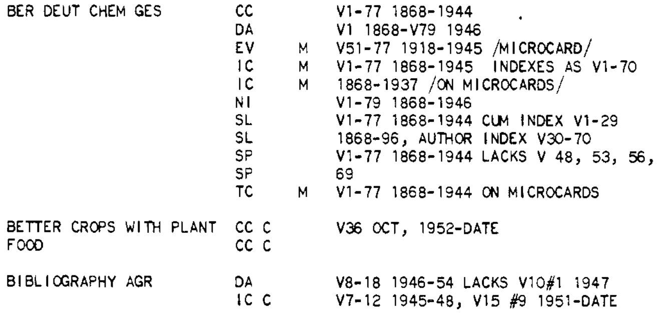

```
BER DEUT CHEM GES         CC          V1-77 1868-1944
                          DA          V1 1868-V79 1946
                          EV     M    V51-77 1918-1945 /MICROCARD/
                          IC     M    V1-77 1868-1945  INDEXES AS V1-70
                          IC     M    1868-1937 /ON MICROCARDS/
                          NI          V1-79 1868-1946
                          SL          V1-77 1868-1944 CUM INDEX V1-29
                          SL          1868-96, AUTHOR INDEX V30-70
                          SP          V1-77 1868-1944 LACKS V 48, 53, 56,
                          SP          69
                          TC     M    V1-77 1868-1944 ON MICROCARDS

BETTER CROPS WITH PLANT   CC C        V36 OCT, 1952-DATE
FOOD                      CC C

BIBLIOGRAPHY AGR          DA          V8-18 1946-54 LACKS V10#1 1947
                          IC C        V7-12 1945-48, V15 #9 1951-DATE
```

Fig. 4.—Typical excerpt from the Union List of Monsanto Serials.

the holdings are on microfilm or microcards. As an appendix to this List of Serials, there is a listing of the major reference works, particularly those that are brought up-to-date by regular reissues. By use of the Document Writer and an IBM 407 printer, this List of Serials is updated occasionally, for wide distribution within the Company.

Conclusion.—By use of modern methods of information handling, we have been able to offer more service to our patrons at considerably less cost. At the present time, we are employing computers and related information-handling machinery to prepare catalogs which are easy to use and have wide distribution. The act of information retrieval is still being done by hand, but, by improved make-up of the appropriate catalogs, it has been made easier and information may be sought in greater depth. As the operations of the Information Center grow, we will be in a position to use machine retrieval as soon as this becomes more economical than retrieval by hand.

Extensive Relations as the Necessary Condition for the Significance of "Thesauri" for Mechanized Indexing*

By MORTIMER TAUBE
Documentation Inc., Bethesda 14, Md.

Received February 20, 1963

I

Recently there have appeared a number of attempts to develop "thesauri" for use with mechanized information systems. H. P. Luhn, in discussing the possibility of auto-indexing, suggested that words could be arranged in natural families and noted the parallel between such families and the arrangement of Roget's *Thesaurus*.[1] One of the mechanical translation groups in England actually attempted to use Roget to supply contextual definition of words.[2] The Armed Services Technical Information Agency group, partly to distinguish the results of its work from the "ASTIA Subject-Headings" from which it started, called its product a "Thesaurus of ASTIA Descriptors." The name was also justifield because a grouping of descriptors into 292 descriptor groups was provided, and this grouping suggests the grouping of words in Roget. The Integrated Engineering Control Group of the du Pont Engineering Department also constructed a thesaurus, which used as raw material a "word association matrix" based on certain suggested procedures developed by Documentation Incorporated. This thesaurus, as finally published by the American Institute of Chemical Engineers, is based only incidentally on the "word association matrix" and primarily on relations of terms determined by consulting dictionaries, handbooks, and usage. In the A.I.Ch.E. thesaurus no over-all grouping of terms is provided, but under each term there are displayed "related terms" and terms which in some unspecified sense are "generic to" and "included in" the heading term. In a recent paper,[3] Vickery has taken sharp exception to the use of the term "thesaurus" to describe authority lists, vocabularies, or subject-heading guides, and to the implication that a new name creates a new type of apparatus. However, the ASTIA organization and several other organizations and individuals have continued to insist that in "thesauri" there are provided structures of connections between words which are necessary to ensure satisfactory operation of mechanized systems of coordinate indexing. It is assumed that the connections of terms and the cross reference structures displayed in a thesaurus are necessary both as indexing and searching aids.

Let it be admitted that a structure of terms is a *sine qua non* of a complete indexing system. There remains to be determined whether such a structure is derivable from and reflective of the indexing operation or whether it can possess some sort of independent validity which is prescriptive, rather than descriptive. In other words, is a thesaurus or any authority list an independent semantic standard for an indexing system set up by a process of lexicography, or is it a description of a particular indexing system as developed from a concatenation of subject competence, the literature being indexed, and the requirements for efficient machine search?

In the history of the subject organization of scientific publications there have been many attempts to establish prescriptive systems—the Dewey Decimal System, the Universal Decimal System, the Library of Congress Classification System, the Bliss Classification System, the

* Presented before the Division of Chemical Literature, 142nd National Meeting of the American Chemical Society, Atlantic City, N. J., September, 1962.

Colon Classification System, the Storey Classification System, the Standard Aeronautical Indexing System, and more recently, the Vocabulary for ARDD Technical Efforts (CATE) and the A.I.Ch.E. Thesaurus. Although some of these systems still have supporters, none of them has been accepted as *the* authority. Further, anyone who aspires to create a new Decalogue must certainly reject all others and thus takes on the burden of explaining why any such new authority will succeed in the face of his own rejection of all other authorities.

Actually, no one is truly comfortable in the role of Decalogue-maker, not even Moses. And perhaps the burden is assumed unwillingly by those who feel that mechanized information searching needs a cross reference structure which only an authoritative, semantically based, prescriptive thesaurus can supply. If this supposition constitutes a correct diagnosis, then the cure is not too difficult to achieve. It is only necessary to point out that a cross reference structure of any complexity can be derived from, and need not be prescribed for, any given indexing system. Further, if as developed in section II, a descriptive structure contains cross references which reflect the extensive relation of classes of items rather than prescriptive semantic relations of terms, there is no real problem of compatibility between descriptive systems. The following section of this paper sets forth the principles of such a descriptive structure for mechanized systems.

II

Class Relations

1.0 In an indexing system, each term or phrase in the system is a class.

1.1 The members of a class (term) are the items indexed by or posted under the term.

1.2 Terms can be nouns, adjectives, gerundives, participles, singulars or plurals, verbs, etc.

1.3 In any growing system it is reasonable to consider as members of a class items which are to be indexed (posted) under a term. This expectation is based on the subject competence of the indexer and his knowledge of the literature and vocabulary of the field being indexed.

1.4 Whether or not different forms of a word are treated as one term or several terms is determined by indexing conventions and is reflected by gathering all items in a single array under a grouping of forms or by dividing items under forms selected to be used separately.

2.0 There are no relations among the terms which are not reflected in a relationship of the items.

2.1 This means that all terms are interpreted extensionally as classes and not intensionally as properties or meanings.

2.2 The product, union, and complement of any terms (classes) in the system are also classes.

2.3 "Airplanes $\cap$ Wheels" constitutes a class;
"Airplanes $\cap$ Fuels" constitutes a class;
"Airplanes $\cap$ Astronauts" constitutes a class (in the last case, it may be assumed to be a class without members).

2.4 The basic philosophy is expressed in the following passages from Quine: "Once classes are freed thus of any deceptive hint of tangibility, there is little reason to distinguish them from *properties*. It matters little whether we read '$x \epsilon y$' as 'x is a member of the class y' or 'x has the property y'. If there is any difference between classes and properties, it is merely this: classes are the same when their members are the same, whereas it is not universally conceded that properties are the same when possessed by the same objects. The class of all marine mammals living in 1940 might be regarded as differing from the property of being a whale or porpoise alive in 1940. But classes may be thought of as properties if the latter notion is so qualified that properties become identical when their instances are identical. Classes may be thought of as properties in abstraction from any differences which are not reflected in differences of instances. For mathematics certainly, and perhaps for discourse generally, there is no need of countenancing properties in any other sense."[4] "Our working ontology is thus pretty liberal. But in mitigation it may now be said that this is the end; no abstract objects other than classes are needed—no relations, functions, numbers, etc., except insofar as these are construed simply as classes. In addition to concrete objects we need recognize only classes having such objects as members, then classes whose members are drawn from the thus supplemented totality, and so on. This is presumably all the ontology that is needed for discourse in general; certainly it is all that is needed for mathematics."[5]

3.0 The class relationships considered are equality, inclusion, and partial inclusion or intersection.

3.1 If all the members of Class A are also members of Class B, then Class A is *included* in Class B.

3.2 "A is included in B and B is included in A" is equivalent to "A = B."

3.3 If some members of A are members of B, the same members of B are also members of A and the classes are said to intersect, or to have a product which is not null.

3.4 The class which is the complement of A is the class which contains as members all items in the universe (collection) not members of A.

3.5 In retrieval operations, the complement is never used except in a product statement, *i.e.*, "$A \cap -B$" (the A's which are not B's).

Cross References

4.0 "See" references are used from a class which has no postings or members to refer to a class which has members.

4.1 In some cases, the "see" reference indicates that the classes so connected are equal; hence, it is only necessary to post physically under one of them.

4.2 In some cases, "see" references are used when the number of members of a class is so small that the class is included in another class, and no postings are made in the smaller class.

4.3 The distinction between 4.1 and 4.2 is sometimes expressed as the difference between "see" and "see under."

4.4 In some systems, it may be desirable to refer from a class term that has so many members that it ceases to be a practical retrieval point to classes that would ordinarily be included in it, *e.g.*, "Computers, *see* digital computers, analog computers."

4.5 The basic point is that "see" references are used from terms which have no postings to those which have postings, whatever the reason for the lack of postings in one case and the presence of postings in the other.

4.6 The number of such "see" references used is determined by the indexer's determination of the degree of redundancy required in a particular discipline or with reference to a particular collection. There is no systematic way to establish this number short of all the English language. The indexer will be guided by his background, knowledge, available dictionaries, and continuing usage in the material indexed.

5.0 "See also" references are used between terms, each one of which has members in the system.

5.1 In noncoordinate (or nonmanipulative) systems, there are usually two different relationships between classes connected by "see also" references and these relationships may not be distinguished. "See also" references are used in such systems to refer from a larger class to an included class (what librarians call a reference from a general to a specific class) and to refer between classes that intersect, *i.e.*, have some members in common.

5.2 In systems of coordinate indexing, in which searches are performed by operating on classes, it is important to distinguish between these two uses in order to perform those machine operations necessary and to avoid useless operations.

6.0 When reference is made from a class to a class it includes, use the expression "includes posting for."

6.1 When reference is made from a class which includes it, use the expression "also posted on."

6.2 In order to avoid excessive multiple posting among a series of included classes, the instruction to perform a summing operation can be substituted for multiple posting. The instruction in such cases might be: "Airplanes (for this class also sum Bombers, Fighters, etc.)."

6.3 The fact that the type of reference used must vary depending on how the items are posted indicates that the reference relates the classes and items extensively and is not dependent on the intension or semantic relation of the terms.

6.4 It is always possible to trade off multiple posting against summing operations as determined by statistics of use.

7.0 Whether or not the standard type of "see also" reference should be used between indirectly related classes depends on the degree of overlap between such classes and one or more additional classes.

7.1 A "see also" instruction in a mechanized system is an instruction to make a logical sum of the classes so related. This operation is significant only in the event that there is at least one other class whose intersection with the summed classes has more members than its intersection with the classes which are involved in the initial summing operation.

7.2 It has been the experience of large library systems that "see also" references established on the basis of anticipation tend to proliferate needlessly and complicate both the indexing and search processes.

7.3 In a mechanized system in which periodic checks of degree of overlap can be run, "see also" references should be based mainly on such checks.

Scope Notes

8.0 The use of scope notes indicates a decision of the indexer concerning the type of items included in the class having the scope note.

8.1 The number of scope notes to use, as in the case of "see" references, is a matter of judgment. All words in the English language having more than one meaning are, in a sense, homographs. Hence, the direction "Use scope notes for homographs" does not remove the element of judgment.

Authority Lists

9.0 The design and display of terms in an authority list or "thesaurus" to be used for both machine searching and as a guide to a published index remain to be determined. Experience may indicate that the two purposes cannot be combined and different code books or "thesauri" may be necessary for the mechanized retrieval portion and the published portion of an information center's apparatus. It also may turn out to be the case that different types of guides and instructions are required for the indexer and searcher.

III

Conclusion.—This derivation of a cross reference structure from actual and anticipated class relations indicates the proper method to be pursued for the present revision of the ASTIA Thesaurus. The terms and postings which have been actually put down in the indexing operation since the appearance of the first edition of the ASTIA Thesaurus should be printed out. From a study of this print-out there can be derived all necessary "see" references and references to indicate the inclusion of one class in another.

For studying the relations which are basic to "see also" references, it is necessary to have a print-out of the tracings. Since the checking procedure in this instance is fairly complex, the nature of "see also" relations should be discussed again in connection with this account of how they can be established.

If any two terms have identical postings, a "see also" reference from one to the other is redundant and wasteful. In such a case, one set of postings should be eliminated and a "see" reference substituted for the "see also" reference. In actual fact, the fewer postings two terms have in common, the more effective will be a "see also" reference between them, provided only that there are other terms which have postings massively related to both the original terms.

Consider the logical expression, $(A \cup B) \cap C$. If A and B have most of their members in common, their union will not produce a significantly larger class to intersect with C. This fact is reflected in the first requirement that the smaller the overlap between A and B, the more effective will be a "see also" reference between them.

But even if A and B have a minimum or no common members, for the "see also" reference to be significant, the class C must have members in common with both A and B. If the membership of $[(A \cup B) \cap C]$ is not significantly larger than the membership of $A \cap C$ or $B \cap C$ individually, again the "see also" reference between A and B is nonsignificant. This fact is reflected in the second requirement that two terms connected by a "see also" reference must both be massively related to one or more additional terms.

Hence, it follows that the validity of a "see also" reference cannot be established by comparing only the postings on two terms. There is required in addition an examination of the tracings represented by the postings to discover whether or not there exist in the system additional terms massively related to the terms connected by the "see also" cross reference.

Because the extensive interpretation of "see also" references and the checking required to establish their significance seem excessively complex and laborious, it may be urged that "see also" references be interpreted intensively and used between "related meanings." "Related meanings" not reflected in postings would not be a significant basis for a searching directive, but it might be said that "related meanings" or related terms displayed to the indexer are an aid to "consistent" indexing.

Suppose an indexer selects a term for a particular document from a thesaurus. Suppose further the thesaurus displays to him "related terms" in the form of "see also" references. How should he utilize this information? He may decide that one of the "related terms" better expresses the meaning of the document than the original term he had chosen. In such a case, he would presumably not use the original term and would use the "related term." But then it seems that the reference should be a "see" reference and not a "see also" reference. But suppose that on occasion the original term turns out, after inspection, to be more suitable than the related term. Then it would appear that the "see" reference should be from the related term to the original term; or, rather, that the terms should be combined and assigned a single code designation in the machine.

If it is now urged that "related terms" sometimes mean the same thing and sometimes mean different things, this is tantamount to saying that they are synonyms, since Fowler is our authority that there are no absolute synonyms in the English language except "gorse and furze." Again, it is usually agreed that the "see" reference, rather than the "see also" reference, is the proper reference between synonyms.

Finally, it may be urged that the indexer notes that two terms A and B are related because although they are not both used for the same document, there will be a group of documents in the collection indexed by A and a set of terms and another group indexed by B and the same set. But this reduces the "see also" reference to the type of extensive relationship of postings set forth previously.

Even authority lists for manual catalogs have difficulty with "see also" references unless they are restricted to references from general to more specific classes; and what emerges from this analysis is the thesis that the notion of "related term" as an explanation of "see also" references is too vague to be significant either to the indexer or the searcher in a mechanized system. It appears that the type of extensive relationship of classes discussed above, namely, $(A\cup B)\cap C > A\cap C$ and $(A\cup B)\cap C > B\cap C$, is at least a necessary condition for the significance of "see also" references.

REFERENCES

(1) H. P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM J. Res. Dev.*, 1, No. 4, October, 1957.

(2) M. Masterman, R. M. Needham, and K. Sparck Jones "The Analogy between Mechanical Translations and Library Retrieval," *Proc. Intern. Conf. on Sci. Inform.*, **2**, Washington National Academy of Sciences, National Research Council (1959).

(3) B. C. Vickery, "Thesaurus—A New Word in Documentation," *J. Doc.*, **16**, No. 4, December, 1960.

(4) W. V. O. Quine, *Mathematical Logic*, Harvard University Press, Cambridge, Mass., 1951, pp. 120–121.

(5) W. V. O. Quine, *ibid.*, pp. 121–122.