinate indexing. Devices currently used range in sophistication from high-speed, large-capacity, random-access, general-purpose computers to comparatively simple devices such as Jonker Business Machines, Inc.'s Minimatrix.

There are those who feel that the most significant lack in mechanized systems is that of an automated input -- even to the extent of the selection of indexing terms.

8.4 Input

Index terms must be entered manually into the system for conversion to machine-manipulable codes or for position-coding, such as on peek-a-boo cards. Even in microfilm systems where text can be entered directly into the system, manual indexing and coding is still required. It is sometimes implied that in a microfilm system complete searching of text is possible. What is actually the case, however, is that once an apparently relevant item has been retrieved via the coded index, the user can scan the immediately available text. Examples are devices such as Minicard, Filesearch, and the Rapid Selector, which will be discussed in greater detail in Section 2d. Another more recent development, reported by National Cash Register, is its high-density document storage system utilizing a photochromic microimage memory.[2] In none of these, however, is manual input or manual coding eliminated; and at the present time all searching devices make use of some kind of code: hole position, numbers, optical patterns, binary codes, etc.

8.4.1 Coding Devices

Codes are prepared by various means. The most common devices are IBM or Remington-Rand punched card or tape machines, Justowriters or Flexowriters (tape-producing typewriters), the Jonker 400 Termatrex card hole-puncher, and similar devices. These are neither high-speed nor "automatic", i.e., all require operators and are

limited by the operators' speed.

IBM is investigating the possible use of "Stenotype" in I.R. Stenotype is used by stenographers; it produces a paper tape in which words are abbreviated or "coded" in shorthand form.

Optical and magnetic coding devices will be discussed in greater detail in Paragraph 8.5; these, too, rely on manual input of data.

8.4.2 Character Recognition

It has been proposed that character recognition devices could be used as input devices -- to eliminate the manual operations of coding input data such as index terms or full text. Most existing readers, however, read only text which has been specially prepared manually. That is, the codes or "characters" are manually imposed on the item; examples of this are the American Banking Association's magnetic-ink character readers,[3,4] or the British "Luton Experiment" phosphorescent-coded-mail sorting devices.[5,6,7] The character readers, then, are simply analogous to punched-card reading equipment.

Optical scanners on the market manufactured by Farrington Manufacturing Company, like the devices mentioned above, can only read specially prepared text, or, at best, only selected type fonts.

Existing character readers do not accomplish either of the two things which could make them useful in I.R.: (1) code automatically (i.e., prepare machine input directly from index terms or full text),

or (2) eliminate coding by permitting direct search of printed or written indexes or texts.

There are, however, several competing firms engaged in research on or development of character readers. In August 1960, the Wall Street Journal reported activities by IBM, NCR, RCA, Remington-Rand, Farrington, and Baird-Atomic.[8] Work is being performed by Bell Telephone, Sandia Corporation, and the University of Michigan.[9] Rabinow Engineering, Philco Research Center, and General Electric Computer Laboratory are also engaged in character recognition studies. This is only a partial list of U. S. companies involved. Some foreign companies engaged in this work are Compagnie des Machines BULL (France), Electronic and Musical Industries (England), and Telefunken (Germany).[10]

One of the more advanced readers, a working developmental model of which has been completed, is Baird-Atomic Inc.'s optical print reader. Although this was designed to meet requirements of the Air Force machine translation program, with very little modification it could be used in indexing, searching, and storing. This equipment generates a magnetic tape directly from text on 70-mm. film. This is, in effect, producing the machine-manipulable codes mentioned above. Because the tape produced carries all printing information (fonts, spacing, etc.), it could also be used to operate output printers for graphic-arts quality printout. As will be discussed in Paragraph 8.6, such printers are still

in experimental stages. In fact, in order to check the magnetic tapes produced by its character reader, Baird-Atomic finds it necessary to punch a paper tape and then read this on a paper-tape reader. The principle of operation of the reader was described in a report to the U. S. House of Representatives Committee on "Science and Astronautics."[11] Other printed reports are not available; therefore, some additional information obtained during a visit to the company is detailed here.

Certain limitations of the system were brought to light, not the least of which is the cost. Final models will probably range from $200,000 to $400,000. A limited market is anticipated because of the speeds available; one man at Baird-Atomic predicted that two machines would handle "all the Russian material". Although speeds vary according to the number of fonts being read, a reported minimum speed is 60 char/sec.

At present, the machine works only from 70-mm. film but could be modified to handle some other film, for example, 16-mm. microfilm, but cannot read opaque material.

The machine could also be modified, and perhaps with some cost saving, to read only one or two type fonts. (This might be useful for handling typewritten abstracts or documents.) Twelve fonts are accommodated in the developmental model. This means that at any one time 12 fonts can be read. The 12 fonts can be chosen from any number

of fonts, and preparation of font discs is relatively simple; one method involves sandwiching a piece of film between two layers of glass. Type sizes, however, are critical, because the appearance of the letters varies from size to size and a font disc is required for each size. There is no way to compensate for this size variation by using, for example, different reduction ratios.

The Baird-Atomic character reader does not "read" graphs, drawings, or vertical or slanted printing.

Despite the limitations enumerated above, Baird-Atomic's print reader represents significant progress in the field of character recognition.

What may be called somewhat more sophisticated techniques of character recognition are being investigated by Rabinow Engineering. Their work has been primarily in pattern matching techniques (as in the Rabinow Universal Reader). Their method differs from other matching techniques in that upon converting the given character into an electronic replica and comparing the replica to a standard set of stored electronic images, a quantitative measure of the best match is obtained. Rabinow is also exploring curve tracing, the character elements either being traced by a spot of light (generated by a cathode ray tube, for example), or a radar tracking technique which is used to follow the lines without the aid of a moving light source. The directions in which the character

elements go are then recorded, and the character is recognized by the sequences and lengths of the curve-traced elements.[12] These techniques are potentially applicable to the reading of handwriting; machines which rely on exact match do not have this potentiality.

In this report, we have only intended to show the level of achievement and to indicate what use coordinate indexing systems may make of character readers. For a full summary of the state-of-the-art of character recognition, the reader is recommended Mary Stevens' report published by the National Bureau of Standards.[13]

In conclusion it may be said that although considerable time and effort must still be expended in perfecting a variable-font, full-page, high-speed print reader, sufficient advance has been made to permit reasonable anticipation of such a device.

8.4.3 Automatic Indexing

Luhn's keyword-in-context (KWIC) indexes are discussed elsewhere in this report; these have proven practical in limited application. The technique is more nearly a means of mechanically producing an index than automatic indexing. Luhn is, however, also working on mechanical methods of indexing and abstracting by "counting" or determining the frequency of occurrence in documents of "notion" words.[14,15] A similar method is used in permuted title word indexing.[16]

That such a method will be suitable for many types of users is probable. This method cannot be adequately evaluated, however, in terms of practicality and economy, until automatic character readers become a reality.

8.5 Store and Search

Once the material of the system has been indexed, it must be incorporated in the total store and must be retrievable. Methods of entering the data into the store have been briefly discussed above and will be more fully described in this section.

In some coordinate indexing applications, both the index and the material indexed (text) are part of the "store"; in others, only the index is the "store". In the latter case, the physical store of material indexed is never actually searched. Actually, this is also true in the first case since, whether the text is stored in the system on microfilm, magnetic tape, or otherwise, it is not searched, but is merely available for print-out or on-the-spot scanning. Thus, for our purposes, only the index store is of interest.

8.5.1 EAM and Manual Systems

Many physical forms of coordinate indexes are well known and will not be detailed here. Some of these are Uniterm cards or printed double-dictionary Uniterm indexes,[17] the rather recent Tabledex indexes,[18, 19] and the familiar peek-a-boo[20] and edge-notched

cards.[21] (The references here are to recent or definitive literature; there is, however, considerable other material available which is referenced throughout the bibliography, Section 12 of this report.)

The NBS Microcite Machine is based on an interesting extension of the peek-a-boo principle. In the usual peek-a-boo system, the position of a hole is interpreted as a document serial number. In the Microcite concept, the searcher views a description of the document at each hole position.[22] The description is on microfilm which is projected and enlarged on a screen.

By means of other coding techniques, coordinate indexes may also be stored on punched cards (aperture cards fall in this class), punched tape, COMAC cards, magnetic tape, discs, or other magnetic media, or on microfilm. From all of these, "hard copy" indexes can be generated. (COMAC, also known as the IBM 9900 Special Index Analyzer, is described in "Studies in Coordinate Indexing," Volume V.)[23]

The search means is, of course, dependent on the storage means. Where the index is stored on peek-a-boo cards, for example, search consists of manually comparing term cards. Where IBM punched cards are used, decks can be compared by machine or manually. Most card systems use EAM (electronic accounting machines) for storage and retrieval, but the COMAC and two other IBM devices, the IBM-9310 Universal Card Scanner

and the IBM-101 with Row-by-Row Scanning Attachment (a modification of the IBM-101 Statistical Sorter), could also be used.

8.5.2 High-Speed Magnetic Tape Computers.

Where the index is stored on magnetic tape, a computer is used for the index coordinations, or a special magnetic tape searcher can be used. Where the index is on microfilm, various photoelectric code-sensing devices are used. Since magnetic tape and other magnetic media and microfilm are of special interest because of potential high density of storage and speed of search, these will be discussed in detail below.

There has been considerable debate over the use of high-speed magnetic tape computers for information storage and retrieval. The strongest adverse argument lies in the cost of this equipment. When one is faced with a choice between a high-speed system which rents for from $1,000 to $300,000 per month and a punched card system whose total cost could be less than a year's computer rental, he will obviously be inclined toward the latter. However, one must keep in mind the fact that computers work so quickly that they are required only for short periods, even for large collections of material. Thus, if the computer can be used on a time-sharing basis, it can be completely practical. It has been proposed that before ADP equipment can be used in libraries, its cost must be reduced by a factor of 10.[24] This statement, however,

does not take into consideration conditions such as time-sharing.

Many computers do not have random access; i.e., the entire file must be searched serially. This can be compensated for by "batching" queries, but many users are reluctant to hold questions for even as little as a day, claiming, and possibly with justice, that a high-speed computer should give not only rapid but immediate replies.

Some users have put their entire reference file on magnetic tape, but use this only for periodic printout of indexes and use the printout for searching manually. This is economical insofar as needless repetition in sorting, filing, etc., is eliminated, and the computer itself updates the system. Obviously, this type of operation would be far more attractive if the computer were also used for processing queries.

Both success and disappointment have been reported in the use of high-speed computers for I.R.[25] In some cases, after review of operations it was found that by making certain changes--batching queries, time-sharing, memory modification--the disappointment could be minimized or eliminated.

Generally, the "permanent" index is stored on magnetic tape. In processing or searching, however, a temporary store or memory (magnetic core or drum) is used to hold the queries and/or program as well as results for the search process. An important factor in I.R. systems is the number of searches which can be run simultaneously. The IBM 7090,

which is used at the G.E. Flight Propulsion Division, Technical Information Center, can handle as many as 1,300 questions simultaneously. This requires, of course, powerful logic and large internal memory, but means that the processing is done in a single run through the magnetic tape files. This is important because, too often, powerful computers are slowed down by tape speeds.

Exhibit A shows several characteristics of computers in the $5,000 to $50,000 monthly rental range. This chart was specially prepared with I.R. applications in mind. Other computer-characteristics charts are available; one of the most complete is that prepared by Adams Associates.[26]

IR systems differ in configuration according to different requirements. An ADP (automatic data processing) system may have to be modified or new ones may have to be designed to fit the I.R. requirements. For example, the requirements for either multiple access or single access will have a strong influence on choice of system. In some cases, large internal storage is required; in others a large temporary memory is demanded for processing either internal or external stores. These problems and others, in any combination, may be present.

It is, therefore, essential that a complete systems study be made before a computer is purchased. In the past, disappointment has been registered in cases where this was done or where a company already had

an ADP system and tried to force it to fit the I.R. requirements. A systems study requires examination of hardware which can provide sufficient capability for all functions. There is more to this than "shopping" for a computer. High-capacity disc files and large memory drums, for instance, are available but may not be inherent in an otherwise appropriate ADP system. Exhibit B shows, by way of example, characteristics of just two large-capacity random-access disc memories. One or the other of these may best suit a particular need. Also by way of example, characteristics of two kinds of magnetic storage drums are shown in Exhibit C. In many systems, high-density tape systems will be required. If one compares the various characteristics of magnetic tape systems on Exhibit A, he can see the advantages of mix-and-match system techniques. The magnetic tape system of the Bendix G-20, for example, may be just as useful with another computer. This is one of the higher-density tape systems available, and its characteristics are outlined in detail in Exhibit D.

Many people have been misled into believing that a computer will not solve I.R. problems simply because they see no computer that <u>has</u> solved them. The need, therefore, for systems study and systems design cannot be overemphasized. Because these criteria have not been met, a <u>best</u> solution for a given problem has not been realized.

EXHIBIT A

CHARACTERISTICS OF REPRESENTATIVE SYSTEMS
IN $5,000 - $50,000 MONTHLY RENTAL RANGE

Computer System	AVERAGE MONTHLY RENTAL	WORD SIZE	INSTRUCTION-ADDRESS	INDEX REGISTERS	RANDOM ACCESS	RANDOM ENQUIRY	INTERRUPT CAPABILITY	INDIRECT ADDRESSING	AVERAGE ADD TIME	MEMORY	AVERAGE ACCESS TIME	CAPACITY	BULK STORAGE	AVERAGE ACCESS TIME	CAPACITY	MAGNETIC TAPE UNITS	TRANSFER RATE (6-bit char/sec)	RECORDING DENSITY	RECORD BLOCK	INTERRECORD GAP LENGTH	INPUT OUTPUT CHANNELS	BUFFERING *	INPUT PUNCHED CARDS cards/min	INPUT PAPER TAPE char/sec	OUTPUT PRINTER lines/min
PHILCO 2000	$30,000	48 bits	1	32	✓	✓	✓		15 μs or 4.5 μs	1-4 magnetic core modules	10 μs or 2 μs	8,192 words/modules	1-32 magnetic drums	17 ms	32,768 words/drum	256 max. units 1" Magnetic tape	90,000	750 char/inch	128 words	INA	16	MRWC	2,000	1,000	900
BENDIX G-20	$20,000	32 bits	1	63		✓	✓	✓	15 μs	1-7 magnetic core modules	6 μs	4,096 words/modules	magnetic disc units	.25 seconds (max.)	10 million char/unit	500 max. units 1" Magnetic tape	120,000	1100 bits/inch	INA	.85 in.	6	MRWC	800	500	1,000
NATIONAL CASH REGISTER 315	$8,500	12 bits	1	42	✓		✓		42 μs	1-20 magnetic core modules	6 μs	4,000 char/modules	1-16 magnetic card cartridges 256 cards per cartridge	170 ms "new card" access time	5,600,000 char/cartridge	8 max. units ½" Magnetic tape	60,000	500 char/inch	1-16,000 characters	INA	7	none	400	1,000	900
BURROUGHS B 5000	$13,500-$50,000	48 bits	4		✓		✓		3 μs	1-8 magnetic core modules	6 μs	4,096 words/modules	1-2 magnetic drums	8.1 μs/char	32,768 words/drum	1-16 max. units	66,660	555.5 char frames /inch	INA	INA	4	INA	800		700
MINNEAPOLIS-HONEYWELL 800	$22,000	48 bits	1	64		✓	✓	✓	24 μs	1-8 magnetic core modules	6 μs	4,096 words/modules	optional			64 max units 1/4" magnetic tape	64,000	1,666,666 words/tape	Variable	INA	16	MRWC	650	1,000	900
CONTROL DATA 1604	$34,000	48 bits	1	6	✓	✓	✓	✓	5 μs	1-2 magnetic core banks	4.8 μs	16,384 words/bank	INA			96 max. units ½" magnetic tape	30,000	200 char/inch	Variable	.75 in.	6	MRWC	1,300	350	1,000
RAMO-WOOLDRIDGE RW-400	$50,000	39 bits	2	✓	✓	✓	✓		24 μs	1-8 magnetic core modules	5 μs	4,096 words/modules	magnetic drums	8.5 ms	8,192 words/drum	24 max. units 1" magnetic tape	62,000	200 bits/inch	Variable up to 1,024 words	INA	9	MRWC	2,000	300	900
IBM 1401	$6,500	alpha-numeric	2	3			✓		230 μs	magnetic cores	11.5 μs	1,400 - 16,000 words	INA			10 max. units	62,500	200 or 556 char/inch	INA	.75 in.		none	800	500	600
IBM 1410	$8,000	alpha-numeric	2	15	✓	✓	✓	✓	110 μs	magnetic cores	4.5 μs	10,000 - 40,000 words	magnetic disc units	500 ms.	100 million characters	20 max. units	62,500	200 or 556 char/inch	INA	.75 in.	2	RWC	800	500	600
IBM 7090	$64,000	36 bits	1	3	✓	✓	INA	✓	4.4 μs	magnetic cores	712 μs	32,000 words	INA			80 max. units	62,000		INA		8	MRWC	250		150
RCA 301	$5,000	alpha-numeric	2	1	✓		INA	✓	189 μs	magnetic cores	7 μs	10,000 - 20,000 words	1-5 magnetic disc units	INA	4.6 million char/unit	12 max. units	75,000		INA		2	RC, WC or RW	600	100	600
RCA 601	$32,000	56 bits	1	8	✓	✓	INA	✓	6 μs	1-4 magnetic core modules	9 - 1.5 μs	8,192 words/modules	INA			63 max. units	120,000		INA		16	MRWC	600	1,000	900
UNIVAC FILE COMPUTER I	$12,000	12 alpha numeric	3	0	✓	✓	INA		8.6 ms	magnetic cores	.9 ms	2,000 words	1-10 magnetic drums	17.6 ms	180,000 char/drum	31 max. units	10,400		INA		10	RWC	240	150	600
UNIVAC III	$20,000	6 decimal digits	1	15	✓	✓	INA	✓	9 μs	magnetic cores	4.5 μs	8,000 - 32,000 words	INA			32 max. units	133,000		INA		5	MRWC	700		700

* R = Reading
W = Writing
C = Computing
MRWC = Multiple reading and writing with simultaneous computing

EXHIBIT B

COMPARISON OF TWO LARGE-CAPACITY RANDOM-ACCESS DISC MEMORIES

	TELEX, INC. MODEL IIA	BRYANT COMPUTER PRODUCTS MODEL 4200
Disc Diameter	31"	39"
Rotational Speed	1200 rpm	900 rpm
Recording Surfaces	128	39
Number of Heads	256	234
Head Positioners	64	1 per disc side
Tracks per surface	256	768
Bit density (max.)	400 bits/inch	273 bits/inch
Track density	25.6/inch	64/inch
Storage Capacity		
Per file	617,644,032 bits	603,857,592
Per surface	4,825,344 bits	15,483,528
Per Track		
Zone 1	12,566 bits	11,575 bits
" 2	25,132 "	15,015 "
" 3		18,427 "
" 4		21,840 "
" 5		25,279 "
" 6		28,665 "
Transfer rate		
Zone 1	251,320 bits/sec	174 kc
" 2	502,640 "	225 "
" 3		276 "
" 4		328 "
" 5		380 "
" 6		431 "
Access Time	42 ms(average)	167 ms(max)
Price	$185,000	$140,041

EXHIBIT C

RANGE OF SPECIFICATIONS OF
BRYANT COMPUTER PRODUCTS'
MAGNETIC STORAGE DRUMS

	MASS MEMORY	GENERAL MEMORIES
Capacity (per drum)	up to 6,210,500 bits	20,000-2,500,000 bits
Tracks	up to 825	40-420
Speed	900-3600 rpm	600-24,000 rpm
Size	18.5" dia. x 34" long	5" dia. x 2" long 10" dia. x 19" long
Access Time	as low as 2.5 ms	as low as 16.6 ms

EXHIBIT D

CHARACTERISTICS OF POTTER INSTRUMENT COMPANY 906 II HIGH-SPEED DIGITAL TAPE HANDLER AND HIGH-DENSITY RECORDING SYSTEM

Bit density	up to 2,000/inch
Tape Speed	up to 150 inch/sec
Number of Channels	up to 20 per inch of tape width
Interchannel Time Displacement	less than 0.2 ms at buffer output
Interblock Gap	as short as 0.3"; 0.75" typical for dual read/write operation at 100 in/sec.
Error Detection	Parity channel provides single error detection
Error Correction	Single parity channel makes possible single error correction
Reliability	
Transient Error Rate	1 in 10^7 to 10^8 max. at 1500 ppi.
Permanent Error Rate	1 in 10^8 to 10^9 max. at 1500 ppi.
Reread time to recover transient errors	less than .005% of on-line time at 1500 ppi.

It has been suggested that certain deficiences must be tolerated in I.R. systems in order to expend only reasonable funds. This was pointed out in arguing for the general-purpose computer as opposed to special devices like the COMAC which, it was said, have limited application.[27] With this we cannot totally disagree, but we must insist that even in using a general-purpose computer, one need not take second best just because it exists, since first best might be effected by system modification and/or time sharing.

This has been but a very brief summary of the possible applications of magnetic tape computers to coordinate indexing systems. In the appended references are cited several items which contain additional or more detailed information.[28,29,30].

In Section 9 of this report the operating experience of several facilities which use magnetic tape computers is detailed.

8.5.3 Other Magnetic-Media Systems

Magnetic Tape Searchers

Several efforts are underway to develop file-searching devices based on magnetic tape systems. These are attempts to make special-purpose devices which will do the same processing for I.R. purposes that a computer does, but at a lower cost. It has been reported[31] that most of these developments are too high priced to compete with several moderately priced computers; an exception noted is Herner's Tape

Searcher (approximately $10,000).

Some of the tape systems under development are:

Logic Processor (Aeronautics)
Index Searcher (Computer Control Co., Inc.)
Univac Tape Searcher (Remington-Rand)
Findafact (Rese Engineering Co.)
Tape Searcher (Herner & Co.)

The GE-250 Information Selector was designed and developed for Western Reserve University's Center for Documentation and Communication Research. However, in order to meet the required installation date, the GE-225 was delivered. The GE-225 is a transistorized general-purpose digital computer with a special programming feature which allows the WRU specifications to be met.[32] Work on the GE-250 has been dropped.

Magnetic Cards

The Magnacard system (Magnavox Company) stores data on 1" X 3" magnetic cards. A single card has a capacity of 1000 decimal digits or 600 alphanumeric characters. There is provision for large-scale processing of file items and for random access to individual items.

This equipment, like special magnetic tape searchers, attempts to perform the same I.R. operations that a computer can but to do so at lower cost.

Magnacards, with microfilm attached, are also being used in the Magnavue system (Magnavox Company).

Manufacturer's literature is available on all these systems.

8.5.4 Microfilm Systems

There are basically three types of microfilm systems used in coordinate indexing: those which use pieces of microfilm inserted in or pasted on cards, those which use microfilm in conjunction with magnetic media, and those which use "chips" or reels of microfilm with photo-optic codes.

The first of these will not be discussed in detail here since they are treated as punched cards (e.g., aperture cards used in systems like Filmsort) or are purely manual systems. The Filmsort Company (Division of Minnesota Mining & Mfg. Co.) aperture cards are punched cards on which microfilm is mounted. These are sorted, collated, etc., on standard EAM equipment. (See Paragraph 8.5.1.)

The second of these is somewhat similar to the first in that the microfilm is attached to a card, but this card is of magnetic material (such as Magnacard) and is processed much like magnetic tape. (See Section 8.5.) An example of this configuration is in Magnavue (Magnavox Company).

There are two types of the third system. In one case, the microfilm system is used only for text storage. This is of interest here only in that some of these systems do have "read" devices. For example, in MEDIA (Magnavox Company.) and in FLIP (Benson-Lehner Corp.), item numbers are coded, and the document can be mechanically retrieved by

this number. These systems at present do not provide mechanical retrievability via an index.

(A great deal of work is being done on microfilm systems for text storage, both in increasing density of storage and in making rapid access to documents possible. Much of this work is being sponsored by Council on Library Resources.[33] The AVCO Corporation is developing, with Council support, a system based on microfilm sheets of documents, with as many as 10,000 on a sheet, with 100:1 reduction of page size.)

In the second case, the microfilm is used both as a storage medium and as a coordinate index. Here, the documents are retrieved via the index. In the Rapid Selector, for example, index terms are encoded on punched EAM cards which are photographed onto reels of film, with the indexed document immediately following. The documents are retrieved by means of a punched interrogation card and a patch panel which specify (a) the search criteria and (b) the logical relationships required. The information store is then moved past the photoelectric cells of comparator circuits. When search requirements have been satisfied, a copy circuit is activated, and microfilm copies of the selected documents are made. The average time for a complete search, including processing, as reported by the Bureau of Ships,[34] is 12 minutes.

There are several other systems based on this principle of photoelectric comparison. Some use different coding means (though all rely on opaque and transparent markings indicating binary information) while others employ microfilm cards rather than reels of film. Some of these other systems are:

Filmorex (Filmore & Co., France)
Minicard (Eastman Kodak Company)
Filesearch (FMA, Inc.).

Manufacturer's literature is available on these devices, as well as on MEDIA and FLIP.

Additional information on microfilm systems can be found in a survey report prepared by System Development Corp.[35]

8.6 Display or Printout

In the publication or other dissemination of indexes, references, abstracts, etc., a major shortcoming has been in the preparation of graphic-arts quality printout from processing devices such as those described in Section 2.

8.6.1 Typewriter-Printers

Output devices of data processing equipment are generally typewriters, paper-tape punches, or card punches. Some systems, like the IBM 870 Document Writing System, use all three. Various converters are used to transfer information from one form to another, e.g., tape-to-tape converters.

Typewriting can be generated on cards, sheets of paper, labels, preprinted forms, etc. Such cards have been widely used in the "shingling" type of copy preparation with Listomatic cameras and for other camera copy.

The disadvantages of typewriter output devices now in use are that only one or a limited number of type fonts can be used and that no provision is made for proportional spacing and other graphic-arts printing requirements.

Typewriter printers operate at speeds up to 1000 lines per minute (tape and card-punches are slower). This is, however, not as fast as many computers can operate. IBM, Remington-Rand, and others are working on the development of faster typewriters and punches.

Examples of kinds of output required from mechanized systems and descriptions of means of producing indexes, etc., are given in Section 9 of this report in which systems operation and operating experience is detailed.

A display system of some interest is not reported on the charts. This is a means for projection of visual data onto a screen either for viewing or for recording (producing hard copy). It is a xerographic technique which may find commercial application in the computer field. It is described in detail by Mott, Clark & Dessauer in Photographic Science & Engineering.[36] This is the technique of the PROXI system

(Projection by Reflection Optics of Xerographic Images) (Haloid Zerox Inc.).

Loewe, Sisson and Horowitz have published a summary of display techniques, including CRT, photographic, electrostatic oil film, and thermoplastic. Basic principles and typical characteristics are given. There is an extensive, comprehensive bibliography. The paper also discusses user requirements.[37]

It should be noted that graphic-arts printers are desirable primarily for printing references, citations, abstracts, or text. They are usually not required for printing out search results (item addresses) or for printing out simple indexes. Therefore, printing of high quality will only be a truly practical endeavor when the cost of input of references and citations, as well as printing instructions, is considerably reduced. (See Paragraph 8.4)

8.6.2 Automatic Composing Machines & Character Generator/Printers

Other techniques are now being widely investigated for graphic-arts quality page composition; they fall into two major categories: (a) automatic composing machines and (b) character generator/printers. While the former meet or exceed graphic-arts requirements, the latter lack clarity and diversity of type fonts. On the other hand, the latter operate at much higher speeds.

Summarized in Exhibit E are characteristics of some representative automatic composing machines. In Exhibit F are summarized some characteristics of representative character generator/printers. These charts list potentially applicable printers, but are not intended to be exhaustive. The information was obtained primarily from manufacturer-provided literature.

It will be noted that automatic composing machines, though not actuated by computer tapes, can be tape-actuated. It may be expedient to use tape translators or to modify the printers' tape acceptance equipment and use several machines in parallel, thus compensating for their slower speeds. This is an economically feasible approach since costs of electronic printing systems such as described in Exhibit F range from $250,000 to $600,000. On the other hand, the cost of an automatic typesetting or photocomposition unit ranges from $20,000 to $60,000. Thus as many as ten of the latter could be used at the same cost.

MIT has reported success in programming its computer for printout on the Photon Photosetter.[38] All printing instructions -- font, spacing, etc. -- must be included in programming.

EXHIBIT E

AUTOMATIC TYPESETTERS

COMPANY	DEVICE	DESCRIPTION	INPUT	OUTPUT	OUTPUT RATE	FONTS	METHOD OF CHANGING FONTS	PROP. SPACING	JUSTIFIED MARGINS
Intertype	Monarch	Keyboardless, tape-actuated, line-casting machine	Perforated tape, wire-service tape	Slugs, Galley	8-14 lines per minute	2,3, or 4 (90-channel) magazines at one time, 5-12 print-sizes	Automatic among 2, 3, or 4, magazines	yes	Automatic
Intertype	Fotosetter	Keyboard actuated photosetter. System consists of circulating matrix in Fotosetter, Fotomat & Camera, Magazines, Line Punch & Correction Box.	Operator at 114-button keyboard triggers matrices from 2 adjacent magazines, thus providing 228 keyboard characters.	Film or Photographic paper for engraver or platemaker.	480 ch/min	4 magazines (117-channel) 18 type sizes from single-font. 150 type faces available.	keyboard triggers matrices from two adjacent magazines	yes	Automatic
Mergenthaler Linotype	Linofilm	Keyboard unit (operator) produces 15-channel Linofilm tape which contains all instructions for photo-unit. System also has tape editor tape corrector, composer and platemaker.	Operator produces 15-channel Linofilm tape. Photo-unit can handle tape from several keyboards.	Film or photographic paper for engraver or platemaker.	15 lines per minute (approx. 720 char/min.)	18 at a time in grid turret. 100 fonts available	Operator punches tape instructions at keyboard	yes	Automatic
Photon	Photosetter	Flexowriter tape-actuated photosetter. (Information on system components not available)	Flexowriter tape	Film, Photosensitive paper, Lead type	8-14 ch/sec.	16 fonts per disc. All hot-metal fonts available	Operator punches tape instruction at keyboard	yes	Automatic

EXHIBIT F

ELECTRONIC DISPLAY-PRINTER UNITS

COMPANY	DEVICE	DESCRIPTION	FONTS	INPUT	MEANS OF GENERATING CHARACTER	OUTPUT	OUTPUT RATE	PROP SPACING	JUSTIFIED MARGINS
A. B. Dick	Videograph	Electrostatic printing tube for high-speed printing of documents from either graphic or digitally coded stores. The videograph process involves the selective charging of a current window in a CRT. Charge pattern appears on application of developing powders.	Variable type face and upper and lower case characters. To customer's specifications. Maximum of 64 possible characters are printed on 8 x 8 matrix in monoscope or aluminum-plated glass target	Hard-copy, Videotape, Perforated tape, Magnetic tape, etc.	Scanning raster in monoscope tube	Alphanumeric video forms which can be printed out or CRT displayed.	20-30,000 char/sec.	yes	yes
CBS	VIDIAC GA-1000	Solid-state video character generation. CRT display for photographic reproduction. Wired core plane character matrix.	512 characters; additional available (within limits of 50 line per character resolution) with plug-in elements.	Computer magnetic tape	Function generator giving a character formed from a line raster	CRT display which is photographed and printed full-size	150,000 char/min.	yes	yes
Recordak Corp.	DACOM	Decoder translates magnetic bits. Bank of 60 (or more) monoscope tubes provide message signal which passes via a video amplifier to a CRT display	Bank of 60 monoscope tubes represents an alphabetic, numeric, or special character in any language desired. Additional banks can be added for additional symbols.	Computer magnetic tape	Electron beam scanning; one letter per monoscope tube.	CRT display which is electronically photographed for microfilm records	20,000 char/sec.	no	(registration)
Data Display Corp.	dd 51	Solid-state video character generation.	43 character alphanumeric repertoire. 63, 127, or special character generators available. Displays 4 sizes at a time.	Computer output	Function generator	CRT Display	150,000 char/sec	no	(registration)

References

1. Taube, Mortimer, & Associates. Studies in Coordinate Indexing. Documentation Incorporated, Washington, D. C., 1953.

   Vol. I - III Studies in Coordinate Indexing
   Vol. IV The Mechanization of Data Retrieval
   Vol. V Emerging Solutions For Mechanizing the Storage & Retrieval of Information

2. NCR Evolving High Density Document Storage System. Electronic News, November 13, 1961.

3. American Bankers Association, Bank Management Commission. Magnetic Ink Character Recognition; The Common Machine Language For Check Handling. Bank Management, Publication, No. 138, New York, July 1956.

4. Magnetic Ink Character Recognition; The Common Language For Check Handling. Banking, Vol. 29, August 1956, pp. 65 - 72. Computers & Automation, Vol. 5, No. 10, October 1956, pp. 10-16, 44. Journal of Machine Accounting, Vol. 8, No. 2, February 1957, pp. 10, 12, 16, 20.

5. Forster, C. F. Use of Phosphorescent Code Marks in Automatic Letter-Facing and Sorting Machines. The Post Office Electrical Engineers' Journal, (London), Vol. 54, Part 3, October 1961, pp. 180 - 185.

6. Pilling, T. and P. Horrocks. Coding Desk and Code-Mark Reader for use with Automatic Letter-Sorting Machines. The Post Office Electrical Engineers' Journal, (London), Vol. 54, Part 2, July 1961, pp. 122-129.

7. Pilling, T. and P.S. Gerard. Automatic Letter-Sorting -- The Luton Experiment. The Post Office Electrical Engineers' Journal, (London), Vol. 54, Part 1, April 1961, pp. 31-36.

8. Penn, Stanley W. Machines That Read. Wall Street Journal, August 25, 1960.

9. Report prepared by B. Adkinson and Staff (NSF) for U. S. Senate Committee on Government Operations. Published in Senate Document No. 113, 1961, pp. 95 - 144.

10. Stevens, Mary. Automatic Character Recognition, A State-Of-The-Art Report. NBS Tech. Note 112, available from OTS, PB 161613, 1961.

11. U. S. Congress, House of Representatives, Committee on Science & Astronautics. Research on Mechanical Translation. House Report 2021, Serial D, June 28, 1960.

12. Rabinow Engineering Company. Character Recognition Machines -- Principles of Operation. Washington, D. C., 1961, 13 pp. unpublished.

13. See 10.

14. Luhn, H.P. Keyword-in-Context Index for Technical Literature (KWIC Index). Yorktown Heights, N. Y., IBM Advanced Systems Development Division, 1959, 16 pp. (ASDD Report RC-127).

15. Luhn, H.P. The Automatic Derivation of Information Retrieval Encodements from Machine-readable Texts. Yorktown Heights, N.Y., IBM Advanced Systems Development Division, 1959, 9 pp. (ASDD Rept. L #438.)

16. Veilleux, Mary P. Permuted Title Word Indexing Procedures for a Man/Machine System. Paper presented for the Third Institute on Information Storage and Retrieval, American University, Washington, D. C., Feb. 14, 1961, 24 pp.

17. See 1.

18. Ledley, Robert. Paper presented as part of Panel Discussion on Published Indexes. ADI Annual Convention, Nov. 5 - 8, 1961, Boston, Mass.

19. Ledley, Robert S. Tabledex: A New Coordinate Indexing Method for Bound Book Form Bibliographies. Preprints of Papers for the International Conference on Scientific Information, Washington, D.C., NSF, 1958, Area 5, pp. 395-417.

20. Thompson, M. S. Peek-A-Boo Index for a Broad - Subject Collection. Paper presented at ADI Annual Convention, Nov. 5 - 8, 1961, Boston, Mass.

21. Hoffer, J. R. Experiences with an Edge-Notched Retrieval System. Paper presented at ADI Annual Convention, Nov. 5 - 8, 1961, Boston, Mass.

22. Stern, Joshua. Extending the Utility of Optical-Coincidence Information Retrieval Techniques. Paper presented at ADI Annual Convention, Nov. 5 - 8, 1961, Boston, Mass.

23. See 1.

24. Alexander, Samuel. The History of ADP: Library Management Impact. Paper presented at U. S. Civil Service Commission Conference on Technical Libraries and ADP, Oct. 26 - 27, 1961, Washington, D.C.

25. U. S. Congress, Senate Committee on Government Operations. Documentation, Indexing, and Retrieval of Scientific Information; a Study of Non-Federal Science Information Processing and Retrieval Programs. Washington, U. S. Gov. Printing Office, 1960. 283 pp. (86th Congress, 2d sess. Senate Document No. 113).

26. Adam Associates. Computer Characteristics Quarterly. Published quarterly in Data Processing Digest. Separate copies, updated, available at subscription rates of $5.00 per year.

27. Hayes, Robert. The Interdisciplinary Character of Information Retrieval. Paper presented at U. S. Civil Service Commission Conference on Technical Libraries and ADP, Oct. 26 - 27, 1961, Washington, D. C.

28. Opler, A. and N. Baird. Relative Merits of General and Special Purpose Computers for Information Retrieval. Proc. of The Western Joint Computer Conference, 1959, pp. 54 - 56.

29. Stanford Research Institute. Bibliographies by Charles Bourne on Mechanization of Information Retrieval. Feb. 1958. Supp. I, Feb. 1959; Supp. II, Feb. 1960; Supp. III, Feb. 1961.

30. Bagley, P. Electronic Digital Machines for High-Speed Information Searching. Master's Thesis, M.I.T. Cambridge, Mass., 1951.

31. Bourne, Charles P. The Historical Development and Present State-of-the-Art of Mechanized Information Retrieval Systems. American Documentation, April 1961, pp. 108 - 110.

32. Report prepared by Clair C. Lasher (G.E.) for U. S. Senate Committee on Government Operations. Published in Senate Document No. 113, 1961, pp. 209 - 212.

33. Brownson, Helen L. Research on Handling Scientific Information. Science, V. 132, No. 3444, December 30, 1960, pp. 1922-1931.

34. McMurray, James P. The Bureau of Ships Rapid Selector System. Paper presented at the Annual Convention of the American Documentation Institute, Nov. 5 - 8, 1961, Boston, Mass. 5 pp.

35. Heller, Elmer and Charles D. Hobbs. A Survey of Information Retrieval Equipment. System Development Corp., Santa Monica, California, SP-642 Series Report, Dec. 15, 1961.

36. Mott, G.R., H.E. Clark & J.H. Dessauer. Quick Processed Bright Displays by Xerography. Photographic Science and Engineering, Vol. 5, No. 2, March - April 1961, pp. 87-92.

37. Loewe, R.T., R.L. Sisson & P. Horowitz. Computer Generated Displays. Proc. IRE, January 1961, pp. 185-195.

38. Barnett, M.P. and K.L. Kelley. Computer Controlled Printing. Paper presented at ADI Convention, Nov. 5 - 8, 1961, Boston, Mass.

9. IMPLEMENTATION OF COORDINATE INDEXING PRINCIPLES -- REPRESENTATIVE FACILITIES

9.1 Introduction

9.2 National Conference on Social Welfare

- 9.2.1 Background and General Information
- 9.2.2 System Elements
- 9.2.3 Vocabulary
- 9.2.4 Time and Costs
- 9.2.5 Evaluation of System

9.3 G. E. (Evendale) Flight Propulsion Division's Technical Information Center

- 9.3.1 Background and General Information
- 9.3.2 Machine Programs
- 9.3.3 Indexing and Quality Control
- 9.3.4 Costs

9.4 Western Reserve University, Center for Documentation and Communication Research

- 9.4.1 Introduction
- 9.4.2 Background and General Information
- 9.4.3 Costs
- 9.4.4 Telegraphic Abstracts
- 9.4.5 Semantic Code
- 9.4.6 Coding
- 9.4.7 Example of Input
- 9.4.8 Searching Procedure
- 9.4.9 Conclusion

9.5 Documentation Incorporated

- 9.5.1 Introduction
- 9.5.2 Index to Chemical Patents
- 9.5.3 Chemical Corps Study
- 9.5.4 Atomic Energy Commission
- 9.5.5 Air Force Office of Scientific Research

9. IMPLEMENTATION OF COORDINATE INDEXING PRINCIPLES -- REPRESENTATIVE FACILITIES

9.1 Introduction

Described in this section of this report are four activities which represent four distinct and different implementations of the principles of coordinate indexing:

(1) a manual, edge-notched card system (uses term on item files) (National Conference on Social Welfare)

(2) a mechanized system using a large-scale general-purpose computer, but having no recourse to aids such as thesauri, roles, links, etc. (item on term) (G. E. Evendale)

(3) a mechanized system using a relatively modest general purpose computer, and using a thesaurus, roles, and links. (term on item) (Western Reserve University)

(4) a mechanically (IBM 1401) prepared index for manual use (item on term) (Documentation Incorporated, Index to Chemical Patents)*

Where possible, size and cost are indicated. Descriptions include details on vocabulary generation and control, on user requirements, and on analysis or studies concurrent with or leading to system implementation.

---

* In addition, brief descriptions of three other studies conducted at Documentation Incorporated are included to illustrate recent developments.

It is worthwhile noting that although each of these groups uses different equipment, different modes of indexing, different indexing aids, and has very different costs, each feels that it is accomplishing what it intends to do in an economical and effective way. This is not to say that any of them is completely satisfied, however. Each strives for improvement and cost reduction.

It must be recognized that each of these groups is facing a particular user requirement and that it strives to meet that particular user requirement.

Each system must be judged, therefore, on its own merits, and not in comparison with another. That each may learn from the other and may profit by the other's errors is, of course, self-evident.

All reports have been corrected and verified by the system operators.

9.2 National Conference on Social Welfare

9.2.1 Background and General Information

The information given herein is a summary of the material contained in the two following publications and is supplemented by information gleaned during a visit to the National Conference on Social Welfare in Columbus by the Documentation Incorporated study group.

> Hoffer, J. R. Information Retrieval in Social Welfare Experience with an Edge-Notched Information Retrieval System. Paper presented at Tenth Annual Meeting of the ADI, Boston, Mass., Nov. 6, 1961.
>
> Hoffer, J. R. Manual for a Hand-Sort Punch-Card System for Indexing Social Welfare Publications, National Conference on Social Welfare, Columbus, Ohio. May 1, 1961.

The projects in which the NCSW hand-sort punch-card system was used were:

1. Indexing of Annual Forum Proceedings 1955-1959. A pilot study to test the system with 283 manuscripts included in the NCSW publications for a five-year period. The Conference has approximately 5,000 documents in printed form.

2. Anatomy of the Twin Cities Annual Forum, May 1961. An analysis of 224 meetings held during the 88th Annual Forum of the National Conference.

3. Indexing of Conference Library Publications. Approximately 300 high utility reports, journal articles, and manuscripts were classified and indexed.

These three projects are varied in content and scope but the problem is identical, i.e., information retrieval by coordinate indexing.

The same "descriptors" were used for all three projects, but the projects were kept in separate files.

9.2.2 System Elements

The NCSW punch card system is built on the Zatoperk System (Zator Company, Cambridge, Mass.). The cards are hand-punched along the edges. At the top, under one row of holes, are two rows of the alphabet, the first letter of the lower row beginning under the fourth letter in the top row. This part of the card is called the name cipher, because it is used to code authors' names and other identifying information. Around the other three edges of the card is a single row of marginal holes, numbered 1 to 65. These are used to code the 50 descriptors and 12 categories which may be used.

Each card represents one item, or document, and its index terms. There is space on the card for an abstract or reference. NCSW is presently using only bibliographic references.

The cards are manually coded with an ordinary hand punch.

An "ice-pick" or "needle" type tool is used in retrieving information. If the pick is inserted through the deck of cards at a particular term "hole", all items indexed by that term will "fall out" of the deck. With two picks a logical product can be obtained, i.e., only cards which have "this term" and "that term" will fall out.

9.2.3 Vocabulary

In such a system, the vocabulary is necessarily limited. In this case there is room on the card edge for only 69 terms. In a system like WRU's or G. E.'s (which are described in the next two sections), the vocabulary is "limitless" and new terms can be introduced with ease. Those systems have approximately 7,000 terms and grow at will. In the NCSW system, however, the vocabulary had to be predetermined and must remain fixed.

The NCSW vocabulary was generated only after considerable debate and discussion among several persons in Social Welfare. The finally generated vocabulary has not been entirely satisfactory, as will be explained in Paragraph 9.2.5.

The Manual, referenced above, goes into considerable detail in defining and limiting the scope of the descriptors. An example is given below to illustrate the amount and kind of information that may be covered by each descriptor.

20. Health and medical

Services designed to prevent and control diseases or to promote health

casework in hospital; chronic illness; health education; maternal care; medical assistance; medical care; nursing; occupational therapy; patients; physical disability; public health;

rehabilitation of the physically handicapped; sanitation; social aspects of illness; social hygiene; social work in secondary settings. See also: #24 - hospitals, residential treatment centers; #27, #36, and #37.

There is also an indexing and searching aid which lists, in alphabetic order, the terms actually used in documents, and indicates the descriptor by which the term is coded. An example is given below.

health and welfare councils
See #8

health education
See #20
See also #46

health insurance
See #33

heart disease
See #20

Hinduism
See #40

The total list of descriptors and categories is shown on the code sheet reproduced on the following page.

EXHIBIT A

NCSW
9/15/61

CODE SHEET

Coded by: _____

Information Retrieval - NCSW Publications

1. Name of Author _____
2. Title of Article _____
3. Publisher _____
4. Year and Source _____

Descriptors and Categories [1/]

Descriptors (Circle all appropriate)

1. Adm. & Org.
2. Adults
3. Casework & guidance
4. Children
5. City & urban
6. Communications
7. Com. developmt.
8. Com. org.
9. Conferencing
10. Corrections
11. Dependency
12. Discrimination
13. Economic factors
14. Educ. -academic
15. Educ. -informal
16. Familial and sexual
17. Financing
18. Governmental
19. Group work
20. Health & medical
21. Historic
22. Human growth & behav.
23. Information retrieval
24. Institutional, and building cent. programs
25. International
26. Leisure & recreational
27. Mental health & mental illness
28. Minority groups
29. National
30. Neighborhood
31. Personnel
32. Philosophic
33. Preventive & protective
34. Private service & practice
35. Professions & relat. fields
36. Psychiatric & psychol.
37. Rehabilitative & mult. serv.
38. Research & studies
39. Rural & agricultural
40. Sectarian
41. Soc. policy & action
42. Social wk. practice
43. Societal
44. Socio-cult. factors
45. State & reg.
46. Teach. &learning
47. Volun. agency
48. Volunteers
49. Youth
50. Omnibus, no or other

Categories (circle only if major importance)

51. Values
52. Knowledge
53. Purposes
54. Methods
55. Auspices
56. Problems
57. Provision & management
58. Services
59. Spec. Prob. Groups
60. Age groups
61. Settings
62. Geog. boundaries

1/ See Manual for a Hand-Sort Punch-Card System, Appendix 3 for Definitions

9.2.4 Time and Costs

The average time to process a document of approximately 3,500 words and to prepare the cards is estimated at 27-30 minutes (coding = 15 minutes) at an approximate cost of $1.00 per document.

9.2.5 Evaluation of System

The three projects so far, according to Mr. Hoffer, indicate that a hand-punch system for indexing social welfare publications with a limited number of descriptors has value for retrieving information in social welfare. This was stated in the ADI paper referenced above, as was the following enumeration of limitations and difficulties and areas requiring further study.

> "The original list of 'descriptors' was not entirely adequate. (Some revisions were made during the projects and have been made since.) The list needs further testing especially with documents published outside the United States, and for high frequency major 'descriptors' and low frequency minor 'descriptors'.
>
> "Limiting the number of descriptors to 50 may be too restrictive for a direct coding system. (In similar systems in some scientific and technical fields from 250 to 300 descriptors are used.) It results in selection of a large number of cards on the first sort, with resulting need for additional sorts to locate sources of data on topics of limited scope.
>
> "The 'categories' were not adequately tested to determine whether they were comprehensive and mutually exclusive, whether they were valid selections and appropriately stated, and whether they should be coded. (Question might be raised

> whether holes assigned to the categories might be better used by enlarging the list of 'descriptors'.)
>
> "The possibilities of using a more complex system than 'direct coding' might appropriately be explored. If seriously considered, both the glossary or dictionary of terms and the type of card used would have to be re-examined.
>
> "How detailed should the indexing become, i.e., deep indexing."

Mr. Hoffer also pointed out in that paper that this system probably has greatest value for a general or generic library or collection such as social planning and public welfare. It has not been determined whether it would have high value in such specialized areas as child welfare, corrections, psychiatric social work or other specializations in which the user may wish greater refinement or technical analysis.

9.3 G.E. (Evendale) Flight Propulsion Division, Technical Information Center.

9.3.1 Background and General Information

Several papers have been published which describe the activities and progress at the G. E. Technical Information Center. The information given herein is a summary of material contained in such papers, especially the following two, and is supplemented by information obtained during a visit to Evendale by the Documentation Incorporated study group.

> Dennis, B. K. Dissemination Via the Automated Technical Information Center. Publication property of American Chemical Society, presented at the 140th ACS National Meeting, Division of Chemical Literature, September 4, 1961.
>
> Dennis, B. K. General Electric's Automatic Information Retrieval System. Presented to Special Libraries Association, Battelle Memorial Institute, Columbus, Ohio, April 4, 1961.

The G. E. Technical Information Center developed from a passive-type technical library into an active information center. This transition necessitated not only that there be means of rapid access to a large file of scientific and technical documents, but also that there be methods of rapidly and effectively disseminating information on both a regular and demand basis.

The Center, therefore, established a manual Uniterm coordinate index. By 1957, however, the index encompassed over 20,000 documents

and was deemed too cumbersome for manual use and was subsequently automated.

Access to the book collection at the Library is still by way of a conventional card catalog.

The information in the mechanized system consists of technical reports and memoranda generated internally and also that obtained from ASTIA, of technical society papers, journal and trade press articles, foreign and U. S. patents, translations, and miscellaneous scientific and technical information.

The system is based on the use of Uniterms and document file numbers. Each document in the system may have 20 or more words to describe it. In addition, a concise (30-50 words) descriptive abstract is prepared for each document in the system.

The index is inverted, i.e., is an item-on-term system.

The document index and abstract are placed on magnetic tapes for use in electronic computing equipment. There are approximately 60,000 document abstracts recorded on six tapes; there are over 7,000 words describing the documents; and there are more than 900,000 access points on the system's master tape. Over 1,000 new documents are abstracted and placed on the tape each month.

Originally, the retrieval system was programmed for the IBM 704. During early 1961, the I.R. programs were rewritten for the IBM 7090. The

7090 is used in conjunction with two 1401's. It should be noted that the principal criteria for choice of these computers were that they were located at the Flight Propulsion Division and that computer time was available to the Technical Information Center.

As part of its dissemination program, the Center publishes a weekly announcement bulletin, TIPS. The same punched cards which are used to update the Center's automated retrieval system are also used to produce the multilith masters from which TIPS is reproduced. Additionally, the punched cards are used for the masters for the Center's conventional catalog cards and for document loan cards. At this time, documents are announced in TIPS in twelve broad subject categories, with a document assigned to only one. The use of the IBM 1401 with a keyword-in-context program for providing an index for the announced material is being considered.

9.3.2 Machine Programs

The machine programs of the Automatic Information Retrieval System are set up in two basic parts. First is the coordination (logical product) part of the search system. Here, key words selected by the searcher are located on tape and their access numbers compared. As many as 1,200 machine questions can be handled on one run.

The Part I magnetic tape file now contains over 5,000 key words (majors) describing more than 60,000 documents. The average depth of

posting is estimated to be between 15 and 20 Uniterms per document. Another 2,000 key words (minors) have been used less than three times and are not yet on magnetic tape. The total inverted file now occupies about 1,200 feet (one-half reel) of high density tape. The input tape contains program instructions, additions to the Uniterm file and/or the search questions, in that order. Search and file maintenance of the Uniterm file may be performed concurrently. However, the file updating is scheduled semimonthly and searches are performed upon request. Whenever an updating run is being made, a new Part I magnetic tape is produced.

Part II of the Automatic Information Retrieval System is an abstract look-up program. Ten thousand abstracts and their citations are filed on one reel of magnetic tape. Thus, for the file of nearly 60,000 documents, there are six tapes of abstracts. During this part of the machine run, abstracts identified by access numbers found during the Part I search are located and transferred to an output tape. The Part II program searches two abstract tapes simultaneously for abstracts, and automatically progresses to another abstract tape when finished with a tape. The program has the ability to edit the Part I results as to groups (searches) or ranges of access numbers, at the searcher's discretion. A complete Part I and Part II search results in an output tape which contains access numbers, abstracts, and customer identification.

information. Contents of the output tape are read and printed by a 1401 computer.

Output from Part II (abstracts) is optional. If only a printed list of access numbers will suffice, the results in Part I may be transferred to an output tape, hence to be printed off-line by the 1401. Or the output from Part I may be used as input to Part II. Another option available in Part I is a "hold", whereby results of the Part I run may be printed out and a summary tape preserved until the printout has been reviewed. Thus, adjustment can be made before going into Part II.

Although the Automatic Information Retrieval System is, to a great extent, a high-speed mechanized version of a manual Uniterm coordinate index, it is at the same time somewhat more versatile than might be implied. For example, in Part I, there is the unlimited ability to relate search questions, thus providing an effective "or" and serves to eliminate duplicate accession numbers. To avoid an unreasonably large output of abstracts on some particular question, the Part II abstract printout can be limited. An important option is the ability to exercise an access-number high-low limits and range control. Since there is a high correlation between access number and date of entry into the file, this control gives the ability to vary chronologically the output of the search. Also, it enables providing a current-awareness service. In a two (or more) term question, should one of the Uniterms

cause the net coordination to go to zero, a "no-blank sort" feature is used to avoid this condition by giving a printout on the remaining terms in the question.

9.3.3 Indexing and Quality Control

Because the index terms are punched in three's on the IBM cards, terms are restricted to 18 characters. Therefore, the vocabulary includes practically no "bound" terms. Indexers (technical abstracters) are generally familiar with the vocabulary of the system and, for the most part, use only terms that actually appear in the document. Synonymity and generic levels are taken into account less by the indexers than by the searchers, who formulate the questions after discussion with the client and with the vocabulary of the system in mind.

It should be noted that in searching, there is no intermediate code look-up required, because the terms are entered on the tapes directly from the alphanumeric punched cards.

Work is in progress in editing the existing vocabulary. The possible generation of a thesaurus is being investigated. There is a hope that a thesaurus, if necessary, may be kept quite simple; this seems likely in view of the adequacy of the present methods, i.e., relying on the searcher's judgment and knowledge of the system. The A.I.Ch.E. thesaurus, as well as others, is being studied as a possible model.

The machine system itself is used for system efficiency. For example, a useful tool for the literature searcher is the machine-tabulated alphabetical list showing frequency of use (posting) of the term.

A key word that is not in the file will not be accepted as a new term unless a prescribed procedure is followed. All such rejected terms are noted on-line during the Uniterm file update. This serves to control the growth of the vocabulary.

9.3.4 Costs

The Center has not made a complete study of input costs, i.e., indexing, abstracting, etc., and is not yet willing to assign a "cost per item" figure.

It should be noted that the Center actually sells its services; most of its customers are in G.E. Sales are negotiated, however, and the customer must be convinced that the results are worth the cost. For typical searches, and where the customer is willing to wait for a few days until his question can be run on the machine with others, a flat-rate price of $75.00 has been assigned.

In the report, "General Electric's Automatic Information Retrieval System", referenced above, Mr. Dennis detailed some of the costs of searching; he has updated the figures for this report.

> "Every minute spent by the IBM 7090 grinding out a literature search costs the Technical Information Center $6.00. However, when one considers just what the machine accomplishes during that minute this cost shrinks into proper relationship ... If we assume that about eight key word questions will be required to describe one customer's question on the machine and if we search only and do not update the file while searching, in about three minutes machine time we can search not only these eight key word questions, but 1,200 such combinations. At the rate of eight key word questions per customer question and assuming an average of 100 searches per run, we can operate with at least twelve customers on a full machine run during the three minute period of time. If we assume further that the literature searcher will require about 30 minutes per customer to set up the machine questions and if we assign a typical engineering rate of $10.00 per hour to the searcher then we find that for a total labor plus machine cost of between $80.00 and $90.00 we can conduct 12 machine literature searches simultaneously. It is probably more likely that in a manual system with a file of comparable size the searcher would not make so many coordinations or so accurately as the machine. In fact, one would probably expect to get less than a third of the coverage at about three times the cost."

9.3.5 Requirements and/or Potential Refinements

In the same paper referenced above, Mr. Dennis enumerated the following plans "to improve the efficiency and effectiveness of retrieval while reducing over-all system cost."

> "Machine programming is under way which will significantly reduce operating costs. After the current phase of machine program development is complete we expect to extend our current system logic to include logical sum and difference. And of course we will always remain alert to possibilities to utilize the capabilities of our machine more fully and more efficiently. In addition to

machine program development we have a great deal of basic vocabulary work to be done... First of all, we are very much in need of a thesaurus for our retrieval system. Although our initial work on the thesaurus is being performed with the searcher primarily in mind, it is entirely conceivable that some day we will build it into our machine system. Secondly, there are problems which would be greatly helped by a post-also routine in our machine system. We see this technique being used in only very special cases since we wish to avoid adding unnecessarily to the length of our inverted file.

'We believe that an automatic Post to -- instead of -- routing by the machine will eventually go a long way toward bringing the synonym problem under control. We have already started this refinement utilizing a mechanical editing process prior to updating of the unifile. We are giving serious consideration to an optional see also program for the machine. This refinement appears to be entirely feasible and will not result in a great increase in the machine operating cost. It appears that the inclusion of a document's uniterms along with its conventional abstracts as machine output would aid the literature searcher considerably. Also since uniterms are selected from the entire document rather than just the abstracts it would help the searcher and his client to more fully evaluate the potential value of documents identified during the search. It was noted earlier in this paper that one of the features of our present system is a 'no-blank sort' refinement. Unfortunately the 'no-blank sort' is on an alphabetical basis. We wish to modify this feature and have an optional priority control. It appears highly desirable to build an automatic cumulative system experience analysis feature into our present machine system. The machine could tell us what key words have been used in questions, their frequency of use, what abstracts have been called for, their frequency, etc. Such an automatic continuing experience analysis would be helpful in future system development."

During the visit to the Center it was noted that records are kept of output of questions. On this record, the technical evaluator who has scanned the material of the output has indicated whether the document was pertinent to the request, somewhat pertinent, or not at all pertinent. When asked whether any count had been made of the non-pertinent output to determine system effectiveness, Mr. Dennis replied that this had not been done because it would in fact not be a test of the system. He pointed out that a "false drop" is very hard to define. For one thing, he says, whether or not the material answers the query is subject to the opinion of (1) the man who indexes, (2) the searcher who formulates the question, (3) the technical evaluator, and (4) the user. A machine, he points out, may correctly retrieve nonpertinent information. It may for example retrieve an item which suits the search in all respects except, say, temperature range or geographical location. It may only be the user who finally makes this decision.

The G. E. system makes no use of roles or links. Mr. Dennis suggested that the problems that may be solved by their use may just as easily be solved by introducing a third term in the search question. He feels that the addition of roles and links may unnecessarily complicate the system, and that perhaps the cost would be unjustifiably increased.

9.4 Western Reserve University Center for Documentation and Communication Research.

9.4.1 Introduction

Although WRU is also engaged in work with the American Diabetes Association, Communicable Disease Center (U. S. Public Health Service), U. S. Office of Education, and others, this report describes only the American Society for Metals Documentation Service, since this is a fully operational information searching system.

Several reports have been prepared by the WRU group for the National Science Foundation (NSF-G-10338) which describe the ASM operation in some detail:

Test Program for Evaluating Procedures for the Exploitation of Literature of Interest to Metallurgists

Part I. Development of an Operational Machine Searching Service.

Part II. Acquisition of Documents for Machine Searching.

Part III. Analysis and Quality Control

Part IV. A Cost Analysis of Abstract Preparation and Processing for an Operational Service

Part V. The Semantic Code Today

The information in this report is a summary of the material contained in those five reports and in a report titled "A Case History for Test Program for Evaluating Procedures for the Exploitation of Literature of Interest to Metallurgists." Although the following report

is almost completely extracted from this literature, it is supplemented by information obtained during a visit to WRU by the Documentation Incorporated study group. However, since the WRU-ASM system is being subjected to careful study and analysis both by outside observers and by the users and operators of the services, this report makes no attempt to evaluate the system.

As mentioned elsewhere in this study, there is considerable disagreement as to the efficiency and economy of term-on-item files. The most common criticism of term-on-item files is that the entire store must be searched, whereas in item-on-term systems only the postings on the question terms need be considered. Inasmuch as WRU processes its files on a GE-225 which can conduct simultaneous searches, and queries can therefore be "batched", the objection to searching the entire file is somewhat overridden. The GE-225 can search as many as 99 questions simultaneously; WRU generally searches 20-30 questions simultaneously.

WRU's ASM system is of special interest to the users of coordinate indexing in that it utilizes several controversial aids: role indicators; punctuation, which may also be called links; and a semantic code, which is analogous to a thesaurus. These devices have been subject to considerable controversy. Some say roles, links, and/or thesauri are entirely unnecessary; others would have only limited use of one or two of these devices; a third group claims that although these

aids may be useful, they are too complicated, i.e., too costly.

It should be pointed out that, as far as could be determined by the visit made to the Center, the WRU group is not enamoured of roles, links, the semantic code, nor the GE-225, per se. These have, however, been deemed valuable to the ASM system. Conscientious revision and monitoring of codes and procedures are continuously performed in an attempt to best meet users' needs. There is much self-criticism and very little, if any, insistence that the ASM system could or should be a model for any other. It was remarked, for example, that in its study for the Office of Education, the group was encountering somewhat different problems in vocabulary and user requirements. (It was pointed out, however, that the basic principles do seem to hold.)

It should also be pointed out that the WRU-ASM system may seem somewhat, perhaps unnecessarily, complicated to those who are more familiar with item-on-term systems than with term-on-item systems. Both for this reason and because of the nature of their subject matter, it is not to be expected that their techniques should apply to all systems. The WRU group, however, believes that because a term-on-item system can be converted to an item-on-term system, or vice versa, many of the principles which have evolved from its work may have general applicability.

9.4.2 Background and General Information

Since 1955 the American Society for Metals has provided financial assistance to the Center for processing the metallurgical literature. The program has consisted of two parts: (1) preparation of short English abstracts(called "conventional")for publication in the Review of Metal Literature and (2) experimental and pilot studies in the preparation of encoded abstracts(called "telegraphic")for use in searching the literature by machine methods. Although these two parts have been conducted concurrently, the first was an operational service when the Center assumed responsibility.

From 1955 to 1957 work was conducted on a semantic code dictionary (a machine-coded dictionary), and the groundwork was laid to establish procedures for transferring recorded information from English-language abstracts to machine-encoded symbols. Methods were developed for training abstracters in the preparation of telegraphic abstracts, for transferring this information to machines, and for testing and evaluating the results of each step.

During 1958 the Center completed 12,000 conventional abstracts for publication in the Review of Metal Literature; telegraphic abstracts were prepared for 4,500 of these. In 1959, of the 12,000 conventional abstracts produced for the Review, telegraphic abstracts were prepared for 7,500. In addition, during the year, it was decided to conduct

some experimental searches. A number of interested users in Government and industry agreed to submit sample questions for continuous searching of the current literature over a given period of time. As a result of the apparent feasibility of the system, ASM inaugurated the Metals Documentation Service in January 1960, with 12,000 conventional and 12,000 telegraphic abstracts scheduled for production in 1960.

During the period of testing and evaluating, it became increasingly evident that there are many scientific areas closely allied to metallurgy (e.g., physics, inorganic chemistry, geology) which had not previously been included. Application was made to the National Science Foundation for funds to support a test and evaluation of a much expanded activity. A grant was received in December 1959 for this purpose. Plans were then made to process an additional 22,000 - 23,000 conventional and telegraphic abstracts, which would increase the 1960 output to about 35,000.

By the end of December 1960 conventional and telegraphic abstracts had been prepared for about 39,000 articles, of which 34,000 were completely processed and put on tape ready for searching. The remaining 5,000 were processed early in 1961.

9.4.3 Costs

In its analysis of input costs gleaned from one year (1960) of operation of the ASM system, WRU has calculated a cost of $6.50 per item. This includes 15% overhead, 4% employee benefits (on personnel costs), cost of acquisition, abstracting, coding, punching, equipment,

supplies, etc. (As mentioned earlier, a report for NSF on output costs is in process.)

The abstracting staff consisted of two full-time and 50 part-time members during the period covered by the report. Part-time abstracters are paid by the piece. The following cost figures were given for abstract preparation: an average of $2.0395 to prepare both a conventional and a telegraphic abstract from a full article; $1.3265 to prepare a telegraphic from a full-length article when the author abstract is used for the conventional; and $0.8017 to prepare a telegraphic from an abstract in an abstract journal. The combined average is $1.5326 per abstract.

A summary of the input costs is reproduced on the following page.

SUMMARY OF ABSTRACTING COSTS

Total and Unit

December 1, 1959 through December 31, 1960

	Number of Abstracts	GRAND TOTAL		PERSONNEL		EQUIPMENT (amortized)		SUPPLIES		SUBSCRIPTIONS & MISCELLANEOUS		FRINGE BENEFITS (4% of personnel)		INDIRECT CHARGES (Overhead - 15%)	
		Total	Per Abstract	Total	Per Abstract	Total	Per Abstract	Total	Per Abstract	Total	Per Abstract	Total	Per Abstract	Total	Per Abstract
TOTAL		$239,345	$6.5030	$152,684	$4.1191	$25,639	$0.7481	$10,675	$0.2924	$13,000	$0.3303	$6,108	$0.1650	$31,219	$0.8481
1. Acquisitions	39,490	25,353	0.6420	6,177	0.1564	357	0.0090	2,765	0.0700	12,500	0.3165	247	0.0063	3,307	0.0838
2. Abstracting	39,490	73,361	1.8577	60,521	1.5326	—	—	850	0.0215	—	—	2,421	0.0613	9,569	0.2423
3. Editing and quality control	36,041	32,371	0.8982	27,066	0.7510	—	—	—	—	—	—	1,083	0.0301	4,222	0.1171
a. Conventional				13,301	0.3691	—	—	—	—	—	—				
b. Telegraphic				13,765	0.3819	—	—	—	—	—	—				
4. Liaison activities	36,041	21,245	0.5895	15,740	0.4367	604	0.0168	1,500	0.0416	—	—	630	0.0175	2,771	0.0769
a. Typing				9,004	0.2498	419	0.0116	1,200	0.0333						
b. Numbering & separating				1,800	0.0500	50	0.0014								
c. Filing				1,135	0.0315	70	0.0020	300	0.0083						
d. Expediting & supervising				3,801	0.1054	65	0.0018								
5. Code making	33,989	14,378	0.4230	11,359	0.3342	625	0.0184	65	0.0019	—	—	454	0.0134	1,873	0.0551
a. Words				4,771	0.1404										
b. Chemicals				4,783	0.1409										
c. IBM listing				1,805	0.0529										
6. Automatic encoding		61,682	1.7908	23,142	0.6691	24,073	0.7039	5,495	0.1574	—	—	926	0.0268	8,048	0.2336
a. Keypunching & verifying	35,030		0.7868	13,452	0.3840	5,100	0.1456	4,875	0.1392	—	—	538	0.0154	3,595	0.1026
b. Other machine operations	33,989		0.6433	8,574	0.2523	10,093	0.2970	—	—	—	—	343	0.0101	2,852	0.0839
c. Card-to-tape conversion	33,989		0.3607	1,116	0.0328	8,880	0.2613	620	0.0182	—	—	45	0.0013	1,599	0.0471
7. Managerial, supervisory & clerical	36,300 (average)	10,955	0.3018	8,679	0.2391	—	—	—	—	500	0.0138	347	0.0096	1,429	0.0393

9.4.4 Telegraphic Abstracts

The telegraphic abstract is one of the essential portions of the system. It is prepared in addition to a conventional abstract and is an "index" to be read, ultimately, by a machine.

A telegraphic abstract is made up of (1) significant words selected from the articles, (2) code symbols called role indicators which fit the selected words into context, and (3) punctuation symbols which separate and group the words and role indicators into various units in somewhat the same fashion as conventional punctuation does.

WRU has made the following assumptions in formulating their indexing procedures for a machine searching system:

1. The names of materials, their properties, processes which they undergo and the conditions of these processes can be used as index terms.

2. Certain roles which the words designating materials, properties, processes and conditions can play in the context of the subject matter are important in indexing. The devices for designating the role of a word in context are the role indicators. (This device is used to avoid the syntactic problems in the popular example: water cooling. By affixing role indicators it can be determined whether (1) the water is used for cooling, or if (2) the water is cooled.)

3. It is useful for an index to show how certain words are grouped together. The devices used to group units of information are called punctuation symbols. (This device is used to avoid the problems found in a set of terms like "gold", "silver", "watch", and "ring". Does the document discuss gold watches and silver rings, or silver watches and gold rings?) Some systems attach a letter or number to the item codes as links, e.g., gold = 11706A; watch = 11706A; and silver = 11706B; ring = 11706B. The WRU indexer uses a double dot (..) to mark the beginning of each associated set of words, e.g., ..gold.watch.. silver.ring. On the magnetic tape these punctuation symbols are numeric codes to indicate the level of grouping, e.g., an "8" may appear at the beginning of each abstract, a "7" may appear at the beginning of large groupings within the abstract, a "6" might indicate a smaller grouping within the larger, etc.

For an illustration of this, see Paragraph 3.6, Example of Input.

4. The index terms can be encoded into an artificial language which will act as a thesaurus to show the "areas of meaning which various words partake of", so that in using the index, if the words of the question mean the same thing as the words in the index, the document will be found. The device used to achieve a thesaurus function for the words selected is called the semantic code. (This will be discussed at greater length in Paragraph 3.4.)

The figure below shows a portion of a telegraphic abstract.

1.	..KEJ,	2.	ROD
3.	.KUJ,	4.	ALLOY
5.	.KUJ,	6.	AL
7.	..KAM,	8.	ANNEALING

The double dots indicate the grouping of the terms rod, alloy, and aluminum. The comma indicates that KEJ is one unit of information and that rod is another which is associated with it. The single dot indicates that KUJ and alloy (which are associated with each other by the comma) are associated with the other terms between the double dots.

KEJ is a role indicator. It shows that the word "rod" is the name of a material which is acted on by a process. When the other role indicators are translated, the information in the sample is as follows: a rod is shown as being composed of an alloy whose major constituent is aluminum and this aluminum alloy rod is being subjected to the process of annealing.

9.4.5 Semantic Code*

Each term (word) in the telegraphic abstract is coded into the semantic code.

---

* Mortimer Taube of Documentation Incorporated has prepared a paper which points out that the WRU semantic code and hence the searching system based upon it can be evaluated by comparing the generic relations embodied in the code with generic relations found in the literature being indexed or coded. The paper, "A Note on the Evaluation of the WRU Semantic Code as an Example of Generic Coding", will be published in the April 1962 issue of American Documentation.

An example is the word "diamond" coded as CERB≠CWRS≠PYPR≠1028, which may be interpreted as "a crystalline form composed of carbon and characterized by hardness." (The codes have been arbitrarily limited to four factors.) Any one of the factors may be searched. That is, the items indexed by "diamond" would be retrieved whether the question called for that specifically or for "things composed of carbon", "hard things", or "crystalline forms"

Actually, CERB≠CWRS≠PYPR≠1028 is not the complete code for diamond, since under the principles of the semantic code this would be true not only of diamonds but of any other, say, hard crystals of carbon. To specify that diamonds and only diamonds are wanted, a further element -- the numerical suffix -- must be given. This is a four-digit figure, the first numeral of which is that of the number of factors in the term. The other three are those peculiar to the particular concept. In this instance the numerical suffix might be 3001. Note that one of the factors, PYPR, is followed by the numerical infix 1028. This 1028 is the identifying numerical suffix for that particular physical property (P-PR), "hardness". Its use as a numerical infix here shows that the particular physical property characterizing (Y) diamonds is hardness. Only numerical suffixes beginning with 1 can be used for infixes. Since 2's, 3's, and 4's indicate that the code for the concept has more than one factor, their use as infixes would make it impossible to particularize specific concepts within the generic framework of a code.

This is further explained by the definitions:

"Semantic factor. By this term is meant the separate units of a code, expressed by the three consonants. In RAML≠RWHT≠TQMS $\frac{+}{0}$1002≠3679, the semantic factors are R-ML, R-HT, and T-MS. Each semantic factor represents one of a number of highly generic concepts. Together they form, as it were, the building-blocks of the code. It should be noted that within a code composed of more than one semantic factor, the separate semantic factors are arranged alphabetically ignoring the infixes.

"Infix. By this term is meant certain symbols used with the semantic factors in a code. In RAML≠RWHT≠TQMS $\frac{+}{0}$1002≠3679, the infixes are A,W,Q, and 1002.

"Alphabetic infix. By this term is meant the infixes represented by alphabetic symbols. They show the analytic relationships of the semantic factors in which they appear to the concept represented by the code.

"Numerical infixes. By this term is meant the infixes represented by numerals following the symbol $\frac{+}{0}$. They show, where used, a degree of particularization in the semantic factor to which they are affixed. Actually, every semantic factor may be thought of as possessing a numerical infix; however, only in certain instances are they explicit, that is, they actually appear in the code. In the majority of instances, they are implicit, that is, they represent a numerical infix '1001' which is not actually printed out.

"Numerical suffix. By this term is meant the particularizing number assigned each individual code to distinguish it from all other codes which, though they represent different concepts, contain the same semantic factors."

The semantic code, then, attempts to eliminate the manual or machine "see also" or "see" references. Furthermore, it attempts to include generic levels, as well as particular characteristics.

The code, of course, varies according to the system requirements. For example, in some systems the following code for diamond may be more

valuable: CERB≠CWRS≠GUMM≠MANR. This may be interpreted as "a mineral in crystalline form composed of carbon and used as a gem".

For comparison, the A.I.Ch.E. thesaurus entry for "diamond" is shown below.

DIAMOND

PO	Carbon
RT	Abrasives
RT	Crystal

PO = Post (also) on, and RT = Related term

It is obvious that it is not necessary in any one system to include all possible meanings, relationships, or generic levels. For example, diamond in the sense of a baseball diamond would certainly be superfluous in a metallurgical system. However, whether or not WRU's four factors are sufficient and whether or not the factors chosen best suit the users' needs are both open to debate. The code is subject to constant evaluation and revision.

9.4.6 Coding

The terms of the telegraphic abstract are punched on IBM cards, one term to a card. When a day's production, about 130 abstracts, have been keypunched, the collected abstracts are considered a "block" and are ready for the next step: matching with the semantic code

dictionary. The average block contains about 5000 cards, with roughly 50% of these cards representing terms, 20% special words such as chemicals and proper nouns, and 30% role indicators and information about levels of logic.

The terms in each block are then separated from the role indicator, punctuation, and title cards, and are sorted alphabetically. They are next compared on a collator with the deck of cards representing the semantic code dictionary. When there is a matching term in the dictionary (and there is, in over 90% of the cards), the proper semantic code is automatically punched in the card with the term. In the 10% of the cases in which there is no match, the card is rejected. When the process is complete, the rejected cards are batched and listed.

There are generally a certain number of spelling, keypunching, collation, and other errors. Almost half of the rejected terms are the result, however, of differences between the abstracters' terminology and that of the dictionary. About a fifth of these terms are caused by the abstracters' use of multiple-word terms which should actually have been broken down into terms which appear individually in the dictionary. The other four-fifths are caused largely by the use of spellings, inflections, and synonyms which, though perfectly correct, do not appear in the dictionary.

This latter factor could very nearly be eliminated by simply assigning to each variant term the code proper to the originally appearing form of the term. Through this, all future appearances of the variant would be automatically encoded, and the manual processing would not be required. This was the procedure originally envisaged and put into effect. With the increasing size of the dictionary, however, the decision has been made that because of the time required by the collating procedure it is better not to increase the number of dictionary entries by including variants except where these are very frequent. This speeds up the automatic encoding procedure at the expense of slowing down the manual encoding procedure, but it is felt that the over-all encoding process is more efficient. Faster matching procedures than those provided by the collator would naturally affect this decision.

When all of this has been done, there remain a number of terms which are in fact new to the dictionary. These must now be encoded and inserted if they are judged acceptable.

In the early stages of the operation, new terms represented a comparatively high percentage of the listings. After the first development of the semantic code dictionary, new terms have come into the dictionary only as they appear in the material being encoded. By now, the dictionary has so increased in size (21,385 terms, excluding chemicals)

that only a very small percentage of the terms in the material encoded are not already in the dictionary. Of eight listings which were analyzed -- representing blocks with a total inclusion of approximately 16,000 terms -- only about 0.6% of one percent of the terms were new to the dictionary. Of course a change to other fields (medicine, law, sociology, etc.) outside the physical sciences would increase tremendously the percentage of new terms.

The items on each listing are corrected by the encoder in accordance with the reason for the appearance of the term on the listing. In most instances nothing is required but to correct the term to fit the entry in the semantic code dictionary. With those few terms which require new codes, however, more elaborate procedures are necessary. The new terms are analyzed both as to meaning and to the aspects of the concept represented by the term which seem most likely to be useful in the searching procedures, and the new code is assigned.

A complete discussion of the analysis required and the procedures of assigning new codes is given in Appendix D of the report referenced previously, The Semantic Code Today.

9.4.7 Example of Input

Reproduced in the following four pages are (1) a sample conventional abstract, (2) the telegraphic abstract of the same item, (3) the semantic codes for the index terms used, and (4) the encoded abstract on tape.

Sample Abstract 3: OK

700-G. High Speed Forming of Metal Plates. Metallurgia, v. 62, Oct. 1960, p. 144-145. B·D

P — Explosive forming of mild steel plates into scale models of the dome end of a pressure vessel. (G-general, T26q; CN, 4-53, NM-k34)

B·D — forming under conditions of

T·U — high speed in telegraphic abstract

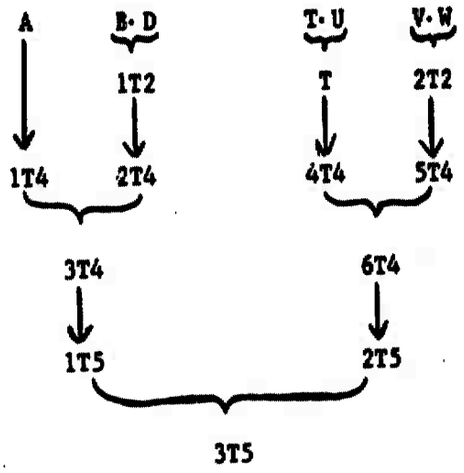
and

V·W

Search requirement satisfied by:

A·B·D·T·U·V·W

Logical Elements of Question Involved



Sample Abstract 3

Do not write in this space

Col. 6-8	Role Indicator (Col. 28-80)		Col. 6-8	Description (Col. 9-27)
1	..KEJ KOV,	(material processed, property given for)	2	PLATE
3	.KUJ,	(component)	4	MILD STEEL
5	.KWV,	(property given)	6	BRIGHT
7			8	
9	..KAM,	(process)	10	FORMING
11	.KAH,	(condition of process)	12	HIGH
13			14	SPEED
15	.KQJ,	(means of process)	16	MILD STEEL
17			18	DIE
19	.KQJ,	(means of process)	20	PRESSURE
21			22	PLATE
23	.KQJ,	(means of process)	24	VACUUM
25			26	PUMP
27	.KQJ,	(means of process)	28	EXPLOSIVE
29			30	
31	..KWJ KOV,	(product, property given for)	32	DOME
33	.KUJ,	(component)	34	MILD STEEL
35	.KIS,	(location)	36	PRESSURE
37			38	VESSEL
39			40	
41			42	
43			44	
45			46	
47			48	
49			50	

Abstracter _____

Sample Abstract 3

Line	Description	Code with English Meaning
2	PLATE	SAHT 1002, sheet, specifically plate
4	MILD STEEL	LALL≠RERN≠CN≠QFQE≠C, alloy + iron + cast iron + Fe as base + C
6	BRIGHT	BART 1008, brightness, specifically bright
10	FORMING	CUNS 1027, fabrication, specifically forming
12	HIGH	RAPR 1089, relative property, specifically high
14	SPEED	GAPR 1041, general property, specifically speed
16	MILD STEEL	LALL≠RERN≠CN≠QFQE≠C, alloy + iron + cast iron + Fe as base + C
18	DIE	MICH≠MYCL≠2009, machine (part) + mechanical
20	PRESSURE	PARS 1001, pressure
22	PLATE	SAHT 1002, sheet, specifically plate
24	VACUUM	TXMS 1004, (lack of) air, specifically vacuum
26	PUMP	FWLD≠MACH≠MQTN≠PQRS≠4001 fluid + device + by means of motion + by means of pressure
28	EXPLOSIVE	PALS 1001, explosive
32	DOME	CYTR≠MAPR≠2004, container (or cover) + material property (shape)
34	MILD STEEL	LALL≠RERN≠CN≠QFQE≠C, alloy + iron + cast iron + Fe as base + C
36	PRESSURE	PARS 1001, pressure
38	VESSEL	CATR 1029, container, specifically vessel

Sample Abstract 3

The File: Encoded Abstract 3 on Tape

8 5 KEJ 1 KOV 2 SAHT 0 1002 2 4 KUJ 2

LALL 1 RERN 1 CN 1 QFQE 1 C 2 4 KWV 2

BART 0 1008 2 5 KAM 2 CUNS 0 1027 2 4

KAH 2 RAPR 0 1089 2 GAPR 0 1041 2 4 KQJ 2

LALL 1 RERN 1 CN 1 QFQE 1 C 2 MICH 1

MYCL 1 2009 2 4 KQJ 2 PARS 0 1001 2 SAHT 0

1002 2 4 KQJ 2 TXMS 0 1004 2 FWLD 1 MACH 1

MQTN 1 PQRS 1 4001 2 4 KQJ 2 PALS 0 1001 2

5 KWJ 1 KOV 2 CYTR 1 MAPR 1 2004 2 4 KUJ 2

LALL 1 RERN 1 CN 1 QFQE 1 QC 2 4 KIS 2

PARS 0 1001 2 CATR 0 1029 2 / 700 G 8

9.4.8 Searching Procedures

The telegraphic abstract, the encoded terms, and the search program taken all together comprise a machine information retrieval system within which the following logical search devices are exploited:

1. The logical product. This means that the machine searching program can require that to answer a question, a document must contain every characteristic specified in the search program.

2. The logical sum. This means that the machine searching program can require that to answer a question, a document may contain any one of two or more characteristics specified in the search program.

3. The logical difference. This means that to answer a question a document must contain one or more characteristics but not a certain other characteristic or characteristics as specified in the search program.

Three pages are here reproduced which show a sample test question, its logical analysis, and its structure.

TEST QUESTION

Original Statement

(Information requested on:)

High velocity deformation of metals, including explosive loading

Additional Comments by Questioner

Further inquiries by the Center revealed that abstracts containing the following information were of interest to the questioner:

1. Impact extrusion
2. Velocity of forging die, stated as a function of width increase of forged die
3. Impact loading for strain tests
4. Explosive hardening

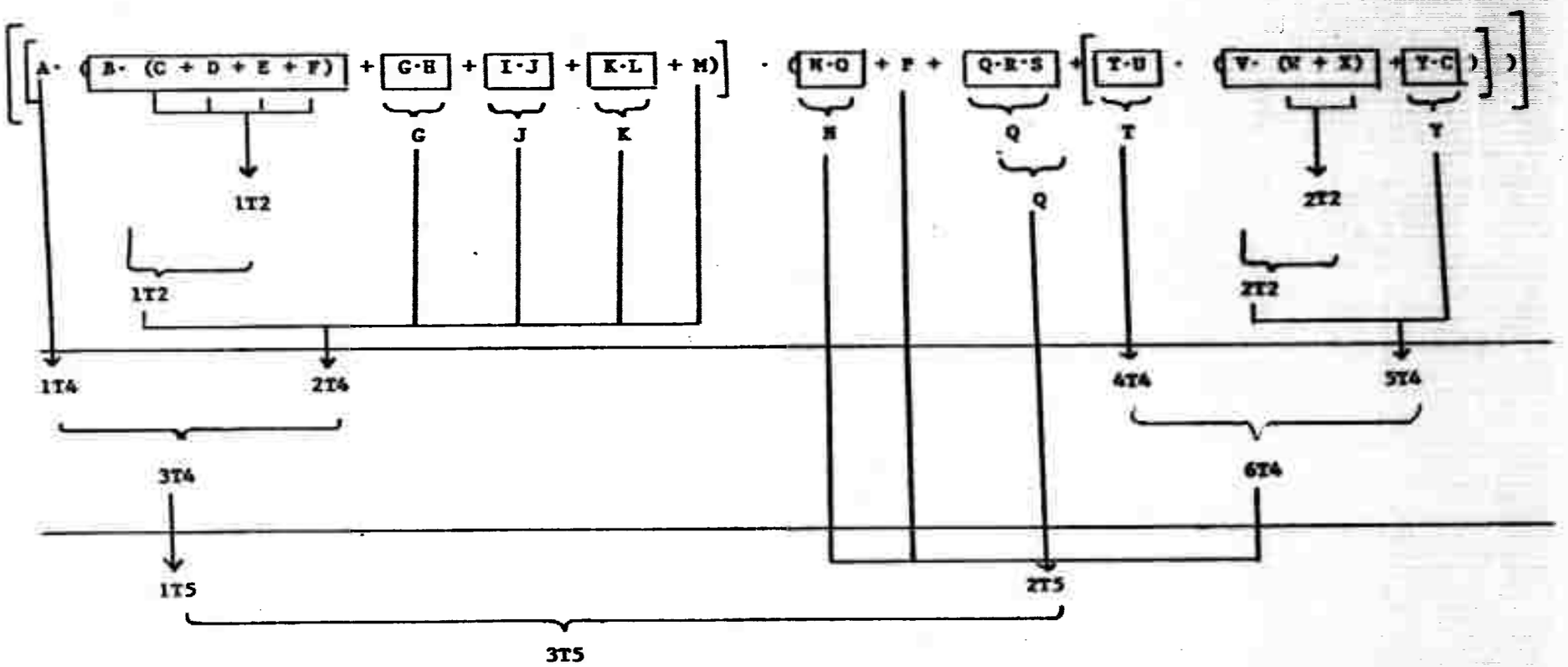
Logical Analysis of Question

$$\Big[\Big[A \cdot \big(\boxed{B \cdot (C + D + E + F)} + \boxed{G \cdot H} + \boxed{I \cdot J} + \boxed{K \cdot L} + M\big)\Big] \cdot \big(\boxed{N \cdot O} + P + \boxed{Q \cdot R \cdot S} + \Big[\boxed{T \cdot U} \cdot \big(\boxed{V \cdot (W + X)} + \boxed{Y \cdot C}\big)\Big]\big)\Big]$$

See attached sheet for structure.

A	KAM	role indicator for process
B	CUNS	- fabrication
C	1001	fabricate
D	1027	forming
E	1030	forging
F	1029	shaping
G	CUNG	- change
H	1076	extrusion
I	M_TN	- motion
J	1008	flowing
K	P_SH	- push
L	1002	pull
M	D_FL	- deflection
N	B_TT	- striking
O	1007	impact
P	P_LS	explosive
Q	BATT	- striking } shock
R	DYDD	- damage } shock
S	2002	} shock
T	RAPR	- relative term
U	1089	high
V	GAPR	- general property
W	1041	speed
X	1042	velocity
Y	NARG	- energy
C	1001	energy

Program Changes (if any)



STRUCTURE FOR TEST QUESTION

9.4.9 Conclusion

In conclusion, then, it can be said that the WRU-ASM system is operational, seems to be meeting users' requirements, and is being subjected to evaluation and improvement.

This system has several aspects which are subject to debate, such as the term-on-file, the semantic code, roles, and links. WRU has published and is preparing several reports of a statistical, analytical nature to show how, how often, and how well the system works for its intended purpose. The debates, it seems, must either wait for the results of these quantitative studies or remain on a purely theoretical basis.

Arguments in documentation have long been on a theoretical basis only, and WRU is to be commended for its efforts to quantitate and analyze its operational data. As long as this is done quite objectively, much can be learned and much can be done toward strengthening or disproving theoretical arguments.

9.5 Documentation Incorporated

9.5.1 Introduction

9.5.2 Index to Chemical Patents

9.5.3 Chemical Corps Study

9.5.4 Atomic Energy Commission

9.5.5 Air Force Office of Scientific Research

9.5 Documentation Incorporated

9.5.1 Introduction

Documentation Incorporated has had extensive experience in information systems design and operation. The projects described in succeeding sections are "case histories" chosen to illustrate specific points.

9.5.2 Uniterm Index to U.S. Chemical Patents

One of the largest coordinate indexing operations sustained for a prolonged period is represented by this Index.

The Index, owned and marketed by Information for Industry, Inc., was initiated in 1955 and has continued through 1961. During this period, Documentation Incorporated prepared indexes for chemical patents issued from 1950 through 1961.

The product is a double-dictionary Uniterm Index, published every two months in updated cumulative form, and sold by subscription. An assignee listing, a patentee listing, and patent excerpts indexed by accession number are included. (The accession numbers are used for coordination and retrieval.)

All patents appearing in the Chemical Section of the Official Gazette were automatically included, and a selection was made from those announced in the General and Mechanical and Electrical sections.

Production procedures have ranged from hand-posting to a machine operation on the IBM 305 (RAMAC). In 1961 the camera-ready copy was produced by means of an IBM 1401. The so-called minor terms, usually names of compounds, were listed separately in the index. (This listing was prepared more conventionally.) This separation permitted ready access to new information, particularly to new compounds.

The initial vocabulary resulted from "free indexing" of the patents; it was not pre-established except as the past training and experience of the indexers affected its structure. The vocabulary was derived directly from the patents; new terms were added where necessary, but each addition was carefully checked. A system of see references was mandatory because of the many synonyms for the names of chemical compounds and the differences in nomenclature.

To illustrate the growth of the vocabulary in terms of the number of patents: In 1955, the 6,065 patents were indexed by 3,700 major terms (13,000 minors); in 1961, the 10,982 patents were reflected by a major-term vocabulary of more than 7,000 terms (50,000 minors, estimated).

Because posting on certain terms was so numerous, and since this was a manual retrieval system, two methods were developed for reducing the bulk of postings. In the first method, see also references were established to indicate that when a term heading was followed by

see also, a complete generic search could only be obtained by additionally consulting all the terms enumerated under see also.

The other method for controlling posting density retained the existing relationship between certain terms such as "oil" and "treatment", or "polymerization" and "catalyst". These terms were pre-coordinated by the indexers and published as bound terms.

The usefulness of the service is att. ted by the rise in the number of subscriptions to more than 100 and by the fact that several companies were actually able to discontinue their patent indexing activities.

9.5.3 U. S. Army Chemical Research and Development Laboratories.

A pilot project was designed to provide a feasibility study and test of the effectiveness and suitability of an indexing system based upon key words and filed on punched cards for storage and retrieval of references by coordination, utilizing relatively low-cost equipment, such as the IBM 9900 Special Index Analyzer. Adaptability of this system to high-speed computer operations was also considered.

A sample based on 2,000 selected internally-generated reports was used for the pilot project. An Ad Hoc Committee representing the various research elements monitored the program on the contract and

conducted a retrieval test based upon 75 especially-prepared questions to test accuracy and depth of indexing and retrieval efficiency.* Evaluation of the results led to the recommendation that the Laboratories should convert the Technical Library to this system of coordinate indexing.

This pilot project included an experiment in the application of a method of "two-level indexing". This method can be described as follows: In the major index, compounds are treated as units, and the full name of the compound is considered a term in the system. Postings for documents appear under such terms, just as they do under the other terms of the system. Supplementing the major index is another index to compounds where the terms are the parts of the compounds, and the compounds become items in the system. The compounds are numbered and posted under appropriate terms. A generic search would then proceed in two steps, one a search in the compound index for all compounds indexed under certain generic terms, e.g., all compounds which are Dichloro derivatives and heterocyclic. Coordination at this level will deliver a set of numbers which represent compounds used as terms in the major coordinate index. In the second step, the basic question with

---

* (It is interesting to note that the vocabulary developed for this contract was unusually large, about 12,000 terms.)

which a searcher is concerned might be: "All such compounds used for insecticides". The numbers delivered by the first search would be summed and intersected with the other term, "Insecticide". This would then yield the relevant documents in the system. In a mechanized system, this procedure is much less complicated than is its description.

9.5.4 Atomic Energy Commission

A pilot experiment was directed toward the mechanized preparation of indexes to Nuclear Science Abstracts. This work was an extension of a technique already in use at Oak Ridge, wherein a "short title" was cited under personal author and corporate author entries.

The methods comprised establishing and keypunching a uniform title line for each report, storing these titles in the IBM 305 (RAMAC), and querying the RAMAC by the terms appropriate to the type of index desired.

Such a program appeared generally feasible. No significant loss of information resulted from the substitution of the uniform title line for the more conventional modifier system used to follow the subject headings. Machine compilation of the index eliminated typing and reduced proofreading time. The appearance of the computer printout (all upper case letters) was markedly improved by the addition of a "boldface-type" feature.

Although Nuclear Science Abstracts does not employ coordinate indexing, the work described here is included because of the applicability of the uniform title line to coordinate indexing systems as evidenced by its use in the NASA Scientific and Technical Information Facility operated by Documentation Incorporated.

9.5.5 Air Force Office of Scientific Research

Investigations have been directed toward the development of a storage and search theory. This theory led to the design of a mechanized system consisting of a corpus of computer-stored data and information, with an index to the store maintained externally. The validity of the concept was put to test as the Experimental Contract Highlight Operation (ECHO), which was oriented toward the documentation requirements of scientists administering Air Force research projects in practically all scientific fields. Documents generated in the normal course of administration of these projects were used during the research and for the experimental operation. Over 3,000 research efforts from 750 organizations were ultimately represented in the store and in the index; the latter attained a size of some 3,600 subject terms and 700 terms in administrative categories.

A new contract is directed toward work in multipurpose information system design with simulation to culminate in the development of IBM 1401 programs. (The original philosophy will be retained in that the

index will continue to be maintained external to the computer store.)

A planned step in the program is an opportunity to conduct, simultaneously with normal operations, experimentation in which the documents will be indexed (by their originators) from a list of 12 categories and a vocabulary restricted to approximately 1,000 key words.

10. OPERATING PROBLEMS DISCLOSED BY OPERATING EXPERIENCE

10. OPERATING PROBLEMS DISCLOSED BY OPERATING EXPERIENCE

Four major problems have been encountered in the operation of coordinate indexing systems.

(1) Cost

(2) Evaluation

(3) Vocabulary and Structure

(4) User Education.

Not listed as a major problem is the mechanization of posting and search, since this only poses problems peculiar to a particular environment. One possible exception is in chemical nomenclature. The length of terms and the complexity of chemical compounds create both a manipulation and a search problem. In many systems the compound is simply treated as a term in the over-all vocabulary (with perhaps some generic posting), assigned a code, or otherwise processed normally.

One of the systems in which the chemical compound problem was solved relatively easily was at Proctor & Gamble where an IBM 101 electronic statistical machine is used for searching. The solution is described below.

> "... At first we gave each chemical mentioned in a report a code number ... We found this impractical, for example, when we encountered a report in which a chemist itemized results on as many as 50 or more compounds he had tested for germicidal propertities. Many times individual chemicals are examined once for a special purpose and, proving unsatisfactory, they are never looked at again. To solve the problem we

> decided that a chemical must be studied (i.e., used as an index term) at least 10 times before it would be given a code as an individual. It could be coded on an IBM card under a chemical family code, a "silver and compounds" or "amines" but not as a specific chemical until it passed the "10-times" test." 1

The printout problem of superscripts, subscripts, etc., can be handled quite simply by devising new conventions.

In some systems a high degree of specificity in chemicals is required. The Patent Office, for example, has outlined the following requirements for one of its operations.

> "To identify a chemical compound for patent searching purposes, it is believed that it is necessary for a system to be able to do several things, especially in the phosphorus art.
>
> 1. To be able to identify each of the fragments comprising a compound.
>
> 2. To be able to identify the number of times each different fragment occurs in the compound.
>
> 3. To be able to find the relationships between these various fragments in the compound.
>
> 4. To be able to ask the search question either very specifically or generically, that is, with any degree of genericity desired."[2]

The Patent Office, R&D, the above report states, will have available detailed instructions on fragmenting and nodalization.

A random access method of searching is being used at the Patent Office, using the RAMAC, as well as a punched card system.

Other groups have, having fragmented and/or nodalized compounds, used other search means, such as peek-a-boo cards.

To a certain extent, cost and evaluation go hand in hand. Where cost is high it may be justified in terms of value received. However, as pointed out in Section 7 of this report, there has been no satisfactory method for determining the effectiveness, let alone the value, of a system. Certain measures are available; for example, a system used solely for general academic research might not justify the expense of detailed input -- deep indexing or introduction of structure.

Where a system is used, on the other hand, to determine whether or not a specific kind of work has already been done, the savings in total development cost to either a contractor or the Government may be so large as to justify a costly system.

Cost is also a factor in determining whether or not structure is imposed on a system at the input or at the output. For example, where a large amount of extraneous material is retrieved during search, and cost is incurred in screening, evaluating, disseminating and discarding quantities of material, it may indeed be worth using roles, links or other such devices at the input.

Here again, no satisfactory method has yet been found of evaluating a system to a point where the mode and value of structuring can be determined.

The means and location of structuring in a system will, to a certain extent, determine the form of the vocabulary and indexing. Where a computer program, for example, can be used to post and search on higher and lower generic levels, there is no necessity for complicated indexing procedures and indexing aids. Of course, thesauri or other posting instructions have to be developed for the machine.

At this time, the vocabulary problem is probably the most critical. Much of the work underway in this area is detailed in Sections 4 and 5 of this report.

There is doubtlessly some limit below which a vocabulary can be totally free and without complications such as roles, links, and thesauri. Where the limit lies and how areas outside that limit can best be handled has not been determined.

Especially with a large vocabulary and with broad system capabilities, user education is of prime importance. It has often been said that the statement of a problem is at least half the solution. That scientists and technicians cannot formulate their questions is therefore not surprising. If, however, they have at least some idea of what they want, the system should be such that (1) the vocabulary will assist them or (2) that an intermediary searcher can assist them. The work of Stiles, Stevens, and others (described in Section 7.4) is toward the very large goal of having a system find material automatically even

though the terms are not even in the question.

Less ambitious, but equally important, are the efforts to translate requests by intermediate searchers. The work of such a person, an "information researcher" , at Esso Research's Technical Information Division is described in the example below.

> "... One request read: 'Please arrange for a patent search of methods of solidifying petroleum fractions by reaction with stearates or stearic acid.' If the search had been made without further discussion of the request, it could have been a monumental task, covering the entire recorded information on greases and other thickened fluids. Actually this request was found to involve only enough general background to enable the questioner to have a very superficial acquaintance with thickened fluids. The request was then handled within five minutes. However, it required about 30 minutes and several phone calls to reach mutual understanding."[3]

The importance, illustrated by this example, of a feedback between user and system cannot be overstressed. Not only can the system be made to function more efficiently if needs are known, but also the user is not left unsatisfied and disappointed. Too often, users criticize information retrieval systems for giving them too much or too little information, without realizing that they are getting just what they asked for.

REFERENCES

1. Schultze, Else L. An Application of Automation in the Library: Indexing Internal Reports. Special Libraries, Vol. 52, No. 2, February 1961.

2. Frome, Julius. Mechanizing Searching of Phosphorus Compounds. Patent Office Research & Development Reports, No. 18, January 3, 1961.

3. Knox, William T. Information, Please. Paper presented at 53rd Annual Meeting, American Institute of Chemical Engineers, Washington, D. C., December 5, 1960.

11. CONCLUSION

11. CONCLUSION

Having completed the study and preparation of a report to this point, Documentation Incorporated recognizes an obligation to set forth its own conclusions concerning the present state of coordinate indexing and recommendations concerning future developments.

As indicated in the report itself, most of the literature and controversy which have developed in this field have been concerned with degrees of freedom or, conversely, with degrees of structure. Although this debate has usually been concerned with connections among terms, it has also involved the initial selection of terms. With reference to the selection of terms, proposals have included automatic selection of words in the title, abstract, or text; the selection of words from a text by clerical workers; the assignment of words by competent indexers or subject specialists; and the selection of terms from rigid authority lists and similar devices. The words or terms have been variously called "Uniterms", "keywords", "descriptors", "selectors", "designators", etc., and structural elements surrounding or added to indexing terms for the purpose of eliminating the "noise" have been called "permuted indexes", "KWIC", and "standard title line indexes", etc.

Although the individuals or companies that have suggested one designation or another have usually tried to distinguish the connotation

of their name for an indexing term from all other connotations, the public in general has had only a hazy idea of the differences, if any, among Uniterms, keywords, descriptors, selectors, designators, etc. There has been a similar haziness in the public's mind regarding the specific nature or the specific differences among thesauri, subject heading lists, authority lists, code dictionaries, etc.; among links, roles, role indicators, interfixes, semantic factors, etc.; and among coordinate indexes, concept coordination, multilevel indexing, and correlative indexing.

One may conclude from this situation that much of the controversy reduces to a proprietary interest in a particular name for a quite ordinary activity, suggested by one individual or another or one firm or another. It is not intended to excuse Documentation incorporated from this general estimation. But when one comes to recognize what has occurred, it is a mark of wisdom, if not of valor, to say with relief, "A plague on all our houses" and to take a fresh look not at favorite names but at the operating problems which require solution. The devices and apparati named can then be seen as a number of available tools of the systems designer for solving special operating problems in special environments. Some environments will call for free indexing; others will call for hierarchical posting; others will require pre-coordinations, role indicators, or links.

Some will permit free redundant indexing; others will require carefully controlled vocabularies with elaborate structures of cross-references. The true expert in information retrieval systems will know how to select the proper apparatus and the proper degree of freedom to design the best system for any particular operating environment.

It is the considered opinion of the Documentation Incorporated study group that linguistic, semantic, and syntactic studies do not satisfy the immediate needs of coordinate indexing. This opinion is supported in the literature* and was found among various individuals interviewed during the survey. This is not a universal belief, however. A glance at the list of research efforts in scientific documentation** discloses a considerable amount of such work. Without attempting to resolve the question, this study group would suggest that a better balance between linguistic studies and systems work be attempted.

To summarize, Documentation Incorporated feels that the following recommendations are in order:

(1) That the I.R. problem be viewed in part as a problem of the optimum design of an engineering system and not solely as a problem of basic research into linguistics

---

* See Bar-Hillel, Yehoshua, "A Logician's Reaction to Recent Theorizing on Information Search Systems," American Documentation, Vol. VIII, No.2, April 1957, pp. 103-113; and "Some Theoretical Aspects of the Mechanization of Literature Searching," Technical Report No. 3, prepared under ONR Contract No. N62558-2214 and under a grant from NSF, April 1960.

** See Current Research and Development in Scientific Documentation, No. 9, National Science Foundation, Office of Science Information Service, NSF-61-76, November 1961.

and meaning. Such systems work would tend to bridge the gap between linguistic research and the actual operation of systems.

(2) That the annual preparation of a critical review encompassing all applicable areas of information retrieval be undertaken or sponsored by some appropriate agency.

12. ANNOTATED BIBLIOGRAPHY

Ahlin, J. T., "The IBM 650 Information Retrieval System," International Business Machines Corporation, September 29, 1959.

Ahlin prepared a program to simulate on an IBM 650 the searching and matching operations of coordinate indexing performed by an IBM 9900. In effect, his paper demonstrates that the fixed matching circuits of the IBM 9900 can easily be programmed in any more powerful computer.

ASTIA, "Multiple Aspect Searching for Information Retrieval," (Conference sponsored by Armed Services Technical Information Agency, Washington, D. C., February 12-13, 1957), 125 p.

Contents:

Review of ASTIA Arrangement with Documentation Incorporated in Development of the Uniterm System of Coordinate Indexing
James L. Ferguson

Why a Documentation System
Robert S. Bray

Use of Concept Coordination in the du Pont Engineering Department
Eugene Wall

Report on Use of the Uniterm System
Nell Steinmetz

Evaluation of Techniques to Control Research Data
Dr. I. A. Warheit

Modification of a Multiple Aspect System for Company Use
John P. Wadington

Application and Retrieval of Information
Dr. John A. Sanford

A Basic Approach to Information Handling
Dr. Harold Wooster

Posting to Cardineer Wheels
Benbow Cheesman

Use of the Dashew Data Poster
A. F. Caprio

A Deep Index for Internal Technical Reports
Dr. Fred Whaley

Use of Uniterms in Large Collections
Isabelle Burtnett

The ONR Program on Multiple Aspect Systems
Commander G. W. Hoover

Matrex Junior and Use in Large Collections
Hans Ullmann

The Peek-A-Boo System - Optical Coincidence Subject Cards in Information Searching
Dr. Joshua Stern and W. A. Wildhack

Bailar, John C. Jr., Heumann, Karl F., and Seiferle, Edwin J., "The Use of Punched Card Techniques in the Coding of Inorganic Compounds," J. Chemical Education, 25, 1948, pp. 142 - 143, 176.

As part of its attempt to set up a mechanized system for correlating molecular structure with biological activity, the Chemical-Biological Coordination Center of the National Research Council felt it necessary to develop codes for expressing chemical compounds in punched cards. This paper describes the development of a code for inorganic compounds. The problems of coding are not directly related to the problems of coordinate indexing, but this paper is included because it does recognize that the cards may be used to correlate different terms, i.e., descriptions of structure and descriptions of biological activity.

Bailey, M. F., Lanham, B. E., and Leibowitz, J., "Mechanized Searching in the U. S. Patent Office," (presented at a meeting of the Division of Chemical Literature, American Chemical Society, February 1951), J. Patent Office Society, 35, August 1953, pp. 566 - 587.

This paper reports further experiments in the Patent Office of the use of punched card equipment to provide "multiple categorization of compositions" and search by one or more categories. It also illustrates the unwillingness of the Patent Office group to depend entirely upon a system of coordinate indexing and their use of

generic coding within each category. Thus, the system they describe is a classification system modeled closely on the present classification system used in the Patent Office. The machine supplements what is possible in the manual search only to the extent of making possible a search by more than one generic code at a time.

Bailey, M. F. and Cochran, S. W., "Patent Searching -- General Files," Punched Cards, Their Applications to Science and Industry, edited by Robert S. Casey and James W. Perry, New York, Reinhold Publishing Corporation, 1951, pp. 367 - 377.

This discussion of patent searching points out the difficulties which arise from single place classification systems even when such systems are provided with cross-references. The authors show a solution to this problem in the use of coding and machines which would make possible searches by intersections of classes. They conclude that standard sorting machines are not adequate to the requirements of chemical coding and multiple-term searching, as required by Patent Office searches.

Bar-Hillel, Yehoshua, "A Logician's Reaction to Recent Theorizing on Information Search Systems," American Documentation, Vol. VIII, No. 2, April 1957, pp. 103 - 113.

An examination of some of the writings of the Western Reserve group, Documentation Incorporated, and Calvin Mooers on the development of new I. R. systems. Bar-Hillel states that the value of new contributions has been exaggerated and that more work should be done to improve traditional systems. The only real contribution he finds in the new theories is the recognition of the simple fact that a set of documents and their terms "form then a Boolean Algebra with respect to the operations of complementing, intersecting, and joining."

Barden, William A., Hammond, William and Heald, J. Heston, "Automation of ASTIA, A Preliminary Report," AD-227 000. Arlington, Virginia, Armed Services Technical Information Agency, December 1959, 50 pp.

"Early considerations in automation" by William A. Barden: The history of ASTIA's experience in planning and implementing the automation of its functions is presented. Different ideas were

examined and discarded in a search for a more efficient method of indexing and retrieving information. In 1953 a preliminary study based on systems concepts embracing all the functions and services of the Agency was conducted, but a full scale study was not possible until 1958. When the final selection of the Remington Rand USS-90 (Univac Solid State Computer) was made, the ASTIA staff devised methods for making optimum use of the equipment in both the business-type and information retrieval functions.

"Automation program by William Hammond: The pre-automation and automated processing of reports through ASTIA and validation of requests of military contractors is described. The three stages by which the automatic data processing system will be put into operation are examined, and the process of compiling mechanized cumulative indexes to the Technical Abstract Bulletin is presented.

"Creation of a Thesaurus of Scientific Descriptors by J. Heston Heald: The main objectives of Project MARS (Machine Retrieval System) are: (1) to prepare a Thesaurus of descriptors; and (2) to assign these descriptors to all AD numbered reports in the ASTIA collection. The ASTIA subject headings and subdivisions were overhauled and the list reduced from 70,000 to about 9,000 headings, now termed descriptors. The scope of subject coverage was divided into about 290 generic categories called display schedules. Procedures were established for the assignment of retrieval terms, both standard descriptors from the Thesaurus and "open-ended terms" which will not appear in the Thesaurus but will provide additional retrieval access points in the form of project names, equipment nomenclature, trade names."
[American Documentation Abstract]

Batten, W. E., "Specialized Files for Patent Searching," Punched Cards, Their Applications to Science and Industry, edited by Robert S. Casey and James W. Perry, New York, Reinhold Publishing Corporation, 1951, pp. 169 - 181.

It is from this paper that the designation "Batten systems" has been derived to designate inverted systems using optical coincidence as a method of search. Batten was primarily concerned with presenting the advantages of an inverted "aspect" system as contrasted with conventional Hollerith systems. He did not recognize explicitly that the coincidence of holes on any two aspect cards indicated those items which were members of a product class. In fact, he proposed that his aspect cards be arranged not as a set

of terms but as a classification system. However, he did realize that his classification system was mobile and that he could compare members in one class with members in another.

Bernier, Charles L., "Correlative Indexes, I - V," American Documentation, Vol. VII, No. 4 (October 1956), Vol. VIII, No. 1 (January 1957), Vol. VIII, No. 3 (July 1957), Vol. VIII, No. 4 (October 1957), Vol. IX, No. 1 (January 1958).

Correlative indexing provides, in a nonmanipulative form, e.g., books or card files, the type of Boolean search permitted by a coordinate index. It does so by printing out under each term not only the number of the item but all the other terms used in connection with a given term to index that item. Essentially correlative indexing provides in book form the same type of revolution of positions of terms developed by the Chemical-Biological Coordination Center. Dr. Bernier also discusses problems of vocabulary control and the type of terms to be used in correlative indexes.

Bibliography in an Age of Science (Louis N. Ridenour: "Bibliography in an Age of Science," Ralph R. Shaw: "Machines and the Bibliographical Problems of the Twentieth Century," Albert G. Hill: "Storage, Processing and Communication of Information"), Urbana, University of Illinois Press, 1952.

The reference to this volume is included in this bibliography because it represents one of the earliest recognitions by the library profession itself of the impact of machines on traditional library activities. The article by Dr. Ridenour emphasizes the compression of storage and the communication of bibliographical information from central depositories. The article by Dr. Shaw describes various punched card devices but is primarily concerned with the Rapid Selector, a microfilming scanning device developed by Dr. Shaw based upon a suggestion of Dr. Vannevar Bush. Although Shaw envisioned the use of the Rapid Selector to store traditional indexes, e.g., the Index to Chemical Abstracts, the potentiality of the Rapid Selector as a coordinate searching device was recognized by many others.

Bohnert, Lea M., "Two Methods of Organizing Technical Information for Search," American Documentation, Vol. VI, No. 3, July, 1955, pp. 134 - 151.

> "Two methods of organizing technical information for search have been distinguished on theoretical grounds. One was the traditional method of library classification. The other method was non-hierarchical and relied on combinations of general terms to characterize specific ideas or terms.....
>
> "So far, the second method alone has been tried either in relatively small collections (10,000 to 50,000 documents) or in new, i.e., marginal, fields of knowledge, such as instrumentation. Its practicality for larger collections (hundreds of thousands of documents), and ones in which most of the fields of science and technology would be involved, is still to be proven."

[Author's Conclusions]

Bracken, R. H., and Tillitt, H. E. "Information Searching with the 701 Calculator," Association for Computing Machinery Journal, Vol. 4, No. 2, April 1957, pp. 131 - 136.

> "The application of a 701 calculator is described, using magnetic tape data storage, for the control of about 14,000 items with a coordinate index. Over 9,600 descriptors are used for subject access, and the system is used for approximately 16 searches three times a week. The total time for the set of 16 searches is 11 minutes. Planned modifications include the use of a new type of tape and the substitution of a core memory for the electrostatic memory."

[AD Abstract]

Brockway, Duncan, "Coordinate Indexing at the University of New Hampshire Library," American Documentation, Vol. X, No. 3, July 1959, pp. 228 - 231.

This account of an experiment in coordinate indexing is of interest because the author gives figures for density and distribution of posting and also figures for "false drops" as a function of the number of terms in a search. The project covered a narrow special field and its results are probably applicable to similar special collections.

Bush, V., (Chairman of the Committee), Report to the Secretary of Commerce by the Advisory Committee on Application of Machines to Patent Office Operations, Washington, Department of Commerce, 22 December 1954.

When this report appeared, it was considered a milestone in the progress towards mechanization of information storage and retrieval. Five recommendations were made by the Committee to the Secretary of Commerce, namely:

1. The Patent Office should put machine searching of compositions of matter on an operational basis.

2. The reclassification of patents should be accelerated.

3. A research and development unit should be established in the Patent Office.

4. The National Bureau of Standards and the Patent Office should undertake a joint program to stimulate and develop machines and techniques specifically adapted to the Patent Office operations.

5. An advisory committee should be attached to the Office of the Secretary of Commerce to stimulate and coordinate the program and related efforts within the Commerce Department.

An estimate of accomplishment in these five areas was made by a Committee of the National Academy of Sciences - National Research Council in 1960. The 1960 report indicates that Recommendation 5 was not carried out and that Recommendations 1, 3, and 4 require re-thinking and modification in the light of experience since 1954. However, the 1960 report did not mention any activities under Recommendation 2 and it is this recommendation which is crucial for the problem of coordinate indexing and mechanized search. The Patent Office group has always felt that hierarchical classification, rather than coordinate indexing, is the required intellectual structure for mechanized information storage and retrieval. It recognized that its existing classification system was not adequate for mechanization but it supposed that the existing classification system could be modified without changing its essential structure. It is possible that whatever activity there has been under Recommendation 2 has operated to the detriment of accomplishment under Recommendations 1, 3, and 4.

Casey, R. S., Bailey, C. F., and Cox, G. J., "Punched Card Techniques and Applications," J. Chemical Education, 23, 1946, pp. 495 - 499.

Although this paper is primarily concerned with describing punched cards and the techniques of coding and handling them, it does recognize what it calls the possibility of using such devices to correlate information. In considering the use of punched cards for chemical information, the paper indicates the possibility of combining a search for a class of materials and a class of properties or uses.

Center for Documentation and Communication Research, School of Library Science, Western Reserve University, "Comments on 'A Logician's Reactions,'" American Documentation, Vol. VIII, No. 2, April 1957, pp. 117 - 122.

Comments on "A Logician's Reaction to Recent Theorizing on Information Search Systems."

[Cf. Bar-Hillel, "A Logician's Reaction to Recent Theorizing on Information Search Systems;" Mooers, "Comment on Bar-Hillel's 'A Logician's Reaction to Recent Theorizing on Information Search Systems'"]

"The Chemical-Biological Coordination Center of the National Research Council," Washington, National Research Council, September 1954.

Although primarily devoted to a discussion of coding and a description of the operations of the Center, this paper does contain a brief statement of the way in which punched card equipment is used for coordinate search:

> "There is at the Center, then, a growing file of punched cards which can be searched mechanically for variables. Although a single criterion can be looked for, such as a test organism or a manner of administration, it is in the facility for search of combinations of ideas that this method affords a major advantage over conventional indexing. Thus, all compounds tested for a specific response from a given organism or group of related organisms under any of the usual variable conditions of testing can be selected from all other compounds not meeting those specifications."

Cherenin, V. P., "Certain Problems of Documentation and Mechanization of Information Search," (mimeographed translation), Moscow, 1955.

This generalized discussion of documentation problems contains a full account of search by the logical intersection, sum, and complement of classes. It indicates that the ability to perform such searches is the basis of mechanizing information search. However, the paper also concludes that there is a requirement for a special "machine language" which will supply grammatical and syntactical relationships of terms in addition to the Boolean operations which form the basis of machine search.

De Grolier, Eric, "Method for the Retrospective Searching of Scientific Documents: A Preliminary Report," Paris, Unesco, August 24, 1955.

> "De Grolier was commissioned to prepare this paper by the Unesco International Advisory Committee for Documentation and Terminology in Pure and Applied Science in consultation with the committee secretariat. Its major sections discuss the scope of the problem; retrospective searching from the user's standpoint; methods of facilitating the retrieval of documents (classification, indexing, filing, automatic selection and codes); and organizational considerations affecting the choice and utilization of methods. Contains bibliographies following each section and sub-section of the report."
> [AD Abstract]

Dunham, B., "The Formalization of Scientific Languages, Part 1. The Work of Woodger and Hull." IBM Journal of Research and Development, Vol. 1, No. 4, October 1957, pp. 341 - 347.

> "The problem of language structure in the mechanical storage and retrieval of information is discussed. The "formalization" of language, as attempted by Woodger and Hull, is examined as a solution to the problem of language structure in mechanical operations."
> [AD Abstract]

Fairthorne, R. A., "Algebraic Representation of Storage and Retrieval Languages," Proceedings of the International Conference on Scientific Information, Washington, NAS-NRC, 1959, pp. 1313 - 1326.

"This paper has outlined a possibly useful method of representation in which vocabularies and hierarchies of vocabularies are regarded as the sum, in any consistent sense, of repetitive dyadic vocabularies, not necessarily clerically realizable. It generalizes and unifies many special methods, such as various algebraic identities used traditionally to demonstrate number representations, and models used in investigating some properties of ordinary language. Here it can be used to discover the sympathetic magic principles (like-produces-like) characteristic of linguistic systems and to apply them usefully to more systematic vocabularies. We have seen that to some extent it can cope, though not simultaneously, with additive and, in general, modular properties such as cost and selective information, and with the partial orderings and looser generalized operations, synonymity and homonymity, that are essential to retrieval."
[Author's Abstract]

Fairthorne, R. A., "Automata and Information," Journal of Documentation, Vol. 8, September 1952, pp. 164 - 172.

This paper is included because of its clear account of automata and their possible application to the storage and retrieval of information in libraries. Fairthorne is clear that the automata devices which can operate on physical strings of information and as such, can perform only a clerical or engineering function in a library, can contribute to the information problem by manipulating "tags" in an index.

Fairthorne, R. A., "Delegation of Classification," American Documentation, Vol. IX, No. 3, July 1958, pp. 159 - 164.

By "delegation of classification," Fairthorne apparently means a scheme whereby the correct assignment of classes can be performed by clerks or by people not directly concerned with making or interpreting the classification. He thinks such delegation is important because:

> "It would be local in time as well as space, because no librarian acts as his own classifier longer than his

> term of office or his term of life, whichever may be the shorter. If you have to do everything yourself, including classification and retrieval, you do not have to know how to do it, you only have to be able to do it. For delegation we have to construct rules for making non-contradictory decisions about relevance."

Fairthorne, R. A., "Information Theory and Clerical Systems," Journal of Documentation, Vol. 9, June 1953, pp. 101 - 116.

A discussion of the library problem in terms of information theory. Fairthorne distinguishes clearly between semantic problems which occur at the level of the selection of terms or the understanding of a question put to the system and the problem of organizing the physical elements of a library, be they documents, terms, or codes. Just as Shannon and Weaver point out the irrelevance of semantic problems to the type of information with which communication theory is concerned, so Fairthorne points out the irrelevance of semantic problems to the design of information systems for libraries:

> "The semantics of any library activity can be settled by practical study, and intelligent anticipation, of clients' behaviour in bibliographical situations. Syntactic problems of coding, and pragmatic problems of matching the codings to operations such as marshalling, selecting, and siteing of documents and tallies can then be considered as clear-cut questions of cutting down average time and labour and cost of concrete tasks."

Fairthorne is also clear that a system of indexing can be interpreted as a Boolean algebra and that the problem of mechanization becomes one of finding the most efficient code and the most efficient instruments for manipulating such codes.

> "The cost and trouble of working clerical systems [information systems] depends not on what we say in them, but how we say it.... No philosophical issues are involved at this level of library action. Information Theory can unify and generalize much empirical knowledge in this field."

Fairthorne, Robert Arthur, "The Mathematics of Classification," Proc. Brit. Soc. for International Bibliography, 9, October 14, 1947, pp. 35 - 42.

Library classification, especially the U. D. C., is presented as a particular and limited application of the algebra of classes. The limitation follows from the emphasis in library classification upon the relationship of inclusion. The advantage of an exclusive concern with a single relationship is that it permits numerical coding of the relationship and, hence, the linear organization of materials on shelves. If classification is considered to be more than a device for arranging materials on shelves, it becomes apparent that classes may have to one another other relationships than the relation of inclusion. Fairthorne points out that the subject which discusses the relationship of classes is Boolean algebra and he indicates that library classification can profit from using other relations described in a Boolean algebra. He points out that Boolean algebra "is the algebra governing networks of switches; that is, the controls of computing machinery, among other things.... As the application of such apparatus to libraries is an urgent topic, the consistency of the laws governing these machines with those governing classification systems need careful consideration, which I am not going to give in this paper, through lack of time and competence. But it will have to be done, if complicated machinery is not to be misapplied." Fairthorne recognizes one difficulty in the application of the algebra of classes to problems of library classification. One of the elementary operations of the algebra of classes is negation and the use of negation as an operation in information retrieval presents very serious problems.

Farradane, J. E. L., "A Scientific Theory of Classification and Indexing and Its Practical Applications," Journal of Documentation, Vol. 6, June 1950, pp. 83 - 99.

Although this paper is primarily concerned with a new approach to the classification of knowledge, it approaches this problem by defining what it calls "isolates" and "operators." An isolate is a term or subject and an operator is a connection between terms or subjects. Farradane calls these operators "logical operators" but in specifying their nature he describes them as "expressing appurtenance, equivalence, reaction, and causation." Prima facie, reaction and causation are not logical operators but terms used to describe factual relations. Appurtenance, which means property, might be considered a logical relation within a predicate logic, but it is not a relationship within the algebra of classes. As for equivalence, Farradane

rules out equivalence as a relation of an item to itself and restricts it only to the part-whole relationship, which is also interpreted factually, rather than as a logical relationship of class inclusion. In terms of the above, it might be supposed that Farradane's work was irrelevant to the development of coordinate indexing, but actually it represents an early effort to insist that in an indexing system, it was necessary to use more than logical relations to express factual relationships among terms. This relates his work to subsequent work on roles and links, which function as empirical connections among terms of an indexing system.

Farradane, J. E. L., "A Scientific Theory of Classification and Indexing: Further Considerations," Journal of Documentation, Vol. 8, June 1952, pp. 73 - 92.

This paper, which departs still further from either a theoretical or a practical concern with indexing and information systems, is a continuation of "A Scientific Theory of Classification and Indexing and Its Practical Applications." Even more than the previous paper, it confuses problems of indexing systems with extraneous philosophical considerations, with child psychology, epistemology, and theories of perception. Further, Farradane has no conception of the type of logical and mathematical considerations which are relevant to the indexing and information problem and which have been set forth so well by Fairthorne. The problem of organizing information in a library is, as Fairthorne has pointed out, an engineering problem and not a philosophical problem.

Francisco, R. L., "Use of the Uniterm-Coordinate Index System in a Large Industrial Concern." Presented before the Metals Division, Special Libraries Association, Philadelphia, October 20, 1955.

> "The Uniterm System has been used for more than a year and a half in indexing technical reports in the Technical Data Center of the General Electric Company, a Center which has over 150,000 reports in its files. The System is described and two pitfalls are discussed; namely, the importance of avoiding the use of synonyms, which will tend to lose information, and also the necessity of avoiding terms so general that they encompass the entire library. In avoiding the use of synonyms, it is helpful to keep a dictionary of terms used. The author states that they have never yet encountered any problem with 'false drops.'"

[AD Abstract]

Gamble, D. T., "A Coordinate Index of Organic Compounds," (presented before the Division of Chemical Literature, 127th Meeting of the American Chemical Society, Cincinnati, Ohio), March 31, 1955.

The major interest of this paper lies in the fact that it is one of the first concrete demonstrations of the convertibility of conventional and inverted systems. The paper describes an experiment in which 3,000 compounds were broken down in a manner similar to that established by the Chemical-Biological Coordination Center, but the data were entered on Uniterm cards. This produced a file of 263 cards representing the terms in the system, on which the 3,000 compounds were posted. Like the CBCC system, the coordinate index described in this paper permitted only searches by logical functions of the terms, and hence some noise might result in a search for compounds having a particular arrangement of the functional groups described by the terms in a search. Although the author used Uniterm cards, he does point out that an identical system could be established using Peek-a-boo cards. Here again is one of the early realizations of the logical identity of different mechanisms used in coordinate search.

Garfield, Eugene, "Preliminary Report on the Mechanical Analysis of Information by Use of the 101 Statistical Punched Card Machine," American Documentation, Vol. V, No. 1, January 1954, pp. 7 - 12.

The Welch Medical Library Indexing Project, sponsored by the Armed Forces Medical Library (now the National Library of Medicine), was concerned with the study of the general problem of indexing the world's medical serials and with developing techniques for both mechanized search and mechanized print-out of indexes. Garfield's paper is based upon his participation in the Welch Medical Library Indexing Project. The Project utilized an IBM 101 and by ingenious wiring of this device, Garfield was able to conduct searches by a many-term question with any specified Boolean function relating the terms. What Garfield achieved was a form of superimposed wiring, logically equivalent to the superimposed coding developed by Calvin Mooers. It is known in the art that superimposed coding requires random codes and the Welch Medical Library Indexing Project used generic codes. With generic codes it is impossible to calculate in any meaningful sense the percentage of false drops which will occur from any given degree of superimposition. Perhaps this is why Garfield does not discuss this problem in his paper.

Herner, Saul, and Meyer, Robert, "Classifying and Indexing for the Special Library," Science, Vol. 125, No. 3252, April 26, 1957, pp. 799 - 803.

The authors propose to correct the recognized disadvantages of general classification systems, not through the use of coordinate indexing, but by preparing special classifications designed for special purposes and special groups of users. A description of such a special classification is given.

Holmstrom, Dr. J. E., "A Classification of Classifications," (Paper presented at Berne Conference of International Federation for Documentation, 1947, and published in its report; since slightly amended). The Royal Society Scientific Information Conference, 21 June - 2 July 1948, Report and Papers Submitted, London, The Royal Society, 1948, pp. 501 - 515.

This paper is one of the earliest to distinguish between library classification, alphabetical subject heading, and what Holmstrom calls mechanical selection. Although Holmstrom recognizes that mechanical devices present "the possibility of sifting out from an accumulated mass of records, at any time, any desired conjunctions [italics his] of information", he does not see, as did Fairthorne, that the conjunction is essentially a product relation as described in the algebra of classes and that it is only one of a number of possible logical functions of classes. Holmstrom discusses two types of apparatus for achieving conjunctions, namely, manual key-sort and the Batten system, which uses optical matrices. These devices can handle logical products much more readily than they can handle logical sums or other types of logical functions. This may explain Holmstrom's exclusive emphasis on conjunction.

Jonker, Frederick, "The Descriptive Continuum: A 'Generalized' Theory of Indexing," Washington, AFOSR TN 57-287, June 1957.*

> "The generalized theory of indexing postulated in this article... looks upon all indexing systems as a continuum, the descriptive continuum. The main parameter of this continuum is the average length of the 'entries'

* This paper also appeared in Proceedings of the International Conference on Scientific Information, (Nov. 16-21, 1958) Washington, NAS-NRC, 1959.

or 'headings' used. At one end of the continuum or 'spectrum' is keyword indexing; subject heading indexing is somewhere in the middle, while hierarchic classifications are at the other extreme. The average length of the headings or descriptive terms used determines the position in the continuum.

"Throughout the continuum, all other parameters behave as functions of the average term-length. Some of these parameters are:

---Potential depth of indexing
---Permutability of indexing criteria
---Degree of hierarchical definition of indexing
---Potential need for a coordinating mechanism
---Retrieval noise
---Size of the access apparatus
---False coordinations
---Capacity for handling semantic indeterminacy.

"These parameters are discussed and explained. They are believed to contain all the considerations basic to the indexing problem.

"The theory indicates that once the main parameter, average term length, is determined, all other properties of the indexing system are fixed. For every information collection there is an 'optimum' position in the continuum, according to which the collection should be organized. This optimum position is determined by the diffuseness of the information in that particular field."
[Author's Abstract]

Joyce, T., and Needham, R. M., "The Thesaurus Approach to Information Retrieval," American Documentation, Vol. IX, No. 3, July 1958, pp. 192 - 197.

The use of a thesaurus is recommended as providing retrieval not merely on a yes-or-no basis but in terms of degrees of relevance. The degrees of relevance are presumably set up by setting up relationships among terms in the thesaurus. So far as the operation of retrieval is concerned, it still proceeds on a yes-or-no basis.

King, Gilbert S., "Applications of Punched-Card Methods to Scientific Computations," Punched Cards, Their Applications to Science and Industry, edited by Robert S. Casey and James W. Perry, New York, Reinhold Publishing Corporation, 1951, pp. 407 - 422.

Although this paper is primarily concerned with handling scientific data and scientific computations on punched card equipment, it is, as has been noted in the text, one of the earliest papers to recognize explicitly that an understanding of machine possibilities could be derived from a study of symbolic logic.

> "The machine carries out operations of symbolic logic, and for proper coding and programming the scientist should know the basic algebra of the black boxes into which he will put cards."

Although Dr. King does not expand on this point, it is clear that if the machine operates by performing logical functions, a search should be designed in terms of such logical functions.

Luhn, H. P., "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, Vol. 2, No. 2, April 1958, pp. 159 - 165.

> "Excerpts of technical papers and magazine articles that serve the purposes of conventional abstracts have been created entirely by automatic means. In the exploratory research described, the complete text of an article in machine-readable form is scanned by an IBM 704 data-processing machine and analyzed in accordance with a standard program. Statistical information derived from word frequency and distribution is used by the machine to compute a relative measure of significance, first for individual words and then for sentences. Sentences scoring highest in significance are extracted and printed out to become the 'auto-abstract'."
> [Author's Abstract]

Luhn, H. P., "The Automatic Derivation of Information Retrieval Encodements from Machine-Readable Texts," Yorktown Heights, N. Y., International Business Machines Corporation, September 6, 1959, 9 p.

"The problem of organizing scientific and technical literature for purposes of information retrieval may be alleviated either by (1) the adoption of a common machine coding language to avoid conflicts and duplication or (2) the development of completely automatic methods for deriving clues to document contents. In arguing for the latter course, this paper reviews some of the drawbacks of the common-language approach and suggests several approaches to automation based on statistical methods. These include the derivation of similarity coefficients from word-frequency lists and indexing by the automatic printout of key words in their context."
[Author's Abstract]

Luhn, H. P., "General Rules for Creating Machinable Records for Libraries and Special Reference Files," Yorktown Heights, N. Y., International Business Machines Corporation, September 30, 1959, 8 p.

"The purpose of establishing general rules for creating machinable records for libraries and special reference files is to insure the greatest possible utility of a basic record in the many phases of documentation, library services, the storage, dissemination and retrieval of information. If, at the first instant of accession of a document, a duly complete machinable record thereof is manually produced, this should be the last time that human effort is expended in view of the facility with which modern information processing devices can adapt this information to all subsequent requirements that might arise.

"In order that such a master record fulfill these requirements, its format must be substantially unbiased with respect to any specific preconceived system it is to serve. The adoption of a standard format of an original record within an organization and, possibly among organizations, will permit the exchange of such records, yielding additional savings in effort, and

will insure complete freedom in devising individual systems most suitable for given situations. Also, in those cases where systems have become inadequate or did not perform as expected, the modification or redesign of such systems may be carried out without renewed manual effort since new types of records can always be derived from the master records by automatic means. Finally, work on the creation of a collection of records may safely be started long before a particular system for processing them has been established."
[From the text]

Luhn, H. P., "The IBM Electronic Information Searching System," IBM Research Center, Yorktown Heights, New York, February 15, 1952.

This paper is one of the earliest and most thorough attempts to describe a completely mechanized searching system. It covers coding problems, input problems, and searching problems, and describes a special machine which has come to be known in the art as the Luhn Scanner. The Luhn Scanner, by its use of complementary coding and the movement of the cards being searched against the question card, eliminated the requirement for fixed field coding and provided for searching by the three Boolean operations of product, sum, and complement. The machine could also be instructed to search by any sub-set of terms without specifying the terms included in the sub-sets. The paper, "A New Method of Recording and Searching Information," constitutes the appendix to this paper.

Luhn, H. P., "IBM Punched Card System for Indexing and Classifying Information and Method of Searching and Analyzing Such Information," March 25, 1948.

This unpublished paper, prepared by Mr. Luhn in March of 1948, is an early statement of the theory of coding and searching which was reduced to practice in the Luhn Scanner and in several earlier prototypes. Mr. Luhn was concerned with developing a method of coding which would permit several terms to be coded on the same card without the requirement that the terms be placed on the card and searched for in any specific order. He was also concerned with the ability to search for material indexed by any combination of terms on the card.

Although the details of coding are not given, the coding is stated to provide economical storage of data along with the possibility of print-out, sorting, and other operations carried out by standard IBM machines. Although Mr. Luhn recognized the importance of free indexing, he did indicate that his coding system could provide for generic relations among the terms of the system.

Luhn, H. P., "The IBM Universal Card Scanner for Punched Card Information Searching Systems," Yorktown Heights, N. Y., International Business Machines Corporation, November 17, 1958, 24 p.

> "An electronic machine that answers the requirements of information retrieval by scanning of punched cards has been developed by IBM. This machine, called the 'Universal Card Scanner' (UCS), scans cards fed through it in a manner similar to that employed on conventional punched card sorters. It is capable of discovering whether any one or several of a given set of patterns are wholly or partly contained in any of the record cards scanned. This function is performed by a 'no-pulse matching' process under the control of a 'question card' which contains prototypes of the patterns sought, likewise represented by punched holes. This is the adaptation of an electronic method to the optical principle of 'matching by black-out', employed in an earlier experimental IBM card scanning machine, frequently referred to as the 'Luhn Scanner'. As was the case in the earlier model, the present machine features the use of a punched IBM card (Question Card) for furnishing the patterns to be searched for in a record file."

[From the text]

Luhn, H. P., "Identification of Geometric Patterns by Topological Description of their Envelopes," Poughkeepsie, N. Y., International Business Machines Corporation, April 23, 1956, 6 p.

> "The use of electronic equipment in the field of literature searching and of correlation of information has pointed up the need of linear notations for multidimensional representations such as chemical structures. This paper proposes a notation for the identification of geometric patterns such as used for representing chemical structures. The notation

is based on a topological description of the envelope enclosing such configurations."
[Author's Abstract]

Luhn, H. P., "Keyword-In-Context Index for Technical Literature (KWIC Index)," Yorktown Heights, N. Y., International Business Machines Corporation, August 31, 1959, 16 p.

"A distinction is made between bibliographical indexes for new and past literature based on the willingness of the user to trade perfection for currency. Indexes giving keywords in their context are proposed as suitable for disseminating new information. These can be entirely machine-generated and hence kept up-to-date with the current literature. A compatible coding scheme to identify the indexed documents is also proposed. In it elements are automatically extracted from the usual identifiers of the document so that the coded identifier yields a maximum of information while remaining susceptible to normal methods of ordering."
[Author's Abstract]

Luhn, H. P., "A New Method of Recording and Searching Information," (Presented at American Chemical Society Meeting September 11, 1951). American Documentation, 4, January 1953, pp. 14 - 16.

By the use of overlapping circles, Mr. Luhn illustrates how any topic can be identified by a set of terms and how related topics which are indexed by sub-sets of the terms can be indicated. The new method which involves both indexing and retrieving by "a plurality of aspects," is contrasted with conventional methods of indexing and classifying. Luhn also points out the manner in which a relatively small set of terms can, in combination, describe uniquely millions of diverse topics. Finally, Luhn proposes that by varying the number of terms used in a search, a search can be made as generic or as specific as one pleases.

Luhn, H. P., "Potentialities of Auto-Encoding of Scientific Literature," Yorktown Heights, N. Y., International Business Machines Corporation, May 1959, 22 p.

"The introduction of mechanical devices for the processing of scientific information raises the question as to the extent to which machines will be able to assist in the selection, storage, dissemination and retrieval of information. In order to appreciate fully the functions that information processing machines are capable of performing in this area a number of typical operations are presented and their potential usefulness to the development phase as well as operational phases of information systems is explored. The solution of particular problems is illustrated by way of examples based on the availability of scientific literature in machine-readable form. The examples cover the compilation of word lists, establishment of word relationships, the preparation of word patterns for retrieval and compilation of dictionaries and thesauri. Some of the results of Information Retrieval Research at the IBM Research Center are presented in the form of machine print-outs such as the keyword-in-context index for bibliographies, the auto-abstract, the word pair matrix, derived code words, and the statistical analysis of a document."
[Author's Abstract]

Luhn, H. P., "Row-by-row Scanning Systems for IBM Punched Cards as Applied to Information Retrieval Problems," Yorktown Heights, N. Y., International Business Machines Corporation, May 1959, 37 p.

"'The row-by-row method of recording obviates the need of superimposed coding and overcomes the disadvantages previously enumerated. Alphabetic or numeric information may be spelled out by character and may therefore be uniquely matched during the scanning process, thereby eliminating incidences of false selection. It is furthermore possible to express relationships amongst recorded items in many ways, to indicate ranges of values, alternative conditions and many other features. Also, the recorded information may always be recovered, which is not possible in superimposed coding schemes'. Provides details on row-by-row coding schemes, typical scanning codes and their

assembly into rows, and preparation of cards. Describes the IBM 101 Electronic Statistical Machine with row-by-row attachment and gives examples of applications."
[AD Abstract]

Luhn, H. P., "Selective Dissemination of New Scientific Information with the Aid of Electronic Processing Equipment," Yorktown Heights, N. Y., International Business Machines Corporation, November 30, 1959, 19 p.

"Improvement of scientific communication is sought through machine assisted dissemination of new information. A service system is described in which a new document is characterized by a vocabulary or pattern of keywords. This pattern is then compared with the vocabularies or profiles characterizing each of the participants of the service. If a given degree of similarity exists between the two, the affected participants are notified by a card carrying an abstract. The recipient signifies whether the information is in fact relevant or not by returning or not returning a stub provided with the card. His affirmative response, which may include his request for a copy of the document, is reflected on his profile by incorporating the pattern of the accepted item. Profiles are kept current by discarding patterns after they have reached a certain age. The feedback includes notification of authors as to the reception of their work. The service also facilitates participants' referral of information to others and generally endeavors to promote interchange of information by personal contact."
[Author's Abstract]

Luhn, H. P., "A Serial Notation for Describing the Topology of Multidimensional Branched Structures." (Nodal Index for Branched Structures)," Poughkeepsie, International Business Machines Corporation, December 12, 1955.

Many users of coordinate systems have felt that the proper encoding of chemical compounds required ordered relationships among the terms, as contrasted with the commutative relationships of a Boolean algebra.

This paper is one of the earliest descriptions of a system which imports order into the relationship among the terms of the system.

> "The systems refer to branched and interconnected arrays of linear elements, such as used in delineating chemical structures, the flow of processes, or the assembly of mechanical and electrical circuit elements. A primary objective of these systems is to derive a linear notation which permits the discovery of inclusion of a given structure in another structure by serial comparison of the elements of the respective notations (serial scanning). Such serial comparisons may be performed by machines without the aid of an internal memory. Another objective is the derivation of a unique notation for purposes of identification."

[Author's Abstract]

A similar scheme was developed at the Bureau of Standards for coding compounds in the patent literature.
[Cf. Ray, L. C., and Kirsch, R. A., "Finding Chemical Records by Digital Computers"]

Luhn, H. P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM Journal of Research and Development, Vol. 1, No. 4, October 1957, pp. 309 - 317.

This paper is largely concerned with the problem of setting up a vocabulary of "notions" and establishing related families of terms for use in both encoding and searching. The relationship of terms is to be determined by a statistical study of and use with one another.

Maron, M. E., "Automatic Indexing: An Experimental Inquiry," Santa Monica, The Rand Corporation, August 10, 1960, 37 p.

> "This inquiry examines a technique for automatically classifying (indexing) documents according to their subject content. The task, in essence, is to have a computing machine read a document and on the basis of the occurrence of selected clue words, decide to which of many subject categories the document in question belongs.

"This paper describes the design, execution and evaluation of a modest experimental study aimed at testing empirically one statistical technique for automatic indexing."
[Author's Summary]

Mooers, Calvin N., "Choice and Coding in Information Retrieval Systems," Transactions of the I. R. E. for 1954, pp. 112 - 118.

"Information retrieval systems are susceptible to treatment by communication theory at the coding and machine level, and there are a number of analogies between retrieval systems and multiplex signalling systems. Historically, retrieval theory has been aided by communication theory. In the other direction, there is reason to believe that developments -- both theoretical and practical -- originally made with retrieval systems may be applicable to the development of signalling systems. For instance, some retrieval practice seems to be ahead of work in asynchronous multiplex signalling. For another thing, techniques for handling semantic information in retrieval -- not discussed here -- may be suggestive for further development in communication theory if and when such matters are undertaken."
[Author's Abstract]

Mooers, Calvin S., "Coding, Information Retrieval, and the Rapid Selector," American Documentation, Vol. I, No. 4, October 1950, pp. 225 - 229.

This paper by Mooers discusses the type of coding suggested in the paper by Wise and Perry, "Multiple Coding and the Rapid Selector." It points out that word coding does not result in random codes and therefore the number of false drops may be much higher than expected. Mooers also objects in this paper to the use of the terms "polydimensional" or "dimensional" to describe coordinate indexing on the grounds that the terms or descriptors in an indexing system do not fall in definite series or dimensions. On this point, see Taube, Mortimer, "The Functional Approach to Bibliographic Organization."

Mooers, Calvin N., "Comment on Bar-Hillel's 'A Logician's Reaction to Recent Theorizing on Information Search Systems,'" American Documentation, Vol. VIII, No. 2, April 1957, pp. 114 - 116.

A reply to Mr. Bar-Hillel's "A Logician's Reaction to Recent Theorizing on Information Search Systems."

Mooers, Calvin N., "Information Retrieval on Structured Content," presented at the Third London Symposium on Information Theory, sponsored by the Department of Electrical Engineering of the Imperial College of Science and Technology and held at the Royal Institution, September 12-16, 1955.

This paper represents Mooers' contribution to the discussion of ordered relations among terms in contrast to a pure coordinate indexing system. Mooers proposes to group descriptors into interlocking sets, which he calls "n-tuples." Each such set involving a number of terms in an ordered relation in effect constitutes one term which can be coordinated with any other set.

Mooers, Calvin N., "A Mathematic Theory of Language Symbols in Retrieval," Proceedings of the International Conference on Scientific Information, Washington, NAS-NRC, 1959, pp. 1327 - 1364.

> "A mathematical model is presented which relates the language symbols of retrieval to the documents retrieved. The model is applied to three families of retrieval systems: those using for language symbols (1) descriptors, (2) characters with hierarchy, and (3) characters with logic. Most information retrieval systems now in use are variations of one of these systems. The similarities and differences between the three systems are displayed by the model. According to the model, a retrieval prescription is represented by a point in a space P. This space can be generated by taking the cardinal product of a repertory of simple partially ordered systems. The output of the retrieval system is a subset of documents, and each of these is represented by a transformation from a point in space P to a point in space L. Two different retrieval transformations are defined. Future elaborations and extensions of the model are outlined."
> [Author's Abstract]

Mooers, Calvin N., "Zatocoding Applied to Mechanical Organization of Knowledge," American Documentation, Vol. II, No. 1, January 1951, pp. 20 - 32.

In addition to its description of the Zator Selector and Mooers' method of superimposed coding, this paper is one of the earliest presentations of the method of indexing by separate descriptors and of retrieving information by searching for a combination of descriptors. Because Mooers used superimposed coding in a single field, he limited his description of searches to products of descriptors. The percentage of false drops in the system went down as the number of terms in a question went up. On the other hand, any question involving a sum of descriptors would so increase the number of false drops as to make the system unusable. It is perhaps for this reason that Mooers restricted his discussion to product searches.

Morris, J. C., "The Duality Concept in Subject Analysis," American Documentation, Vol. V, No. 3, August 1954, pp. 117 - 146.

This paper is a defense of traditional library subject headings analysis against the claims of coordinate indexing. It argues that the elimination of connectives, word order, and grammatical variants of indexing terms and the exclusive use of terms and Boolean functions will lead to the scattering of information in an index and a high rate of noise in retrieval. Much of the later criticism of pure coordinate indexing followed the arguments set forth in this paper.

Morris, J. C., "Evolution or Involution? Notes Critical of the Uniterm System of Indexing," J. Cataloging and Classification, 10, July 1954, pp. 111 - 118.

> "The Uniterm system of coordinate indexing is an innovation in subject analysis. Some of the features inherent in the system belie the claims made by its proponents, particularly as to dependability for subject retrieval and as to speed or ease of searching. Some of the administrative directives and rules for setting up Uniterm indexes appear almost certain to lead to superficial rather than intensified subject indexing. These two factors taken together indicate that such a system set up as recommended would be its own

worst enemy and would be self-defeating in the long run."
[Author's Abstract]

The author points out that the four major devices for the retrieval of information in libraries have been card catalogs, indexing systems, classification schemes, and subject bibliographies. These are interdependent and each has met "with some degree of success the logical criteria of subject analysis which have evolved."

The major object of the paper, however, is to review critically "statements and implications" which have been made in the Installation Manual for the Uniterm System of Coordinate Indexing (Armed Services Technical Information Agency, Dayton, O., 1953). Such matters as matching numbers, the basic vocabulary of Uniterms, the density of postings, relationships between concepts or ideas, and the "false-drop" problem are discussed.

Nolan, J. J., "Information Storage and Retrieval Using a Large Scale Random Access Memory," (Presented before the American Chemical Society, April 15, 1958). American Documentation, Vol. 10, No. 1, January 1959, pp. 27 - 35.

"Describes the application of the IBM 305 RAMAC as an information retrieval tool, demonstrating how random entry to such a large-capacity memory, and the associated programmable features, may make coordinate searching of large collections practical by offering the possibility for overcoming such problems as the recognition of specific-generic relationships and false association between search terms."
[AD Abstract]

Opler, Ascher, "Dow Refines Structural Searching," Chemical and Engineering News, Vol. 35, No. 33, August 19, 1957, pp. 92 - 96.

"Staff at Dow Chemical Company has been developing a system for searching coded chemical compounds for desired structural features, using a high-speed digital computer (the IBM 704). A general searching program has been written to take care of 90 percent of the searches requested by Dow chemists; the other 10 percent can be handled by writing special programs or by modifying the general program. 10,585 compounds have been coded and recorded on magnetic tape so far. Experience

during the past two years has shown the feasibility and accuracy of machine searching and the capabilities and limitations of the code. The staff has developed the multiplexing principle for conducting simultaneous searches. This approach consists of taking a number of searches in a group and comparing them with a number of structures in a group. With the IBM 704, a group of searches can be compared with a group of 120 structures at a time. The search criteria are considered to be four hurdles or acceptance tests which each compound under examination must pass in answering the search."
[AD Abstract]

Paden, B. R., "Information Retrieval on Automatic Data Processing Equipment," Special Libraries, Vol. 50, No. 4, April 1959, pp. 162 - 165.

"The first in a series of four articles by a senior mathematician programmer with the IBM Corporation explaining some fundamentals underlying mechanized retrieval methods, presented simply and with examples to help clarify each point. Discusses the breaking down of subject headings into descriptors suitable for coordinate searching, emphasizing that 'the development of a set of descriptors adequate to a particular application is a major part of the battle.'"
[AD Abstract]

Paden, B. R., "Information Retrieval: Punched-Card Equipment," Special Libraries, Vol. 50, May-June 1959, pp. 197 - 200.

"The second in a series of four articles by an IBM senior mathematician programmer, this one being devoted to the functions of the IBM keypunch, sorter, accounting machine and collator, as they relate specifically to retrieval methods."
[AD Abstract]

Paden, B. R., "Information Retrieval: Punched Card Techniques and Special Equipment," Special Libraries, Vol. 50, No. 6, July-August 1959, pp. 244 - 249.

> "The third in a series of four articles, this one dealing with the look-up and compare technique using a collator and unit-record cards which have been sorted in descriptor files. Included are brief explanations of coordinate descriptors, superimposable numeric coding and scanning, the universal card scanner, and the special index analyzer."
> [AD Abstract]

Peakes, Gilbert L., "Report Indexing by Machine-Sorted Punched Cards," Punched Cards, Their Applications to Science and Industry, edited by Robert S. Casey and James W. Perry, New York, Reinhold Publishing Corporation, 1951, pp. 115 - 136.

Although concerned primarily with coding techniques and sorting techniques, this paper does contain a recognition that punched card devices can be used to search for an intersection of headings. It presents this notion by pointing out that one of the advantages of punched card systems is that such a system can reduce the number of cards which must be employed in a manual system when any item is indexed by more than one heading. The system described by Mr. Peakes had up to seven headings selected from seven different categories, i.e., product, customer name, raw materials, processing, etc. A search could be made by any combination of terms punched on a card. Since Mr. Peakes proposed using simple sorting equipment, he recognized that such a search would have to be linear, that is, a search for a second term within a deck selected from the total file by a search for the first term.

Perry, James W., "Indexing, Classifying, and Coding the Chemical Literature," Industrial and Engineering Chemistry, 40, May 1948, pp. 476 - 477.

Although in this and in many subsequent papers Perry continued to insist that mechanization must await the solution of problems of nomenclature and semantics, this paper does contain an excellent statement of the manner in which mechanical, i.e., coordinate, search differs from standard indexing and classification systems:

"We now have available mechanical tools, both simple and complex, which offer promise for escaping the limitations previously imposed on indexing and classifying systems. For example, punched cards are able to register separately and independently a fairly large number of criteria which may characterize any single entity among a group of its fellows. Punched cards, furthermore, permit us to use any desired combination of such criteria in carrying out a search to isolate certain entities characterized by the desired combination of criteria. Owing to the limited number of holes that may be punched in a given card, considerable ingenuity may be required successfully to cope with the large number of criteria necessary to characterize chemical subject matter. Such mechanical difficulties may perhaps be avoided by using other mechanical or electronic devices."

Perry, James W., "Information Analysis for Machine Searching," American Documentation, Vol. I, No. 3, August 1950, pp. 133 - 139.

After commenting on the difficulties which arise in the use of conventional classification and indexing systems, Perry proposes that machine methods will make possible the search by various combinations of indexing terms. There seems to be no realization that the machine will search by products, sums, or complements of classes but it is stated that machines will make possible searches according to the following possibilities:

(A or B) + (C or D)
(A or B) + (C - D)
(A + B) or (C + D)
(A or B or C) - (C + D + E)
(A or B) - (C or D)

There is, of course, some analogy between this schema and a Boolean schema but it was not until some time later that Perry realized this fact. On the other hand, the failure to appreciate the logic of machine search, while it led to certain errors in symbolism and particular descriptions of the search process, does not detract from Perry's early empirical feel for how the machines operated and the searching techniques they would make possible.

Perry, James W., "Punched-Card Coding -- Some Practical Suggestions," Punched Cards, Their Applications to Science and Industry, edited by Robert S. Casey and James W. Perry, New York, Reinhold Publishing Corporation, 1951, pp. 267 - 275.

In this paper, as in many others by Perry, there is a strong emphasis upon the necessity for the proper sorts of indexing and coding before any great utility from the machine can be realized. He argues that

> "... a punched-card code must be based on well-defined, carefully selected terminology, and must be used in a consistent, standardized fashion when incorporating new items into the file."

Certainly this is true. Perry does not emphasize, however, that mechanical selection, while requiring as much rigor in selection of terms as any other indexing system, does free the indexer from the necessity of decisions concerning the order of the terms to be used in a search.

Perry, James W., "The Utilization of Scientific Knowledge," Scientific Monthly, 66, May 1948, pp. 413 - 417.

In this report of the work of the Punched Card Committee of the American Chemical Society, most of the discussion concerns abstracting and coding problems. With reference to coordinate searching, Mr. Perry in this paper is less sanguine than in his paper, "Indexing, Classifying and Coding the Chemical Literature", and concludes that efficient mechanical search of large files will not be possible until new coding systems are developed for chemical compounds.

Rakov, B. M., and Cherenin, V. P., "Machines for Retrieving Information in the U. S. S. R., "Unesco Bulletin for Libraries, Vol. 11, No. 8-9, August-September 1957, pp. 192 - 197.

> "The theory and construction of the experimental information-retrieving machine, EIM, is described. The machine, which is based on the C-80-1 analyzing computer, was built in 1954 by the Institute of Scientific Information of the Soviet Academy of Sciences. It uses 80-column, 12-line punched cards on which information may be entered by position, non-position, superimposition, or direct codes. Multiple codes

may be combined on a single card by dividing the card into code zones. The modified standard alphabetical puncher used can be set to switch codes automatically as the appropriate zone is reached. For retrieval, questions are fed through a standard switchboard or through a special panel with one switch for each spot on the card. Cards are scanned once at a rate of 7 per second; characteristics of question and answer are compared; and the cards are sorted into accepted and rejected slots on the basis of logical exclusion principles."
[AD Abstract]

Ray, L. C., and Kirsch, R. A., "Finding Chemical Records by Digital Computers," Science, 126 (3278), October 25, 1957, pp. 814 - 819.

By treating chemical structures as spatial arrangements of atoms, a group at the Bureau of Standards developed a way of searching for chemical structures by numbering the position of the atoms, starting from any arbitrary point in the structure. Presumably the system provided for the possibility of generic search for compounds having any portion of their structures in common. The method provides not only a search by coordination, but search in terms of the order of terms, that is to say, it uses more than Boolean functions of terms. The system resembles that described by H. P. Luhn in "A Serial Notation for Describing the Topology of Multidimensional Branched Structures."

Rockwell, Harriet E., Hayne, Robert L., and Garfield, Eugene, "A Unique System for Rapid Access to Large Volumes of Pharmacological Data; Application to Published Literature on Chlorpromazine," Federation Proceedings, Vol. 16, No. 3, September 1957, pp. 726 - 731.

"Use of IBM punched cards with multiple coding for information retrieval in the pharmacological field."
[AD Abstract]

Schultz, C. K., and Shepherd, C. A., "A Computer Analysis of the Merck Sharp and Dohme Indexing System." (ONR Contract Nonr-2297(00) NR 048-116). [n.d.]

This report describes "an empirical study of a satisfactorily functioning punched card system." Although a great deal of descriptive information was derived, it was concluded that the results of the study did not permit any evaluation of the system described or of other systems.

Shannon, Claude E., and Weaver, Warren, The Mathematical Theory of Communication, Urbana, The University of Illinois Press, 1949.

This volume is relevant to the development of coordinate indexing in two respects. Shannon's equations for the capacity of communication channels are relevant to the design of efficient information systems and coding for such systems. In the second place, Shannon's very definition of information as the logarithm of the number of available choices from a set of messages, when applied to information systems, has the result that the freer the system, the more information it contains. In other words, a system of coordinate indexing without roles and links and without any requirement for categorizing or ordering terms and in which all the terms could be coordinated with one another would have more information in it than any system having one or another of the above constraints.

The relevance of Shannon's work to problems of information storage and retrieval has often been questioned. However, this relevance can best be presented by considering together the following statements from the book:

> "The semantic aspects of communication are irrelevant to the engineering aspects."
>
> "This does not mean that the engineering aspects are necessarily irrelevant to the semantic aspects."

If one properly understands these two statements, one can also understand why mechanized systems and coding can contribute to the semantic aspects of information storage and retrieval systems and why semantic considerations cannot contribute to the solution of problems of mechanization (engineering aspects). Suppose one wished to develop a high-fidelity system for the reproduction or transmission of music. Such a high-fidelity system, properly engineered, might

convey a good violin tone, i.e., the engineering would contribute to the esthetics. On the other hand, whether or not violinists in general played sweet or sour notes would make no contribution to the development of high-fidelity systems, i.e., esthetics would not contribute to the engineering. We are only interested in storage and retrieval systems because individuals can index material, although some index poorly. Whether the indexing is good or bad does not contribute to the engineering aspects or the mechanization of storage and retrieval systems. On the other hand, good mechanized systems can convey the results of good indexing.

Shera, Jesse H., "Classification: Current Functions and Applications to the Subject Analysis of Library Materials," The Subject Analysis of Library Materials (Papers Presented at an Institute, June 24-28, 1952, under the Sponsorship of the School of Library Service, Columbia University, and the A. L. A. Division of Cataloging and Classification), Edited and Introduction by Maurice F. Tauber, New York, Columbia University School of Library Service, 1953, pp. 29 - 42.

After a discussion of traditional library classification and its limits, Shera introduces symbolic logic as providing a new type of class order which is non-hierarchical. Shera supposes that such new types of order and new types of indexing are in some sense derived from the logic. Actually, the logic serves only to describe the type of order and the type of indexing. Shera also supposes that the class relations described in a Boolean algebra are a species of the genus classification, which also includes hierarchical classification. Most logicians have argued that hierarchical classification is a special classification which utilizes exclusively the relation of "inclusion" between classes.

Taube, Mortimer, "Functional Approach to Bibliographic Organization: A Critique and a Proposal," (Presented before the Fifteenth Annual Conference of the University of Chicago Graduate Library School, July 24-29, 1950), Bibliographic Organization (edited by Jesse H. Shera and Margaret E. Egan), Chicago. University of Chicago Press, 1951, 57-71.

A major portion of this paper is concerned with a critique of traditional hierarchical classification and alphabetical subject heading systems. It concludes that such systems can never be the

basis of national or international systems of bibliographical organization. As a third possibility, it recommends the construction of categories of terms and the use of a set of terms to index any document, such a set to be constructed by selecting one term from each category. Although there is here an implicit recognition of the intersection of terms to index a document, the logic of this intersection was not set forth and the paper emphasized primarily the arrangement of the terms in categories.

Taube, Mortimer, "Specificity in Subject Headings and Coordinate Indexing," Library Trends, Vol. 1, No. 2, October 1952, pp. 219 - 223.

Whereas in a subject heading system specificity of indexing is provided by subdivision, in a coordinate indexing system such specificity is provided by intersecting terms or by determining product classes. In principle, the use of enough subdivisions in a subject heading system could provide the same degree of specificity as a system of coordinate indexing. However, a subject heading system which limits by convention the number of subdivisions will also limit by convention the specificity of any individual heading. The Science and Technology Project of the Library of Congress, by convention, used only one subdivision of any given heading and found it necessary to use many different headings and subdivisions in order to express the contents of an item being indexed. The limitation on degree of subdivision in an alphabetical system is usually imposed because of the difficulties with alphabetization and permutation created by subdivision. These difficulties disappear in a coordinate index.

Taube, Mortimer and Associates, "Storage and Retrieval of Information by Means of the Association of Ideas," American Documentation, Vol. VI, No. 1, January 1955, pp. 1 - 18.

What this paper calls "association of ideas" later came to be part of the technique of thesaurus building employed by Wall and Costello for the du Pont Company and H. P. Luhn at IBM. Any search by a Boolean function of classes may lead to a negative answer because the class created by an intersection of classes might have no members. In order to provide for mechanized systems some indication of the nature of the material in the system corresponding to the ability to browse in a manual index, it was felt that a mechanized system should display to the searcher the class intersections which had members. The technique chosen to achieve this end was as follows: For every term

in the system there was created a logical sum of all other terms used with that term to index any document. Each term, then, headed a subset of terms in the system "associated" with it. The searcher could input any term and have displayed to him this subset of terms which could guide him in making subsequent intersections for searching. The technique chosen was limited to the co-occurrence of two terms only.

Taube, Mortimer and Associates, "Studies in Coordinate Indexing." Washington, Documentation Incorporated, Vol. I - V, 1953-1959

Taube, Mortimer, Gull, C. D., and Wachtel, Irma S., "Unit Terms in Coordinate Indexing," American Documentation, Vol. III, No. 4, October 1952, pp. 213 - 218.

This paper was one of the first public announcements of the Uniterm system. Besides describing the type of card and the method of posting, it discussed the creation of vocabularies by breaking up standard subject headings and classification systems into sets of terms. It also discussed the problem of multiple-word indexing terms and presented a suggested rule for combining or separating words in an indexing term. The rule has become known in the literature as the rule for "free" and "bound" terms. Finally, the paper presented a set of rules for organizing a Uniterm system.

Taube, Mortimer, "The Coordinate Indexing of Scientific Fields," (Read before the Symposium on Mechanical Aids to Chemical Documentation of the Division of Chemical Literature, September 4, 1951), unpublished.

The substance of this paper has been presented in the text.

Taube, Mortimer and Wooster, Harold, "Information Storage and Retrieval, Theory, Systems, and Devices," (Air Force Office of Scientific Research Symposium), New York, Columbia University Press, 1958. (Number Ten, Columbia University Studies in Library Service).

Thorne, R. G., "The Efficiency of Subject Catalogues, and the Cost of Information Searches," Farnborough, England, Royal Aircraft Establishment, April 1955, 21 p.

"An expression for the efficiency of a subject catalogue or index is derived from the probability of success when using the catalogue, and the cost of making and using the catalogue, compared with the cost of finding material in the library stock when no subject catalogue is available.

"The method, developed primarily for assessing the efficiency of subject catalogues or indexes, can be applied also to author catalogues and other bibliographic aids.

"Numerical examples illustrate application of the method to data from tests of the Catalogue of Aerodynamic Data developed by the National Aeronautical Research Institute Amsterdam, the Uniterm System of Coordinate Indexing, and the Universal Decimal Classification Catalogue of the R. A. E. Library."
[Author's Summary]

Tyler, A. W., Myers, W. L., and Kuipers, J. W., "The Application of the Kodak Minicard System to Problems of Documentation," American Documentation, Vol. VI, No. 1, January 1955, pp. 18 - 26.

Although essentially concerned with the description of a device, this paper is an example of the movement away from pure forms of coordinate indexing. The Minicard Selector was designed to have the capacity to select by several different groups of terms. Within each group the terms were related by the regular Boolean functions but the Selector could operate on several groups at the same time. The relation between the groups was presumably of a different type from the relation between the terms in any one group. This provided the possibility of "multi-level" searching.

Through its use of high-reduction unitized film, the Minicard System also departed from simpler types of coordinate indexing by posting the total text, rather than a number designating the text, under each term.

Research and Development Reports, No. 1 - 20, 1956-1961. U. S. Patent Office. Washington, D. C.

This series of reports, prepared under the direction of Mr. Don Andrews, Chief of the Office of Research and Development, United States Patent Office, and by the Bureau of Standards group working with the Patent Office in accordance with the recommendations of the Bush Committee, describes various attempts to code chemical compounds and to utilized punched card equipment and the Bureau of Standards CEAC for mechanizing searching of patents. In addition, the number of papers by Newman attempt the creation of a new language, which he calls "Ruly English." In this language, every idea would have one word and every word would express only a single idea. To the extent that the creation of such a language is a sine qua non of mechanizing the storage and retrieval of information, it can be assumed that this goal is impossible. The ad hoc experiment carried out by the Patent Office and Bureau of Standards groups has not led to any real advance in the art or pointed out the direction in which the mechanization of Patent Office search is to be achieved. The Kelly Report on The Role of the Department of Commerce in Science and Technology sums up the work described in these reports as follows:

> "On the other hand, there is a possibility that here research has proceeded in the wrong order, in that components of the program have been designed before sufficient thought was given to the operation of the over-all system. Thus there may be here a warning that when research is bound too closely to production, there may be a pressure to show that something specific is being worked on; since initially the only specific things are components, there may result a misdirection of effort."

Vickery, B. C., Classification and Indexing in Science, London, Butterworth's Scientific Publications, 1958.

A defense of the traditional British view that classification is prior to indexing, in this case, that classification is prior to coordinate indexing.

Vickery, B. C., "The Function of Classification in Information Retrieval," ASLIB Aeronautical Group, Fourth Annual Conference, Cranfield, April 1954.

This paper is one of the earliest expressions of a complex of views which has characterized British work in the I. R. field. This view is reflected in the Cranfield Project and has received fuller expression in Vickery's volume, Classification and Indexing in Science.

Vickery, following Robert Thorne of RAE, set himself the task of comparing different indexing systems. This task is, of course, the basis of the research project which Cleverdon has carried out under National Science Foundation auspices at Cranfield. Vickery distinguishes four types of retrieval systems: the alphabetical subject index, the coordinate indexing system, the classified index, and automatic selection. Automatic selection is obviously not a system of indexing like the other three, but a method which can be used with the other three. The paper concludes that whereas a coordinate indexing system may be satisfactory for the actual indexing operation, such a system must be supplemented with a classification of terms. The classification of terms provides clues to the use of the indexing system for anyone unfamiliar with its vocabulary. What Vickery calls the classification of terms has been referred to in other papers as categorization of terms, although the English have attempted to set up very rigorous schedules of terms, following both the UDC and the type of "faceted" analysis developed by Ranganathan. Vickery does not provide any evidence for the basic question he raises, namely, whether categorization of terms is possible for general systems. Although many people have suggested the development of such systems, no one has actually produced a general categorization or classification of a total vocabulary. A distinction is being made here between such a total vocabulary and a categorization of terms in a limited area, e.g., instrumentation.

Vickery, B. C., "Problems in the Construction of Information Retrieval Systems," Journal of Documentation, Vol. 14, No. 3, September 1958, pp. 136 - 143.

> "Speaking for the Classification Research Group, the author lists unsolved problems that ought to be tackled systematically. He states the basis of an information retrieval system to be a lattice of terms with potentially unlimited interconnections, from among which every system must select certain interconnections for display. The selection is

based on postulates concerning the semantic level of terms, the categories of terms to be used, the generic, coordinate, and conjunctive relations to be displayed, and the types of search operation to be conducted. The first problem is to explore the variety of postulates that can be made, and to assess their situations. Other problems are detailed in similar manner, followed by discussions and extracts of the CRG meetings."
[AD Abstract]

Vickery, together with the CRG, which he dominates, has been one of the strongest forces opposed to free or even relatively free I. R. systems. They have insisted on the fundamental importance of tightly structured hierarchical systems.

Vickery, B. C., "Some Comments on Mechanical Selection," American Documentation, Vol. II, No. 2, April 1951, pp. 102 - 107.

This paper is one of the earliest criticisms of coordinate indexing based upon the presumed superiority of a classification system to suggest the terms to be used in both indexing and searching. Vickery recognizes that specific subjects can be both indexed and retrieved by combinations of terms but he feels that both the indexer and the searcher need more than an alphabetical list of such terms in order to use a system. He suggests that any system of mechanical selection be supplemented by an alphabetical index to a systematic classification from which, in turn, the terms used in indexing could be derived.

Wachtel, Irma, "A Punched Card Index for Nuclear Data," American Documentation, Vol. III, No. 1, January 1952, pp. 56 - 57.

This paper describes one of the earliest applications of the Batten or optical coincidence principle to the indexing of a special field of information. The index was designed to enable physicists "to determine quickly and easily which nuclides possess specified combinations of properties." A card is set up for each property and each punching position on the card represents a particular nuclide. Each nuclide has the same position on every card. The search for any nuclide or nuclides having a specific combination of properties is made by superimposing the selected property cards on one another and noting those areas which have holes on all the superimposed cards. This method of searching delivers only product classes. It is a characteristic of optical matrix systems that they are very efficient for product searches and cannot readily be used for sum or complement searches.

Wall, Eugene, "A Practical System for Documentation," Library Journal, Vol. 85, No. 5, March 1, 1960, pp. 889 - 897.

"This paper is concerned with a review of the fundamentals and principles of building an information system. Discusses problems of viewpoint, generics, semantics, and syntactics and their solution through prescription of vocabulary or through redundancy in storage or in retrieval. The technical thesaurus is considered as a means of solving semantic and generic problems. Unit terms and resulting syntactical problems, role indicators, arrangement of units in the system, and abstracts are other aspects discussed."
[AD Abstract]

Warheit, I. A., "Evaluation of Library Techniques for the Control of Research Materials," American Documentation, Vol. VII, No. 4, October 1956, pp. 267 - 275.

This paper is an examination of proposed mechanized systems, including especially the Uniterm System of coordinate indexing, as described in the early papers. It is quite critical of the claims made for coordinate indexing.

Weinstein, Shirley J., and Drozda, Raymond J., "Adaptation of Coordinate Indexing System to a General Literature and Patent File: Machine Posting," American Documentation, Vol. 10, No. 2, April 1959, pp. 122 - 129.

"Describes a procedure for using IBM punched-card equipment to tabulate document numbers for a Uniterm index."
[AD Abstract]

Wildhack, W. A., Stern, Joshua, and Smith, Julian, "Documentation in Instrumentation," American Documentation, Vol. V, No. 4, October 1954, pp. 223 - 237.

This paper contains both a description of a special peek-a-boo device constructed by the Office of Basic Instrumentation of the Bureau of Standards and an indexing system to be used with the

device. The indexing system is characterized as "The OBI Multi-Aspect System." It is a system of coordinate indexing which eliminates connectives and word order and also variant grammatical forms of the indexing terms. However, because the system covers only a very special field, namely, instrumentation, the designers of the system found it possible to set up the indexing terms in a set of categories, e.g., Physical Property, Principle of Measurement, Name of Instrument, Field of Application, etc.

Wise, Carl S. and Perry, James W., "Multiple Coding and the Rapid Selector," American Documentation, Vol. I, No. 2, April 1950, pp. 76 - 83.

Like many of the early papers on coordinate indexing, this paper emphasizes coding problems, largely because the devices suggested, namely, edge-notched cards and even IBM cards, had limited coding areas. This paper suggests the use of an alphabetical code and indicates that such a code developed for keysort cards could also be applied to the proposed Rapid Selector. The major contribution of the paper to coordinate indexing is its recognition that

> "In constructing an index, it is not practical to provide separate entries for every combination of entities, concepts and operations mentioned in the material being indexed. If this were attempted, the index would be too bulky. Nor is it practical to establish as separate classes and sub-classes every possible permutation of all basic criteria used in classification. If this were attempted, the resulting complexity of the system would defeat its own purpose."

However, the bulk of the paper is concerned with superimposition of word codes and the appearance of the paper resulted in a reply from Mr. Calvin Mooers.

Wise, Carl S., "A Punched Card File Based on Word Coding," Punched Cards, Their Applications to Science and Industry, edited by Robert S. Casey and James W. Perry, New York, Reinhold Publishing Corporation, 1951, pp. 93 - 113.

Only the introduction of this paper points out that searches by mechanical sorting methods "may be directed to combinations of words

and phrases." The balance of the paper is concerned with the use of letter codes having mnemonic qualities as opposed to numerical codes; and a consideration of the mathematics of coding.

Wise, Carl S., "Multiple Word Coding vs Random Coding for the Rapid Selector. A Reply to Calvin N. Mooers," American Documentation, Vol. III, No. 4, October 1952, pp. 223 - 225.

In this reply to Calvin Mooers, Wise defends the use of letter coding as opposed to the random number coding recommended by Mooers. Wise admits that word coding may lack the efficiency of random number coding but he thinks this lack of coding efficiency is more than made up by the mnemonic character of word coding which might eliminate the necessity of dictionary look-up at both input and output. Wise argues further that letters can be randomized just as numbers can be randomized and that if one were willing to give up the mnemonic characteristic of word coding, one could use letters just as efficiently in coding as one uses numbers.

BIBLIOGRAPHIES

Bourne, C. P.

Bibliography on the mechanization of information retrieval. Menlo Park, Calif., Stanford Research Institute, 1958. 22 p.

----Supplement I, 1959 25 p.
----Supplement II, 1960. 14 p.
----Supplement III, 1961. 27 p.

Casey, Robert S.

Annotated bibliography on uses of punched cards. In Robert S. Casey, and others, eds., Punched cards: their applications to science and industry. 2nd ed. New York, Reinhold, 1958. p. 638 - 72.

Columbia University. Watson Scientific Computing Laboratory

Bibliography on the use of IBM machines in science, statistics, and education; compiled at the Watson Scientific Computing Laboratory by Joyce Alsop, Anne T. Flanagan, and Eric V. Hankam. New York, International Business Machines Corporation, 1956. 81 p.

Engineering Societies Library, New York

Bibliography on filing, classification, and indexing systems for engineering offices and libraries. Rev. ed. 1960. 33 p.

Jacobstein, J. M.

Indexes and indexing: a selected bibliography of periodical articles. Indexer, v. 1, Sept. 1958. p. 48 - 49.

James, P.

Literature on information retrieval and machine translation: bibliography and index. Yorktown Heights, N. Y., IBM Research Center, Sept. 1958.

Loftus, Helen E., and Kent, Allen

Automation in the library -- an annotated bibliography. American Documentation, v. 7, No. 2, April 1956. p. 110 - 126.

Service Bureau Corporation

Literature on information retrieval and machine translation: bibliography and auto-index. 2d ed. New York, 1959. 38 p. Z699.2.S4 1959

Steiner-Prag, E. F.

Indexes and indexing: a selected bibliography of books and pamphlets. Indexer, v. 1, Sept. 1958. p. 48.

U. S. Armed Services Technical Information Agency

Documentation; a report bibliography prepared by ASTIA. Arlington, Va., December 1961. 278 p. (AD 267 000)

Wayne, Jean M.

Indexing, with emphasis on its technique; an annotated bibliography, 1939-1954. New York, Special Libraries Association, 1955. 16 p. Z695.9.W29

Western Reserve University. School of Library Science. Center for Documentation and Communicati Research

Documentation and information retrieval: a selected bibliography. Cleveland [1961] 7 p.

UNCLASSIFIED