PAGES _____

ARE

MISSING

IN

ORIGINAL

DOCUMENT

# LANGUAGE ORIENTED RETRIEVAL SYSTEMS

# ASTIA



February 1962

ARMED SERVICES TECHNICAL INFORMATION AGENCY

• The Scientific Information Center of the Department of Defense
• An Activity of the Air Force Systems Command

# LANGUAGE ORIENTED RETRIEVAL SYSTEMS

by

Paul H. Klingbiel

February 1962

CONTENTS

## PROLOGUE

This series of papers has been in preparation for well over a year. The initial effort was conceived during the creation of the Thesaurus of ASTIA Descriptors, and the time span of this effort (since early 1960) is indicative of the incubation period for the ideas expressed as much as of the procrastinative tendencies of the author.

I am no longer enamored with the Lattice model of L-O Retrieval Systems. I have long since been disenchanted with the thought that any form of Boolean algebra would function as an appropriate model. The Lattice model taught me only one new fact: the ideal nature of coordinate searching. Since this aspect of coordinate indexing can be communicated easily without the need for Lattice theory jargon, my discovery of this fact through this particular medium is significant only as personal history.

All mathematical models I have seen suffer from the same basic defect: they lack predictive content. A model which merely describes is a scientific bauble. This does not mean that a mathematical theory with predictive power is unattainable for L-O Retrieval Systems. All I wish to convey is my belief that an adequate model lies in the future.

I hope to show that many aspects of L-O Retrieval Systems can be simulated by Monte Carlo methods applied to a suitable Urn problem. If this venture is successful, a model of a sort will have been attained, but I anticipate that the successful conclusion of this project will provide more in the way of data upon which to cerebrate than anything in the nature of a definitive product.

It is not too widely known that an acceptable definition of mathematics does not exist. It is probably even less widely known that there is no acceptable definition of language. If we had a postulational definition of language, and if our research were advanced enough for us to know the properties of the systems obtained by deleting one or more of the initial postulates, then perhaps we would understand what it is we have in the wide variety of current systems of retrieval terms, their specific uses, and their specific limitations. All of this, too, is in the future. We progress toward our goal by explaining to each other what it is we do not know.

# Part I: SYNTAX

CONTENTS

## INTRODUCTION

The decision made by ASTIA in 1959 to automate its essential functions marked the beginning of a new era in documentation. Good librarianship henceforth demands more than a nodding acquaintance with electronic computers and their peripheral equipment. More--now that a new era is here--the need for a thoroughgoing mathematical theory of information retrieval systems becomes urgent.

Mathematical descriptions of retrieval systems are not new; some of them will be discussed in Chapter 3. None of the current theories is adequate; a common fault is that they attempt too much and most vacillate between a mathematical treatment and semiphilosophical or semantic considerations. Such considerations are valid and should be discussed, but in the initial attempt such discussions must be kept separate from the purely mathematical statement.

This mathematical description is restricted to Language-oriented (L-O) Retrieval Systems, i.e., such systems as Subject headings, Uniterms and Descriptors. Minicard, Peek-a-boo and other types of retrieval systems are included only insofar as they depend upon a dictionary of authorized terminology.

Some philosophers of language see three distinct components in the analysis of language: pragmatics, semantics and syntax. My primary concern is with syntax, i.e., a system of formal rules which determine certain formal properties and relations of L-O Retrieval Systems. Considerations of semantics and pragmatics, as they relate to L-O Systems, require separate discussions of their own—discussions which, however, must be influenced by the syntactical structure demonstrated here. Only two aspects of L-O Retrieval Systems are analyzed: the code book and the word sets derivable from it. Considerations beyond these lead into semantics and pragmatics.

The natural vehicle for a discussion of syntax is mathematics. Since all of the mathematics utilized in this presentation is available in standard textbooks, no proofs are presented. The organization of the paper is primarily heuristic; definitions and theorems are stated, and key portions are interpreted (given a realization) in terms of specific examples.

Commentaries such as recently given by Bar Hillel[1] go much too far in a negative direction, and in some instances I believe his criticisms are not valid. In particular, the discussion which follows is concerned with partially ordered sets and lattices as the relevant structures for analyzing the properties of the set D, the set of retrieval terms available to an analyst in categorizing documents, and the power set P(D), the bed of all possible retrieval prescriptions obtainable with D as basic vocabulary. D is a collection of partially ordered sets; P(D) is a lattice. The third set of interest is P'(D), a set derived from P(D), which corresponds to the machine memory and for which no mathematical structure is posited. This latter set will receive its most complete analysis on a semantic rather than on a syntactic level.

Partially ordered sets and lattices have been used before, particularly in the ICSI papers of 1958. I must agree with Bar Hillel that these efforts leave something to be desired, but I disagree with the conclusion that the concepts are useless. Mooer's paper[2] is correct in its essentials. I differ with him in details of emphasis and in the technique of theory construction: the right ideas are sometimes applied to the wrong problems.

In any theoretical investigation of information retrieval as presently realizable by electronic hardware in the L-O field, two sets of hard fact must be accounted for: the repertory of retrieval terms (basic list, authorized vocabulary, thesaurus, etc.) and the retrieval prescriptions derivable from this source. That these elements are analyzable in terms of such simple concepts as partially ordered sets and lattice theory rather than in terms of more arcane disciplines may be psychologically disappointing. But the hard fact is there: we are required to deal with a given set, some of whose elements are class terms, and the power set of the given set. Now, the power set of any given set, among other things, is a partially ordered set, a lattice, an Abelian group, an associative semi-group, a Boolean ring, a Boolean algebra, and a topological space. The theorist is faced with the problem of picking from this welter of permissible structures that one structure which most illuminates the retrieval process.

A final, canonical theory has not been developed for the whole complex field of documentation, but the lack of such a theory is ground for neither despair nor derision. The first discovery of the cellular structure of living matter did not abolish all of the ills to which flesh is heir, nor is the second law of thermodynamics all of physics. It is possible to find models which describe particular libraries, and it is possible to automate some of the functions normally performed by librarians in their reference work. Having forged a hammer with which to pound nails, let us not despair because the same tool will not saw wood.

## Chapter 1

Given a nonempty set D of elements a, b, c, . . . :

1. Def. The power set of D, denoted by P(D), is the set of all of the subsets of D including D itself, and the empty (or null) set $\emptyset$.

2. Def. Given two nonempty sets E and F, the Cartesian product of E and F is that set whose elements are all ordered pairs $(x, y) \ni x \in E,\ y \in F$. This set is denoted by $E \times F$. If $E \equiv F$, then the product is denoted by $E \times E$. If E and F each contain a finite number of elements, say p and q, respectively, then $E \times F$ contains pq elements (x, y).

3. Def. A relation $R_1$ on a set E is a subset of $E \times E$, and $R_1 \in P(E)$.

4. Def. A relation $\mathcal{E}$ is an equivalence relation if $\mathcal{E}$ satisfies the following properties:

   (i) reflexive $\quad a \mathcal{E} a$
   (ii) symmetric $\quad a \mathcal{E} b \Rightarrow b \mathcal{E} a$
   (iii) transitive $\quad a \mathcal{E} b$ and $b \mathcal{E} c \Rightarrow a \mathcal{E} c$.

5. Def. Given a nonempty set D and an equivalence relation $\mathcal{E}$ on D, a subset $G \subset D$ is a residue class modulo $\mathcal{E}$ if

   (i) $G \neq \emptyset$
   (ii) $G \times G \subset \mathcal{E}$
   (iii) $(a, b) \in (G \times D) \cap \mathcal{E} \Rightarrow b \in G$.

   If the collection of all residue classes modulo $\mathcal{E}$ on D is denoted by $\mathcal{S}$, then $\mathcal{S} \subset P(D)$.

6. Def. Let $x \in D \ni \exists\, G \in \mathcal{S} \ni x \in G$. D is the union of all the elements of $\mathcal{S}$, and this <u>covering</u> is represented by $D = \bigcup_{G \in \mathcal{S}} G$.

7. Theorem. An equivalence relation $\mathcal{E}$ on a nonempty set D of elements a, b, c, . . . , partitions D into disjoint, nonempty subsets which cover D.

**THESIS**

L-O oriented retrieval systems have, <u>to a first approximation</u>, the structure given by 7. That is, an L-O system is mathematically representable by a set of abstract elements upon which an equivalence relation is defined which generates residue classes that cover the set.

Example 1: The ASTIA descriptor system

Identify the authorized set of descriptors with D
The Schedule Groups constitute a set of 292 residue classes which cover D
(The Descriptor Fields are another set of residue classes which cover D)
Identify $\mathcal{E}$ with a subject matter relationship

Example 2: The Uniterm system

Identify the posted set of Uniterms with D
Identify $\mathcal{E}$ with a subject matter relationship
Each Uniterm is its own residue class

Note: Compare examples 1 and 2 with the sets of integers J mod (n) and J mod ($\infty$), respectively. The Uniterm system represents a limiting case.

Example 3: The ASTIA subject heading system

Let H represent the set of major headings
Let S represent the set of subdivisions
Identify D with $H \times S$

Note: Subject headings as used were in the Uniterm form, i.e., each ordered class $(h, s) \in D$, $h \in H$, $s \in S$, was its own residue class.

The use of the symbols $\cup$ and $\cap$ in 5 and 6 require comment. The use of these symbols is not in any sense synonomous with the concept or presence of a Boolean algebra. D is not and cannot be a Boolean algebra. To make this point very clear, consider the following axiomatic definition of a Boolean algebra.

8. Def. A Boolean algebra B is a set of elements a, b, c, ..., which satisfy the following conditions:

(i) B has two binary operations cup and cap ($\cup$ and $\cap$) which obey the idempotent, commutative, associative, and distributive laws;
(ii) B has a binary relation, symbolized by $\subseteq$, which is reflexive, antisymmetric, transitive, and satisfies the consistency principle;
(iii) B contains two elements $\emptyset$ and I which are universal bounds and which obey the intersection and union laws; and
(iiii) B has a unary operation of complementation which satisfies the laws of complementarity, dualization, and involution.

Unfortunately mathematicians tend to write only for each other, and consequently the unwary documentalist can take the set D and apparently verify each of the requirements stated in the definition. One crucial phrase is omitted from the definition (which was copied from a standard text[3]) because all mathematicians assume it. Condition (i) implicitly contains the statement that

B is closed under two binary operations . . . .

What does this mean? Simply that if a and b are any two elements whatever of B, then $a \cup b$ and $a \cap b$ must also be elements of B. Clearly this is not the same case for the set D, either abstractly or in any of the realizations given in the three examples above. Specifically, the union (or intersection) of any two descriptors, Uniterms, or subject headings is not again a descriptor, Uniterm or subject heading.

There is another reason why D is not a Boolean algebra which is prior to the requirement of closure. Only one author[4], to my knowledge, has pointed to this particular trap which, almost without exception, every documentalist has fallen into who has attempted to employ Boolean algebra. This trap might be called the inappropriateness of the operation for what is desired of it. A set not only must be closed under a proposed operation, but that operation must be appropriate (make sense) in the context under consideration. This is simply a verbose way of saying that the operations of cup and cap are operations on classes.

A check of statements 1 through 7 will verify that nowhere is it required or postulated that the elements of D are classes. Statements 5 and 6 which employ the cup and cap notation are stated in terms of sets and subsets, i.e., classes.

Finally, the operations of cup and cap are operations peculiar to set theory—a topic of somewhat wider scope than Boolean algebra! What has been glibly labelled as Boolean algebra in most discussions of retrieval systems would better be classified as a hypnotic regard for Venn diagrams.

It is clear, then, that the further development of a mathematical model for L-O Retrieval Systems proceeds along lines which bypass the notion of Boolean algebra. However, further analysis cannot proceed beyond the model given in the thesis immediately following 7 without bringing in the notion of class.

The requirement that the elements of D be classes is not an unreasonable one. It is true that such systems as the ASTIA descriptors, Uniterms and subject headings contain terms which are not classes, but it is equally true that many of their terms are classes. Clearly the notion of class as used in this context is in need of clarification. Many notions or definitions of classes exist. The development in the sequel is based on the notion of class as utilized by formal logicians: classes in extension as opposed to classes in intention.

9. Def. A partially ordered set is a system consisting of a set C and a relation $R_2$ satisfying

(i) a $R_2$ b and b $R_2$ a $\Rightarrow$ a = b
(ii) a $R_2$ b and b $R_2$ c $\Rightarrow$ a $R_2$ c

For this analysis, $R_2$ is to be interpreted as class inclusion. There are four ways in which two classes J and K can be related by inclusion:

(i) $J \subseteq K$ and $K \subseteq J$ or $J \equiv K$
(ii) $J \subseteq K$ but not $K \subseteq J$ or $J \subset K$
(iii) $K \subseteq J$ but not $J \subseteq K$ or $K \subset J$
(iiii) Neither $J \subseteq K$ or $K \subseteq J$. In this case J and K are said to be incomparable.

10. Def. If every pair of elements of a partially ordered set C are comparable, then C is said to be linearly ordered or a chain.

When the partially ordered set is finite, the relation $\subset$ can be expressed in terms of covering.

11. Def. K is a cover for J if $J \subset K$ and there is no element M such that $J \subset M \subset K$.

12. Def. If $J \subset K$ in a finite partially ordered set, then there is a chain $J = J_1 \subset J_2 \subset J_3 \subset \ldots \subset J_n = K$, in which each $J_{i+1}$ covers $J_i$.

This definition enables one to represent any finite partially ordered set by a diagram.

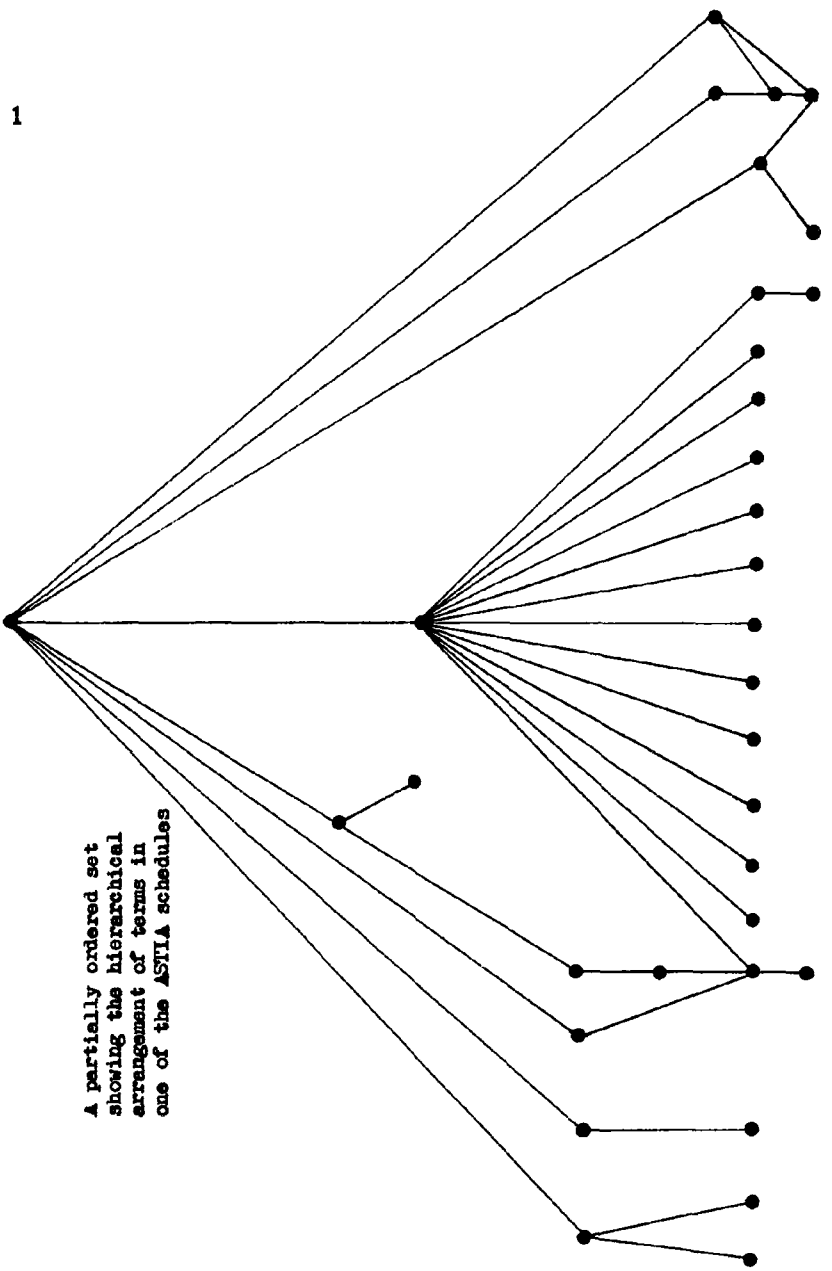
Clearly the set D is a partially ordered system. Some of the immediate consequences of this fact are:

(i) The internal consistency of D (or more usually, of the equivalence classes of D) can be checked;
(ii) Relationships between terms are made visible and help define the scope of related terms;
(iii) Retrieval can be attempted on the basis of known term relationships as shown by the partial orderings.

THESIS

The analysis of the set D demonstrates that L-O Retrieval Systems, to a second approximation, may be categorized as a set of abstract elements upon which an equivalence relation is defined in such a manner that disjoint residue classes are generated which cover D. The residue classes constitute a measure of the degree to which D is classificatory. The fine structure of the residue classes, that of partially ordered systems, is a measure of the degree to which D is hierarchical.

Figure 1



A partially ordered set showing the hierarchical arrangement of terms in one of the ASTIA schedules

P(D) is the power set of D. It is a class of classes, as distinguished from D which, at least initially, as a class of elements. By implication, those elements of P(D) which correspond to the elements of D must also be classes. The precise statement of the relation between D and P(D) is that P(D) is a class which among its elements contains a collection of subsets isomorphic to the individual elements of D, i.e., D is said to be <u>imbedded</u> in P(D). The analysis of L-O Retrieval Systems, then, moves on to a larger context and, insofar as D participates in the properties of that context, the following analysis is a continuing analysis of D.

The analysis in Chapter 1 showed that D is closed under neither the $\cup$ nor $\cap$ operation. P(D) is closed under both. In effect, P(D) has the property that for any sets $A, B \in P(D)$, $(A \cup B) \in P(D)$ and $(A \cap B) \in P(D)$.

13. Def. The least upper bound of A and B is denoted by $A \cup B$ and the greatest lower bound of A and B by $A \cap B$.

14. Def. A partially ordered set in which any 2 elements have a least square upper bound and a greatest lower bound is called a <u>lattice</u>.

It is easily verified that P(D) satisfies Def. 9 and is a partially ordered set. Then by Def. 14 P(D) is a lattice.

15. Def. A lattice is complete if any finite subset $X = \{A_\alpha\}$ has a least upper bound $\cup A_\alpha$ and a greatest lower bound $\cap A_\alpha$.

16. Def. A lattice for which $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ is called distributive.

17. Def. A lattice is called modular if for $B \subset A$, $A \cap (B \cup C) \equiv B \cup (A \cap C)$.

18. Def. An element 1 of a lattice is called an <u>all</u> element if $A \subseteq 1$ for every A in the lattice. Similarly, an element $O = \emptyset$ is called a <u>none</u> element if $\emptyset \subseteq A$ for every A.

Since D is a finite set, P(D) has a finite number of elements. Such finite lattices satisfy two chain conditions:

(i) Descending chain condition - There exists no infinite properly descending chain $A_1 \supset A_2 \supset A_3 \supset \ldots$.

(ii) Ascending chain condition - There exists no infinite properly ascending chain $A_1 \subset A_2 \subset A_3 \subset \ldots$.

Note that if A is a fixed element of P(D), then the subset of elements $X \ni X \subseteq A$ is a sublattice. If $B \subseteq A$, the subset of elements of P(D), $X \ni B \subseteq X \subseteq A$, is also a sublattice of P(D).

19. Def. Given $B \subset A$, we say we have a composition chain connecting A and B if there is a finite sequence $B = A_1 \subset A_2 \subset \ldots \subset A_n = A$ in which each $A_{i+1}$ is a cover for $A_i$.

P(D) is a lattice with an all and a none element, 1 and $\emptyset$. For such systems there exists a composition chain which connects those two elements, and such systems are said to have finite length. The number of intervals in this chain, which is uniquely determined by the system, is called the length (or dimension) of the system. If A is an element of P(D), the sublattice $P(D)_A$ of elements $X \subseteq A$ satisfies the same conditions imposed on P(D); in particular, the two chain conditions hold. A is the all element of $P(D)_A$. The length of $P(D)_A$ is called the rank, r(A) of A. Now if $B \subseteq A$, then $r(A \cup B) = r(A) + r(B) - r(A \cap B)$. This formula is called the fundamental dimensionality relation for modular lattices.

20. Def. A subset A of the lattice P(D) is called a principal ideal (a) if A consists of all $X \in P(D) \ni a \subseteq X$ for fixed $a \in P(D)$.

21. Def. A lattice P(D) with $\emptyset$ and 1 is said to be complemented if for every $A \in P(D)$ there exists an $A' \ni A \cup A' = 1$, $A \cap A' = \emptyset$.

22. Def. If A is any element in P(D) with $\emptyset$ and 1, an element $A' \ni A \cup A' = 1$, $A \cap A' = \emptyset$ is called a complement of A.

23. Def. If $B \subseteq A$, an element $B_1 \subseteq A \ni B \cup B_1 = A$ and $B \cap B_1 = \emptyset$ is called a complement of B relative to A. (This means that for every $A \in P(D)$, the sublattice $P(D)_A$ of elements contained (or equal) in A is complemented.)

24. Def. In a modular lattice P(D) with $\emptyset$ and 1, a finite set $A_1, A_2, \ldots, A_n$ of P(D) is called join independent if $A_i \cap (A_1 \cup \ldots \cup A_{i-1} \cup A_{i+1} \cup \ldots \cup A_n) = \emptyset$ for $i = 1, 2, \ldots, n$.

25. Def. An element p of a lattice with $\emptyset$ is called a point if p is a cover of $\emptyset$.

26. Theorem. If P(D) is a complemented, modular lattice that satisfies both chain conditions, then the all element of P(D) is a least upper bound of independent points.

The mathematical (or syntactical) analysis is now practically complete. There remains the task of identifying some of the mathematics with L-O Retrieval Systems. Additional discussions of these identifications and their implications are given in the next chapter in connection with commentaries on other papers.

At first thought P(D) would seem to represent the computer memory (insofar as that memory is concerned with information retrieval) because of the elements of P(D) are sets of retrieval terms, and documents are categorized and retrieved precisely by means of matching such sets against sets evolved from request actions. But this first impression is incorrect in several ways: first, P(D) does not represent the computer memory in any explicit sense, but rather represents the totality of all possible sets of retrieval terms assignable with D as generator; secondly, P(D) is more than just this totality becuase P(D) contains D, the generating set (more precisely contains a set isomorphic to D), and these elements, singly, categorize none of the documents in the collection. P(D) must, then, be interpreted as representing the entire mathematical universe or syntactical bed of L-O Retrieval Systems inasmuch as it contains not only the code book D, but also every

potential use of that book in terms of sets of retrieval terms available with which to categorize documents. A more precise mathematical formulation is given in the final statement of the thesis at the end of this chapter.

The analysis just concluded demonstrates that the mathematical structure of L-O Retrieval systems is that of a modular lattice (Def. 13-17). This lattice has an all element and a none element (Def. 18). For practical retrieval purposes neither of these elements is useful. Their presence is needed only for the theoretical aspects of the system. The lattice is finite, and therefore the two chain conditions require no mathematical comment. They are, however, of practical importance in the actual retrieval process, particularly since chains are involved in the concept of composition chains (Def. 19) and these, in turn, are (practically speaking) derived from certain lattice ideals (Def. 20).

As the discussion in the next chapter shows in some detail, the maximal search pattern defined on the basis of a request for information is precisely the lattice complex generated by the retrieval term statement of the request. After the primary retrieval ideal (retrieval lattice, since every ideal is a sublattice) has been identified by the reference analyst, it is extended "downward" to generate one or more additional ideals. From these, in turn, composition chains are chosen which define the actual search pattern to be conducted to fulfill the initial request.

Later concepts (Def. 21-25) are needed for a theorem (Theorem 26) which is needed for the mathematical formulation of P(D) as stated in the final thesis. The fundamental dimensionality relation for modular lattices and the concept of completeness (Def. 15) are referred to in the next chapter in a discussion of some of the statistical aspects of P(D) and of retrieval processes.

THESIS

The analysis of the set P(D) demonstrates that L-O Retrieval Systems, in the third and final approximation, may be categorized as a complemented modular lattice which satisfies both chain conditions and in which the "all" element in the least upper bound of independent points. The distinguished set of independent points is isomorphic to the set D. The disjoint residue classes which cover D indicate the degree to which D is classificatory. The fine structure of the residue classes indicates the degree to which D is hierarchical.

## Chapter 3

The papers to be discussed in this chapter were chosen primarily because they illuminate in interesting fashion the preceding analysis. Before proceeding to this discussion, two other papers are singled out for special notice. The first of these papers is by Bar Hillel[4]. This paper should be required reading for anyone doing any serious work in either the pragmatic, semantic, or syntactic features of L-O Retrieval Systems, for it introduces much needed fresh air into a densely smoke-filled topic. The second paper is by C. Mooers [5]. His analysis is in no way incompatible with the analysis presented here. The problem discussed by Mooers is prior to the discussion in this paper in the sense that he builds a mathematical theory of epistemology which terminates with the concept of the descriptor. The present analysis takes the descriptor (retrieval term, subject heading, Uniterm, . . . ) as given and proceeds to draw necessary consequences from that base. Mooers paper may be said to deal with the microstructure of L-O Retrieval Systems: the present analysis is concerned with the macrostructure of those same systems.

The first paper we wish to discuss at some length was authored by Dr. R. C. Buck[6] of the University of Wisconsin. His analysis was motivated by the desire to control by computer (actually the preparation of a cumulative subject index) the specialized collection of some 70,000 documents reviewed for Mathematical Reviews (the largest eventual figure mentioned is 150,000 documents). Buck essentially begins with a set $\Sigma$ which he says is "somewhat ill-defined and chaotic." The set $\Sigma$ differs from P(D) in that $\Sigma$ represents just those sets of retrieval terms which are actually assigned to documents rather than to the mathematical structure in which such sets must of necessity be imbedded. (The corresponding set in the system proposed here is labelled P'(D) and it will be discussed in due course.) By working with the actual set of retrieval terms instead of with P(D), Buck fails to perceive the lattice structure which underlies L-O systems. The "Uniterm complexes" which he recovers from $\Sigma$ are sets in their own right in P(D) and do not require an intersection process for their recovery or isolation. Indeed, the whole idea of intersection and of coordinate indexing requires clarification. The retrieval process is one of matching one list or set of terms against another list, not one of intersecting sets or lists to find common elements.

A point almost universally overlooked is that the relation of any document in the system to all other documents in that same system is determined the moment the retrieval terms categorizing that document have been selected and stored in the computer memory. Whether or not that document is selected in a properly conducted search has nothing to do with the class intersection operation. The document is related through its assigned retrieval terms by inclusion (in one of the four possible ways mentioned in Chapter 1) to all of the other documents in the system: the intersection of its set of retrieval terms with any other set of retrieval terms signifies nothing unless a null result of such an operation is interpreted as a "miss," a result giving the original set a "hit," and anything in between as ambiguous. But certainly this is an obtuse way of describing a simple matching process. (The intersection process in retrieval is a result of thinking oriented solely to term deck arrangements.)

The pseudo topology by which a covering is obtained for $\Sigma$ bears interesting relationships to the lattice ideals and composition chains which are the natural retreival vehicles in the lattice model given here. A feature of the nested sets defined by Buck is their intended use to specify the depth of retrieval desired to satisfy a given request, and this usage of nested sets is a clever approach to the problem of retrieval depth. This aspect of L-O systems has received only passing mention in this syntactical investigation because the problem of retrieval depth is more properly analyzable in terms of a semantic context. A searching semantic analysis of what constitutes depth of retrieval would quite surely have syntactical ramifications, and in that sense there is considerable interplay between discussions on the semantic and syntactic levels. The syntax present here must influence any semantic discussion, but the primary location of the discussion of the particular problem of retrieval depth is semantic. Apart from the feature of retrieval depth, Buck's nested sets correspond to the concept of the composition chain which is the end product of the reference analysts procedure to prepare a request for machine search. That these nested sets are the last step in a process derivable from the mathematical structure of L-O systems is a fact which Buck misses precisely because he missed the basic underlying mathematical structure when he began his analysis with $\Sigma$ instead of with P(D).

The code construction suggested by Mr. Buck is more ingenious than practical. He acknowledges the practical difficulty of the character (or word) size of his code. It also appears to be impractical as a tool to be used by the average librarian, although admittedly this was not one of Mr. Buck's concerns since in the very beginning of his paper he states that after all the easiest answer to the whole problem might be to subsidize mathematicians to act as living retrieval systems. This solution is not available for discussion in the present context, so the practical difficulties inherent in the code remain. Part of that difficulty is that the code is intended to serve the multiple ends of machine access and of retrieval depth significance. Since the coding problem is inseparably connected with retrieval depth in Buck's system, no further comment is warranted at this time.

In discussing actual retrieval problems, Mr. Buck sees the possibility of the utility of programming for partial solutions to requests rather than scanning the entire collection for every request. As he points out, the resolution of this problem is dependent on the size of the collection and on the capabilities of the machine. Another very practical related matter not mentioned by Buck is that of available machine time.

To sum up, Mr. Buck's contribution fails in its intent because the underlying mathematical structure of the system is overlooked, and becuase too much is attempted with too little in terms of what can be expected of a numerical code at the present time. The paper is nevertheless important because the role of nested sequences of sets in the retrieval process is at least partially recognized as well as the possible role of partial solutions to retrieval searches.

Before turning to the discussion of the final paper, it is advantageous to spell out in some detail the retrieval process as derived from the mathematical structure of P(D). To keep this example within bounds, a full discussion is given only for the case of a simple retrieval problem determined by six retrieval terms. Implications for searches of lesser and of greater magnitude will be given as the argument progresses.

The elements of P(D) are classes or sets of retrieval terms and these have been indicated here by capital letters, thus, $A \in P(D)$. For purposes of illustration, the internal components of A, the individual retrieval terms, need to be displayed. This is easily done by identifying the various class elements of P(D) with suitable n-tuples: if A is a class of six retrieval terms, set A = (a, b, c, d, e, f) or, more simply, (abcdef), where the lower case letters indicate the individual retrieval terms which make up the set A. Such n-tuples are initially assumed to be unordered. Given a retrieval request based on precisely the six retrieval terms denoted by (abcdef): this set is certainly an element of P(D), but it may not be a member of the computer memory, and the disparity between P(D) as an idealized mathematical structure and P'(D) as the actual computer memory must be kept in mind.

The retrieval problem is first defined in terms of the theoretical structure, P(D). The set (abcdef) exists in P(D). This set is also imbedded in every element of the ideal generated by A. Therefore the maximal search pattern for retrieval (the set of elements which the reference analyst must consider in preparing a request for machine search) and the actual search pattern required to satisfy the request seemingly coincide. (To repeat, the maximal search pattern is the ideal generated by A. Note, too, that in this theoretical case, the problem of partial solutions, i.e., of sets of retrieval terms derived from (abcdef) by deleting certain terms, does not arise.) But this does not mean that every document selected [and in P(D) every element of this ideal would be selected in this search since every element exists in P(D)] is a bonafide answer to the request. Some of the sets in which A is imbedded will contain A only accidentally, not essentially. This is the problem of noise - or combinations of retrieval terms which only abstractly satisfy the request. The noise problem is deferred for the moment except to point out that it can be controlled to some extent in this case by not accepting as solutions sets of retrieval terms containing (abcdef) which are larger than nine elements, say, under the assumption that larger sets would contain (abcdef) only incidentally and not essentially in terms of the request in hand.

We now consider (abcdef) as a request to be satisfied in P'(D). The first step in processing this request is to enlarge the domain of interest to P(D) and to construct again the ideal generated by A. This previously considered maximal search pattern is only partially realizable in P'(D), i.e., most of this ideal is vacuous. Practically, an upper bound for the search is attained if it is known that no document in the collection is categorized by more than n terms. The ideal generated by A obviously can extend no further - that is, there is a natural cutoff point. But there is an additional problem in P'(D) which has no counterpart in P(D) and that is the possible nonexistence of A = (abcdef) itself! This would mean that even that portion of the ideal generated by A which we might expect to find in P'(D) on the basis of the known properties of P'(D) is also vacuous. That is, the ideal generated by A is totally vacuous in P'(D), and we can at most satisfy the request only in part. The search pattern required for this case is the dual of the ideal used previously. The first ideal was the ideal generated by A. We now form the ideal of those elements contained in A. Since every ideal is a lattice, we list the elements of the lattice generated by the six retrieval terms of A. These elements are:

$\emptyset$ (the none element for this lattice)

a, b, c, d, e, f

ab, ac, ad, ae, af, bc, bd, be, bf, cd, ce, cf, de, df, ef

abc, abd, abe, abf, acd, ace, acf, ade, adf, aef, bcd, bce, bcf, bde, bdf, bef, cde, cdf, cef, def

abcd, abce, abcf, abde, abdf, abef, acde, acdf, acef, adef, bcde, bcdf, bcef, bdef, cdef

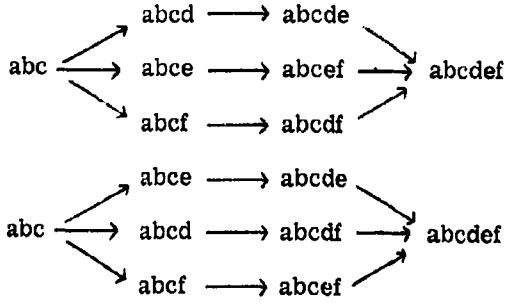
abcde, abcdf, abcef, abdef, acdef, bcdef

abcdef (= A, and the all element for this lattice).

These 64 sets must be considered in terms of partial solutions. Not every one of these sets is a valid partial solution, but every set must be carefully considered since any relevant set overlooked is a retrieval possibility which remains unrealized. Within this lattice (the statement is true without this qualification also) every element of the set of 64 generates an ideal, so there is a maximum of 64 ideals to consider. However, since the interest is in possible partial solutions, there must be certain terms among the six defining terms (a, b, c, d, e, f) which are of less importance than others in the sense that documents lacking these terms may still be acceptable solutions to the retrieval problem. This implies that a, b, c, d, e, and f can be graded in order of importance—or, in short, that they can be ordered. Let the terms be ordered from left to right, and let us further assume that documents characterized by less than three of the six terms are of no interest for this particular search—i.e., if the first three terms in order of importance are not present in the set which characterizes a document, that document is rejected in the search. For simplicity we will assume that the three most important terms in order are a, b, and c. This means we are interested in the ideal generated by (abc). This ideal contains every set of the lattice of 64 elements generated by a, b, c, d, e, and f which contain (abc). These elements are:

abc, abcd, abce, abcf, abcde, abcdf, abcef, abcdef.

The elements of the ideal permit a choice of six composition chains:




or, separately written, and in descending order so that the element on the right is always a cover for the element on the left, we have:

abcdef ⊃ abcef ⊃ abce ⊃ abc

abcdef ⊃ abcde ⊃ abcd ⊃ abc

$$abcdef \supset abcdf \supset abcf \supset abc$$

$$abcdef \supset abcde \supset abce \supset abc$$

$$abcdef \supset abcdf \supset abcd \supset abc$$

$$abcdef \supset abcef \supset abcf \supset abc.$$

Careful checking of the retrieval terms at this point may show that some of these combinations are not useful, and this may eliminate certain sets from the chain or may even eliminate complete chains. The remaining elements (chain or chains) together with the ideal generated by A = (abcdef) is the basic search pattern for this particular request. One further proviso is required. Just as in the case of the theoretical problem in P(D) where we suggested a cutoff to keep the ideal generated by A within bounds, so too a cutoff should be provided here in terms of numbers of retrieval terms allowed for the members of the chain or chains to be employed. Unless such a cutoff is provided the program would count as acceptable the ideals generated by each member of the chain(s) in the search pattern.

To summarize, the mathematical theory of L-O systems identifies the maximal search pattern for any set of descriptors or retrieval terms, and provides a theoretical solution for the retrieval problem in terms of lattice ideals and composition chains. Practical devices for assessing each combination and for cutting down the elements of the search where a large number of retrieval terms are involved are not given as a consequence of the theory but are a matter of technique. A few such techniques will be mentioned in the discussion to follow on the probability and statistical aspects of P(D).

The paper by H. M. Wadsworth and R. H. Booth[7] presents the statistics of pure random retrieval and of random retrieval under various side conditions including searches restricted to a portion of the entire collection where that portion is obtained by probability distribution considerations. The formula n/S is given as the probability of picking (in a random search) one pertinent entry from a collection of S entries which contains n pertinent entries; that is, given a collection of S of 200,000 documents and n 50 pertinent documents pertaining to a search problem, the probability of finding one of those pertinent entries in one random selection from the collection is 50/200,000 or 0.00025. This figure can also be interpreted in terms of the amount of information (or density of information -- but do not push this interpretation too far!) lodged in each of the 200,000 entries relative to the particular search in question. The figure may also be interpreted as a measure of the efficiency of a search involving the whole collection when n items at most can be expected to be pertinent.

The authors give the usual definitions for the probability of disjoint events in which at least one of the events (pertinent items) is retrieved and also the more general formula where the events are not necessarily disjoint. This latter formula,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

has the same form as the fundamental dimensionality relation for modular lattices. This is not too surprising when it is noted (and apparently it has not been noted previously) that P(D) is a $\sigma$-algebra. This fact is a consequence of the

theory (Def. 15) already presented. The use of these formulas is valid provided they are used in connection with multiple searches which may have elements in common, or to subsearches of a given search. They are not valid (or at best are meaningless) if applied to a single search.

Wadsworth and Booth also give some consideration to searching only a selected area of the whole collection in order to increase the "efficiency" of the search, and their primary concern is that the baby is not thrown out with the bath. Such a concern is always legitimate, but in this case its primary cause is the lack of a clear concept of the mathematical structure involved in the memory, whether in terms of P'(D) or P(D).

Several lines of thought converge at this point in connection with partial searches, searching only a part of the collection, large searches, and the efficiency of searching. These problems can best be discussed against the backdrop of the following kinds of working tools:

(i) a frequency count of the use of each retrieval term;
(ii) a frequency count of the number of documents categorized by a given number of retrieval terms;
(iii) a cumulative frequency diagram derived from (ii).

P(D), whatever the generating set D may be, is strongly binary, and on that basis a binomial distribution is given for P(D) lattices of $2^{16}$, $2^{17}$, and $2^{18}$ elements, respectively. This chart shows the number of documents categorized by sets of 2, 3, 4, . . . , retrieval terms each if that distribution is binomial. One glance at the chart is convincing evidence that most collections would not be binomially distributed. The peak is much too sharp, and the tails of the distribution are unrealistic. A guess at the distribution for a document collection the size of the current ASTIA Document collection (200,000) is also given. This is followed by a cumulative frequency distribution based on that guess, except that the cumulation is the reverse of that usually presented. The intersection of any ordinate and abscissa of this chart shows the number of documents in the collection with a retrieval set whose size is equal to or greater than the cardinal of the abscissa.

A frequency count of the use of each retrieval term used in conjunction with a cumulative frequency diagram of retrieval sets would enable the reference analyst to place numerical bounds upon the number of expected hits for a given search, including the case of partial searches when it is evident or probable that the original set of retrieval terms defining the search may not exist as such in P'(D).

The theory of P(D) makes no special provision for extremely large searches of a bibliographic nature which may involve 25 or more retrieval terms. Nevertheless, some guidance can be obtained from the general background developed in this paper. Given a frequency count of the number of documents categorized by a given number of retrieval terms, the probability can be fairly accurately estimated of whether or not documents exist in the collection which by themselves satisfy the search. A search characterized by six retrieval terms would have a fairly good chance of success in terms of single documents each of which contain the six terms in question because the chart (guess) shows that there are about 20,000 documents with precisely six retrieval terms in the collection. Note, too, that the bulk of the curve is to the right of this abscissa. This is verified by

Figure 2

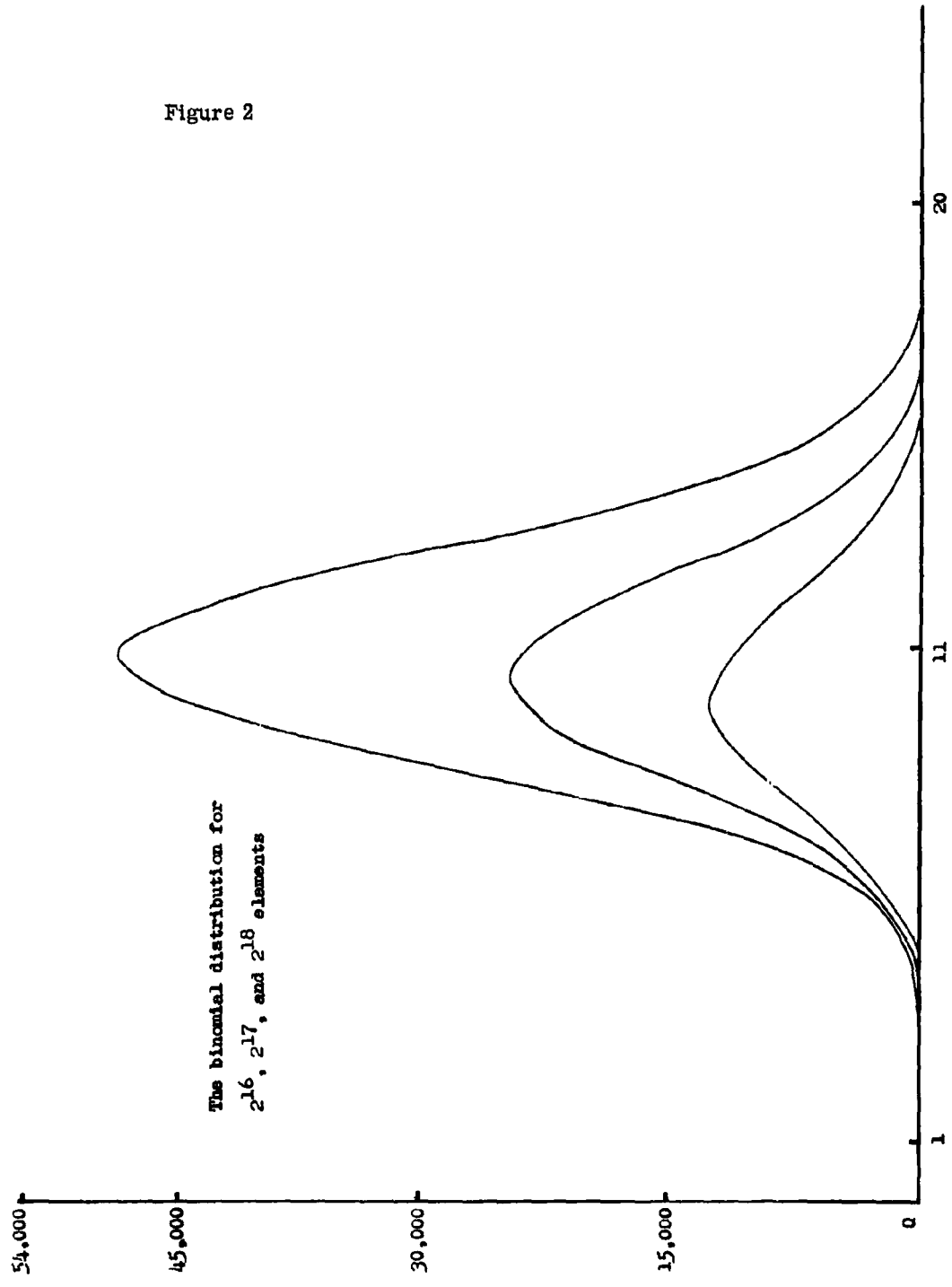

The binomial distribution for
$2^{16}$, $2^{17}$, and $2^{18}$ elements
54,000
45,000
30,000
15,000
0
1
11
20

Figure 3

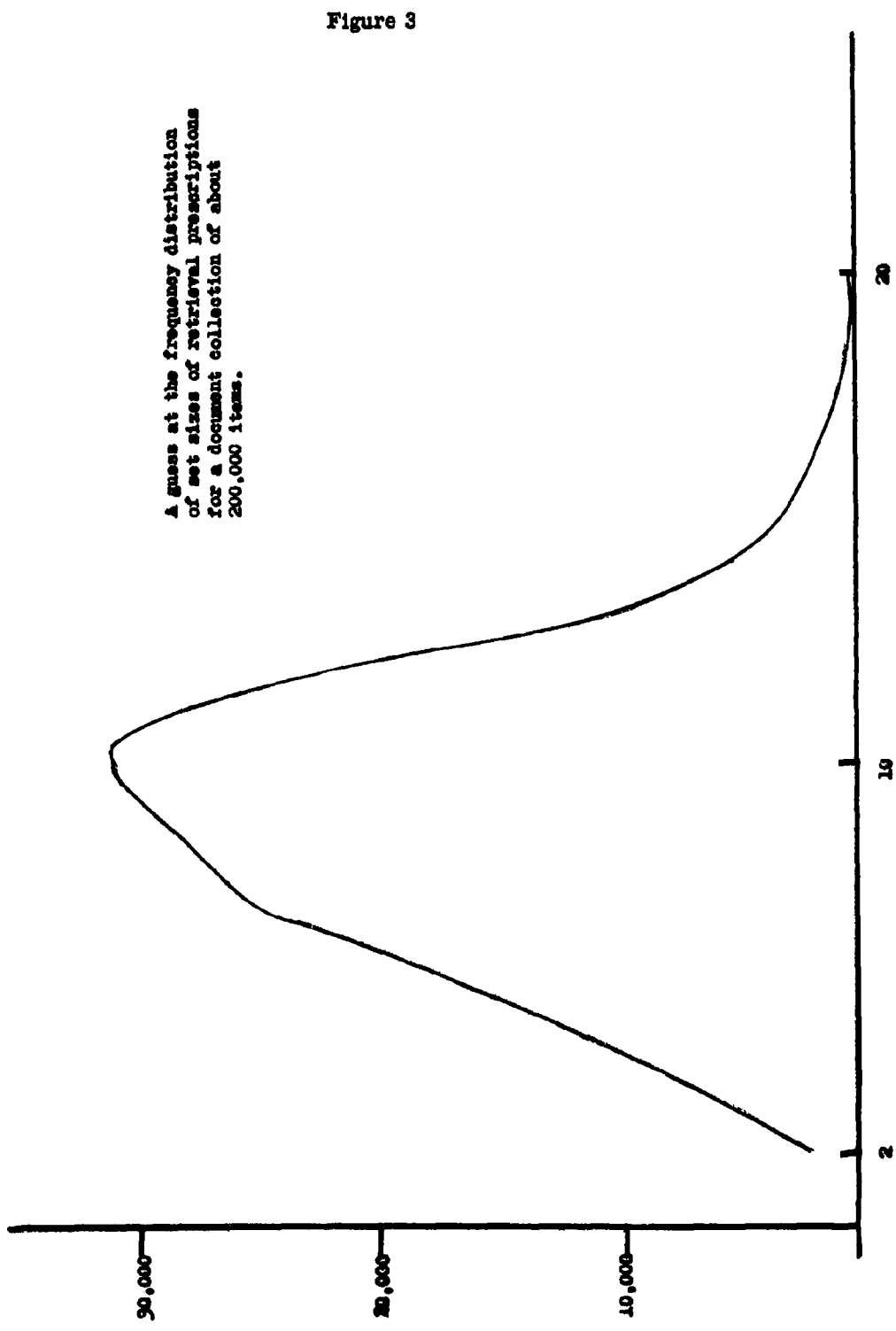
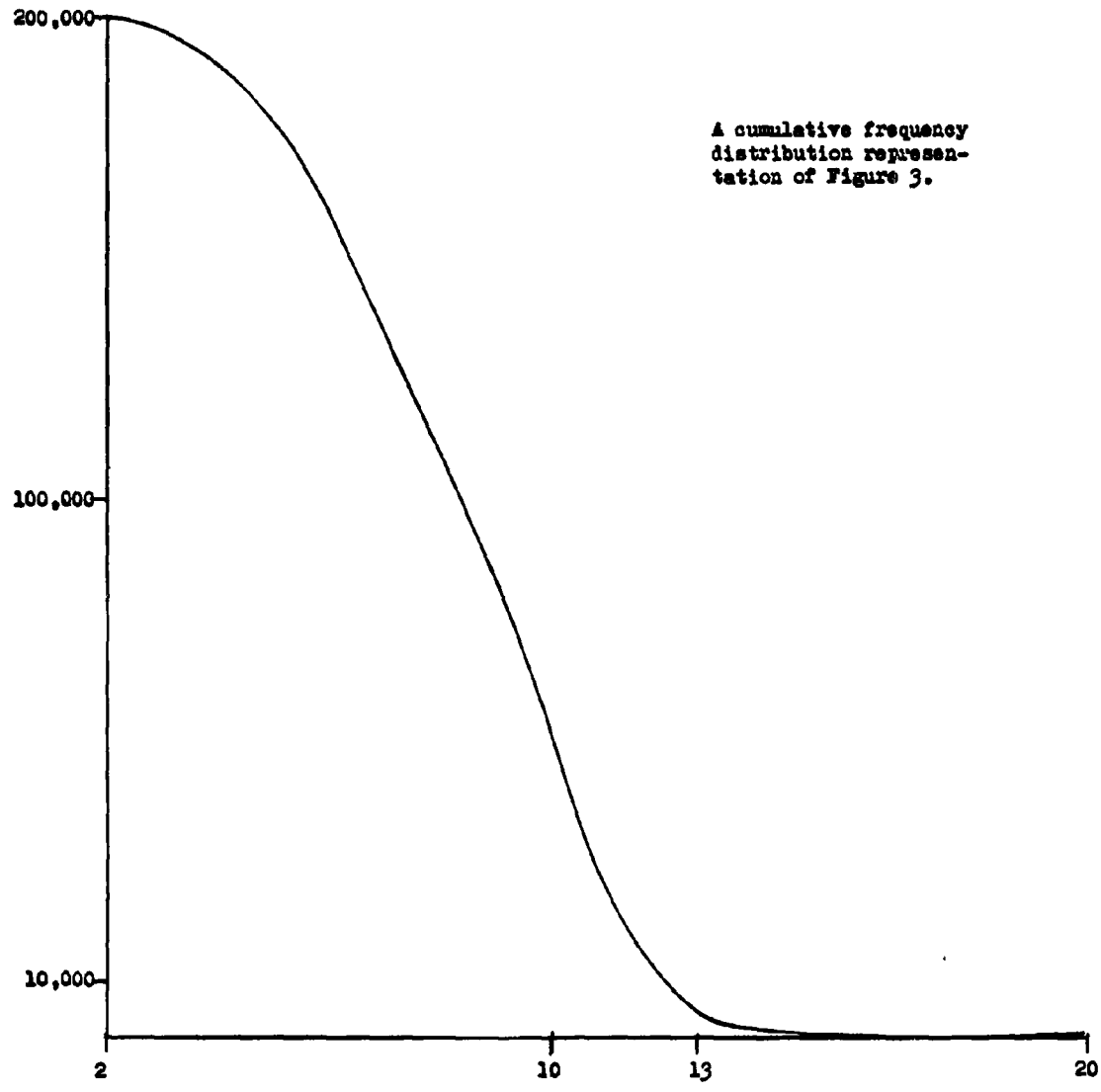

A guess at the frequency distribution
of set sizes of retrieval prescriptions
for a document collection of about
200,000 items.
30,000
20,000
10,000
2
10
20

Figure 4




A cumulative frequency distribution representation of Figure 3.

consulting the cumulative frequency distribution chart which shows that there are about 168,000 documents which are characterized by six or more retrieval terms in a collection the size of the ASTIA Document collection. With these figures in mind, the ideal generated by A = (abcdef) looms as a rather formidable collection of documents. The magnitude of the figures involved adds a convincing argument to the concept of a cutoff point beyond which it will be assumed that (abcdef) is present only accidentally.

Consider the case of a bibliographic search involving 25 or more retrieval terms. The frequency distribution of retrieval sets shows that there is no single document in the collection with that many terms. This fact alone (there are obviously other factors from which the same conclusion is deducible) shows that either a very general bibliographic subject is being requested or that the subject is many-faceted or intra-discipline. The first stage in any such proposed search is a careful analysis of the given retrieval terms to locate relationships which tend to group the 25 terms into smaller subsets. If no such tendency is apparent, then—at least in the ASTIA collection—careful consideration should be given to the schedule designation of those terms. If the bulk of the terms falls into a single schedule, it may be feasible to conduct a preliminary search which would isolate that schedule in toto onto a working tape and to later fine search just this portion of the collection. If the 25 or more terms do fall into subclasses, the schedule search is still a possibility, especially if the frequency charts indicate a heavy response, or the subgroups can be programmed as separate searches, and the results of these searches could then be matched to eliminate duplicates. The actual method to be employed, in terms of search efficiency and of machine time, must be worked out for each special case, and it would be hoped that experience would indicate optimum procedures to be followed.

## Conclusion

Neither the mathematical model presented in the first two chapters nor the discussion of some of its implications as brought out by contrast with other analyses as discussed in the third chapter is grounds for considering the contents of this paper as a panacea for solving all of the problems connected with L-O Retrieval Systems. As pointed out in the introduction, the present analysis is at best only one third of the picture; companion discussions of the semantic and pragmatic aspects of L-O Retrieval Systems are required before any claim to near completeness can be made. However, within the self-imposed limitations of the preceding analysis, certain definite conclusions seem warranted.

First, a mathematical structure has been presented which is more than descriptive. Certain consequences follow from the mathematical structure as to the maximum search pattern for retrieval, and mathematical correlates are found for the ideas of hierarchy and classification. The dictionary is recognized, mathematically, as a distinguished set of points within a total complex. Buck's paper[6], discussed in Chapter 3, missed the underlying mathematical structure, although some of the consequences of that structure, such as the role of nested sequences of sets, are discovered from other considerations. The statistical analysis presented in the next paper[7] is true so far as it goes, but again several points have been missed because of a lack of an underlying structure, and a more pertinent analysis is possible once that structure is known. The role of a $\sigma$-algebra was overlooked in that paper, yet this is of primary importance in a discrete probability context. Much ado was made of getting around the necessity of searching the whole collection for each request; again, more pertinent comments are possible with a firm model in mind. Selected parts of the collection need not be chosen on the basis of some probabilistic distribution with the inescapable loss of information that such a procedure implies. With proper coding any collection which has been referenced by a code book split up into suitable residue classes can easily select from the total collection one or several relevant residue classes with every expectation of fulfilling the motivating request with a probability of 1, providing the information asked for exists within the collection. Even if residue classes had not been utilized or were not a part of the system, relevant parts of the collection can be separated out for further study on the basis of lattice ideals.

Second, the fine structure of residue classes has been uncovered. This advances the theory of retrieval terms beyond the notion of the residue class, just as the residue class concept is an advance over that of a simple alphabetical listing. Any attempt, no matter how small or abortive, to display the fine structure of residue classes in terms of specific diagrams of partially ordered sets will pay large dividends. The process involved in attempting such a display calls for a searching re-examination of the definition of each term, its intended usage, and its relation to other terms.

The analysis as presented is very close in spirit to Mooers' ICSI paper[2], but there are sufficient differences between our respective treatments and procedures so that my results are not mere restatements of Mooers' work. In the abstract to his paper Mooers states that "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." My position denies the uniqueness of these systems. Any language system has hierarchical elements which are

discernible even on the syntactical level, and consequently a nonhierarchical system must be a deliberate construction containing no class terms. Characters with logic do not qualify as a separate category because such systems simply illustrate one possible way of programming an L-O system for retrieval.

Mooers investigates the properties of two spaces: P, the space of all possible retrieval prescriptions (identical to my P(D)) and L, the space of all document subsets. This latter space could be generated from P'(D) as P[P'(D)] in my terminology. The set D is not analyzed by Mooers as such, but is simply referred to as "the repertory."

Mooers considers P a partially order set generated by the cardinal product of more elementary partially ordered sets, while L is categorized as a Boolean lattice. I believe that partially ordered sets are relevant only to the set D (the repertory), and that P = P(D) is not distinguished as a partially ordered set but by its lattice structure (Mooers later recognized this structure). The recognition of certain retrieval terms as classes in extension by-passes most of the analysis of Mooers "characters with hierarchy" and simply and objectively answers most of the syntactical problems connected with hierarchy. The semantic problems require a completely separate treatment.

Another substantial difference in our analyses is my investigation of the retrieval process within the confines of P(D) while Mooers conducts his inquiry in terms of L = P[P'(D)]. The space L is well defined, but I am not convinced that it is the appropriate space in which to study retrieval, or, if it is, that its lattice structure is pertinent for that analysis.

A concluding statement is required as to the relation between L-O Retrieval Systems and Boolean algebra. And at this point a deception perpetrated upon the reader (mathematicians excepted) must be admitted. The nature of this deception is given by:

27. Def. A Boolean algebra is a lattice with a none and an all element that is distributive and complemented.

In brief, P(D), is a Boolean algebra.

The deception was deliberate, but by no means a prank. The unfortunate fact is that the Boolean nature of much of L-O Retrieval Systems was discovered far too early in the history of the discipline. The intuitive leap that led from the conventional subject heading concept to broader retrieval concepts was too perceptive; this stifled further advance. Given that retrieval systems seemed to be Boolean algebras, practically nothing of any consequence could be drawn from that fact. Apparently the only definition for Boolean algebras known was one akin to that given in Chapter 1, and retrieval was thought of almost exclusively in relation to term decks. Consequently, theorists busied themselves intersecting and coordinating everything and anything at hand with the result of much commotion but little advance in the state of the art.

To substantiate these remarks and to demonstrate that the identification of P(D) as a Boolean algebra is not the supreme cap to a long analysis, the mathematical analysis is extended for a few more steps.

28. Def. Let B be a Boolean algebra. The composition $a + b = (a \cap b') \cup (a' \cap b)$ is called the symmetric difference of a and b.

It can then be shown that B is an abelian group relative to the operation +. In addition, B is a semigroup relative to the operation (which we now denote by $\cdot$). Then $(B, +, \cdot)$ is a ring.

29. Def. A ring is called Boolean if all of its elements are idempotent.

30. Theorem: The following two types of abstract systems are equivalent: Boolean algebra, Boolean ring with identity.

This theorem implies that the analysis of Boolean algebras can be conducted in an equivalent system, Boolean rings with identity. Now rings are fairly well understood mathematical structures. They represent generalizations of the notion of a field. The analysis of ring structure is conducted by means of ideals. But ring ideals are quite different structurally than lattice ideals -- so different in fact, that ring ideals appear to have no relation to the retrieval problem, whereas, as we saw, lattice ideals are the essence of that problem. This lack of application of ring ideals to the problem at hand is strong presumptive evidence that the Boolean structure is not really relevant to the retrieval problem -- whereas the lattice structure is. Again, what is important is not that P(D) is a Boolean algebra, but that the Boolean algebra involved here is a certain kind of lattice.

Finally, if the Boolean aspect of L-O Retrieval Systems is to have any great importance in the mathematical analysis of such systems it will be through considerations hitherto not investigated; that is, the relation of Boolean rings and $\sigma$-algebras to measure theory. If any strong connection can be established here, then P(D) may be subjected to a rather sophisticated statistical analysis.

The syntactic theory has reached its present capabilities. Further advances depend upon an interplay of experience and theory, the one suggesting advances to the other.

## Bibliography


1. Some Theoretical Aspects of the Mechanization of Literature Searching
   Yehoshua Bar-Hillel
   ONR Contract N62558-2214, Technical Report No. 3
   AD-236 772

2. A Mathematical Theory of Language Symbols in Retrieval
   Calvin N. Mooers
   Preprints of Papers for the International Conference on Scientific Information, Washington, D. C., 1958.

3. A Survey of Modern Algebra
   Garrett Birkhoff and Saunders Mac Lane
   The Macmillan Co., N. Y.

4. A Logician's Reaction to Recent Theorizing on Information Search Systems
   Yehoshua Bar-Hillel
   American Documentation, Vol. 8, No. 2, pp. 103-113
   AD-139 565

5. Some Mathematical Fundamentals of the Use of Symbols in Information Retrieval
   Calvin N. Mooers
   Zator Company, Cambridge, Mass.

6. On the Use of Gödel Indices in Coding*
   R. Creighton Buck
   The Mathematics Center, The University of Wisconsin
   AD-207 325

7. Some Statistical Sampling Concepts Applied to the Information Retrieval Process of Documentation Systems
   H. M. Wadsworth and R. H. Booth
   Center for Documentation and Communications Research
   Western Reserve University
   AD-201 864


Prepared 9 May 1960

*Also printed in American Documentation, July 1961.

# Part II: SEMANTICS

CONTENTS

# INTRODUCTION

Perhaps the first problem to be settled in any discussion of semantics is the problem of "whose." Is the semantics under discussion of the type associated with Ogden and Richards[1] Meaning of Meaning, is it of the Korzybski[2] Etc. variety, or is it semantics as a formal logician such as Carnap[3] sees it in his Introduction to Semantics? These approaches are quite distinctive, and a discussion of semantics in terms of any one of them would lead in a direction quite different from the direction given the same discussion in terms of either of the other two.

Since the topic of discussion is L-O Retrieval Systems, Part II: Semantics, the same general outline followed in Part I will apply: first a discussion of the set D, then of sets P(D) and P'(D). In none of these discussions have I felt constrained to follow a given school of semantics. Certain techniques of the various schools suggested somewhat similar techniques to me in the retrieval context, but my usage of these techniques is probably sufficiently unorthodox for each school to reject me as one of theirs. The structural differential of Korzybski[2] occurred to me as adaptable to a discussion of hierarchy[4], but as the paper developed, I did not actually use the concept. Syntactically, the hierarchy problem is still an open one. Carnap's[3] state description discussion led me to use the term in describing P'(D), but there are obvious differences in our usage of the term. The concept as Carnap uses it is restricted to language systems of declarative sentences: a list of retrieval terms is not a language and contains no sentences. Indeed, the whole "meaning" idea is restricted by formal logicians to apply to the truth value of sentences displayed within state descriptions of whole language systems. In a retrieval context, "meaning" adheres either to individual words or to word sets (structured or unstructured depending upon whether or not role indicators or relationship designators are used), and here my whole approach has been extensional rather than intensional.

In any L-O Retrieval System the set D is a collection of words which has been used, is used, or will be used to characterize the content of documents (or books). It is a set which has been used in the Uniterm type system where words are added as they appear (and appear significantly, at least to the indexer) in the document or book at hand. The words are arranged in term deck fashion, and a list in the sense of an alphabetical list of words used to a given date may not exist in any explicit fashion. In this case meaning enters only as the indexer understands each word in its context and by a judgment of significance decides either to enter or to ignore that term for that document. By and large this means meaning as normally understood, as defined by dictionaries, general and technical, as understood in day to day conversation. This layman's approach to semantics determines (see Chapter 3) that the uncontrolled Uniterm indexing approach to information retrieval will exhibit a maximum false drop as compared with other more controlled systems.

The set D is an alphabetical listing of authorized terms in retrieval systems based either on subject headings or descriptors. The list may consist of used terms only, or of permissible terms some of which may never be used. A subject heading list is the authority to which the analyst must refer in the categorization of new acquisitions. Additions to the list of used words are screened before admittance and cross-referenced within the existing subject heading list. The cross-referencing in terms of "Also see" defines the referenced words by the association of ideas or the association of related, but different, words. "Includes" references give some identification of the scope of the word referenced. However, the system suffers from the limitations inherent in an alphabetical listing as that listing grows. Subject related words are not grouped, synonomous terms infiltrate the system, and the average analyst learns a "vocabulary" and assigns terms within those limitations.

A thesaurus of descriptors retains the advantages of the traditional subject heading index and eliminates some of its shortcomings. The arrangement of words in schedules of subject related words displays at once all of the authorized terms in a given area, and, by context, defines a word more fully than the use of "Also see" and "Includes" references can do. Synonomous terms are more easily kept out of the system, and because the use of schedules affords a quick review of the available terms in a context, the analyst is not now restricted to a "learned" vocabulary and consequently may more easily handle unusual reports in fields other than his own specialty.

A feature of the schedule system not available in other schemes is the option of displaying terms hierarchially.[5] That is, given a set of discipline-related words one can inquire as to the semantic interrelationships between these words, particularly as regards the subordination of one word to another. This is a further refinement in pinpointing meaning and context. On the syntactic level such subordination was represented in terms of class inclusion, and this relationship was taken as a syntactic definition of hierarchy. On the semantic level this approach leads to the kind of display shown in Figure 1. The relationship so shown is unambiguous. All of the terms are class terms, and the extension of these terms decreases as one moves down the diagram.

There are many terms in any system which simply are not hierarchical. Such

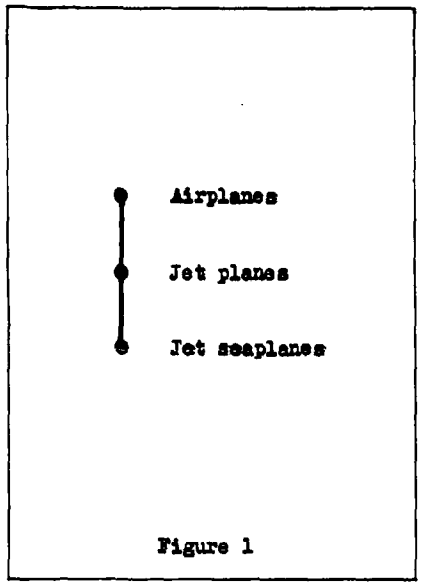



Figure 1

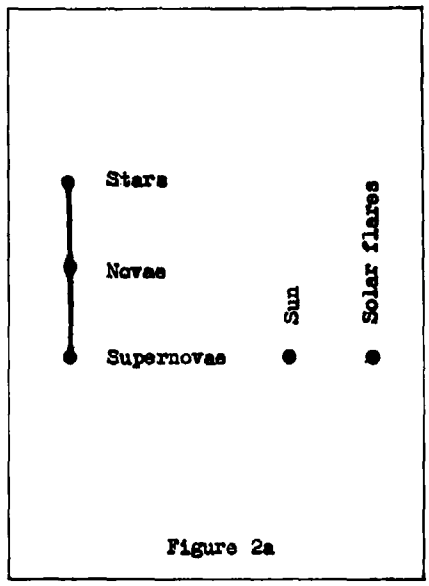



Figure 2a

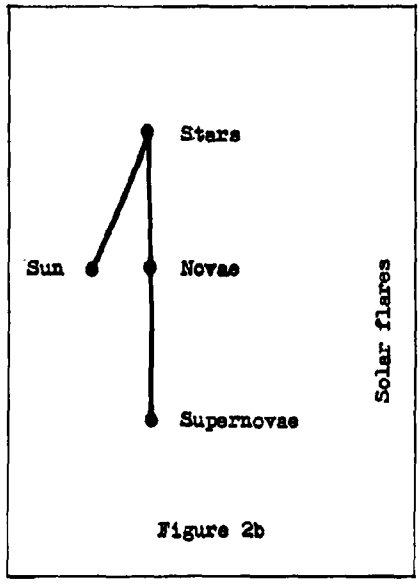



Figure 2b

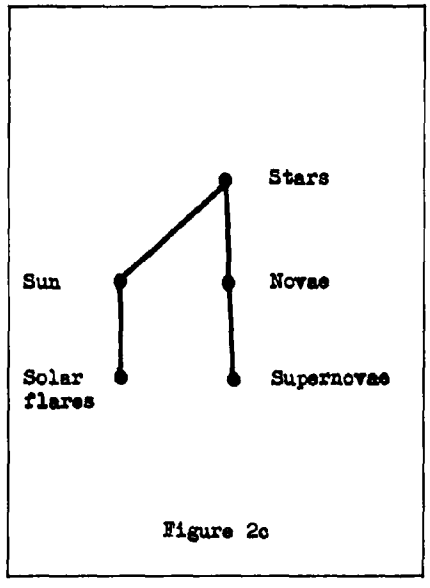



Figure 2c

terms are not class terms. A few illustrations are Alaska, air to air, analysis, and volume. But between the extremes of complete hierarchy and complete independence on a syntactical level, there appears on the semantic level an intermediate position in which terms appear to be hierarchically related, but which do not fulfill the requirements of class inclusion. Figure 2a shows five closely related terms only three of which can be related hierarchically by class inclusion. Now certainly the term sun is extensionally subordinate to stars and one would like the arrangement shown in Figure 2b. The argument can be advanced that this is indeed a proper hierarchical display. But sun is not a class term except by such mental gymnastics as "it is the class of stars which is our primary and therefore is a class containing only one member, itself." Less artificially, however, sun is related to stars not by class inclusion but by class membership. Figure 2b, then, shows without distinction two kinds of relationships: class inclusion and class membership. One further addition seems reasonable in that solar flares is subordinate to sun and the diagram of Figure 2c is suggested as the correct display of the hierarchical arrangement for all five terms. Now sun is not a class term as stated above, but solar flares is. How then can the arrangement of Figure 2c be justified? This difficulty can be met by focussing attention not on the class defined by solar flares, but on the property designated by solar flares. This formulation avoids the embarrassment of subsuming a class term under an individual term, but it necessitates a third kind of relationship, that of "property of." Therefore, what has appeared on a syntactic level as a simple matter of class inclusion appears on the semantic level to involve at least three different kinds of relationship: Class inclusion, class membership, and property of. The concepts of class inclusion and class membership, though, are still closely related as witness the following set-theoretic theorem: if $x \in A$ and $A \subset B$, then $x \in B$. This may be paraphrased as Socrates $\in$ {Men} and {Men} $\subset$ {Mortals}, then Socrates $\in$ {Mortals}. Semantically this multiplicity of relationships is of no consequence. Considering the meaning of the terms by intension, Figure 2c is a justifiable hierarchical display. What is needed is a syntactic (extensional) correlate to the semantic situation, or, hierarchy can not be fully delineated on a syntactic level. If it can be argued that as one moves up (or down) the diagram one consistently becomes more (or less) abstract, then the syntactic formulation would better be stated in terms of "greater than" ($>$) rather than in terms of class inclusion ($\subset$) or class membership ($\in$). Whether or not this can be consistently done is not a problem to be solved here. What is clear is that the demands of hierarchy on a semantic level transcend the simple notion of hierarchy previously advocated on the syntactic level.

The movement as outlined above from an unstructured and unlisted set of words, through an alphabetical list, to a termination in hierarchical displays is a movement toward tighter and tighter control of the indexing terminology. Since all of these systems have realization in actual practice, the movement is motivated by matters of presumed retrieval efficiency and by personal preference. A discussion of these matters is a discussion in pragmatics and will not be pursued at this time. However, insofar as the motivation is a movement toward tighter control of the semantic content of the retrieval terminology, this trend should logically terminate in the use of role indicators or pattern indexing. There are several such schemes in operation or under development and undoubtedly there are more to come. One comment pertinent to the present context is that role indicators, which seemingly are the epitome of word control, can be applied at any point in the spectrum from an unstructured set to a rigid thesaurus. This is so because role indicators are applied to sets of terms attached to separate documents and are therefore independent of the structure of the basic indexing list.

P(D) is the power set of D and consists of all possible subsets obtainable from D. P'(D) consists of just those sets of P(D) which have actually been assigned to documents in a collection. P'(D) is a much smaller set than P(D). P'(D) contains multiplicities of the same set: P(D) is comprised of unique sets. Structurally (syntactically) P(D) is best examined in terms of lattice theory. Semantically, the structure of P(D) is not important: it may be considered an unstructured set of sets. Syntactically P'(D) was structureless. Semantically, P'(D) may be structured in a variety of ways.

We concern ourselves first with a model of P'(D). One such model is a matrix array of index terms and documents which may be called a state description[6] of P'(D). It appears as Figure 3. This matrix qualifies as a state description because it displays for view the status of all authorized indexing terms in relation to any document or document set in the collection. Given such a display, any retrieval question which is pertinent to the collection may be answered by mechanically checking the appropriate terms against each document. A display of this type is not practical for any but very small collections. For example, the current ASTIA collection would require at least a 7000 x 250,000 matrix. It should be noted that the rows of the matrix form the much touted inverted index files. The columns correspond to manual catalog files where each catalog card contains all of the indexing terms relevant to that particular document. If the columns are separated and the indexing terms wrapped around the edge of a card, one has notched cards, just another way of displaying the total status of each individual term relevant to each document in the collection.

Bibliographies correspond to submatrices of the state description. The ideal (lattice ideal) nature of the documents within a bibliography can be shown by a proper permutation of the rows and columns of the bibliographic matrix. A series of such transformations, beginning with an actual bibliographic submatrix is illustrated in Figures 4 through 6.

The terms used in retrieving this bibliography are written in upper case as INDEX TERMS. The DOCUMENT numbers refer to pertinent acquisitions found in the retrieval process, and are also designated in upper case as DOC 64, for instance. Terms assigned to the documents retrieved, but not used in this retrieval operation are shown in lower case as index terms, while lower case document number designation (doc 4) represent nonpertinent documents (false drops) retrieved in this bibliographic operation. The series, Figures 5 through 7, represent an actual machine processed bibliography: considerations of space make it impractical to show these charts in full detail (for example, Figure 5 should be a 64 x 148 matrix). The pertinent statistics relevant to this bibliography are:

The documents were retrieved from a collection of about 40,000 items;
11 TERMS were used for retrieval;
43 pertinent DOCUMENTS and 21 nonpertinent documents resulted from this search;
137 different index terms were assigned to these 64 documents in addition to the 11 TERMS used for retrieval;

Figure 3

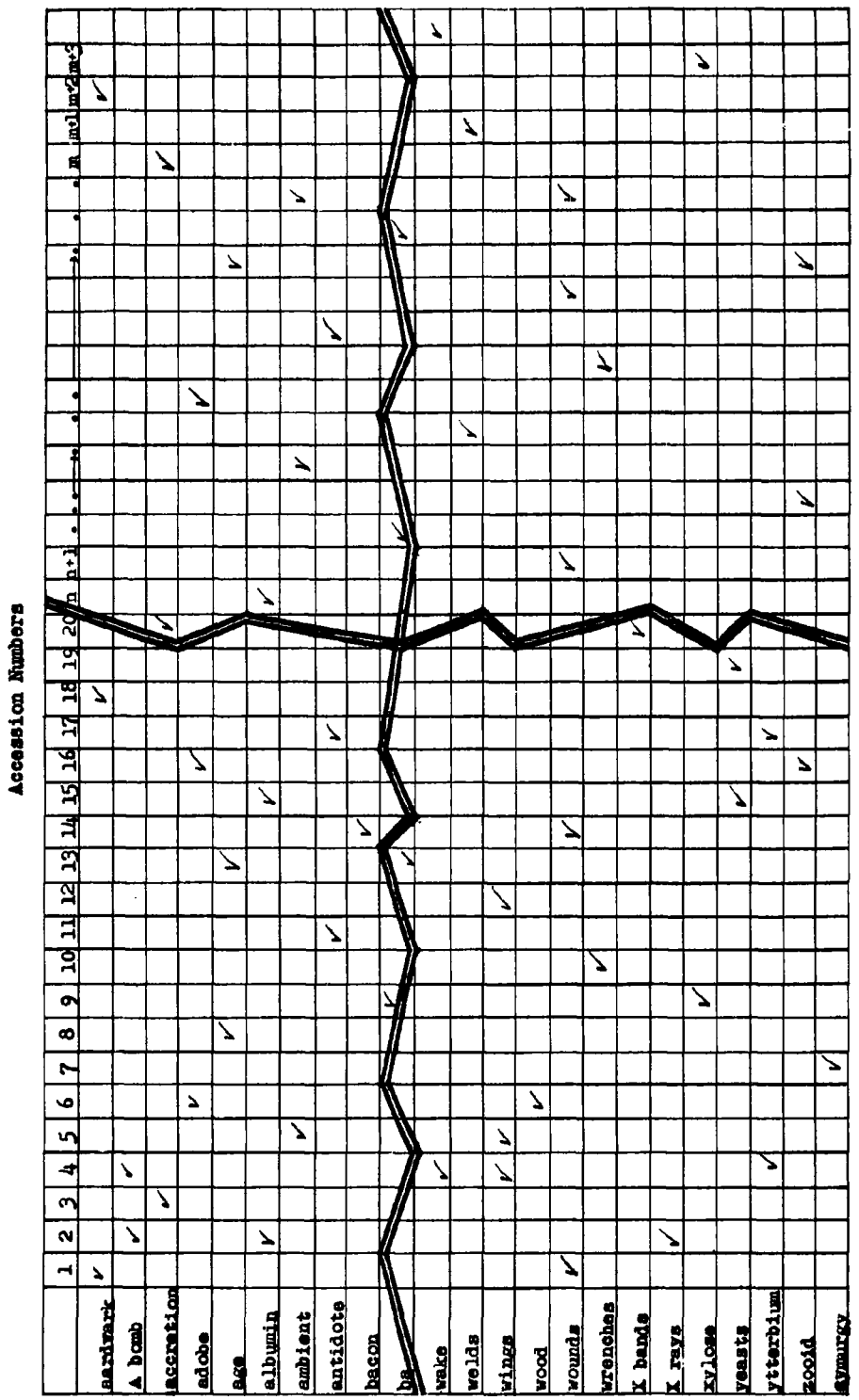

Accession Numbers
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 n n+1 m m+1 m+2 m+3
aardvark
A bomb
accretion
adobe
ace
albumin
ambient
antidote
bacon
be
wake
welds
wings
wood
wounds
wrenches
X bands
X rays
xylose
yeasts
ytterbium
zooid
zymurgy

Figure 4

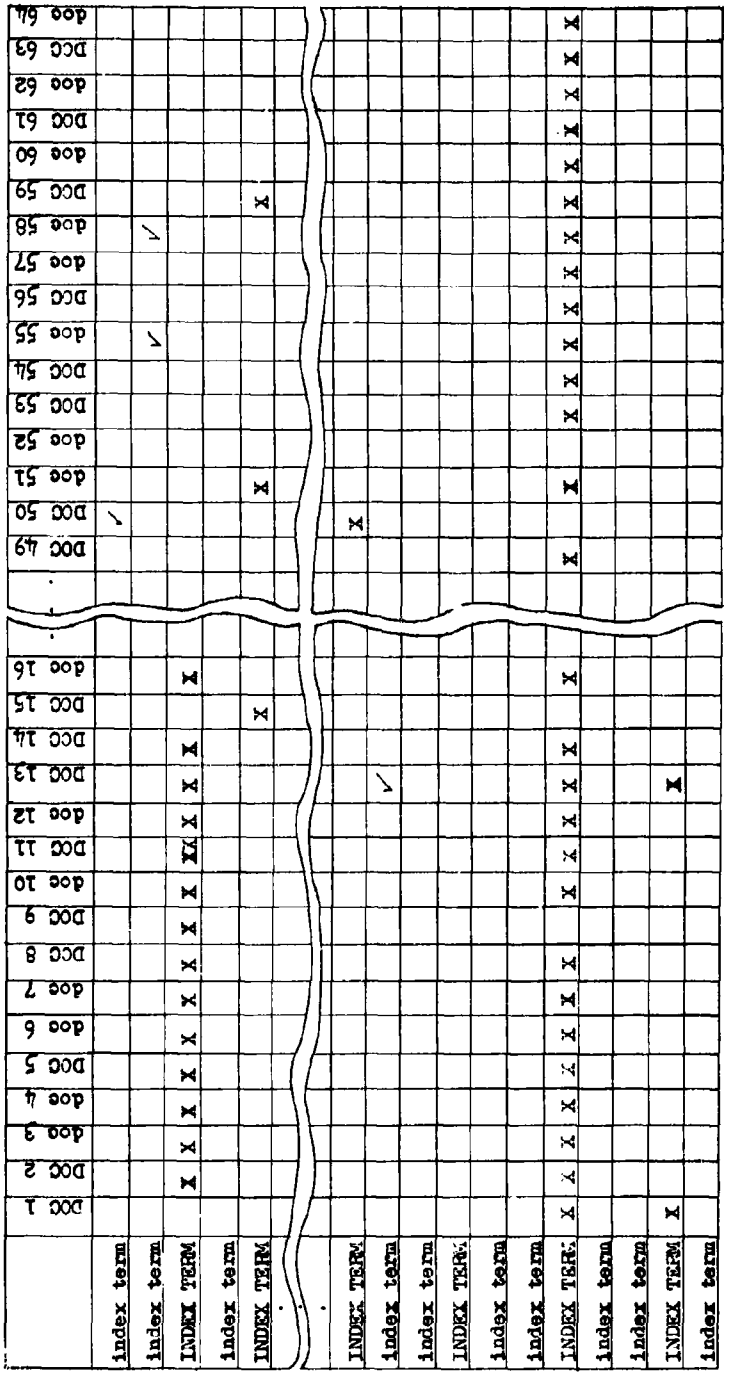

index term
index term
INDEX TERM
index term
INDEX TERM
INDEX TERM
index term
index term
INDEX TERM
index term
INDEX TERM
index term
index term
INDEX TERM
index term
doc 1
DOC 2
doc 3
doc 4
DOC 5
doc 6
doc 7
DOC 8
DOC 9
doc 10
DOC 11
doc 12
DOC 13
DOC 14
DOC 15
doc 16
DOC 49
DOC 50
doc 51
doc 52
DOC 53
DOC 54
doc 55
DOC 56
doc 57
doc 58
DOC 59
doc 60
DOC 61
doc 62
DOC 63
doc 64

Figure 5

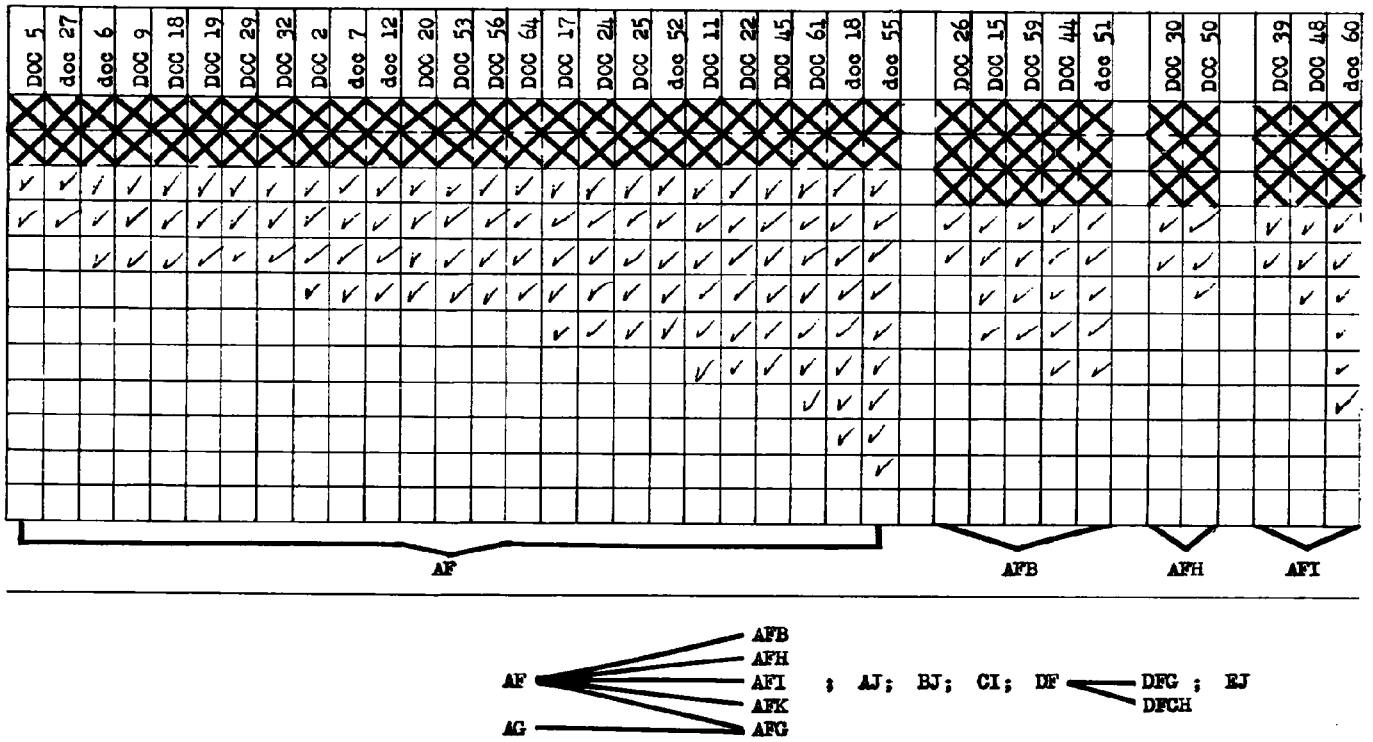

DOC 5
doc 27
doc 6
DOC 9
DOC 18
DOC 19
DOC 29
DOC 32
DOC 2
doc 7
doc 12
DOC 20
DOC 53
DOC 56
DOC 64
DOC 17
DOC 24
DOC 25
doc 52
DOC 11
DOC 22
DOC 45
DOC 61
doc 18
doc 55
DOC 26
DOC 15
DOC 59
DOC 44
doc 51
DOC 30
DOC 50
DOC 39
DOC 48
doc 60
AF
AFB
AFH
AFI
AF
AFB
AFH
AFI
AFK
AG
AFG
; AJ; BJ; CI; DF
DFG ; EJ
DFGH

Figure 6

DOC 38
DOC 49
doe 40
doe 62
DOC 37
doe 46
DOC 63
DOC 43
DOC 8
DOC 13
doe 57
doe 3
doe 10
doe 42
doe 43
doe 4
doe 16
DOC 14
DOC 47
DOC 1
DOC 31
DOC 23
DOC 34
DOC 54
DOC 33
DOC 21
DOC 28
doe 35
doe 36
AFK
AG
AFG
AG
DF
DFG
HDGH
GI
BJ
KJ

Figure 6

The average number of terms assigned to each document in the bibliography was 6.55.

If the rows and columns of the state description for this bibliography (Figure 4) are permuted so that the 11 TERMS constitute the first 11 rows (with the remainder of the terms arranged in alphabetical order following them) and the pertinent and nonpertinent documents are also grouped, we obtain the display shown in Figure 5.

Although not apparent in this abbreviated form, the arrangement of Figure 5 indicated that the request TERMS represented 3 separate schedules and 3 separate fields. The associated terms represented 56 schedules and 14 fields. The average set size for the pertinent DOCUMENTS is 6.32 and 7.00 for the nonpertinent documents.

Figure 6 shows the same material arranged by ideal. The TERMS were coded in alphabetic sequence from A through K. These terms are denoted by the large X in the document column to which they apply and are also indicated by letter below the appropriate ideal grouping. The associated terms are shown by a check (✓) mark. The significance of this display for the retrieval process will be further discussed under the topic of pragmatics.

We turn back momentarily to P(D). Using the ASTIA system as example, P(D) contains about $2^{7000}$ (or about $10^{2100}$) elements whereas P'(D) consists of about $2.5 \times 10^6$ sets -- and these are not all distinct. It would seem, then, that by far the largest part of P(D) is without real significance to the theory of L-O Retrieval Systems. This is not correct. P(D) contains every set formable with descriptors. This means that P(D) contains as distinctive elements each schedule (there are now 292 such), and any grouping of these, including that arrangement now called a Field (there are presently 19 of these). It also contains as distinctive sets each bibliographic request which can be stated in retrieval terms, many of which are satisfied only by a group of documents (the union of individual document prescriptions). In addition, P(D) contains another kind of set which is very important to the theory and practice of L-O Retrieval Systems: These are the sets associated with a given descriptor by usage. The analysis of these sets, called context sets, which exist in their own right in P(D), but only as the union of certain sets in P'(D), will be conducted in the next chapter.

## Chapter 3

Given any indexing term T, its usage in a L-O system is associated with retrieval prescriptions $A_1$, $A_2$, $A_3$, . . . , $A_n$ which in turn are associated with documents $D_1$, $D_2$, . . . , $D_k$. The associated set[7], A(T), of T is given as

A(T) = $\cup A_i$, i = 1, 2, 3, . . . , n in which multiplicities of the same term are permitted and counted.

In addition, we shall be interested in the following concepts:

P(T) = the principal associated set derived from A(T) by considering only those terms in A(T) of high multiplicity, such as the ten words most often found in a given A(T).

$P_o(T)$ = the reduced principal associated set derived from P(T) by suppressing multiplicities.

Note that the size of P(T) varies with T, but $P_o(T)$ is a fixed set size independent of T.

C(T) = the context set of T derived from A(T) by suppressing multiplicities.

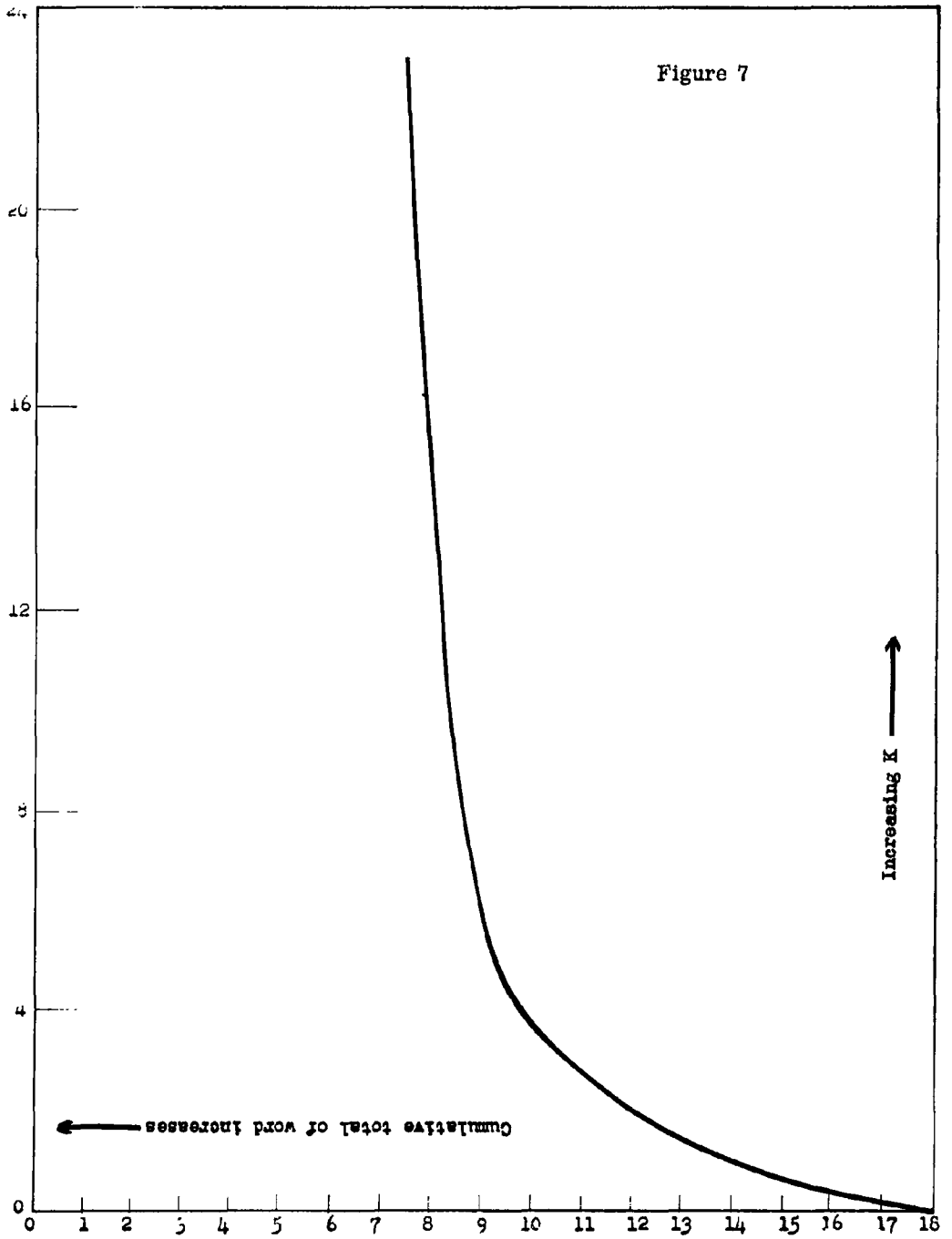
g(C) = the growth function of C(T).

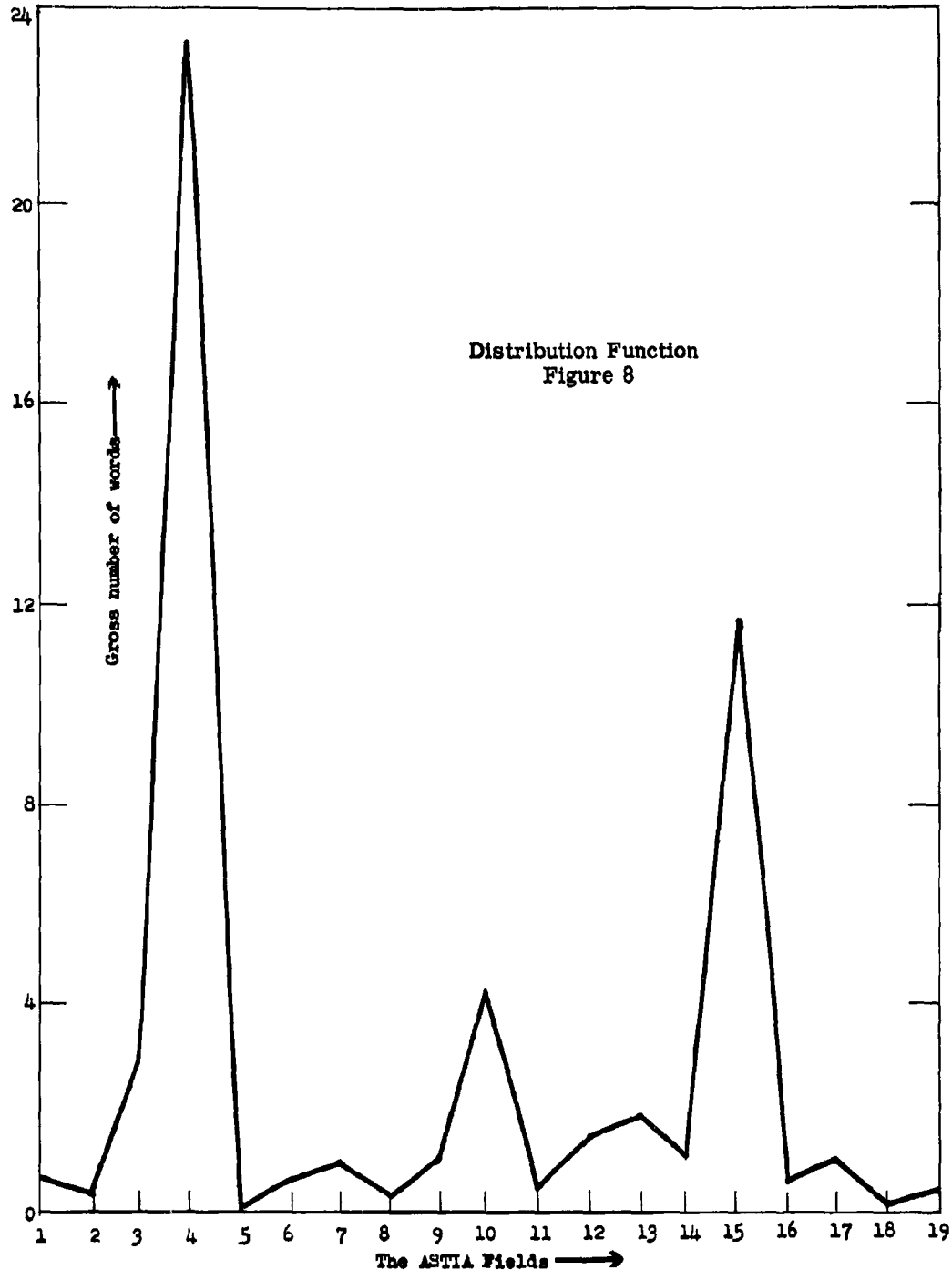
Because the number of available indexing terms is finite, and because T will tend to be associated with terms subject-wise related to it (this fact will be graphically displayed), the growth function will look like the curve depicted in Figure 7. The point at which the curve flattens out is the point at which more words are repeated than there are new words added.

A measure of the semantic intensity or discriminatory power of T is given by the lowest order of multiplicity required to encompass 50% of the words in A(T). Clearly, if words of no lower than multiplicity 5 are required to cover 50% of $A(T_1)$, but words as low as multiplicity 2 are required to cover 50% of $A(T_2)$, then $T_1$ is more discriminatory than $T_2$. The size of the multiplicity, then, can be used to rank retrieval terms, $T_1 > T_2$ (since $5 > 2$), but one cannot and should not conclude that $T_1$ is 2 1/2 times as discriminatory as $T_2$.

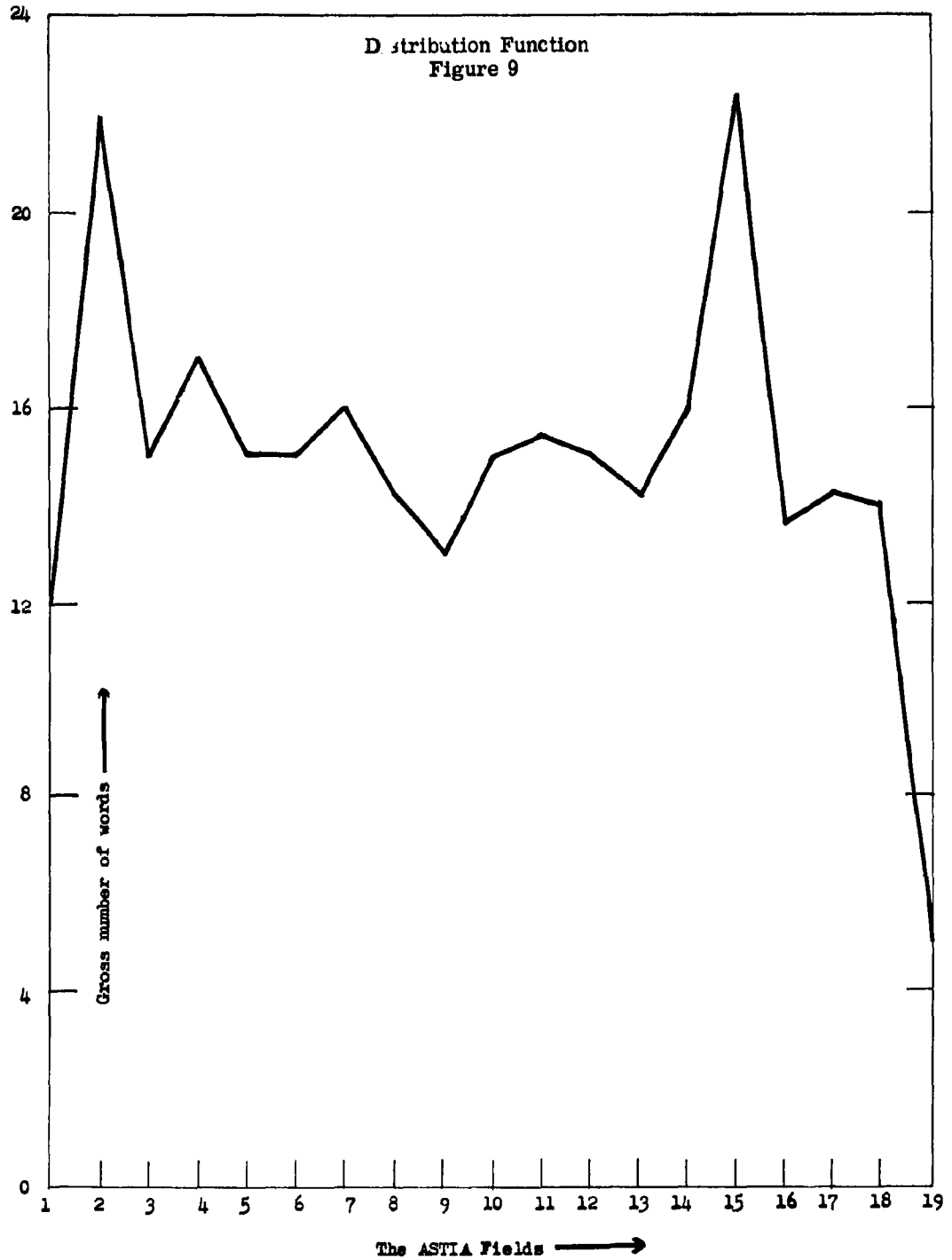
A second way of showing the discriminatory power of T is by means of a distribution function concept. (This method is particularly applicable to controlled vocabularies in which individual words are associated with groupings as in ASTIA's Field and Schedule system.) The distribution function shows distinctive peaks which indicate the concentration of the associated terms in the various subject fields. Note the contrast between the distribution function as given in Figures 9 and 10. (These are idealized curves. Actual examples are given in the Appendix.) The word whose context set[8] is displayed in Figure 9 is associated with a particular scientific discipline and its associated words reveal that fact. The word whose context set is shown in Figure 10 is not so associated and this is evidenced by the lack of sharply pronounced peaks.

Figure 7


Increasing K
Cumulative total of word increases
0
4
8
12
16
20
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18




Distribution Function
Figure 8


D stribution Function
Figure 9
24
20
16
12
8
4
0
Gross number of words
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
The ASTIA Fields

The significance of the peaks in the distribution function as depicted in Figure 9 requires further elucidation. What are some of the reasons for multiple peaking? Let us recall that A(T) represents the sum of all the words used in conjunction with T, and that T represents a technical term associated with some scientific discipline. Therefore, multiple peaks may reveal different usages of T as the words associated with T defined one way may not coincide with the words associated with T when T is used with a different connotation. (The indifferent use of multiple meaning terms is a fertile source of false drops in machine retrieval. Under controlled vocabulary conditions it is also a measure of inconsistent cataloging.) Second, even if T is single-disciplined, there will be words in A(T) which appear not so much because of their relation to T, but because T itself is associated with W (consequently $W \in A(T)$) and these words are strongly associated with W. If W is a word of moderate multiplicity in A(T), and X, Y, and Z are words of high multiplicity in A(W), then X, Y, and Z will be words of low multiplicity in A(T). If, however, X, Y, and Z are field inter-related they will cumulate and peaking will occur in the frequency distribution curve of T.

Which of these factors is predominant in a given instance can be determined only by a careful analysis of the words in A(T) and the documents to which they are attached. What is important in this situation is the realization that the members of A(T) do not all attach to T with the same degree of relevance: what we are actually witnessing when viewing the totality of A(T) is a dense core of words primarily associated with T plus a diffuse envelope of words which T shares with other words, $W_i$. This fact provides the possibility of the practical construction of a distance function which is the topic of the next chapter.

Bar Hillel[9] has stated

> "As a basic prerequisite for (a mathematical theory of literature searching) adequate measures of distance between topics and of relevance of documents to topics in the form of certain functions of the sets of index terms on one hand and of the sets of topic terms on the other hand . . . have to be defined."

The premise is doubtful, and I would be reluctant if I were called upon to substantiate that any proposed mathematical theory of literature searching had totally failed of its purpose unless adequate distance functions were a part of that theory.

In the first place, with the present development of automation in this field, nothing should be asked which is not available in manual form. Librarians do not now have a distance measure: they have very subjective notions about the "closeness of terms" and the "relatedness of documents." Secondly, if a distance function is defined, it is, strictly speaking, not a part of the mathematical theory of information retrieval, but a part of the semantic theory of information retrieval.

The distance concept, as a mathematical problem, was virtually unanalyzed prior to 1900. Since then, distance has been subsumed under the more general notion of measure, and as such distance concepts are applied to situations not formerly considered appropriate. The elementary and classic example of the generalized notion of distance is the question of the "length" of the unit interval after the set of rationals has been removed. Measure theory is now a specialization in its own right, and metrics are imposed on spaces which are a far cry indeed from the intuitive Euclidean space and Euclidean length. Indeed, a metric space is defined as any set of elements such that to each pair of elements x and y there is associated a non-negative real number $\rho(x, y)$, called the distance between x and y, with the properties:

$$\begin{aligned} &\text{(i)} && \rho(x, y) = 0 \Leftrightarrow x = y, \\ &\text{(ii)} && \rho(x, y) = \rho(y, x), \text{ and} \\ &\text{(iii)} && \rho(x, z) \leq \rho(x, y) + \rho(y, z). \end{aligned}$$

A set of retrieval terms does not constitute a language, nor is the argument which follows a misguided attempt at a new or original analysis of the language problem in any of its many guises, but it is helpful to enlarge the scope of the discussion in order to say something about meaning and language in relation to the distance concept. First, meaning as defined by logicians is analyzed in terms of sentences (and still almost invariably in terms of declarative sentences) and within a given language state descriptions are set up which in principle determine the meaning of a declarative sentence formable in the language. Meaning as it needs to be discussed here is in terms of individual words (where air to air or doppler effect are considered instances of individual words). Second, it should be kept in mind that there is no such thing as LANGUAGE (except, perhaps, insofar as there is such a thing as Plato's GOOD), but only languages -- and these exist only in terms of activity between people, whether vocal and immediate or written and delayed. Linguistics is a frozen slice of this activity and in that static stage rules of

grammar, and the rest of the paraphernalia which the study of language is bejargoned with, can be codified. This subject is deep, involved, and complex, and chairs are set up in universities to provide a means for pontification on the subject. We sould not change this, although one wonders at times, in the spirit of the provincial tourist, how it is that mere infants in France are able to speak French so well.

Let us move on to basics. How does a child learn to speak English, or French, or Mandarin Chinese? The over-all answer to this is clear, only the details are complex. Language for a child is just one of the many components of James' "Buzzing, blooming, confusion" which the child sorts out and codifies. As with any other learned activity, the talent is acquired in a context which inhibits in some directions and permits in others. Just as there is no LANGUAGE, not even an ENGLISH LANGUAGE, so too there is no CONTEXT. The interplay between changing contexts and changing verbal activity is a measure of the difference between Chaucer's English and Brooklynese.

New words and new meanings are added to one's repertory at a later stage through the printed word which includes dictionaries. An unfamiliar word may appear on the printed page with sufficient context that its probable meaning is clear without recourse to a dictionary. If that context is not sufficient, the dictionary supplies the missing context, sometimes in terms of a sentence (context) showing an instance of the proper usage of the term.

Synonymous terms are synonymous because they are appropriate to the same context. The weather is (fine, good, agreeable, excellent, pleasant, salubrious). But in a different context, He paid his fine, the other parenthetical terms are no longer synonymous, i.e., they do not fit into this context. Indeed, the word itself has changed meaning, and multimeaning terms are troublesome terms in information retrieval. Synonymity, then, is not so much the fact of interchangeability, but interchangeability in a given context. All of this is oversimplified and elementary and is brought out here only because these factors are relevant to the construction of a distance function between terms.

Distance is definable in the last analysis only in terms of a given system. A generalized distance function, $\rho(x, y)$, relating two "points" x and y, is system independent only because the symbols are contentless. Given a realization of the symbols, x = chair, y = boy, then $\rho(x, y) = d$ is system dependent: the English language as represented by the words in an unabridged dictionary; the same language as represented by the words in an abridged dictionary; Basic English; or the words as part of a given, finite, indexing set. The numerical value of "d" as a measure of the "distance" between "boy" and "chair" will be different in each instance even with the same distance "formula." Moreover, even within the confines of one of these word sets, the distance will be a slowly (under suitable conditions) changing value. However distance is defined for information retrieval usage, and alternate definitions are certainly possible, distance between terms will not be a fixed quantity.

Bar Hillel's criticism of previous attempts at defining measure (see Notes & Bibliography for references) may or may not be justified, but the example he uses to demolish such previous attempts is, in his own words, ". . . unsupported and floating in thin air . . . " because he, as much as anyone he criticizes, thinks of the distance concept in fixed terms.

> "Assuming now that a certain classification of all animals is fixed, e.g., one of the standard biological classifications . . . "

As I see it, this initial premise misses the point or mis-states the problem in at least two ways. First, the whole concept of information retrieval must be taken out of the Comte context of a universal scheme for the categorization of all knowledge. Retrieval makes sense only for a given library, and distance between terms is meaningful only for a given library: if $\rho(x, y) = d$, that "d" is valid only for that library and then only for a given date.

We cannot assume that ". . . a certain classification of all animals is fixed..." but we state that within a given system of retrieval terms certain names of animals are used and these may be arranged in some hierarchical fashion or according to some classification scheme. Hierarchic or not, structured or not, the distance between cats and dogs, under this premise and provided both are within the system, is fixed only until such time as the system is enlarged (either by the addition of more documents, more terms, or additional use of given terms), and if domestic animals is a bonafide term of the enlarged system it will bear some unambiguous relationship to the terms "cats" and "dogs", respectively: if domestic animals is not a term of the system, no item has been cataloged with that term, consequently no retrieval problem exists. To insist that some document in the system is about domestic animals and unless this document is retrieved there is a deficiency in the retrieval process is to insist either upon omniscience or to confuse the domains of applicability of distance functions, depth of indexing, cataloging procedures and the like. This leads to the second manner in which the illustration misses the mark. A classification should not be imposed upon a system from without, but should faithfully reflect the contents of the system which it is intended to serve -- and that collection only. In short, I find the whole of Bar Hillel's argument at least as confused as the schemes it is intended to refute.

Two further statements deserve passing comment. Foskett's[10] broadside, which Bar Hillel appears to quote with approval, against some proponents of non-traditional literature searching schemes

> ". . . attempts to disguise (their) commonplace notions in weird and sometimes self-invented pseudoscientific jargon, supported, albeit unnecessarily by masses of impressive mathematical diagrams and calculations . . ."

also misses the mark. It is an opening gambit in an effort to "sell" the faceted classification scheme, and it says nothing about the problem of literature control. The second statement is by Bar Hillel[9]:

> ". . . whether the use of lattice-theoretical or topological terminology might have a clarifying effect on the ways of thinking of librarians and documentalists who are not used to it - is a moot question . . . ."

This reminds me of statements made to me both by medical students as well as possessors of M.D.'s when confronted with the disparity between the old-fashioned, family doctor approach (in which the doctor was available at any time and for whom it was not unusual to sit with a patient during a crisis) and the modern aseptic approach (in which the doctor is a businessman whom one sees for five

minutes and that by appointment only): what good did sitting up all night with a patient do? It only made the patient (and/or his family) feel better!

To return to Foskett[10]: as to the "weird and sometimes self-invented pseudoscientific jargon", that depends pretty much upon where one sits and at best is a personal value-judgment. Is it not a little early to raise up sacred cows in the documentation field? The unappreciated datum is that some people, recognizing real problems, try to analyze them. It is an open question still whether any traditional mathematical structure will serve as an adequate retrieval model or whether new and "weird" systems are required. We should remember that negative results in documentation research circulated among interested parties may be just as valuable as blank sorts are to a state of the art survey. More to the point is that classification schemes to encompass all sciences, or more modestly, a classification scheme for a given library imposed from without and taken from such a universal scheme is an excellent example of medieval thinking at its best, and should have been decently buried with Auguste Comte (1798-1857) with annual resurrections only for the edification of and warning to new crops of graduate students. The faceted scheme of classification is distinguished only by its complexity. Classification, if it is to be sensibly done at all, should be done for a local collection on the basis of that collection -- almost by a factor analysis technique. This will be clarified after a distance function has been defined.

I propose to measure the "distance" between any two words of a given indexing list by considering the degree of overlap of their respective context sets, where those context sets are determined solely by usage.

Intuitively, two words, $T_1$ and $T_2$, become "closer" to each other as their contexts become more and more alike. The first intuitive concept will not do, however, because if one simply counts the words common to two context sets

$$\rho(T_1, T_2) = C(T_1) \cap C(T_2)$$

then the closer the terms, the larger the distance between them*. The reciprocal of this quantity will not do either because one would like to have the first two postulates for a metric space satisfied, and the reciprocal does not permit $\rho = 0$.

The next approach is to take

$$\rho(T_1, T_2) = C(T_1) \cup C(T_2) - C(T_1) \cap C(T_2).$$

Now as the words in common increase, the numeric measure decreases provided that

$$C(T_2) \cup C(T_2)$$

is fixed. However, the union grows as fast as or faster than the intersection until

*The notation as presented is ambiguous. $\rho(x, y)$ is a number: $A \cap B$, where A and B are sets, is a set, and clearly a number is not equal to a set. We should write $\rho(x, y) = N[C(T_1) \cap C(T_2)]$ where N indicates that we count the number of the terms in the intersection. Then $\rho$ is defined to be numerically equal to that number. The "N" is suppressed throughout the discussion for simplicity.

the growth function g(T) has stabilized. But even under that circumstance the measure is not adequate because if

$$C(T_1) = 3 \text{ and } C(T_2) = 3$$

and $$C(T_1) \cap C(T_2) = 2$$

then $$\rho(T_1, T_2) = C(T_1) \cup C(T_2) - C(T_1) \cap C(T_2) = 4 - 2 = 2$$

which implies that if we had in addition

$$C(T_3) = 10 \text{ and } C(T_4) = 10$$

we would need

$$C(T_3) \cup C(T_4) = 11 \text{ and } C(T_3) \cap C(T_4) = 9$$

before

$$\rho(T_1, T_2) = \rho(T_3, T_4).$$

This, in effect, penalizes the large set sizes, and we seek a measure which is more set-size independent. The formula I propose is

$$\rho(T_1, T_2) = \frac{C(T_1) \cup C(T_2) - C(T_1) \cap C(T_2)}{C(T_1) \cap C(T_2)}$$

or, equivalently,

$$\rho(T_1, T_2) = \frac{C(T_1) \cup C(T_2)}{C(T_1) \cap C(T_2)} - 1.$$

In terms of our previous example we now have

$$\rho(T_1, T_2) = \frac{4 - 2}{2} = \frac{2}{2} = 1$$

$$\rho(T_3, T_4) = \frac{11 - 9}{9} = \frac{2}{9}$$

a result which is much closer in accord with our intuitive feeling for distance.

This definition will also take care of a growing union under any conditions which are valid for measure considerations. If g(T) has not levelled off, then the distance between $T_i$ and any other term $T_j$ is valueless because of its rapid change. But if g(T) for both terms has progressed to the flat portion of the curve, then the distance function is valid and provides a numeric indication of the closeness of the terms measured relative to the collection at hand.

The relation of a distance measure to a tree structure or hierarchical display of terms is dependent upon two things: a stable growth function and the philosophy of indexing. If the indexing is consistently done in a hierarchical fashion, then two adjacent terms should be closer together than they would be if the same two terms were separated by a third in the tree structure. If the indexing is not consistently

hierarchical then this relationship need not, and probably will not, hold. However, on the basis of the distribution function implication of how related terms cluster one would expect that terms in one tree would be closer to each other than to terms in any other tree.

My previous remark about a factor-analysis like approach to classification can now be made more specific by pointing out that instead of taking a tree as given and asking whether the distance between terms as given by some suitable measure satisfying our intuitive feeling really holds, we could prepare a matrix of distances of each term from every other term in the index set, and group those terms which by actual usage were closest to each other.

As to the problem of changing a given request to a closely related request, normally this is a matter of "see" and "also see" references, and, if the cataloging is being consistently done (a topic more appropriate to a discussion of pragmatics) then these will also be terms "close" to each other. The distance function alone, as defined here, is not a sharp enough tool to allow mechanical changes in request terminology.

The context of a given term has at least two dimensions. We have discussed one of these, the number of distinct terms in the context set. The other dimension is the frequency with which each of these terms occurs in the context. This brings us back to the concept of the principal terms of a context, and the set involved is P(T). The distance measure as just defined did not include the multiplicity factor. The definition can be restated in terms of P(T), and the choice between alternatives is not a question of a different numerical value obtained, but simply one of consistent usage in a unified system.

The multiplicity parameter of the context illuminates the problem of the seeming relatedness of all terms which has apparently defeated some previous attempts to define distance. All of the terms listed in Webster's Unabridged Dictionary are related to each other in a trivial sense that they occur in the same context -- that dictionary. They are not all equally related in a usage context. This kind of discrimination can be made either in terms of A(T), where multiplicities are taken into account, or in terms of C(T) where multiplicities are suppressed. Working with only the derived sets, P(T), one can form chains and clusters of sets by means of which one may move from one term, T, to another on the basis of relatedness. A question of some theoretical interest is the minimal set of terms whose P(T) sets cover the index. This would be far from unique. The distribution of such sets of terms subject-wise might, however, be of interest. Study of such clusters of P(T) sets may provide a clue to the feasibility of programming for the enlargement of a request.

At the present stage of the development of the "distance" idea for information retrieval (i.e., circa 1960-1961), no discussion would be complete without reference to the recent work of Maron and Kuhns[11] and cf Parker-Rhodes and Needham.[12]

Maron and Kuhns advance the concept of probabilistic indexing in which the indexer-analyst weights the indexing terms chosen for a document on a relevance scale between 0 and 1. The bibliographer-analyst similarly weights an indexed form of a request statement. Proper manipulation of these weights leads to

(i) Dissimilarity measures between documents

(ii) Significance measures for index terms
(iii) Closeness measures between index terms
(iv) Relevance numbers for documents

which together with other statistical data enable a search routine to be devised which generates a list of ordered documents which satisfy the request and which are ranked according to the probable relevance of the documents.

This analysis can not be brushed off as just one more attempt to derive numerical measures for closeness, relevance, and self-enlargement of request procedures. Criticism, if criticism it is, is not to be directed toward the specific measures proposed -- others can be generated from the same base as these authors point out -- nor to the adequacy of these measures to do to a very large degree what one would expect "sensible" closeness measures to do. Criticism begins when Maron and Kuhns state that conventional indexers work on a go, no-go basis. This is not correct as stated. Conventional indexers do not work on a go, no-go basis. They are quite aware that not all terms assigned a document apply on an equal basis. Consequently conventional indexers also of necessity index probabilistically. What they fail to do is to indicate on a conventionalized scale their judgment of the degree of applicability of each term. Conversely, it could be maintained that there is a sense in which probabilistic indexers operate on a yes-no basis. They either assign a given term or they do not. The probabilistic indexer differs from the garden variety by adding the step of publishing his value judgments. With this clarification firmly in mind, the mathematical results of Maron and Kuhns follow.

Additional comments on the Maron and Kuhns paper now consist of two kinds: (1) discussions keeping to the spirit of their inquiry, but changing their specific measures for others motivated by questions of practicality, and (2) discussions of the reliability to be expected from probabilistic index techniques considering the human element. Both avenues of discussion quickly become matters of pragmatics, and consequently no lengthy discussion will be presented here. However, one crucial question may be indicated:. with present generation computers, how practical is the probabilistic scheme in actually evaluating measures? Given a set of 7000 retrieval terms covering a library of 250,000 documents with some retrieval terms used on the order of 25,000 to 50,000 times, what is the order of magnitude of the job of filling in a chart which requires for any two terms the number of times they have been used together, the number of times the first has been used without the second, the second without the first, and finally, the number of instances in the collection when neither of the terms was used? At what combinations of vocabulary and collection size does this program become impractical? Do we have here another instance of the brilliant solution to the submarine menace?

Parker-Rhodes and Needham approach the distance concept motivated by a machine classification scheme where the classification is designed to optimize the selection or retrieval process. Their approach formalizes to a large extent the few brief remarks I have made about a factor analysis approach to classification. It is also compatible with my remarks on context sets. Again, the practicality of their scheme for a really large library is a question, but efforts in the spirit of their analysis are much needed.

The tenor of my thinking in utilizing a distance function for classification was to match such a derived classification against the original, partly as a check of the

consistency of cataloging practice and partly for indications of possible improvements over the original in hopes of gaining additional consistency and efficiency of retrieval.

## Conclusion

It should come as no surprise that more questions have been raised than answered. If it were not so, either nothing of interest was said or the millenium would have arrived. I am sure that the second alternative is not so.

There may also be some question as to whether or not the preceding discussion was really about semantics. Devotees of "schools" will insist that the whole point has been missed. Others will have looked for but not found rules of -- or at least arguments for -- procedure. How should this or that word be assigned? Such discussion has been avoided deliberately. The context of such an essay is as much pragmatics as it is semantics, and I believe it is very much dependent upon the library and librarians involved. The time for generalizations in that area is not yet.

I am firmly convinced that a detailed examination is demanded of the context set concept. The last paragraph of Chapter 3 is the basis for the discussion of consistency and why ASTIA chose to put a given retrieval term in only one place in the Thesaurus as explained in Automation of ASTIA - 1960.[13] That discussion and the discussion in Chapters 3 and 4 here, if it appeals at all, is more apt to tantalize than satisfy. Therefore, a few comments are in order here.

First, I believe that a mathematical model for the situation described is statable as an urn problem. Given an urn with N white marbles, and an operator who is to take "r" marbles at a time, inspect them, and if "s" ($s \leq r$) of these are white, color them black and return all "r" to the urn. The value of "r" is to depend upon a given probability curve. Determine as a function of the number of selections made, the probability that there are white marbles in the $i^{th}$ selection. The cumulative number of white marbles drawn should approximate the curve given in Figure 8.

Second, detailed curves tied to specific retrieval terms for the concepts illustrated by means of Figures 8 to 10, inclusive, are grounded in specific collections and the analysis of such curves requires a knowledge of the collection discussed, the philosophy of indexing involved, and the nature of the indexing terms available for use. Such an analysis would require papers with titles such as The . . . Retrieval System: A Case Study. Clearly such a discussion is beyond the scope of this study. Just as clearly, such studies would pay high dividends in terms of insights to the retrieval process.

In neither this paper nor its predecessor have I referred to Mortimer Taube by name, though Uniterms have been discussed in several instances. The absence of his name from the bibliographic listing is indicative of neither dislike nor neglect. However unfair it may be to Taube's position now, the original concept of the Uniterm and of coordinate indexing represent a convenient straw man against which to compare other schemes in that the original concept represented the epitome of unstructured, uncontrolled indexing. A lesser degree of control is not possible, for that next degree is complete chaos. Nonetheless, when the definitive history of information retrieval is finally written, no matter how information retrieval may look by then, Taube and Uniterms will be recognized for their immense seminal effect.

A final word about the distance concept and Bar-Hillel. In my more lucid

moments I must agree with much of what he says. But he, too, has the happy facility of functioning as a convenient scapegoat and although I many times agree, at least in principle, with his position, his emotionally loaded style tends automatically to trigger an emotional response.

Calmly now, I do not think that a distance function is as important to information retrieval as he thinks, and if the idea is more closely analyzed, I think it will be found to represent a poorly disguised wish for an impossibly simple answer to most of the sticky problems of electronic information retrieval. Its practical utility (again circa 1960-1961) is on a par with the concept of having a machine program itself for the most efficient search pattern possible for the request at hand.

Philosophically I seem to be a pragmatist and a relativist, hence I am in principle opposed to classification schemes imposed on collections from without. But a pragmatist is constrained to face facts. Schemes exist - and work - which have classifications imposed from without and there are some which have virtually no classification scheme at all in any conventional sense. In the face of such facts I can do no better than misquote Ben Johnson[14]:

> "Sir, a retrieval system without benefit of classification is like a dog walking on his hind legs. It is not done well; but you are suprised that it is done at all."

## Appendix

Philosophically it is comforting to think of the universe as representing more than local order. If science is to be an adequate mapping of that universe it must also be a unity. Yet, a seamless garment may show the individuality of the artisan in the warp and woof of the threads, in the choice of color and design. Similarly, science is of, for and by men, and shows its origin in terms of fields, areas, and disciplines which were convenient to its formulators. Historically the major scientific disciplines existed as autonomous entities and each acquired its own distinctive terminology. Scientific advances, however, are ever antagonistic to Comptean classifications, and rigid boundaries crumble and fall. Old words are used with new meanings, and despite the fact that the English language is famous for its use of multiple meaning terms (We eat what we can, and what we can't eat we can) this facility is a stumbling block not only to the foreign student of English, but to the documentalist as well.

The mere fact that an engineer or scientist states with some vehemence that a given technical term has an unambiguous meaning for him, is by no means adequate or sufficient evidence that his meaning is what the scientific community understands by the term. Indeed, it is an error to think of the scientific community as an unstructured unity. The scientific community is structured just as science is, and within this structure the engineer and scientist is just as prone to parochialism as any other specialist. Hence Plasma means one thing to a nuclear physicist, quite another to a medical doctor or biologist. If an engineer from Boeing speaks of Stability, he is almost certainly talking about Aeronautical stability. A psychologist friend of his uses the same word, and his conferees do not need to be told that Mental stability is what is being talked about.

Now, then, if we poll the delegation, what does Stability mean? I submit that the delegation has already been polled and the results have been published. Moreover, the publication is periodically revised. Any reader interested in the results of the poll need only consult any standard non-abridged dictionary.

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. If his library is a homogeneous, single-discipline library, meaning is of little concern. Words mean whatever they mean within the context of the discipline his library services. He will be concerned with the adequacy of indexing, subject-area breakdowns or classification schemes, and the currency of his vocabulary.

Multi-discipline libraries do present a semantic problem for now a sufficient number of words occur which mean quite different things to different people. The librarian or documentalist can take one of two different approaches. He can ignore semantic differences and list all varieties of meaning together. This is the Fibber Magee approach to information retrieval. Open the closet door and you are inundated with information out of which you must paw and claw that which is pertinent at the moment. The other approach recognizes semantic differences and makes use of them to pinpoint searches. This may even involve arbitrary decisions (to the outsider) on the part of the librarian as to what certain terms are to mean. We should not be horrified when this happens. Documentation is just as much a

technical field as engineering, and to insist that library practices be such that a satisfactory bibliographic search is immediately available without instruction is no more reasonable than to complain that doctors had no right to make medicine so difficult - thus preventing the non-doctor from practicing medicine. If the problems of mass documentation were that simple, all of the problems would have been solved long since.

But documentation is a recognized specialty and the specialist should be concerned with retrieval efficiency. One aspect of such efficiency is to make semantic differences an aid rather than a liability. This can be done in a structured system along the lines laid down in Chapter 3.

The term Electrolytes was used 370 times over a period of eight years in the ASTIA system. Obviously this word was not assigned to documents in isolation: 749 other descriptors were used in the 370 instances of documentation requiring Electrolytes. Figure 1 represents the distribution function for this technical term. That is, it shows the distribution of the 749 words in the context set for Electrolytes. This figure shows that these 749 terms represent 18 of the 19 Fields given in the Thesaurus of ASTIA Descriptors. In the order of their representation (based on the ratio of words used to the total number of words available in a given Field) they are:

| Field Number | Discipline | Percentage |
|---|---|---|
| 4 | Chemistry | 27.5 |
| 15 | Physics & Mathematics | 17.6 |
| 8 | Human Engineering & Psychology | 15.7 |
| 12 | Medicine | 15.7 |
| 10 | Materials & Metals | 14.1 |

The remainder of the fields were represented by less than 10% of the total available terms. Field 19, Space Technology, was not represented.

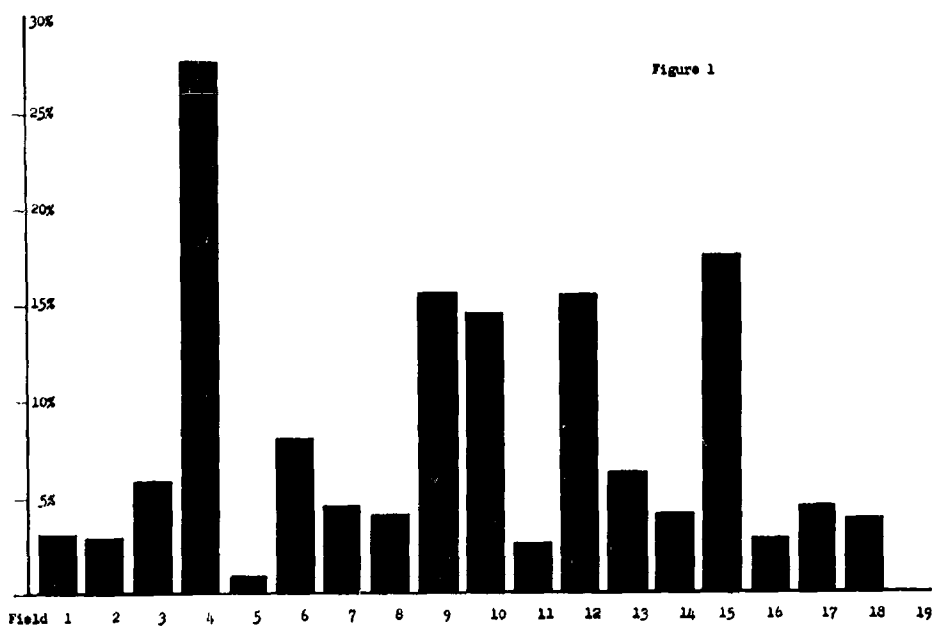


Figure 1

Actually, ASTIA has distinguished between two uses of the term: Electrolytes (Physiology), and Electrolytes. In the first instance the terms is in Field 12, Medicine, the other usage is in Field 6, Electronic and Electrical engineering.

Electrolytes (Physiology) was used 64 times and its context set consists of 246 terms. Its distribution is given by Figure 2. Note that the highest percentage of the terms in the context set falls within the same field as the primary term, Electrolytes (Physiology). In order of magnitude we have:

| Field Number | Discipline | Percentage |
|---|---|---|
| 12 | Medicine | 14.57 |
| 3 | Bio-Sciences | 5.71 |
| 8 | Human Engineering & Physiology | 4.24 |
| 15 | Physics & Mathematics | 3.80 |
| 4 | Chemistry | 3.72 |

Note that Field 12, the Field of the basic term, is represented almost three times as often as its nearest neighbor - and its nearest neighbor, Bio-Sciences, is clearly closely related to Medicine. This shows the power of definition and the degree to which a well-defined term will carry related terms with it.

Electrolytes, Field 6, is shown in Figure 3. It was used 306 times and its context set consists of 565 terms. Again in order of magnitude we have:

| Field Number | Discipline | Percentage |
|---|---|---|
| 4 | Chemistry | 23.80 |
| 10 | Materials & Metals | 14.06 |
| 9 | Industrial Methods | 13.86 |
| 15 | Physics & Mathematics | 13.85 |
| 6 | Electronic & Electrical Engineering | 7.65 |

Both the graphs and the tabulated data show that definition is not an idle gesture, but both graphs and tables give a global, nonspecific view. Let us look at the context words themselves. The principal associated set for our primary terms are as follows:

Electrolytes (Physiology)

(19) Metabolism; (18) Physiology; (15) Body fluids; (15) Excretion; (15) Kidneys; (14) Water; (13) Potassium; (12) Sodium; (11) Pathology; (10) Therapy.

Electrolytes

(67) Electrochemistry; (57) Design; (51) Electrodes; (44) Conductivity; (37) Solutions; (32) Storage batteries; (31) Temperature; (30) Tests; (29) Electrolytic cells; (26) Polymers.

These sets do not have a single word in common! (The numbers in parentheses indicate the multiplicity of the terms within the context set.) Perhaps the top ten words to not constitute a fair sample: the eleventh word and succeeding words might be held in common. In each instance we add the next ten words according to rank:

Figure 2

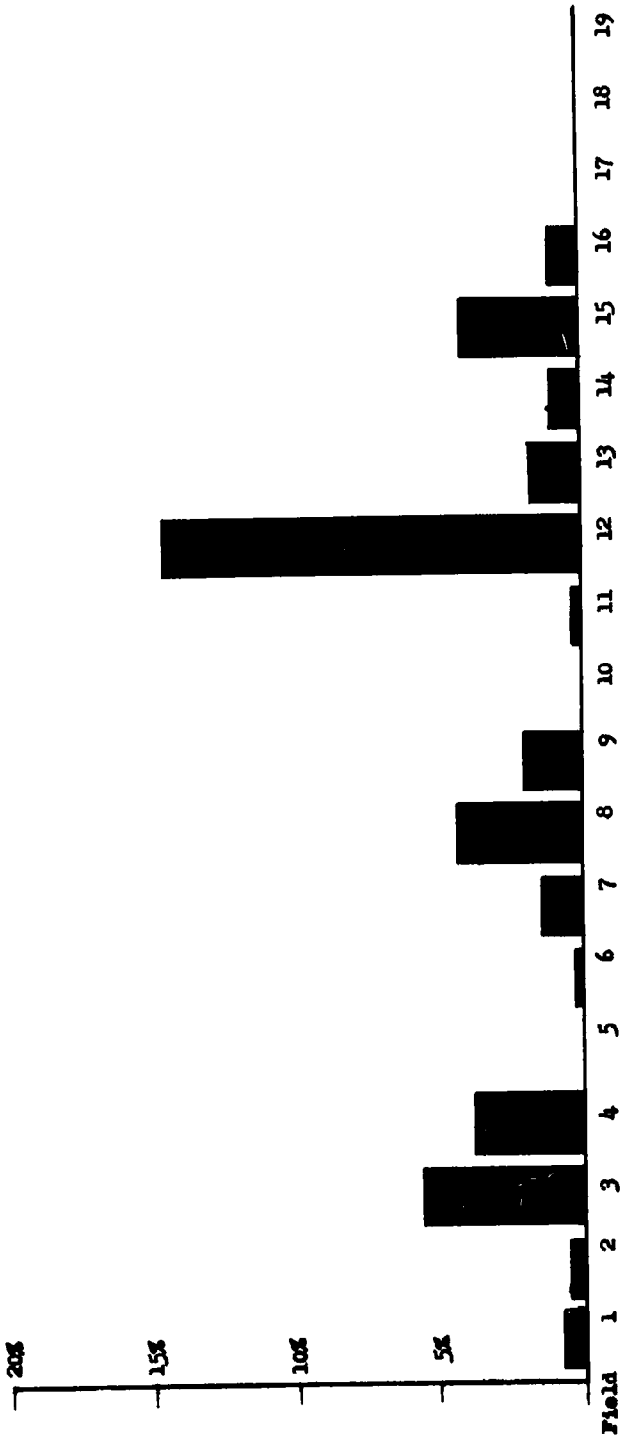
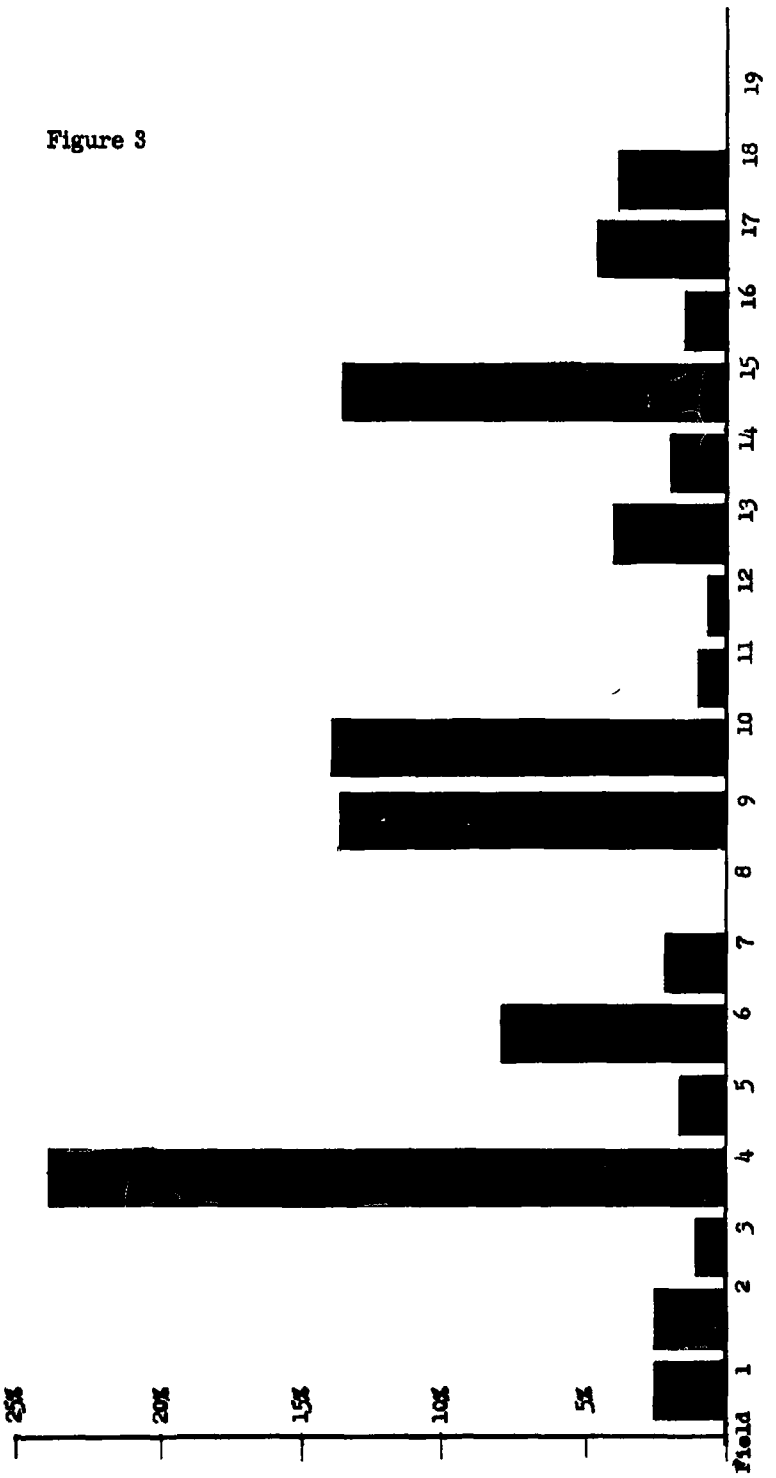


Figure 3



Electrolytes (Physiology)
(9) Biochemistry; (9) Urine; (8) Blood; (8) Chlorides; (7) Blood plasma; (7) Liver; (6) Cholera; (6) Corticosteroids; (6) Heart; (6) Muscles.

Electrolytes
(25) Cathodes (Electrolytic cell); (25) Ions; (25) Oxides; (24) Alkaline cells; (24) Dry cells; (24) Thermodynamics; (23) Polarization; (12) Materials; (22) Power supplies; (21) Theory.

This means that the 20 most frequently associated words with the given descriptors are disjoint sets! Obviously there are words common to the two context sets. The first such common word is Chlorides. It occurs as the 14th most frequent word associated with Electrolytes (Physiology) and as the 24th most frequent word associated with Electrolytes. The next word common to the two context sets is Measurement. It occurs as the 25th ranking word associated with Electrolytes (Physiology) and as the 30th ranking word associated with Electrolytes.

Comparison of the two context sets reveals 61 common terms. The complete context sets follow as well as the set of common terms. The terms in each set are arranged by order of frequency, the number in parentheses referring to the frequency of occurrence of the words which follow it.

CONTEXT SET: Electrolytes (Physiology)

(19) Metabolism; (18) Physiology; (15) Body fluids; Excretion; Kidneys; (14) Water; (13) Potassium; (12) Sodium; (11) Pathology; (10) Therapy; (9) Biochemistry; Urine; (8) Blood; Chlorides; (7) Blood plasma; Liver; (6) Cholera; Corticosteroids; Heart; Muscles; Nitrogen; Tissues (Biology); (5) Biophysics; Cells (Biology); Measurement; Proteins; Skin; (4) Bicarbonates; Chemical analysis; Determination; High altitude; Inhibition; Radiation injuries;

(3) Acidosis; Alkalosis; Blood circulation; Cholinesterase; Dehydration; Diuretics; Drugs; Electrocardiography; Exposure; Glucose; Hormones; Intestine; Membranes; Metabolic products; Perspiration; Radioactive isotopes; Rats; Sodium compounds; Surgical trauma; Temperature; Toxicity; X rays;

(2) Absorption; Adrenal cortical extract; Aldosterone; Anoxia; Blood transfusions; Body temperature; Body weight; Bullet wounds; Cardiac muscles; Countermeasures; Diet; Digestive system; Electric currents; Electric potential; Electrical properties; Fluids; Gamma rays; Hydrogen ion concentration; Hypertension; Intravenous feeding; Ions; Labeled substances; Lipoproteins; Mammals; Mathematical analysis; Muscular trauma; Nerves; Neuromuscular transmission; Pituitary hormones; Plasma volume; Posture; Potassium compounds; Radiation effects; Secretion; Serum; Separation; Simulation; Statistical analysis; Surgery; Test methods; Tests; Thailand; Theory; Traumatic shock;

(1) Acetic acids; Adrenal glands; Albumins; Alternating current; Altitude chambers; Amines; Amino acids; Ammonia; Analog systems; Analysis; Anaphylaxis; Aorta; Arteries; Ascites; Atomization; Atropine; Azoles;

Belgium; Bibliography; Bile; Blood pressure; Blood sugar; Blood volume; Body; Brain; Burns; Calcium; Carbon dioxide; Carboxylic acids; Cardiac glycosides; Chemical reactions; Chemical warfare agents; Chemical warfare injuries; Circuits; Clamps; Climatic factors; Colombia; Computers; Control; Cooling; Culture; Cytochemistry; Cytochromes; Dehydrogenases; Density; Design; Detection; Dextran; Diabetes; Diathermy; Dielectric properties; Diseases; Dogs; Dosage; Dose rate; Electrical conductance; Electrodes; Electronic equipment; Electrostatic capacitance; Embryos; Endocrine glands; Epidemiology; Erythrocytes; Excitation; Exercise; Fats; Fatty acids; Flames; Flash burns; Fluid flow; Gall bladder; Germany; Glands; Glycerols; Glycogen; Growth; Head injuries; Heat; Heat tolerance; Hemorrhage; Heparin; Histamines; Histological sections; Hyperthyroidism; Hypothalamus; Hypothermia; Impedance; Injection; Injuries; Insulin; Inulin; Iodine; Legs; Leukopenia; Lipids; Machines; Man; Mucous membranes; Netherlands; Nitrates; Nucleotides; Nutrition; Oxidation; Oxygen poisoning; Pancreas; Paralysis; Parathormone; Peritonitis; Pharmacology; Phosphates; Penetration; Peptic ulcers; Phosphorus; Phosphorus compounds; Photometers; Poisoning; Polarization; Polymers; Porosity; Pressure; Projectiles; Pyridines; Radiography; Recovery; Resistance; Respiration; Rheumatism; Sea water; Semipermeability; Spectrophotometers; Starvation; Steroids; Stress (Physiology); Survival; Test equipment; Thermodynamics; Thiocyanates; Thyroid hormones; Tritium; Tumors; Urea; V agents; Veins; Viability; Virus diseases; Viruses; Voltage.

CONTEXT SET: Electrolytes

(67) Electrochemistry; (57) Design; (51) Electrodes; (44) Conductivity; (37) Solutions; (32) Storage batteries; (31) Temperature; (30) Tests; (29) Electrolytic cells; (26) Polymers; (25) Cathodes (Electrolytic cell); Ions; Oxides; (24) Alkaline cells; Dry cells; Thermodynamics; (23) Polarization; (22) Materials; Power supplies; (21) Theory; (20) Wet cells; (19) Primary batteries; (18) Anodes (Electrolytic cell); Chlorides; Fuel cells; Salts; (17) Chemical reactions; Electrical conductance; (16) Electric potential; (15) Measurement; (14) Diffusion; Metals; (13) Density; Magnesium; Organic solvents; (13) Production; Zinc; (12) Bromides; Corrosion; Electrical properties; Sodium compounds; Solvent action; Viscosity; (11) Physical properties; Silver compounds; Water; (10) Ammonia; Battery separators; Colloids; Electrodeposition; Hydrogen; Ionization; Lithium compounds; Methanol; Potassium compounds; (9) Aluminum; Cadmium; Catalysts; Chemical properties; Electrolysis; Hydroxides; Surface properties; (8) Dioxides; Electrolytic capacitors; Manufacturing methods; Nickel; Polarographic analysis; Sulfates; (7) Absorption; Benzenes; Dielectrics; Electrical double layer; Eutectics; Low temperature batteries; Magnesium compounds; Oxygen; Solubility; Sulfur compounds; Titanium; Voltage; (6) Capacitors; Carbon; Conductors; Dielectric properties; Impurities; Industrial production; Manganese compounds; Melting; Molecular structure; Perchlorates; Radioactive isotopes; Solids; Steel; Synthesis;

(5) Absorption; Alcohols; Bibliography; Chemical analysis; Ethylenes; Fluorides; High temperature research; Hydrogen compounds; Iodides; Lead compounds; Liquids; Manganese dioxide electrodes; Nickel compounds; Organic compounds; Polymerization; Preparation; Pyridines; Silver; Stability; Sulfuric acids; Vinyl radicals;

(4) Acids; Anodes; Coatings; Cobalt; Containers; Copper compounds; Decomposition; Determination; Formamides; Fuels; Hydrocarbons; Low temperature research; Mathematical analysis; Membranes; Methyl radicals; Oxidation; Oxygen electrodes; Pellets; Peroxides; Plasticizers; Plastics; Processing; Resins; Semiconductors; Stereochemistry; Sulfides; Surface tension; Surfaces; Thin films; Ultrasonic radiation; USSR; Vapor pressure; Vapors; Zinc electrodes;

(3) Aging; Alkyl radicals; Alloys; Ammonium radicals; Brass; Butyl radicals; Carbides; Carbonates; Carboxylic acids; Cathodes; Ceramic materials; Chemical bonds; Chromium; Copper; Corrosion inhibition; Crystal structure; Deterioration; Effectiveness; Ethanol; Gases; Halides; Heat exchangers; Hydrochloric acid; Impregnation; Ion exchange resins; Iron; Life expectancy; Light; Magnesium electrodes; Mercury; Mercury electrodes; Mixtures; Molecules; Nickel electrodes; Nitrates; Nitrogen compounds; Nylon; Oxidation-reduction reactions; Phase studies; Physical chemistry; Platinum; Polymer solutions; Pressure; Radiofrequency; Scattering; Seals; Single crystals; Soaps; Sound; Styrenes; Sulfonic acids; Thiocyanates; Titanium compounds; Tracer studies; Transport properties; Ultrasonics; X-ray diffraction analysis;

(2) Alkali metal compounds; Alternating current; Amines; Ammonium picrate; Boric acid; Borides; Cadmium compounds; Carbon compounds; Carbon dioxide; Catalysis; Cathodic protection; Chromium alloys; Chromium plating; Communication equipment; Complex ions; Copper coatings; Corrosion research; Cyanides; Deposits; Detergents; Dipole moments; Electric discharges; Electroerosive machining; Electrolytic polishing; Electron beams; Electrons; Electroplating; Electrostatic capacitance; Electrostatic fields; Foils; Gas generating systems; Graphite; Heat resistant alloys; Heat treatment; Hydrazines; Hydrogen electrodes; Hydrogen ion concentration; Impedance; Iodine; Ion exchange; Iron alloys; Lattices; Liquefied gases; Machining; Magnesium alloys; Magnetic fields; Manganese; Mercury alloys; Metal coatings; Metallic compounds; Metallurgy; Microstructure; Mine fuzes; Nitro radicals; Phenyl radicals; Phosphates; Phosphoric acids; Plasma physics; Porosity; Potentiometers; Powder metallurgy; Radiation damage; Rectifiers; Reduction; Refractive index; Sea water; Sealing compounds; Separation; Silicon compounds; Silicones; Silver electrodes; Sintering; Sound transmission; Spectrographic analysis; Stabilization; Sulfur; Test equipment; Thermochemistry; Thio radicals; Time delay fuzes; Valence; Velocity; Vinyl alcohol; Volume; Wire; Zirconium;

(1) Acetates; Acetone; Acetonitriles; Acoustics; Acrylic resins; Additives; Aerosol generators; Aerosols; Air; Aircraft; Alkaline earth compounds; Alkaline earths; Aluminum compounds; Amides; Ammonium compounds; Analysis; Anthracenes; Antimony compounds; Antisubmarine warfare; Atomic orbitals; Auxiliary power plants; Beryllium; Beryllium compounds; Betatrons; Bismuth alloys; Bonding; Bromine; Bubbles; Butanes; Calorimeters; Camera shutters; Canada; Casting; Cellulose; Chelate compounds; Chemical equilibrium; Chemical milling; Chemicals; Chromates; Chromatographic analysis; Circuits; Cleaning; Climatic factors; Clock delay mechanisms; Coagulation; Coal; Cobalt alloys; Colorimetry; Complex compounds; Conferences; Construction; Coolants; Corrosive liquids; Crystallization; Crystals; Culture; Culture media; Cyanates; Cyclohexanes;

Cysteine; Cystine; Data processing systems; Dental materials; Detection; Deuterium oxide; Diamonds; Diodes; Discharge tubes; Distribution; Dissociation; Drops; Economics; Elasticity; Electric arcs; Electric currents; Electric power production; Electric propulsion; Electricity; Electromagnetic fields; Electrophoresis; Electroplating solutions; Energy; Enthalpy; Entropy; Escherichia; Esters; Ethanes; Ethylene oxide; Ethyleneamines; Extrusion; Fatty acids; Field emission; Floating docks; Flotation; Fluorine; Fluosilicic acids; Formaldehyde; Formates; Formic acid; Fracture (Mechanics); Freezing; Frequency; Fuel tanks; Fuzes; Galvanometers; Gas discharges; Gas ionization; Generators; Germanium; Germany; Glycerols; Glycols; Gold alloys; Grains; Growth; Guided missiles; Gun barrels; Halogens; Handling; Heat; Heat of formation; Heat of reaction; Heat of solution; Heat production; Heat transfer; Heating; Heavy water; High altitude; High frequency; Hydrides; Hydrofoil boats; Hydrolysis; Hydrophones; Imines; Impedance bridges; Industrial equipment; Infrared detectors; Inorganic substances; Instrumentation; Intensity; Interferometers; Internal combustion engines; Ion beams; Ionic current; Iron compounds; Isotopes; Italy; Kerosene; Lead; Lead alloys; Lithium; Low frequency; Machine tools; Magnetic susceptibility; Magnetohydrodynamics; Magnetostriction; Mechanical properties; Mercury compounds; Metal films; Methanes; Microorganisms; Microspectrophotometers; Microwaves; Mines; Moisture; Molding; Molecular association; Molecular isomerism; Molybdenum alloys; Molybdenum compounds; Monoxides; Motion; Naphthacenes; Naphthalenes; Napthyl radicals; Naval research; Neutron bombardment; Niobium; Nitrobenzenes; Nitromethanes; Nuclear magnetic resonance; Nuclear power plants; Numerical analysis; Optics; Oscillation; Packaging; Palladium catalysts; Particles; Phosphate coatings; Phosphorus compounds; Photographic filters; Plasma oscillations; Polycyclic compounds; Porous materials; Porous metals; Powders; Power; Precipitation; Pressure gages; Pressure vessels; Programming; Propagation; Pumps; Radiosondes; Raman spectroscopy; Reaction kinetics; Reagents; Recombination reactions; Reflection; Refraction; Refractory materials; Regeneration; Reliability; Resistance; Resonance absorption; Safety devices; Sea water batteries; Sedimentation; Semipermeability; Shock waves; Silicon alloys; Simulation; Sintered alloys; Sodium; Sodium alloys; Solvates; Sonar; Space charges; Specifications; Spheres; Stainless steel; Storage; Structures; Sulfonates; Sulfones; Switches; Tankers; Tantalum; Tantalum capacitors; Tantalum compounds; Tellurium alloys; Test methods; Thallium alloys; Thiazoles; Thiols; Thiourea; Thorium; Thorium compounds; Time switches; Tin; Tin alloys; Titanium alloys; Titration; Transistors; Tungsten; Tungsten compounds; Ultrasonic properties; Underwater explosions; Uranium compounds; Urea; Vanadium; Vibration; Voltage regulators; Water activated batteries; Water vapor; Wave transmission; Waveguide slots; Weapons; Zinc alloys; Zinc coatings; Zinc compounds; Zirconium compounds.

Terms Common to the Context Sets for Electrolytes and Electrolytes (Physiology)

Absorption; Amines; Ammonia; Analysis; Carbon dioxide; Carboxylic acids; Chemical analysis; Chlorides; Circuits; Climatic factors; Culture; Density; Design; Determination; Electric currents; Electric potential; Electrical conductance; Electrical properties; Electrodes; Electrostatic capacitance; Fatty acids; Germany; Glycerols; Heat; High altitude; Hydrogen ion concentration; Impedance; Iodine; Ions; Mathematical analysis; Measurement;

**Membranes; Nitrates; Oxidation; Phosphates; Phosphorus compounds; Polarization; Polymers; Porosity; Potassium compounds; Pressure; Pyridines; Radioactive isotopes; Resistance; Sea water; Semipermeability; Separation; Simulation; Sodium; Sodium compounds; Temperature; Test equipment; Test methods; Tests; Theory; Thermodynamics; Urea; Voltage; Water.**

1. Meaning of Meaning
C. K. Ogden and I. A. Richards
Harcourt, Brace & Company

2. Science and Sanity
Alfred Korzybski
The International Non-Aristotelian Library Publishing Company

3. Introduction to Semantics
Rudolf Carnap
Harvard University Press

4. Abstraction ladders are mentioned in this context at least as early as 1951. See: Electronic Digital Machines for High-Speed Information Searching
Philip Rutherford Bagley
MIT

5. For a more complete discussion of this topic see:
Hierarchy, Role Indicators and Retrieval
Doretha A. Bebbs
ASTIA (unpublished)

6. This kind of display is far from original and may be found in such advertising literature as:
IBM Information Retrieval
(undated and unsigned)

7. The importance of this set was pointed out to me by Mr. Martin Brooks during discussions on false drops and efficient searching procedures and tools. If each word T were printed out with its associated set, this compilation could be checked prior to a machine run and would prevent searches of unconnected terms. Such an arrangement was manually accomplished in the master Subject Heading File in ASTIA for Subdivisions. Each Subdivision had listed the main headings for which it had been authorized.

8. A similar technique is given on Page 16 of Automatic Indexing: An Experimental Inquiry
M. E. Maron
The RAND Corporation
AF 49(638)700, RM-2601
AD-245 175

9. Some Theoretical Aspects of the Mechanization of Literature Searching
Yehoshua Bar-Hillel
ONR Contract N62558-2214, Tech. Rept. No. 3
AD-236 772

10. The Construction of a Faceted Classification for a Special Subject
D. J. Foskett
Preprints of Papers for the International Conference on Scientific Information, Washington, D. C., 1958

11. On Relevance, Probabilistic Indexing and Information Retrieval
M. E. Maron and J. L. Kuhns
Journal, Association for Computing Machinery, 1960

12. The Theory of Clumps
A. F. Parker-Rhodes and R. M. Needham
Cambridge Language Research Unit
Cambridge, England, Feb. 1960

13. Automation of ASTIA - 1960
AD-247 000

14. The correct quotation is, "Sir, a woman's preaching is like a dog's walking on his hind legs. It is not done well; but you are surprised to find it done at all."

Additional references relating to a distance function:

(a) Correlative Indexes III: Semantic relations among semantemes -- The Technical Thesaurus
C. L. Bernier and K. F. Heumann
Amer. Doc. 8, No. 3, July 1957

(b) Delegation of Classification
R. A. Fairthorne
Amer. Doc. 9, No. 3, July 1958

(c) Retrieval by the Method of Proximity Transformations
C. N. Mooers
Unpublished, 1958

(d) Some Mathematical Fundamentals of the Use of Symbols In Information Retrieval
C. N. Mooers
Presented at UNESCO Paris Conference, June 1959

(e) The Structure of Information Retrieval Systems
B. C. Vickery
ICSI, Area 6, Washington, D. C., 1958

Prepared 8 December 1960

# PART III: PRAGMATICS

CONTENTS

## INTRODUCTION

In the previous sections of this paper L-O Retrieval Systems were discussed as though they were linearly arranged with respect to each other, with Uniterms as originally proposed as one extreme or boundary point and with a thesaurus of controlled retrieval terms as the other terminus. It is not surprising, therefore, that the same arrangement is pertinent in a discussion of Pragmatics, specifically, the human factor in the use of these systems.

This linear arrangement is meant to portray the degree of control exercised over the retrieval terminology. Even though subject headings, in effect, provide a degree of pre-coordination (i.e., Indexes -- Preparation) which is largely lost in a thesaurus approach, subject headings do not constitute as completely a controlled vocabulary as that provided by a thesaurus.

There is no great issue here. Other aspects of retrieval would indicate a different ranking for these systems. I am interested in a scale which shows the movement from an uncontrolled to a more and more stringently controlled vocabulary, because, as I interpret the literature, both theorists and practitioners are moving in the direction of controlled vocabularies for machine use. I believe that this movement is inevitable and will increase, but I am not sure whether further increases in control necessarily imply such techniques as links and role indicators. Alternatively, if such controls are required, techniques for their implementation are required which are far simpler than any devised to date.

I think that as time goes on a common ground and much information of mutual interest will be discovered between librarians and their controlled machine-oriented terminology and linguistic techniques now being developed as tools for the machine translation of one language to another. Since this development is for the future, there is no discussion of this presumed connection in the text. Nevertheless, it is my firm conviction that theorists should begin to examine more closely the techniques and linguistic analyses of machine translation for the illumination they may provide to the problems of a controlled vocabulary -- which, in the final analysis, is one of the earmarks of language.

## Analyst Requirements

For simple Uniterm systems, technically trained personnel are not required, and indeed this is one of its selling points. Clerks, instructed to pick out "important" terms in the text, function quite adequately as analysts. No decisions of a semantic or technical nature are required of the clerk-analyst, nor are they required to evaluate the importance of the "important" words they choose. In short, the clerk-analyst indexes each report, and each is indexed within the framework of its own terminology. Such an arrangement leads to a very rapid growth of the number of different retrieval terms used.

Very quickly this naive system gives way to a more sophisticated position in which some attempt is made to specify rules involving plurals, the consolidation or continued separation of such clusters as hardness, hardenability, hardening, etc. The end result is a system as described by J. C. Costello, Jr.[1] At this stage we are no longer on the boundary point of the retrieval system continuum but are somewhere within. The chief distinction between Uniterms, Subject Headings, Descriptors, and other varieties of retrieval terms becomes a distinction between the two major philosophies of document categorization: indexing and subject analysis. Consequently, succeeding remarks on personnel requirements also fit the modified Uniterm systems.

Subject headings as retrieval terms (as utilized by large libraries of which the Library of Congress is the prime example as well as large document libraries such as ASTIA under manual operation) are no longer a matter for clerk-analysts. An apprenticeship is required and subject specialists are sought - specialists to the degree that the beginning salary will attract, a not too happy situation. The length of apprenticeship depends, of course, on the mental agility of the trainee as well as on the complexity of the system.

Subject headings may be generated on some philosophical basis, such as a faceted scheme, or they may be tailor-made to fit a given collection and grow like Topsy. An example of the latter is the fourth edition of the ASTIA Subject Headings which contained about 70,000 entries and which contained many surprises for the novice.

> Sheets was never used in the sense of Bed linen; Grille was used for a cross-hatched effect or a type of construction, and was distinguished sharply from Lattice which was used as Crystal lattice and never as Lattice in the context of mathematics. There was also a definite rule which limited Stabilization systems to non-aircraft uses. Aircraft contexts required use of the term Control systems.

Consequently this was not the easiest system to learn to use. The philosophy was that of categorizing the document in broad outline with (usually) four or less subject headings displayed as follows:

1. Scientific reports -- Classification

2. Indexes -- Effectiveness
3. Dictionaries -- Preparation
4. Data storage systems -- Effectiveness

The most complicated scheme -- from the analysts' point of view -- is represented by the ASTIA system of descriptors. It would not be much in the way of exaggeration to compare this system with a dialect. To quote in part from Webster's New Collegiate Dictionary (1958)

> "Dialect. . . applies chiefly to a form of language persisting in the locality or among a group and marked by peculiarities in vocabulary, pronunciation, usage, etc."

In the ASTIA system words with one or more dictionary meanings are given an in-house meaning which prescribes their usage within sometimes narrow bounds. Each technical term (about 7000) has been placed in only one group. This group membership gives a context to the term and gives a clear indication of usage.

Albedo
(Nuclear reactor technology)

Beaches
(Hydrology)

The concept of assigning each term to but one given area has been a source of irritation to some, and a source of bewilderment to others. Much unnecessary confusion would be eliminated if the group designator (that within parentheses immediately below the descriptor -- i.e., "Nuclear reactor technology" and "Hydrology" in the examples above) were mentally put in what heretofore has been their traditional location: Albedo (Nuclear reactor technology) and Beaches (Hydrology) means exactly what Albedo (Nuclear reactor technology) and Beaches (Hydrology) means.

This innovation pays dividends. It not only defines a context for each term as its more traditional forbears did, it is also an access point to the group or schedule designated by the parenthetic instruction where other related terms may be scanned. It also serves as a warning that this is a controlled vocabulary and asks the user if he is using the terminology in an appropriate way.

Further clarification of the "extent" of the word may be given by Includes designations.

Beaches
(Hydrology)
Incl: Coasts
Seashore

Conveyors
(Transportation)
Incl: Tramways
Transfer trays
Trays

## Rules to Live By

Analysts in any system operate under rules. Such rules may be informal and unwritten, or written, formal, and specific. Uniterm indexers operate under a general rule to post all important words given by the document, where important may be defined rather closely and specifically as indicated by the Costello[1] article already referred to. The key word in the last sentence is indexers. The philosophy implied by this word is what distinguishes the activity of the Uniterm analyst from analysts of other systems. The Uniterm analyst is an indexer, and commonly chooses from 3 to 50 terms (or more) per document. Some insight into the statistics of this facet of information retrieval can be obtained from the National Science Foundation's reports on Nonconventional Technical Information Systems in Current Use[2].

The Subject heading or retrieval term analyst, by contrast, is a cataloger. He commonly uses far fewer than 20 terms per document, and some of those that he does assign may not appear in the document at all. For comparison we display instances of Uniterms, Subject headings, and Descriptors assigned by the same group of people (but at different times extending over a period of several years) to the same documents (labeled 1 and 2). The subject headings by their nature, and augmented by the manner of their display, tend to tell a story (this is a good test of the adequacy of Subject heading cataloging). Uniterms and Descriptors break this pattern and allow a free association of terms, but the pattern must be reconstitutable. In general more Uniterms than Descriptors are assigned, though a large number of Descriptors may be assigned from time to time (Documents 3 and 4).

Document 1

| | |
|---|---|
| 1. Microwaves -- Propagation | Solid-state |
| 2. Plasma oscillations -- Excitation | Millimeter-wave |
| 3. Microwave amplifiers -- Design | General |
| 4. Semiconductors -- Electron transitions | K-band |
| 5. Resonance absorption (RF) -- Applications | Magnetic Resonance |
| Microwaves | Spectrograph |
| Propagation | Semiconductor |
| Plasma oscillators | Cyclotron |
| Microwave amplifiers | Absorption |
| Design | Plasma |
| Semiconductors | Electron |
| Electron transitions | Oscillation |
| Resonance absorption | |

Document 2

| | |
|---|---|
| 1. Microwave oscillators -- Development | Maser |
| 2. Microwave amplifiers -- Development | Amplifier |
| 3. Resonance absorption (RF) -- Applications | Noise |
| | Atom |

Microwave oscillators
Design
Microwave amplifiers
Resonance absorption
Radiofrequency

Device
Effuser-test
Oscillator
Wave
Molecule
Resonance
Generator

The words following the double dashes are words supplementary to the main term which precedes the dashes, and each succeeding line is of lesser importance than the one preceding it. In printed form, this is a method of role indication. Unless specifically provided for, these aids are lost in machine systems.

What are the rules for assigning descriptors to documents? Certainly the first rule is to be true to the document. The descriptors are meant to describe the scientific content of the document in terse form. Second, each descriptor assigned must be used only in the context allowed by the controlled vocabulary. Third, the descriptors assigned must be reviewed to determine whether or not they are adequate to retrieve this document as part of a search within an area of which this document is a logical part.

Documents 3 and 4 illustrate descriptor assignments which follow these rules. Document 5 contains a mistake.* Document 6 is correct.

Document 3

Cerenkov radiation; Microwave amplifiers; Design; Radio astronomy; Hyperfine structure; Magnetrons; Microwave equipment; Molecular beams; Molecular spectroscopy; K band.

Document 4

Microwaves; Radiofrequency generators; Frequency multipliers; Diodes; Ferrites; Waveguides; Microwave amplifiers; Transmission lines; Production; Detection; Communication systems; Propagation.

Document 5

Diseases; Schistosoma; Physiology; Diet; Pathology; Antibodies; Antigens; Histology; Metabolism; Proteins; Nutrition; Hematology; Schistosomiasis; Molluscacides; Anemia; Folic acid; Bone marrow; Sprue; Electrolytes; Diagnosis; Therapy.

Document 6

Microorganisms; Separation; Flotation; Escherichia; Culture; Culture media; Sodium compounds; Chlorides; Water; Salts; Sea water; Test methods; Phosphates; Electrolytes; Carbonates; Nitrates; Sulfates; Reagents.

*If the reader cannot locate the error, he is requested to re-read the Appendix to Part II: Semantics.

Cataloging or indexing, by whatever name, is not an end in itself. It is done solely for retrieval. The fact that no document can be retrieved by means of a term not on it is too often forgotten. However, the play-it-safe attitude of assigning multitudes of terms to each document can be almost as bad as assigning an insufficient number of terms, for the increase of possible false coordinations as new terms are added is very high.

Machine capabilities open up possibilities for retrieval in depth. This concept is not synonymous with a mere increase in the number of retrieval terms per document, though this certainly is one of the more apparent consequences. Far more important is the maintenance of quality (or pertinency) with increasing depth of analysis. The mere presence of a technical term in a document is not by itself sufficient evidence that the term should be used as a retrieval term. Context is important here. Consider a document on the chemistry of the halogens in which fluorine has been excluded. Fluorine may be mentioned many times in the document by way of exception. Yet fluorine is not properly a retrieval term for the document. Clearly no unequivocal rule can be given. We note, however, certain facts: (1) The simple Uniterm analyst need make no decision. Every technical term in the document is grist for the mill, and each word chosen has EQUAL WEIGHT. (Similarly, statistical techniques, no matter how weighted in the process, result in a list of words of EQUAL WEIGHT. A list of descriptors falls into the same category. An exception is probabilistic indexing , but whether this scheme is really feasible is not known.) (2) Depth of indexing increases the mental demands placed upon modified Uniterm and retrieval term indexers, for although the words chosen are of EQUAL WEIGHT, they must be chosen only in those contexts which the specific cataloging system allows. Hence indexing in depth demands of the indexer full mastery of his allowable terminology and extensive knowledge of the subject field in which he is indexing.

Hierarchical analysis is sometimes advocated as a means of analysis in depth. Hierarchical charts have been displayed in other portions of this paper, and a study of such charts should instill proponents of this kind of analysis with caution. Hierarchical analysis only leads to bedlam unless the strictest rules of procedure are followed. One must start to index at the lowest level permitted by the document. One then builds a generic structure over the basic terms along the lines permitted by the hierarchical chart for that discipline, and which of several permissible lines will be followed in a given case must be determined by the document. Consequently, hierarchical indexing, except for the most elementary kind, cannot be considered an automatic machine job. The key to hierarchical indexing - as it always must be - is the document itself. But no matter what the excellence or import of the document, hierarchical indexing is successful only to the degree that the analyst is knowledgeable and consistent.

A few examples of analysis by category and by depth (with hierarchy) may be of some value.

Example 1

Title: Concepts of Automatic Data Storage and Retrieval in the Simplex System[4].

| Analysis by Category | Analysis in depth and with hierarchy |
|---|---|
| Data storage systems | Data storage systems |

Data processing systems
Combat information centers
Army operations
Digital computers
Automatic
Design
Packaged circuits

Data processing systems
Military communications
Communication systems
Communication equipment
Army operations
Military operations
Digital computers
Computers
Packaged circuits
Circuits
Data transmission systems
Combat information systems
Automatic
Design

Example 2

Title: Research on various phenomena for the performance of circuit Functions [5].

Author abstract: An investigation of various solid-state phenomena and effects which might have an application in electronic systems was made. Technical articles on fifty-two phenomena and effects are presented in this report. Each article includes a definition, a detailed description, and a short bibliography. Eighteen other related effects are covered. A subject index in which the effects are grouped according to a general classification scheme and an alphabetical index are included.

Categorizing descriptors

Circuits
Dictionaries
Electromagnetic theory
Electronics
Physics
Solid state physics
Semiconductors
Scientific research
Subminiature electronic equipment

Indexing in depth

Hall effect
Metals
Germanium
Silicon
Magnetic fields
Thermodynamics
Photoelectric cells
Photoconductivity
Dielectric properties
Dielectrics
Luminescence
Polarization
Selenium
Thermionic emission
Superconductivity
Magnetostriction
Ferroelectricity
Antiferroelectricity
Electric fields
Crystals
Electrets
Magneto-optic rotation

Ionic current
Nuclear magnetic resonance
Paramagnetic resonance
Electrons
Lattices
Electronic circuits
Electromagnetic effects
Magnetic effects
Magnetism
Resistance
Thermoelectricity
Transients
Electronic systems
Surface properties
Light
Photopotentials
Refractive properties
Transistors

Indexing in depth is largely motivated by the fear that significant material in the collection will be missed by topical cataloging. The concern for completeness is quite legitimate, but indexing in depth does not automatically guarantee 100% coverage. Several factors are at work here. First, topical or categorical analysis, perceptively and consistently done, with retrieval based on the same considerations and with a full knowledge of the cataloging system, should function at a high level of efficiency. The fear of missing information usually is stated in terms of detailed information, not always in the primary field of interest of the document in question. As an example, consider the cataloging of this report:

Title: A Study of Restricted Random Walk[6]

Category analysis

Numerical analysis
Numerical methods and procedures
Digital computers
Theory
Tests

Depth indexing (and hierarchy)

Numerical analysis
Numerical methods and procedures
Applied mathematics
Statistical analysis
Probability
Sampling
Digital computers
Computers
Programming
Theory
Tests

This report is a detailed account of the simulation of random walk by computer. Random walk is not a retrieval term in the ASTIA system, it is covered by the concept of Monte Carlo methods, which in turn is included in the descriptor Numerical methods and procedures. Within the document a short history of the problem of restricted random walk is given dating back to 1934. We have the following statement:

"Among the physical properties which can be derived from a random chain model are light scattering, elastic properties of rubber-like chains, viscosity effects, diffusion, sedimentation, and birefringence of polymers in the stretched state."

Note that in neither topical nor depth indexing did any of these terms appear. They are incidental to the message of the report, and descriptor references to Viscosity or Polymerization would definitely be misleading. (Consequently it is not sensible to expect to retrieve this document in a bibliographic request on, say, Polymerization.) This may be noncontroversial and it may be admitted that any reasonable statistical approach to cataloging should not pick up these words. Each is used with frequency one. However, consider the following quotation from the same report.

"Although it is difficult to take account of a memory that remembers every step of a restricted random walk, it is possible to give a form theory for a random walk that will remember a fixed number of steps. Montroll (J. Chem. Phys., 1950, p. 734) gave the first exact treatment for a random walk in a two dimensional rectangular lattice, that had a four step memory, that is, four step overlaps were forbidden, but higher order overlaps were permitted. Montroll's treatment, subsequently extended by Frisch, Collins, and Friedman (J. Chem. Phys., 1950, p. 1402) enumerated all possible step configurations of a restricted random walk, considering each configuration as a state in the Markov chain."

Question: Should Markov chains be picked up in the cataloging? In the ASTIA system this term is included under Probability, which was picked up in the analysis in depth, but not because of this usage of Markov chain! This raises at least two questions: Pertinency of cataloging and the cataloging terminology. The problem of pertinency arises in all cataloging: it is simply aggravated in indexing in depth. Cataloging terminology is another matter. There are generic terms like Electronics, and there are specific terms like Battery separators, generic terms like Biology and specific terms like Liver. Hierarchical indexing requires a generic term one level above the most specific term used, depth of indexing requires a great multiplicity of specific terms. This raises the often asked, but never answered, question of the number of retrieval terms required to adequately cover a library of a certain size. Certainly that number will depend upon the philosophy of cataloging. Indexing by category requires a preponderance of generic type terms and the terminology would be relatively small. Indexing in depth for specific details requires many, many specific terms and the total number of retrieval terms could be quite large. (The ASTIA system probably falls somewhere between these two limits.) If hierarchy is to be superimposed upon indexing in depth, then the total number of retrieval terms may grow to be very large indeed. Each of these systems will present its own problems for machine retrieval.

Chapter 2

## Machine Preparation of Bibliographies

What is a good bibliography? More modestly, what is a bibliography? In its final form it is a list of citations. This may consist of acquisition numbers, of titles, or catalog cards which include title, author, abstract, etc. A good bibliography is not good in the abstract; it is good for somebody. On this basis, only a recipient of a bibliography can judge its merit. This however is not equivalent to, nor an excuse for, offering unedited machine runs as bibliographies.

Bibliographies are made up for a variety of reasons, some demand broad area coverage, some specific coverage. Clearly, the broader in scope the bibliography, the more permissive is peripheral material. As the specificity of subject matter increases, the amount of tolerable peripheral material contracts. Probably no bibliography can be made so specific as to eliminate all peripheral material to everyone's satisfaction. We are thus faced with a real, but fuzzy problem of the allowable false drop percentage in raw machine output if such a percentage is to be a measure of the adequacy of the retrieval system.

However, the raw output of EDPS equipment may not constitute a suitable bibliography, where suitable is defined as more than the allowable number of non-pertinent references. Illumination for this problem is obtained by comparing the methods of bibliographic search in manual and machine systems. For this purpose we suppose we have a bibliographic requirement on the subject of "Methods of indexing and classifying reports for retrieval." We assume a manual, subject heading context, and we will assume that the first three of the four subject headings given at the beginning of Chapter 1 will cover our needs. These headings are:

1. Scientific reports -- Classification
2. Indexes -- Effectiveness
3. Dictionaries -- Preparation

The bibliographer proceeds to the files, finds the header card labeled "Scientific reports", and within this file locates those cards which were assigned the heading "Scientific reports -- Classification." The analyst now proceeds to read these cards in some particular fashion. He certainly will read the title, and depending upon the title will scan or read word for word the abstract if there is one. In addition he will check the other subject headings listed on the cards filed under "Scientific reports--Classification."

Upon reading the cards -- let us assume there are five cards in the file with the given subject heading -- the bibliographer determines that only three out of the five are pertinent to his present search. He therefore records only three acquisition numbers. However, the headings on one of the three that he keeps suggests that the following heading might be worth searching: "Data storage systems -- Effectiveness."

Note that this heading was not on his original list. The bibliographer has therefore enlarged his original research on the basis of citations already discovered. Let us suppose that he finds ten cards with this heading and that five of these are pertinent. He records these acquisition numbers, and proceeds to the second subject heading on his original list: Indexes -- Effectiveness.

We will suppose that the following tabulation covers the results of his work.

| | Cards in file | Pertinent cards |
|---|---|---|
| Subject heading 1 | 5 | 3 |
| Suggested heading | 10 | 5 |
| Subject heading 2 | 50 | 35 |
| Subject heading 3 | 6 | 4 |
| Totals | 71 | 47 |

The bibliography as sent to the customer contains 47 citations, or 66.2% of the references searched (in a manual system this figure is almost never available). Before commenting further, let us look at this same bibliographic request as it might be processed by machine.

| Primary descriptors | Secondary descriptors |
|---|---|
| 1. Scientific reports | a. Classification |
| 2. Indexes | b. Effectiveness |
| 3. Dictionaries | c. Preparation |

Again we tabulate the results where 1a indicates a match between primary descriptor "1" and secondary descriptor "a."

| Match | Gross hits | Pertinent |
|---|---|---|
| 1a | 5 | 3 |
| 2b | 50 | 35 |
| 3c | 6 | 4 |
| 1b | 10 | 1 |
| 1c | 8 | 0 |
| 2a | 0 | 0 |
| 2c | 5 | 0 |
| 3a | 5 | 1 |
| 3b | 10 | 0 |
| Totals | 99 | 44 |

Actually, this tabulation is a little misleading: coordinations such as 1b, 1c, etc., will contain a certain percentage of gross hits which simply reproduce hits already recorded under 1a, 2b, and 3c. Therefore, let us imagine that the number of net hits is 71 of which 44 are pertinent. Then the machine search provides us with 62% pertinent hits. This compares favorably with the 66% under the manual system.

There are several points to make. First we look at only machine results (i.e., the comparison with a manual search is never made) and we tend to think that 62% pertinency implies that there is something wrong with the cataloging. (This may be so, but we will also illustrate that there can be something wrong with the way the machine is approached.) The second point is that the machine missed entirely the citations under "Data storage systems -- Effectiveness."

During a manual search discrimination and judgment are employed by the bibliographer at each step. He not only chooses only pertinent documents from the full file of cards, he may also enlarge or decrease the scope of his search on the basis of what he has already found in the file. Automatic self-enlargement of requests have been mentioned, but I know of none that has been implemented. A good reason for this is that several meanings can be attached to the idea of "self-enlargement."

In contrast with a manual search, a machine prepared bibliography is simply a pre-selected file dump regardless of the number of coordinations involved. This is fast, but the only discrimination and judgment allowed the bibliographer is that judgment which he uses in selecting his terms in the initial instance. His judgment of pertinency becomes effective, if it is allowed at all, only by screening the end results of the machine dump, and at that time he may also enlarge the search.

## Approaching the Machine for Bibliographic Citations

Five different examples of machine coordination for bibliographic citations will be given. The subject matter of the searches will not be spelled out in detail because these examples are meant primarily to illustrate the scope of the method of coordination. This is not an exhaustive or definitive study, and criticism is intended in terms of machine coordination rather than criticism as to the handling of a specific request in detail.

**Example 1: Rocket motor reliability**

| Primary level | Secondary level |
|---|---|
| Rocket motors | Reliability |
| | Life expectancy |
| | Quality control |
| | Failure (Mechanics) |
| | Sensitivity |
| | Climatic factors |
| | Vibration |
| | Shock resistance |

This search pattern represents coordination with Rocket motors and each of the terms on the secondary level. In subject heading terminology this would represent searching for combinations such as:

Rocket motors -- Reliability
Rocket motors -- Life expectancy . . . etc.

This search involves eight coordinations.

**Example 2: Wave velocity in metals**

| Primary level | Secondary level |
|---|---|
| Metals | Detonation waves |
| Alloys | Sound transmission |

Wave analysis
Wave characteristics
Wave transmission

This search pattern means that Metals will be coordinated with each of the five terms on the secondary level, and Alloys will also be coordinated with each of these five terms. This search involves ten coordinations. The "waves" referred to are nonelectromagnetic and might be mechanically induced by sound. On that basis the terms Wave analysis and Wave characteristics are ill advised. Both terms are meant to apply to the area of Optics and Spectroscopy. We therefore have four out of ten coordinations which should result primarily in false drops. In their stead the terms Vibration mechanisms and Ultrasonics would have been more meaningful. These coordinations of course were not accomplished, and any literature on wave velocity categorized by these terms will have been missed except in those instances when other terms of the search are also present on the same document. The search pattern is a valid one, but the choice of terms for coordination is questionable.

Example 3: Chemistry of the upper atmosphere

| Primary terms | Secondary terms |
|---|---|
| High altitude | Chemical analysis |
| Ionosphere | Chemical elements |
| Upper atmosphere | Chemical properties |
| | Chemical reactions |

This is the pattern for the coordination of three terms each with four terms for a total of 12 coordinations. The use of High altitude is questionable. Its primary usage is intended as a modifier to aircraft, balloons, and general usage in an aeronautical sense. Hence the four coordinations involving this term will probably produce mostly false drops if the coordination can be made at all.

Example 4: Circuit design for two-way communication devices, transistorized

| Primary terms | Third level terms |
|---|---|
| Communication systems | Butterfly circuits |
| Communication systems equipment | Circuits |
| Intercommunication systems | Clipped circuits |
| Radio equipment | Coupling circuits |
| Radio receivers | Delay circuits |
| Radio telephone | Differentiating circuits |
| Radio transmitters | Electrical networks |
| Secret communication systems | Electronic circuits |
| Telephone communication systems | Inverted circuits |
| Underwater radio transmission | Oscillator circuits |
| Voice communication systems | Printed circuits |
| | Scaling circuits |
| Secondary terms | Switching circuits |
| Transistors | Timing circuits |
| | Trigger circuits |

Triggered gates
Tuned circuits
Tuning circuits
Wiring diagrams

Fourth level terms

Design

A pattern for four level coordination involves coordinating the first two levels, and coordinating the accessions discovered with the third level terms. The accessions resulting from this coordination are finally coordinated with the fourth level terms. This is our first example involving more than two level coordination, and as such it illustrates the "Ideal" nature of coordination (Syntax: Chapter 3). If we label first level terms by "a," second level terms by "b" and third and fourth level terms by "c" and "d" respectively, then we may symbolically write:

$$abcd \supset abc \supset ab \supset a.$$

This symbolism illustrates one possible meaning of enlarging (or decreasing) the scope of a search. Actually, machine coordination provides the largest possible search first, then successively narrows the scope of the search with each coordination. This means, given the above terms as adequate coverage for the bibliographic request in question, that a deck search of the "a" terms gives the widest possible answer to the subject matter of the bibliography. Obviously such an unscreened deck search would contain many nonpertinent items. The scope of the subject matter area is decreased by proceeding to a second level coordination; and we, likewise, eliminate a good percentage of the nonpertinent material. In theory, each succeeding coordination should further refine the subject matter area, and this process is self-limiting. At some point there will be no documents containing all of the terms asked for by the coordination pattern. (Another kind of self-enlargement of a search is obtained by changing the initial set of descriptors.)

This particular search pattern requires 11 coordinations for the first two-level search. If each of these coordinations involves only one document, then in effect the third level search requires 11 times 19 or 209 coordinations. If each of the third level terms is involved at least once, then there are an additional 19 coordinations at the fourth level search. A very modest estimate, then, of the number of coordinations required by this search is 11 plus 209 plus 19 or 239 coordinations. For a library of any given size, there may be many, many more than 239 coordinations involved, though the final number of citations may be small.

Analyzing the terms used in this search, it would appear that the primary term Underwater radio transmission is not a good choice. This term is part of the terminology reserved for Wave propagation. However, it may be argued that this term is worth coordination with since there is no descriptor such as Underwater radio transmitters, that is, there is no piece of equipment which covers this field. If this reasoning is valid (and it is very doubtful validity; such documents should be descriptorized as Radio transmitters and Underwater to underwater), then Underwater radio receivers should have been included. Underwater telephones seems to have been missed completely.

The secondary term is unambiguous in light of the request statement, but why was not Electrolytic transistors included as a secondary term?

Wiring diagrams does not appear to be an appropriate third level term. There certainly is a connection between Wiring diagrams and Circuits, but not in connection with Design, the fourth level term. Wiring diagrams in this coordination would imply some such descriptor set as Communication systems, Wiring diagrams, Design. This is obviously incomplete, and adequate descriptorizing would require Communication systems, Circuits, Wiring diagrams, Design. At best it would seem that Wiring diagrams, for this particular search, is redundant.

This particular bibliographic search is also illuminating in terms of the utility of hierarchy in retrieval. Consider the group of primary terms: Radio equipment is generic to Radio receivers, Transmitters, and Telephones; Communication systems is generic to the remainder of the terms with the exception of Underwater radio transmission which is probably an error in any circumstance; Transistors is generic to Electrolytic transistors. Of the third level terms, Circuits is generic to all terms except Wiring diagrams which is probably misplaced, as stated, Electrical networks and Triggered gates. Hence hierarchial descriptorizing would have permitted the following simple coordination for this bibliography:

| Primary terms | Third level terms |
|---|---|
| Communication systems | Circuits |
| Radio equipment | Electrical networks |
| | Triggered gates |
| **Secondary terms** | **Fourth level terms** |
| Transistors | Design |

Example 5: Ejection of electrons from surfaces by ions, metastable atoms, and neutral atoms

Search one:

| Primary terms | Secondary terms | Exclude result of coordination when document contains any of the following terms |
|---|---|---|
| Atoms | Beams | |
| Electrons | Ejection | |
| Helium bombardment | Metals | |
| Ion beams | Surface properties | Field emission |
| Ion bombardment | Surfaces | Secondary emission |
| Molecular beams | | Thermionic emission |
| Particle beams | | |

Search two:

| Primary terms | Secondary terms | Exclude result of coordination when document contains any of the following terms |
|---|---|---|
| Electrons | Excitation | |
| | Motion | |
| | Production | Field emission |
| | Sources | Secondary emission |
| | | Thermionic emission |

Note how this bibliographic request is handled as two separate searches. The reason for this is clear if the secondary terms of the second search are merged with the secondary terms of the first search for coordination with Electrons which is a member of the primary terms of the first search. Obviously such a merge will give false drops of the kind Ion beams -- Production, a combination not pertinent to this search. Search two requires no further comment.

Search one of Example 5 will lead to trouble. The primary terms are suitable. However, Beams is a member of the group devoted to Structural members and forms, hence these could be wood beams as well as metal beams. On the other hand if structural forms are to be included, then, in addition to beams, we should list such items as Angle bars and Box beams.

Although the term Ejection is in the statement of the bibliographic request, it is not a suitable secondary term, given the primary terms. Coordinations such as Ion beams -- Ejection, will probably result in nonpertinent information. A more promising first search would look as follows:

| First level | Second level | Third level | Fourth level |
|---|---|---|---|
| Atoms | Metals | Electrons | Ejection |
| Helium bombardment | Surface properties | | |
| Ion beams | | | |
| Ion bombardment | | | |
| Molecular beams | | | |
| Particle beams | | | |

These four coordinations are to be followed by the "Exclude instructions" already given.

Even a fourth order coordination need not result in a high degree of pertinency of hits, and I refer to my previous remark that in theory, each succeeding coordination should further refine the subject matter area. The reason the theory does not always work is clear if we list the cataloging possibilities for those documents which have successfully passed the four level coordination criteria.

| | | |
|---|---|---|
| Atoms | Atoms | Atoms |
| Metals | Surface properties | Surfaces |
| Electrons | Electrons | Electrons |
| Ejection | Ejection | Ejection |
| . . . | . . . | . . . |
| Helium bombardment | Helium bombardment | Helium bombardment |
| Metals | Surface properties | Surfaces |
| Electrons | Electrons | Electrons |
| Ejection | Ejection | Ejection |
| . . . | . . . | . . . |
| Ion beams | Ion beams | Ion beams |
| Metals | Surface properties | Surfaces |
| Electrons | Electrons | Electrons |
| Ejection | Ejection | Ejection |
| . . . | . . . | . . . |

| | | |
|---|---|---|
| Ion bombardment | Ion bombardment | Ion bombardment |
| Metals | Surface properties | Surfaces |
| Electrons | Electrons | Electrons |
| Ejection | Ejection | Ejection |
| . . . | . . . | . . . |
| Molecular beams | Molecular beams | Molecular beams |
| Metals | Surface properties | Surfaces |
| Electrons | Electrons | Electrons |
| Ejection | Ejection | Ejection |
| . . . | . . . | . . . |
| Particle beams | Particle beams | Particle beams |
| Metals | Surface properties | Surfaces |
| Electrons | Electrons | Electrons |
| Ejection | Ejection | Ejection |
| . . . | . . . | . . . |

The three dot (. . . ) notation is to indicate that each list is (possibly) only part of the set of terms assigned to the documents retrieved by this four-level coordination. The use of the term Surface properties as a second level term leads in each instance to descriptor lists which have a good chance of being nonpertinent even after four coordinations, i.e.,

Ion bombardment
Surface properties
Electrons
Ejection

simply does not make as much sense as, for instance,

Ion bombardment
Metals
Electrons
Ejection.

It is true that, as the dots indicate, other retrieval terms assigned to the same document might make Surface properties meaningful. But this is of no help because it says that Surface properties has nothing to do with either Electrons or Ejection. Consider the prescription

Ion bombardment
Metals
Surface properties
Electrons
Ejection.

This is interpretable as the Ion bombardment of Metals with the Ejection of Electrons dependent upon the Surface properties of the Metals. Note, however, that this prescription and interpretation is obtained without coordination of the term Surface properties by the machine. Consequently Surface properties is of no help to the machine search for this bibliography, because when relevant it is picked up by the help of other terms, and when it is not relevant it probably contributes to

false drops. If the bibliographer is of the opinion that really relevant material must contain the term Surface properties, then the coordination required is as follows:

| First level | Second level | Third level | Fourth level |
|---|---|---|---|
| Atoms | Metals | Surface properties | Electrons |
| Helium bombardment | Surfaces | | |
| Ion beams | | | |
| Ion bombardment | | | Fifth level |
| Molecular beams | | | |
| Particle beams | | | Ejection |

This five-level coordination provides us with documents with the following cataloging:

| | |
|---|---|
| Atoms | Atoms |
| Metals | Surfaces |
| Surface properties | Surface properties |
| Electrons | Electrons |
| Ejection | Ejection |
| . . . | . . . |
| Helium bombardment | Helium bombardment |
| Metals | Surfaces |
| Surface properties | Surface properties |
| Electrons | Electrons |
| Ejection | Ejection |
| . . . | . . . |
| Ion beams | Ion beams |
| Metals | Surfaces |
| Surface properties | Surface properties |
| Electrons | Electrons |
| Ejection | Ejection |
| . . . | . . . |

plus six more combinations or a total of 12 five-level coordinations. Continued analysis would suggest that if there are documents in the collection cataloged as

Atoms
Surfaces
Surface properties
Electrons
Ejection

then the cataloging may not be adequate because we ought to be talking about the Surface and Surface properties of something. Therefore, really adequate cataloging would require the following kind of prescription:

Atoms
Metals
Surfaces
Surface properties
Electrons
Ejection

To insure this combination of retrieval terms on every document of the search would require a six-level coordination! It is not suggested that this degree of coordination is always or even frequently necessary. The analysis was carried through to show the implications of and results to be expected from a given bibliographic request pattern. Nor is this illustrative merely of the analysis required by the bibliographer before he approaches the machine. It is also indicative of the high order of consistency and competency required of the document analyst in assigning retreival terms.

Cataloging for Retreival: II

The above analysis of bibliographic requests suggests that depth of indexing may refer not merely to how many descriptors the analyst assigns to a document, but to the degree of coordination and refinement which the cataloging system permits in the way of retrieval. It is suggested that a six-level coordination is, in a sense, retrieval in depth which cannot be approached by two- or three-level coordinations. However, just as not every document permits of depth of indexing in the sense that 30 or 40 terms may validly be assigned to it, even with hierarchy, so not every bibliographic run requires sixth-level coordination. What is required is that the assignment of retrieval terms to documents must be done with retrieval in mind.

The bibliographer is in the position of second guessing the customer (unless he has voice communication with him), and to require that he also second guess the document analyst is to unnecessarily burden an already sufficiently complicated job. The bibliographer cannot retrieve what has not been descriptorized, and he can retrieve what was inadequately descriptorized only at the penalty or expense of increasing to a high value the ratio of pertinent documents to false drops.

Cataloging for retrieval then, has not so much a requirement of indexing in depth as it has a requirement for indexing for completeness and adequacy. This is illustrated by returning to the cataloging examples which represented the minimum cataloging for fourth level coordination: are these examples illustrative of completeness and adequacy if it is assumed that this cataloging is attached to documents which satisfy the bibliographic requirement for information on "Ejection of electrons from surfaces by ions, metastable atoms, and neutral atoms?"

It is not valid to argue that these are not pertinent examples because the dot notation ( . . . ) indicates that other terms belong to each descriptor set. This argument is not valid because those missing terms, whatever they may be, and however many there may be, do not enter into the machine coordination. If the remaining terms are significant to material satisfying the bibliographic request they must be included in the retrieval terms fed to the machine. If the reader will review these examples of descriptor sets until he finds a retrieval prescription which satisfies him as representative of a valid answer to the bibliographic request, then the reader will have determined for himself the required degree of completeness of descriptorizing necessary for machine retrieval.

One further point: quite possible there are no documents in the collection that specifically answer the request. This means that the request must be answered in terms of documents which deal only in part with this particular bibliographic question. Nevertheless, if retrieval is to take place, a group of retrieval terms similar to those given by the preceding examples of coordination must appear in the collection. This means that the possibility for retrieval of information rests

with the analyst. Retrieval can take place only if sufficient retrieval terms have been assigned to cover each aspect of the document. A multi-discipline or multi-topic document is retrievable only to the extent that each part is adequately descriptorized. This simply returns the argument to the requirement for a retrieval prescription which looks like this:

Atoms
Metals
Surfaces
Surface properties
Electrons
Ejection
.
.
.

where the vertical dots indicate that the document in question contains more retrieval terms in its prescription -- perhaps 30 more. In the latter case, the ejection of electrons from metal is probably only one of several topics covered in the document. But this information is retrieved simply because the portion of interest is adequately descriptorized. If any four of the six terms listed are left out when the document is analyzed on the assumption that it is not important, then the material has been lost as effectively as if this portion had not been descriptorized at all. This leads to the observation that if there is information in a document of sufficient worth for retrieval, then that subject matter must be descriptorized in complete detail. If the information is not worth retrieval -- or misleading for retrieval -- (as the case of Polymerization in the report on Monte Carlo methods), then that portion of the information should receive no descriptorization whatsoever.

## The User or Customer

> "Perhaps the most important and least considered factor in the design of information storage and retrieval systems is the user of such systems. Regardless of what other parameters are considered in the development of a storage and retrieval mechanism, it is necessary to consider its potential use and mode of use by the persons or groups for whom it is intended; it is necessary either to fashion the system to suit the user's needs, habits, and preferences, or to fashion the user to meet the needs, habits, and preferences of the system. Both approaches are possible, but the second one, involving the education and re-education of the user, is evolutionary and futuristic."[7]

A picture of ASTIA's users is usually given in terms of "Military", "Grantee", or other similar designation, but this is not much more helpful than calling him Mr. Smith or Mr. Jones. The important categorization for our purposes is by wants and needs. In this regard the study of Mr. Herner already referred to above

> ". . . shows the contrasting need of the pure scientist for mere references to information on the one hand and of the applied scientist for direct access to actual information of the other. The applied scientist . . . requires the services of large storage and retrieval programs manned by highly trained personnel, while the pure scientist is best served by a conventional library . . . arranged on the basis of a classification scheme . . . reflective of the scientist's customary association of subjects . . . ."

How, then, does ASTIA stand? ASTIA's customers are predominantly applied scientists, since this is the predominant portion of DOD sponsored research. Consequently, the recent trend toward indexing in depth is probably a step in the right direction. However, the real benefits derived from a collection indexed by depth are not realizable unless retrieval techniques grow in sophistication. In this regard we will refer again to Mr. Herner and his discussion of the automated library of the Smith, Kline and French Laboratories.

> "In terms of what can be done with the available information, by way of detailed searches and correlations of diverse facts and data, the system is extremely sophisticated. But the thing that makes it truly sophisticated is the fact that it is designed around the clearly defined needs and interests of its users. To illustrate this point, the strictly pharmacological information in the system . . . is indexed in such a way as to produce bibliographies rather than actual data. The reason for this is that the pharmacologists in the company were found to prefer to do their own reading, correlation, and synthesis, and all they want are references to the pertinent literature; they do the rest. On the other hand, the clinicians in the company . . . prefer to receive actual data, and, if possible, they want it correlated and tabulated for them. Therefore, clinical information is entered into the system in such a way as to permit routine correlation and tabulation."

ASTIA, too, is now in the position of having basically two different kinds of customers: (1) the applied scientist and his need for scientific information; (2) the DOD personnel responsible for initiating research efforts who require management type information as to funding in given areas, projects in being, etc. The applied scientist at present is supplied with a bibliography: the management specialist may be supplied with a bibliography, with tabulated data, or with a combination of both. The applied scientist acquires information from the ASTIA document collection, the management specialist acquires information from the RDT&E collection with possible cross referencing to the document collection.

Nevertheless, much remains to be done in the area of consumer research, not only at ASTIA, but in every library, automated or not. Studies are required on user language vs. retrieval language and bibliographic-request language vs. indexing language. Studies are also required on the relation of document requests to TAB announcement (stated this way, this is, of course, an ASTIA problem), the time span of the bibliographic coverage requested by the user, and much, much more.

This portion of pragmatics should probably be the longest portion. It is, in fact, the shortest. I can do no better than to end this brief discussion with the same words with which it was opened:

> "Perhaps the most important and least considered factor in the design of information storage and retrieval systems is the user of such systems."

The Information Retrieval Complex

Basically, information retrieval and its success are not just a matter of Uniterms, subject headings, defined vs. undefined terms, or machine vs. manual systems. Information retrieval is an Analyst-Bibliographer-User complex, each of which acts upon the other. This complex can be simplified to the extent that the Analyst and Bibliographer merge, and we may then speak of Information retrieval as a Library-User complex. These two cannot be merged, but thier relationship to each other can be improved by utilizing feedback for refinement, or as some of the Cyberneticists would say, by achieving homeostasis.

The librarian in charge of a manually operated library does not sit at a bare desk surrounded by shelves of books stacked in alphabetical order. The library data -- books, manuscripts, documents, film -- are organized for search purposes. This organization may take the form of the Dewey Decimal System, UDC, or some other scheme. The kind of system is not at issue: the fact of library organization is what is important. In addition, the librarian is provided with certain searching aids: a subject catalog, an author catalog, etc.

Can a machine-controlled library operate on less? Indeed, is a very library really machine-controlled if it provides less in the way of tools than are available to the conventional librarian? It is a serious mistake to suppose that an electronic computer, data on cards or tape, and a code manual can function as an adequate reference library. This is playing the game blindfolded and with one hand tied behind one's back. Challenging perhaps, but not very efficient -- or sensible.

An electronic computer, by manipulation of the data tapes or cards, can and should provide the bibliographer librarian with the following kinds of desk aids:

(1) Descriptor frequencies

   (a) as assigned to documents by analysts
   (b) as used by bibliographers in answering requests.

(2) Low frequency words with documentation (AD) citations. (Requests involving just these terms -- or primarily these terms -- need not be processed by the computer. They can be handled most efficiently on a manual basis.)

(3) Very specific terms and the generic terms which are hierarchical to them.

(4) Descriptor groupings other than the present 292 -- not as a replacement, but as a supplementary aid. These would take the form of a listing of all adjectival words, all descriptors ending in "equipment", all descriptors which are names of equipment, etc.

(5) Special word groupings as required by the bibliographer on the basis of requests processed.

(6) Full context sets for descriptors used 100 times or less.

(7) Partial context sets for descriptors used more than 100 times

   (a) words used 101 to 1000 times: list the 100 words most frequently occurring in their context sets.
   (b) words used 1001 times or more: list the 50 words most frequently occurring in their context sets.

(8) An author listing

(9) A permuted title index: this is not an external publication, but an in-house tool.

(10) A contract file

(11) A source file

(12) Combination files, such as Source-Contract, Contract-Source, Author-AD, etc.

One of the indispensible aids to the mathematician, scientist, and engineer is the Handbook for Physics and Chemistry. It is time that librarians, particularly those with EDPS equipment, demanded a Handbook for Bibliographers and Indexers. The complete set of data described in items 1 through 12 will not fit into a volume the size of the Handbook for Physics and Chemistry, but even if this data required a six-foot shelf, its utility would be far greater than the over-rated six-foot shelf of Harvard Classics.

One of the above bibliographic aids (item 7) is partially illustrated in a brief appendix. The material is now out of date, and only three of the descriptors represented complete data at the time these examples were prepared. All of the data

were compiled by hand: Quite obviously this is a machine-type job. This material provides some indication of the semantic cohesiveness of retrieval terminology even in a system where word definition was more talked about than adhered to. With data cleanup and a return to a controlled vocabulary, this semantic cohesiveness can be increased, and can be made to function as an aid to efficient machine retrieval.

## Appendix

Partial context sets are presented for eight different descriptors. The eight descriptors are given in upper case and in alphabetical order. The number to the left of each of the upper case descriptors represents the frequency of use associated with that term at the time this sample was prepared. The number immediately to the right of the upper case descriptors is the number of usages sampled to prepare the data. Only three terms were investigated in complete detail: Electrolytes, Electrolytes (Physiology), and Stars. Indented beneath each of the eight descriptors is a listing of the ten terms most frequently used with the referenced descriptors. This listing is by order of magnitude, beginning with the largest, and the number following each word in the partial context set is the number of times that this word occurred in the sample.

(216) ACETYLENES (76)

Liquid rocket propellants (10); Combustion (9); Synthesis (9); Oxygen (8); Polymers (8); Solid rocket propellants (8); Decomposition (7); Flames (7); Propellant properties (6); Rocket fuels (6).

(253) ALASKA (62)

Measurement (6); Meteorological data (6); Canada (5); Glaciers (5); Temperature (5); Weather forecasting (5); Arctic regions (4); Aurorae (4); Climatic factors (3); Geological survey (3).

(198) AZIDES (100)

Lead compounds (28); Explosives (16); Crystals (13); Synthesis (12); Chemical reactions (10); Sensitivity (10); Crystal structure (9); Decomposition (8); Detonation (8); Organic azides (8).

(306) ELECTROLYTES (306)

Electrochemistry (67); Design (57); Electrodes (51); Conductivity (44); Solutions (37); Storage batteries (32); Temperature (31); Tests (30); Electrolytic cells (29); Polymers (26).

(064) ELECTROLYTES (PHYSIOLOGY) (064)

Metabolism (19); Physiology (18); Body fluids (15); Excretion (15); Kidneys (15); Water (14); Potassium (13); Sodium (12); Pathology (11); Therapy (10).

(1976) JET PLANES (1000)

Jet fighters (495); Fighters (342); Jet bombers (251); Flight testing (247); Design (195); Bombers (191); Tests (187); Airborne (133); Great Britain (127); Stability (108).

(943) MOLECULAR STRUCTURE (381)

Chemical reactions (52); Polymers (32); Synthesis (30); Theory (23); Temperature (21); Boron compounds (20); Infrared spectroscopy (20); Crystal structure (19); Methyl radicals (17); Molecules (17).

(126) STARS (126)

Measurement (22); Spectrographic analysis (19); Detection (15); Sky brightness (15); Atmosphere (13); Light transmission (13); Light (12); Radio astronomy (11); Turbulence (11); Instrumentation (10).

It may be of some interest to note that of the 80 words listed in these partial context sets, 68 occur only once. Nine words occur more than once as follows:

| | | |
|---|---|---|
| 2203 | Polymers | 3 |
| 2821 | Synthesis | 3 |
| 6474 | Temperature | 3 |
| 2450 | Chemical reactions | 2 |
| 1060 | Crystal structure | 2 |
| 882 | Decomposition | 2 |
| 37 774 | Design | 2 |
| 7886 | Measurement | 2 |
| 28 141 | Tests | 2 |

The number on the left refers to the frequency of use of these terms in the ASTIA collection at the time this data were prepared. The number on the right shows the number of times this word occurs in our partial context sets. These nine words are among the 300 most frequently used words of the 7000 available for use in the Thesaurus of ASTIA Descriptors, and each represents a quite general concept. In this small sample, eight out of ten most frequently associated terms are highly specific to the referenced descriptor.

## Bibliography


1. Uniterm Indexing Principles, Problems and Solutions
   J. C. Costello, Jr.
   American Documentation, Jan. 1961

2. Nonconventional Technical Information Systems in Current Use
   National Science Foundation, 1959

3. On Relevance, Probabilistic Indexing and Information Retrieval
   M. E. Maron and J. L. Kuhns
   Journal, Association for Computing Machinery, 1960

4. Concepts of Automatic Data Storage and Retrieval in the Simplex System
   MIT, AD-245 472

5. Research on Various Phenomena for the Performance of Circuit Functions
   WADC, AD-257 864

6. A Study of Restricted Random Walk
   University of Maryland, 1957, AD-136 722

7. The Relationship of Information-Use Studies and the Design of Information Storage and Retrieval Systems
   S. Herner, 1958, AD-213 781


Prepared 1 August 1961

## Epilogue*

I am not one who believes that it is any necessary virtue in the philosopher to spend his life defending a consistent position. It is surely a kind of spiritual pride to refrain from thinking out loud and to be unwilling to let a thesis appear in print until you are prepared to champion it to the death. Philosophy, like science, is a social function, for a man cannot think rightly alone, and the philosopher must publish his thought as much to learn from criticism as to contribute to the sum of wisdom. If, then, I sometimes make statements in an authoritative and dogmatic manner, it is for the sake of clarity rather than from the desire to pose as an oracle.

---

*A. W. Watts, Supreme Identity, Noonday Press, New York, 1957.