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

DIVISION OF CHEMICAL LITERATURE
AMERICAN CHEMICAL SOCIETY
CINCINNATI, OHIO
JANUARY 14, 1963

COMPILING A TECHNICAL THESAURUS

T. L. Gillum
Armed Services Technical Information Agency
Arlington Hall Station
Arlington 12, Virginia

ABSTRACT

A thesaurus is defined as a device for controlling and displaying an indexing vocabulary. The vocabulary is controlled in the sense that the individual terms are carefully chosen and appear as distinct, though not inseparable, entities; it is displayed in such a way as to provide ready access from a given entry to related words that may be needed to index a paper or phrase a search question. Some factors that may be considered in formulating the vocabulary for the thesaurus are discussed in terms of experience in compiling The Thesaurus of ASTIA Descriptors. Among the basic criteria to be examined are (1) the volume and scope, in terms of scientific disciplines involved, of the collection to be indexed, (2) the technical competence of those indexing and retrieving the information, (3) the resources, including time, money, and equipment, to be devoted to the retrieval system as a whole and (4) the demands, in terms specificity of information, and speed of service that must be satisfied by the system. Problems that must be dealt with in evaluating the individual terms include (1) synonyms, (2) homographs, (3) generic relationships, (4) logical relationships and (5) proper degree of specificity. Effective display of the vocabulary encompasses format of cross references, methods of delimiting meanings of terms, and graphic displays of related terms.

INTRODUCTION

The development of mechanical methods of storing and retrieving scientific information has brought about an evolution in the form of the traditional indexing authority list. One of the new methods of organizing an indexing vocabulary for efficient use by indexers and searchers has been called the thesaurus concept. A thesaurus is a device for controlling and displaying an indexing vocabulary. The vocabulary is controlled in the sense that it is prescriptive, the individual terms being carefully chosen and appearing as distinct, though interrelated entities. The vocabulary is displayed in such a way that the user may select all appropriate terms rapidly and accurately.

In the following discussion an attempt will be made to explain how control and display of an indexing vocabulary can most effectively be attained. No effort will be made to compare or contrast systems or vocabularies that use the thesaurus concept with those that use other approaches to indexing. Neither will any attempt be made to defend or condemn the use of the word "thesaurus" in this connotation although there has not been unanimous acceptance of this usage. Furthermore, in this discussion the words "indexing terms" or "terms" will be used to mean words or combinations of words that resemble as nearly as possible the scientific terminology that is normally encountered in the literature. This is done with the knowledge that indexing can be and often is performed using single words only, phrases or combinations of several words, or vocabularies in which words or word combinations are arbitrarily defined to represent concepts that may not coincide with popular interpretation.

CONSIDERATION OF THE RETRIEVAL SYSTEM

Since the thesaurus or vocabulary authority is only one facet of a

retrieval system, consideration must be given to several variable factors that will influence future decisions concerning the vocabulary. Among these variables are the volume and scope of the collection, the users of the system, the resources to be devoted to the system, and the results required.

There probably exists some minimum collection size below which the use of machine search methods is superfluous. This size is not known but it is believed to be in the range 5,000 to 10,000 documents. Beyond this, the collection size will dictate to some extent the vocabulary size and the specificity of terminology required since densely posted terms are apt to give voluminous search results. The scope of the collection, in terms of scientific disciplines covered, presents a more complex problem. Where many disciplines are dealt with, ambiguities in terminology are frequent and must be reconciled individually. Moreover, the encompassing of several disciplines drastically increases the size of the vocabulary, making necessary special devices for organizing and displaying the terms. When a single discipline is treated, the problems of ambiguity and of displaying the vocabulary are lessened. However, there seems to be a marked tendency to expect greater reliability from such systems, so effort is made to incorporate more specific and more abstract concepts into the vocabulary. This, in turn, creates many problems in the selection and evaluation of terms.

Some consideration must be given to the potential number of users of the vocabulary over a period of several years, their technical backgrounds, and even their geographical locations. The development of an aid for a small group of indexers who will index a large accumulation of reports will present different problems with respect to vocabulary control than would

the compilation of a thesaurus to be used by several hundred indexers to process current literature over a period of several years. A thesaurus compiled for the use of subject specialists would obviously require less explanatory material for the various terms than for an internal indexing tool.

The resources, i.e., time, money, equipment, and manpower, that are available to operate the system will impose limitations not only upon actual thesaurus construction and indexing procedures but also upon such allied activities as testing, experimentation, and research. Refinements and sophistications that may be desirable from a retrieval point of view often are too costly in terms of time and manpower. Also, machine capability or availability may create obstacles.

A careful study must be made, and a clear understanding reached, of the requirements of the retrieval system, including the specificity of retrieval desired, and the amount of noise that can be tolerated. The advantages and limitations of various indexing schemes must be weighed against these requirements to determine an optimum approach to the vocabulary problem.

COMPILING THE THESAURUS

In compiling the thesaurus, the first step is to evaluate any aids that may have been used previously in indexing material of a similar nature, or, if none exists, prepare lists of terms taken from the material to be indexed or sample search questions if any are available. The terms should be chosen on the basis of (1) frequency of appearance, (2) anticipated value in retrieval, and (3) acceptability as scientific terminology. These three criteria were discussed at some length in a previous paper so will not be dwelt upon here. Frequency of appearance gives quantitative evidence

of a term's utility. This is weighed against considerations of utility founded on examination of the concept conveyed by the term, i.e., whether or not it unambigiously conveys a concept concrete enough to be easily recognized and consistently applied by indexers. Acceptance by others of the term as evidenced by its appearance in papers by other authors, dictionaries, etc. will help ensure that the term will appear consistently in the material to be indexed and more easily understood by users of the thesaurus.

When a list of prospective terms has been assembled each term must be further examined to determine its relationships to the others on the basis of (1) synonomy, (2) homonomy, (3) genericity, and (4) logical associations. These relationships are displayed by a system of cross references.

Where exact synomyms occur, it is desirable to indicate a preference, as is done in standard library practice, e.g., Columbium use Niobium. Inclusion of a specific concept within the meaning of a more general term may be shown in the same way, e.g., Valine use Amino acids. Further, where it becomes necessary to group concepts under one indexing term to insure consistent usage, an effective synonomy is imposed, e.g., Electrical resistance use Electrical conductivity.

In these cases the reverse reference should also be made to aid in defining th scope of the term. Thus the references Niobium includes Columbium, Amino acids includes Valine, and Electrical conductivity includes Electrical resistance. The choice of words in these and all other cross references is an arbitrary one. Any terse expression, when used consistently, is not likely to lead to confusion.

In cases where two or more terms are used to represent closely related concepts which are to be distinguished in indexing, a parenthetical expression or brief explanatory statement may be helpful. For example, to dis-

tinguish diodes utilizing semiconductor devices from those that take the form of electron tubes, two terms, Diodes (semiconductor) and Diodes (Electron tube) can be established, or where both Radiochemistry and Radiation chemistry are used for indexing, the distinguishing features of the two concepts should be explained.

Terms having more than one commonly accepted meaning should be accompanied by definitions to the extent that such practice is consistent with the degree of specificity expected of the system.

Homonyms, words whose meanings change completely when applied to different disciplines, must be distinguished when used as indexing terms. The most common means of distinction is the use of parenthetical expressions, e.g. Cells (Biology) and Cells (Electrolytic). In multidiscipline vocabularies it may be desirable to exhibit cross references from the homonym to the prescribed indexing term e.g. Corona (Electrical) use Electrical corona and Corona (Solar) use Solar corona. The unqualified homonym is rarely a reliable indexing term in a multidiscipline vocabulary. Some means should be taken to insure that the intended meaning cannot be mistaken by the user.

In any vocabulary of reasonable size there will exist generic relationships among terms. That is, some terms will be used that represent general concepts which, by definition, entirely include the concepts represented by one or more other terms. This situation is easily recognized -- the problems arise in the treatment of these relationships in compiling the vocabulary and in indexing.

For example, if an indexing vocabulary contains as bona fide terms Chlorine, Halogens, and Chemical elements, the generic relationships among these are immediately evident to one trained in the discipline; the treatment hinges mostly upon the requirements of the system.

It seems most desirable to indicate generic relationships by cross references. For a given term, such references show what, if any, more general or more specific terms are available in the vocabulary. This device would presumably show a user who had entered the vocabulary on the basis of some one word whether or not he had attained the proper degree of specificity for his purposes. Some criteria must be established to prescribe the number of levels to be displayed for a given generic family. This may vary widely among different families in a multidiscipline vocabulary.

The analysis of generic relationships should be approached after the terminology has been chosen so that the situation is one in which relationships that exist in a carefully prescribed vocabulary are recognized and displayed and not one in which terms are evaluated on the basis of their compatibility with a hierarchical structure. Moreover, care should be taken not to distort concepts represented by a well-established terms merely to allow their inclusion in a generic family. Nor should indexing viewpoint replace basic definitions of terms in determining generic relationships, e.g. no true generic relationships exist between materials and their applications or properties. Since terms are chosen primarily on the basis of their value in the vocabulary it is to be expected that most generic trees will not be symmetrical and many terms will have no generic relationships to any other. There is no reason to feel that such a situation is, of itself, indicative of deficiencies in the vocabulary. However, great care should be taken to indicate all defined relationships. It may be necessary to define some very general terms in such a way that they are excluded from what appears to be an acceptable hierarchy. For example, if Chemical elements is a term that must be used for indexing in extreme

cases, it can be defined in such a way that it need not be included as a reference from each specific element nor need all the elements be listed at the entry Chemical elements. Defined generic relationships can, when analyzed properly, indicate where ambiguous or overlapping terms exist or where terms that are too specific or too general in scope have been established.

Indications of logical associations among terms are of great assistance to users of a vocabulary but are very difficult to evaluate consistently, hence are usually established more or less intuitively. The logical relationships take the form of See also or some equivalent cross reference and are commonly found in vocabulary listings. No rules can be set forth for the establishment of logical relationships, but a thorough knowledge of both the vocabulary and the material to be indexed permits generally acceptable choices of references.

In some vocabularies, particularly those of several thousand terms encompassing several disciplines, it may be necessary to devise additional display media. These may take the form of subject classifications of terms, displays of generically or etymologically related terms, specialized thesauri, etc. These should be designed to fit the needs of the retrieval system and, as with any vocabulary tool, proficiency of design and use will improve with experience and experimentation.

MAINTENANCE OF THE VOCABULARY

In any thesaurus that is designed for indexing current literature, some provision must be made for continuous refinement. New terms will be added to the vocabulary and some means must be available for reindexing documents to reflect these changes. Moreover, experience in indexing and retrieval will indicate that some concepts should be redefined or that

some terms should be abandoned entirely. Refinements can be readily made in a machine retrieval vocabulary, but the need for this capability must be recognized at the outset and resources set aside for this purpose.

The requirement for new terms will become evident over a period of time as accessions are indexed and search questions processed. Proposed terms should be evaluated as carefully as were the terms in the original list. Great care should be taken to insure that new terms represent concepts that are concrete and meaningful enough to permit consistent indexing over a period of time. The use of prestige words and cliches should be avoided. If possible, some mechanical means should be devised to permit the assignment of new terms on a trial basis so a more accurate evaluation can be made of their usefulness and acceptability. Human recollection or intuition alone are rarely reliable.

In cases where valuable concepts have been overlooked in compiling the vocabulary but are the object of searches, provision should be made, and resources set aside, for establishment of appropriate terminology in the thesaurus, again subject to the evaluation criteria set forth, and the required reindexing. Users of a thesaurus must realize, however, that no matter what precautions are taken, there will invariably be instances in which queries cannot be phrased precisely or even approximately with the thesaurus vocabulary. This is a characteristic inherent to the subject indexing concept and is not necessarily indicative of shortcomings of the system.

Mechanized systems provide a variety of means for evaluating terminology, indexing consistency, retrieval accuracy, etc. Statistics can and should be kept to indicate the frequency of use of terms and the associations among various terms in indexing and searching, the number of relevant

and irrelevant responses to queries, the number of terms used in indexing and in formulating searches, etc. The compilation and analysis of these data represent another requirement for resources but such research is necessary in developing an efficient system.

The optimum parameter for the density of posting of a given term will depend upon the variables of the system. However, if a general assessment of this problem can be made, consider that an optimum parameter in a collection of several thousand documents is 25 to 500 postings. In other words, terms posted about 25 or less times in indexing several thousand documents probably represent concepts too specific for efficient machine retrieval. If their relationships with other terms permit, they should be combined with a more general or closely related term to create a concept more suited to machine methods. Any term that has been posted more than 500 times, even in a very large collection, is apt to be of too general or vague a nature to give suitable results in retrieval. Moreover, searching or coordination of such terms is awkward and time-consuming even by machine. Heavily posted terms should be broken down into more specific terms, redefined to limit their use in indexing, or abandoned entirely.

Words that appear together often in indexing or in searching should be examined to determine whether or not there is overlapping in the concepts represented, failure of users to distinguish among terms, or simply synonymous terminology. If overlapping exists, the terms in question should be examined to ascertain the difference in the concepts represented and the need for expressing the separate concepts. This may lead to the combination of terms or the redefining of concepts. If users fail to distinguish terms that actually represent separate and necessary concepts, definitions or indexing rules should be displayed. If synonyms exist,

a preference should be indicated. There exists the additional situation in which two concepts can be combined to give a useful concept, e.g. if the two terms Crystals and Growth appear often in combination, perhaps the term Crystal growth should be established.

Consistent association of a given term with two groups of terms, each representing an entirely different discipline, may indicate that a homograph exists and that separate terms should be established. For example, if the term Precipitation were frequently associated with such terms as Climate, Clouds, Temperature, Humidity, Solutions, Chemical reactions, and Solubility it is evident that two separate concepts are being indexed as one, therefore two terms should be established or the one term redefined.

Statistics should not take precedence over human judgment in the evaluation of vocabulary but these studies and others provide the basis for some useful decisions.

The building and maintenance of a thesaurus represents a tremendous investment of resources, but the successful use of a retrieval system employing even the most perfectly developed vocabulary is contingent upon diligent study and proper understanding of the art and practice of indexing. Users of the system must be made to realize the complexity of the indexing situation; indexers themselves must become students of their art and be aware of their part in the operation of the retrieval system. Indexing for even the simplest system cannot be performed satisfactorily by untrained, unskilled, or unconcerned persons.

Many discussions, arguments, theories, and dissertations have been put forth in an effort to determine what good indexing is. No attempt will

be made to enlarge upon this subject here except to reinterate, especially as regards the thesaurus concept, some of the things indexing is not.

Indexing is not the mere transcription of words from text to worksheet through the intermediary of the thesaurus. Furthermore, deep indexing or indexing in depth is not the mere transcription of a great many words or terms from text to worksheet. Quality, throughness, or completeness of indexing cannot be measured as a function of the number of terms assigned unless this number is viewed in the light of the other variables of the system.

The strategem of amassing a long list of terms with the theory that they "can't miss" conveying a desired concept is as totally unacceptable an indexing practice as it is a searching technique. The thesaurus cannot be considered more than an aid to the indexer and searcher; there is no substitute for a thorough knowledge of the requirements of the system, familiarity with the subject matter, and an appreciation of the finer points of the indexing and retrieval processes.

CONCLUSIONS

In conclusion, the thesaurus is viewed as an evolutionary development of the indexing authority list concept long employed by librarians. Thesauri are designed primarily to be used in conjunction with systems that utilize mechanical concept coordination indexing schemes. The terminology must be chosen with full cognizance of the requirements and limitations of the system, including sufficient provisions for anticipated growth. The vocabulary is displayed in such a way that the user can review rapidly all available indexing terms that may be useful and choose those most appropriate for his purposes. Resources, in terms of, money, manpower, time, and equipment, must be programmed realistically. The tremendous effort

required to build and maintain an information system is often underestimated.

The thesaurus concept, in the present state of its development, does not represent a panacea for documentation problems. Thesauri do, in general, reflect the efforts of documentalists toward resolving some of the complexities of semantics, syntax, and changing terminology that beset previous efforts in the field. The present state of the art reveals only the most heavy handed approaches to language problems and, in general, machine applications that are cumbersome and inefficient. The fact that several operational systems now employ the thesaurus concept of vocabulary control and that reports are indexed and searches performed every day with the aid of this tool gives reason for optimism as to its future position in documentation.

REFERENCES

1. Thesaurus of ASTIA Descriptors, Second Edition, December 1962.

2. Caponio, J. F. and T. L. Gillum, Practical Aspects Concerning the Development and Use of ASTIA's Thesaurus in Information Retrieval. Presented to the American Chemical Society Division of Chemical Literature, September 14, 1962.

3. Holm, B. E. and L. E. Rasmussen. Development of a Technical Thesaurus. American Documentation, 12 (1): 184 - 189.

4. Gillum, T. L., P. H. Klingbiel, C. N. Mooers, and Eugene Wall. The Philosophy and Guidelines for the Revision of the Thesaurus of ASTIA Descriptors. ASTIA November 1961.

5. Wall, Eugene. Final Report on the Revision of the Thesaurus of ASTIA Descriptors. Final Report. AF 18(600)1944, 6 August 1962 AD-278 168.

UNCLASSIFIED