Correlative Indexes. IX. Vocabulary Control

By CHARLES L. BERNIER
The Defense Documentation Center for Scientific and Technical Information
(formerly ASTIA), Cameron Station, Alexandria, Virginia
Received October 14, 1963

Vocabularies used in subject-indexing are controlled to reduce scattering of like information and to help indexers and searchers find words. Control is of:

(1) subject area
(2) cross references
(3) organization of terms
(4) definition and differentiation of terms
(5) number of indexing terms in vocabulary
(6) hospitality for new terms
(7) specificity of terms chosen for indexing
(8) semantic *vs.* statistical use and syntax of terms
(9) compatibility with other vocabularies
(10) choice of terms used to represent subjects indexed
(11) subject importance of the index terms chosen
(12) roles and links

Vocabulary terms are of two general types: (a) indexing—those used in the index (manual or mechanical), and (b) cross reference—those that guide the indexer, user, and machine to an indexing term.

Subject Area.—Control of the subject area of the vocabulary is determined by the scope of the collection of documents. This scope is usually not determined by the indexer. A preselected vocabulary of indexing terms is especially affected by subject area.

Cross References.—Relationships displayed among indexing and cross-reference terms give the principal forms of vocabulary control.[1] Synonymy is the most important relationship to be controlled. Synonyms are joined by "See" (or "Use") cross references. "**Dextrose**. See (or Use) *d-Glucose*." is an example. Indexes and vocabularies not having control of synonymy force users to compile lists of synonyms and to search under each of these as the only way of making a complete search. It is less costly and time-consuming for one indexer to provide cross-reference bridges between terms than to have all users do it for every search.

Terms opposite in meaning need cross references. Antonyms are usually joined by "See also" (or "Also see") cross references, *e.g.*, "**Fluidity**, (see also *Viscosity.*)." Often in a generic vocabulary, and sometimes in specific vocabularies, antonyms can be combined by "See" (or "Use") cross references. An example is: "**Thermal resistance**. See *Conductivity, thermal.*" Indexing of antonyms under the same index heading resembles the rhetorical trope, irony.[2,3]

Relationships other than those of synonyms and antonyms are also shown among vocabulary terms by means of "See Also" cross references. These relationships include: group, genus, or class relationship to kind, species, or subclass; part to whole; mathematical function; cause to effect; and product to use. Examples of cross references indicating, respectively, these kinds of relationships are: "**Glass**, (See also *Vitreous materials.*)"; "**Cement**, (See also *Concrete.*)"; "**Divisor**, (see also *Dividend.*)"; "**Vision**, (See also *Light.*)"; and "**Food**, (See also *Nutrition.*)." All of these relationships are examples of the definitive relationship, which helps to define an indexing term. The relationship of genus to species is the most common of all. Most specific terms have at least one term that is more generic; most generic terms have many that are more specific. For example, "Silver" is a kind (species or variety) of: chemical element, metal, specular reflection, coinage, tableware, coating process, commodity, alloy, gray, bucaneer, cation, etc. "Tree" is the "genus" for aspen, fir, oak, pine, etc. In an index built to the maximum specificity, only "See also" cross references are used between genus and species terms in the subject field of the index. For terms outside of the field, sometimes "See" cross references are used to refer from the more specific term to the more generic,[2] *e.g.*, "**Calculus**. See *Mathematics.*"

In trope indexes (*e.g.*, generic indexes),[2] the species is usually indexed as the genus and the part as the whole. "See" (or "Use") cross references may tie these terms together. Examples are "**Cobalt**. See *Transition elements.*" and "**Hearths**, use *Furnaces.*" In generic indexes there may also be "See also" cross references between generic terms and the next level of terms even more generic than these.

The very large number of "See also" relationships that could be shown in a thesaurus or published index makes limitation common. The user is expected to contribute enough relationships from his background so that most cross references from species to genus and from whole to part can be omitted. This is so because most informed users can recall the genus when the name of the species is seen, and can recall the parts when the name of the whole is displayed.

Organization and Displays.[4]—The hierarchical display of related terms is one form of organization of terms in a vocabulary. In the hierarchical display the species or part can be indented under the genus or whole, respectively.

Arrangements other than hierarchical are possible, *e.g.*, groups of related terms can be listed alphabetically rather than by meaning. Displays of terms arranged by meaning help the indexer and searcher to find terms more rapidly than do alphabetical lists and enable the user, in most

cases, to select the most appropriate generic term without the help of cross references.[5]

Alphabetical arrangement has the advantage of ease of location of terms provided that one can find the first term representing a subject. From this point on, the index (manual or mechanical) will help through cross references and displays.

Terms organized by frequency of use are valuable in lexicography, linguistic studies, and vocabulary- or thesaurus-building, but not usually in subject-indexing. Frequency-of-use lists of descriptors[6] also help in index-searching by machine so as to avoid blank sorts.

Differentiation of Indexing Terms.—Words, terms, and phrases that are used as index headings and descriptors are so indefinite that the context represented by them is largely lost. Definition, differentiation, and supplementation of indexing terms and descriptors are usually required. Homographs like "lead" can easily be differentiated by adding a term or phrase, and by use of cross references, *e.g.*, "**Lead** (electrical)"; "**Lead** (electrical). See *Electrode*."; and "**Lead metal**."

Scope notes[7] are used to differentiate descriptors most closely related in meaning. Vocabulary control in this way is usually necessary to increase precision and accuracy in selection of indexing terms by indexers and index users. For example, the memnonic label for the descriptor "**Food**," might be "**Food**," "**Nutriment**," "**Nutrition**," or "**Edibles**." The scope note differentiating the descriptor labeled "Food" from other closely related descriptors will specify whether foods for animals and bacteria are included or excluded, what kinds of human foods are included and which are excluded, whether "dietetic" foods are indexed by "**Food**," "**Pharmaceuticals**," or yet another descriptor. If precise scope notes are not used, then indexing subjects by descriptors as well as index searching may lose accuracy. If the indexer or index-user is uncertain as to which of several terms or descriptors to choose, he is usually instructed to use all.[8] This practice of "indexing when in doubt" leads to additional entries, and to "noise" or false drops in searches, since more irrelevant references will be recovered than had the descriptors been precisely defined and used. Control of vocabulary terms by scope notes and cross references can be tightened as experience is gained in use of the vocabulary. Ambiguity and uncertainty in selection of descriptors are warnings that scope notes are needed or that existing scope notes need revision.

Vocabulary Size.—Control of vocabulary size[9] is accomplished by: (1) coalescing closely related descriptors into a more generic one or one with a broader scope note and (2) by splitting a broad term into several more specific ones along with addition of new and more-specific scope notes. Frequency of assignment of a descriptor in indexing is used in vocabulary-size control.[6] If a term has been used very infrequently, then correlation of this term with others in searching will be inefficient because it would be as rapid to examine all of the references for relevance as to correlate.

Terms used to index too many documents in a collection can be narrowed in meaning. For example, the descriptor concept "**Aircraft**" can be split into "**Airplanes**," "**Balloons**," "**Dirigibles**," "**Helicopters**," and "**Ground-effect machines**," each with clearly defining scope notes, and the broader term "**Aircraft**" be reserved for general studies in which the subject covers several or all kinds of aircraft.

So far as we know now, there is no critical size for a vocabulary of generic descriptors. The reasons for holding the size to a minimum that still gives adequate selectivity are: (1) to speed finding of terms, (2) to reduce the number of scope notes, (3) to give more effective displays of related terms, (4) to reduce the number of changes as new knowledge is indexed, (5) to include more of the closely related documents in a search, (6) to reduce the number of cross references, and (7) to reduce the number of blank sorts[10] caused by the terms being too specific. What constitutes adequate selectivity is unmeasured and probably not critical either.

Hospitality.—A vocabulary can be controlled as to number of terms and yet be open-ended—that is, hospitable to needed, new indexing terms. As new concepts emerge and need indexing, new indexing terms and cross-reference terms may be needed.

A subject-heading list can be tightly controlled by cross references and yet be open-ended with no limitation in number of terms. Indexing by rule,[11] as is done in subject-indexing by *Chemical Abstracts*, rather than by terms chosen from an authority list can be controlled by changing the rules and by making new cross references and notes. This indexing is open-ended, *i.e.*, without limitation as to number of index-heading terms.

A new index term introduced must represent a new concept or one that has not been indexed before in the collection of documents. Some new concepts come in suddenly and complete with new terms. Examples are popular or "generic" names of new chemicals and substances, names of equipment, such as computers, and trade names in general. In these cases, revision of earlier indexing is unnecessary because the new terminology and concepts arrive at the same time and the first document in the collection on the new subject is indexed by the new term or descriptor.

Often the ultimately accepted terminology arrives later—sometimes years—after the concept. New concepts may be labeled in several ways until one term finally comes to be generally accepted. Examples of terms that came later than the concepts are: "radio," "Raman effect," "transistor," "programmed instruction," and "powder metallurgy." For these cases, the earlier indexing can often be revised so as to make it consistent with the later. If this is impracticable, the earlier collection of documents can be searched with the older vocabulary, and the later collection with the revised vocabulary. Collective indexes usually require revision of the earlier indexing. A second technique of harmonizing vocabularies is to build cross-reference bridges into the past.

Another kind of hospitality is accomodation to changes in meaning of terms. Examples of terms that have become ambiguous are: cell, plasma, inflammable, lead, heat, engine, ablation, petrochemical, bit, light, noise, and sound. Examples of obsolescent terms are: accumulators (storage batteries), wireless telephony, thermofragility, thermal repulsion, phlogiston, metal ceramics, infrared scattering by molecular vibrations, and magnetolysis. The more generic or broader terms used in trope indexing[2] usually have a lower rate of obsolescence than do the

more specific, narrow terms used in subject indexing to the maximum specificity. This property makes generic indexing especially attractive.

Indexing Specificity.—Indexing is done usually to the maximum specificity consonant with the subject(s) reported by the author. This holds for both generic and specific subject-indexing.[12] In generic indexing, the most specific of the generic terms is chosen. For example, if there is choice of using "**Carbohydrates**" or "**Hexoses,**" the latter is chosen as the more specific and accurate provided that the author reported studies of hexoses that did not apply to carbohydrates in general. Had the author reported work on carbohydrates in general, such as their mechanism of thermal decomposition, then the more general term would be chosen as index heading or descriptor to represent, in part, the report.

The use of generic terms in addition to the specific ones in generic indexing of a given document usually pushes the author beyond the scope of the subject that he reported. Accuracy in indexing can be attained by use of these more generic terms only when authorized as novel subjects by the author. It is a simple matter to tie genus-species terms together in machine indexes so that documents correctly indexed by specific terms can also be retrieved by use of the more generic terms.

If indexing is done to the maximum specificity and synonyms are controlled, there is little choice left to the indexer. This removal of choice improves accuracy and precision of indexing. However, it is always necessary that the indexer understand clearly, at least in a general way, the subject matter indexed.

Semantics *vs.* Statistics and Syntax.—If the indexer (human or machine) does not understand the subject matter indexed, then indexing is unintelligent. Indexing has been tried by syntax—the arrangement and relationships of words in groups—and (or) by statistics—the frequency (or infrequency) of terms in the text—and (or) by associations of words in requests, profiles, and documents. Indexing of words, with or without understanding the subject indexed, has been termed "word-indexing."[11]

Choice of indexing terms to represent effectively subjects based solely upon frequency or infrequency of use of terms in the document or field indexed is, at present, not adequately supported by experience or experiment. Relationships among microsemantics (semantics of units of communication smaller than sentences, such as clauses, phrases, words, prefixes, suffixes, symbols, punctuation marks, and bits) and macrosemantics (semantics of sentences and larger units of communication, including parts of paragraphs, paragraphs, pages, chapters, sections, letters, reports, papers, documents, books, collections, and libraries) have not been completely analyzed as yet. Macrosemantics is related to context, concept, gist, substance, purport, macro-meaning, subject, theme, inuendo, implication, indirection, hint, subtlety, irony, miscommunication, and, in general, "what it's all about." Microsemantics contends with problems of taking things out of context and loss of subjects among words. Just how a microsemantic unit can and does represent a macrosemantic unit, subject, or concept is the problem. This problem has been solved in a practical way by indexers, index users, telegram senders and receivers, headline writers, and classifiers. That symbols and arrangements of symbols can be used effectively to stand for things and for macrosemantic units is beyond doubt.

There is not so much information in an index entry or vocabulary term as in the document or part of a document that it represents. Because of the greater context and meaning of an index entry including heading and modification (modifying phrase) than of a term or word, the complete index entry serves more effectively as a guide to the information than does a single word or term, such as those used in manipulative systems.

It is usually not regarded an economical function of indexes to give information about the subject—only to lead the user to the subject.[11] Context controls subject-indexing.

Uniqueness or novelty of subjects might seem, at first thought, to be related in some way to infrequency of association of indexing terms. A frequency of zero should mean no subject. A frequency of one might mean maximum novelty. Subjects seem largely unrelated to word frequency in a document; this is especially true for the novel subjects that indexers select.

Use of associations of words (*e.g.*, syntax) in phrases, clauses, sentences, paragraphs, chapters, sections, papers, and books to choose indexable subjects is also of doubtful significance. Human indexers do not examine word associations in sentences, paragraphs, pages, and the like. It may well turn out, on further study, that there is no valid relationship between the fact that two words appear in the same sentence, paragraph, etc., and their use to represent new subjects reported by an author. Choice of terms and index entries to represent novel subjects depends upon macrosemantics—upon total (or close to total) context.

The fact that many of the same words appear in both a document, in an essay-type request for documents, or in the complete description of a user's interests would seem, on the surface, to be a better criterion of a relationship between document and request or user than would the other mechanically generated criteria discussed in the preceding paragraphs. Let us assume, in the discussion in this paragraph, that the problem of synonymy has been eliminated by table look-up in a computer, the problem of paraphrasing has been removed by mechanical translation into "standard English," and the problem of new terms (those unavailable in the table look-up) has been solved by automatic referral to a human lexicographer. Now, is there any reason why the presence of many identical terms in such a normalized document and in a normalized request or profile of a user is not a valid criterion of value of the document to the user? The answer is, "yes." The identical terms in the document and request or profile may help to describe a subject that is old or well-known to the user and one that he is definitely not interested in reading again. In order to solve this problem, it would seem necessary to program into the document-selecting machine the total educational and experience background of the searcher in the subject area of the search.

Compromises in quality of indexing and the multitude of possibilities for human-machine interaction for indexers and users are avenues that are being studied more fully.[13-15] The cost of indexing is high and the supply of indexers low. Continuing research on mechanical or machine-aided in-

dexing is certainly justified.

Actual problems faced by the human indexer and by the designer of an automatic indexing system can be realized more fully through hypothetical tests using contrived and sometimes distorted abstracts to be indexed.

(a) The abstract has internally contradictory data. The machine indexing matches the manual; both machine and indexer point out the inconsistencies.

(b) The abstract has data inconsistent with the external world of facts. For example, the atomic weight of chlorine might be given as 34. Machine indexing matches manual. Machine and indexer point out the suspected or detected extrinsic inconsistency.

(c) The abstract is factually correct but grammatically incorrect. The machine and manual indexing match, if indexing can be done at all, and grammar difficulties are pointed out.

(d) The abstract has ambiguous statements. Neither machine nor human may be able to index completely without resolving the ambiguity. Both machine and human point this out.

(e) The abstract is rewritten to be completely meaningless (unintelligible but not gibberish). Neither machine nor man can index and both will point this out.

(f) The abstract conveys only well-known information—no novelty. The machine and man refuse to index and point out the lack of novelty as the reason.

(g) The abstract contains only one novel fact with the rest old information. The machine and man pick out, and index alike, only the novel fact and ignore the old information. This abstract comes closest to being the normal one.

(h) The abstract has novel facts so startlingly at variance with existing thought as to seem incorrect. The machine and man index alike. Both question correctness of facts and ask that acceptance of indexing be contingent upon further verification.

(i) The abstract is made absurdly incorrect. The machine and man refuse to index until the statements have been checked.

(j) The abstract is given a title that is a riddle or cute (*e.g.*, by paronomasia). Both machine and indexer ignore the title and index alike from the abstract.

(k) The abstract is rewritten as an allegory, parable, paraphrase, or other complete variation from the original; man and machine index original and variation alike.

Compatibility.—Terms from other vocabularies are incorporated into the chosen vocabulary by giving the terms machine codes (*e.g.*, numbers uniquely identifying terms in the system) identical with those used for terms of most closely related meaning in the chosen vocabulary.[16] A term in one vocabulary can sometimes be represented by two or more in the chosen vocabulary and *vice versa*. Scope notes are used to enlarge or modify the meaning of descriptors to make them hospitable. Cross references guide the searcher to the new terms incorporated.

Some vocabularies can also be made compatible by rules rather than by joining names in lists. For example, organic compounds,[16,17] electron tubes, polymers, and enzymes can be named or symbolized by rule.

Compatibility of indexing vocabularies is important in enabling interfiling of index entries on paper or magnetic tape. It is less time-consuming to search one consistent subject index or magnetic tape than two or more smaller, inconsistent indexes or tapes covering the same material indexed.

Incompatibility may occur in an organization employing several indexers or among several organizations doing indexing in the same subject area because of the time required to communicate choices of new indexing terms and changes in old terms. Boards of lexicographers seem the best way to resolve incompatiblity from these causes.

Terms to Represent Subjects.—Control in use of an indexing vocabulary involves choice of terms to represent subjects indexed.[18] Just how the human indexer selects terms to represent subjects is something of a mystery. Experience, however, shows that well-trained, experienced subject indexers can agree with experienced index checkers, with a consistency of 90% or better, upon terms to be used as subject headings (not complete index entries) in a specific index built by rule and by use of last-year's index as a guide. Selection of terms from a relatively small generic thesaurus with adequate display of terms, should prove to be even more accurate because there are fewer choices to be made and terms to be remembered.

Choice of terms and entries to represent subjects indexed is not so precise as is, *e.g.*, author-indexing.

As the indexer works he finds documents, or concepts in documents, for which no remembered terms seem to provide adequate indexing. In these cases he has recourse to the vocabulary and display to refresh his memory. If a term suitable to index the novel concept is still not found, then he has two choices: (1) fit the new concept under the most appropriate term(s) in the vocabulary, possibly make a "See" reference, and possibly modify the scope note or definition; and (2) generate a new indexing term complete with scope note and interlocking cross references.

In indexing to the maximum specificity, for book-form indexes, so many new terms come in that it is better not to have a subject-heading list that demands continual and frequent consultation and revision, but to index by rule[11] and to use earlier indexing as a guide only, not as a rigid requirement. Editing of the index with application of a master cross-reference file helps to eliminate inconsistencies and errors.[19]

Generic indexing terms or descriptors can be used in nonmanipulative (*e.g.*, published) indexes in pairs or triplets.[12, 20] This precorrelation of two or three terms greatly increases specificity and selectivity without increasing the size of the index. Systematic permutation of indexing terms is undesirable because it gives many entries of little use in guiding searchers to novel subjects.

Subject Importance.—All of the words in a subject-index entry do not have the same value in leading the user to a novel subject. In the example, "**Benzene**, sulfate excretion from absorption of" "Benzene" and "sulfate" are far more effective in guiding the user to the subject represented by this index entry than are the remaining words. In a machine system, this effectiveness or importance can be symbolized by use of asterisks,[20] *e.g.*, by "***Benzene**" and "***Sulfate**." The complementary index entry for the above example would be "**Sulfates**, benzene absorption effect on excretion of." The words "effect," "from," "of," and "on" are totally ineffective

in leading users to this subject. The words "absorption" and "excretion" are not very effective. The author was not studying "absorption" or "excretion" as novel subjects. He could have been. Had novel features of absorption or excretion been studied, then these terms or their synonyms would be needed as headings or asterisked terms. This enables the searcher to avoid old information that lacks pertinence.

Roles and Links.—Display of relationships among index headings and new subjects in nonmanipulative systems (*e.g.*, bound indexes and other manual systems) is by titles or by modifications (modifying phrases). For example, under each of the descriptors, Aerosol, Electric Field, Generation, and Precipitation, the title could be used, *e.g.*, "Aerosol Generation and Precipitation by Electric Field." The title or modifying phrase is much more effective in helping users to select a pertinent abstract, extract, or report than is any single descriptor or disorganized combination of them.

In a manipulative (*e.g.*, mechanized) correlative index the relationships are difficult to express. "Roles" and "links" have been used to express relationships. In the above title, "Aerosol," plays "role" of product, material treated, and waste disposed. The use of these roles in a mechanized system may increase selectivity. Problems in use of roles include accuracy of assignment by indexer and searcher. Inaccurate assignment may cause failure to select relevant information.

Many documents report heterogeneous material. When this material is indexed by descriptors without use of links, titles, or modifications, incorrect cross-correlation of descriptors may occur upon search. For example, an abstract may report new data on sodium chloride and potassium nitrate. If the four words in the names of these two chemical compounds are used separately as descriptors, then false correlations may occur, to give, *e.g.*, "sodium nitrate" and "potassium chloride." Links between, *e.g.*, "sodium" and "chloride," and "potassium" and "nitrate" prevent such confusion. Links often take the form of code numbers added to the codes for the descriptors. Different descriptors codes are united by the same added code links.

Roles and links are unnecessary in manual systems and in manipulative systems that give adequate selectivity (*i.e.*, yield searches with an acceptable minimum of peripheral material). Some roles can avoided by precorrelation of descriptors. For example, the descriptor "Aerosol" can be made more specific by precorrelating it to give the descriptors "Aerosol generation" and "Aerosol precipitation." "Generation" and "precipitation" become roles attached to the descriptor "Aerosol." Need for links can also be reduced by separating heterogeneous reports (preferably at the source) into monographs. This process makes, in effect, the report number a more useful link among descriptors.

REFERENCES

(1) C. L. Bernier, and K. F. Heumann, *Am. Doc.*, 8, 211 (1957).
(2) C. L. Bernier, *ibid.*, 8, 47 (1957).
(3) V. W. Clapp, *ibid.*, 14, 4 (1963).
(4) C. L. Bernier, *ibid.*, 11, 277 (1960).
(5) "ASTIA Thesaurus of Descriptors," 2nd Ed., Defense Documentation Center for Scientific and Technical Information, Cameron Station, Alexandria, Va., December, 1962, pp. A3-80.
(6) "ASTIA Thesaurus Code Manual," Defense Documentation Center for Scientific and Technical Information, Cameron Station, Alexandria, Va., June, 1961.
(7) C. N. Mooers, in "Information Retrieval Today," Institute of Library School and Center for Continuation Study of Minnesota, University of Minnesota, Minneapolis, Minn., 1963, pp. 21-36.
(8) C. L. Bernier, and E. J. Crane, *Ind. Eng. Chem.*, **40**, 727 (1948).
(9) "The Philosophy and Guidelines for Revision of the Thesaurus of ASTIA Descriptors," Defense Documentation Center for Scientific and Technical Information, Cameron Station, Alexandria, Va., November, 1961.
(10) C. L. Bernier, *Am. Doc.*, **9**, 32 (1958).
(11) C. L. Bernier, and E. J. Crane. *J. Chem. Doc.*, **2**, 117 (1962).
(12) C. L. Bernier, G. M. Dyson, and H. J. Friedman, *ibid.*, **2**, 93 (1962).
(13) H. P. Luhn, in "Modern Trends in Documentation," Pergamon Press, New York, N. Y., 1959.
(14) H. E. Maron, J. L. Kihns, and L. C. Ray, "Probabilistic Indexing," Ramo-Woldridge, Data Systems Project Office, Technical Memorandum No. 3, June, 1957.
(15) H. E. Stiles, *J. Assoc. Computing Machinery*, **8**, 271 (1961).
(16) J. Frome, and P. T. O'Day, *J. Chem. Doc.*, **2**, 248 (1962).
(17) "The Naming and Indexing of Chemical Compounds from *Chemical Abstracts*," introduction to the subject index of Volume 56, Jan.-June 1962, The American Chemical Society, Washington, D. C.
(18) Ref. 5, entire volume.
(19) "*CA* Today—The Production of *Chemical Abstracts*," American Chemical Society, Washington, D. C., 1958, pp. 62-67.
(20) DDC Technical Abstract Bulletin, Subject Index, Defense Documentation Center for Scientific and Technical Information, Cameron Station, Alexandria, Va.