Indexing Consistency: The Input/Output Function of Thesauri

Phyllis Reich and Erik J. Biever

This study measures inter-indexer consistency as determined by the number of identical terms assigned to the same documents by two different indexing organizations using the same thesaurus as a source for the entry vocabulary. The authors derive consistency figures of 24 percent and 45 percent for two samples. Factors in the consistency failures include variations in indexing depth, differences in choice of concepts for indexing, different indexing policies, and a highly specific indexing vocabulary. Results indicate that broad search strategies are often necessary for adequate search yields.

The purpose of this study is to determine how well subject authority lists control indexing vocabulary. Successful retrieval of stored information depends, to a significant extent, on consistent—and therefore predictable—representation of subject matter in the retrieval system. Thesauri are subject authority lists designed to bring this consistency and predictability to the information storage and retrieval process. These thesauri have two interdependent functions. One is to introduce order and language standards into indexing terminology—the indexing consistency function. The other is to serve as a source for searching vocabulary—the retrieval function. F. Wilfrid Lancaster defines a thesaurus as an input/output device whose purpose is to "bring the language of the searcher into coincidence with the language of the indexer."[1]

A number of indexer consistency studies exists. Michael R. Middleton obtained match rates of 22 percent and 18 percent when comparing terms assigned to the same references by different education indexes.[2] Lawrence E. Leonard reported widely divergent results in his survey of inter-indexer consistency studies.[3] This wide range in the consistency figures can be attributed, at least in part, to the criteria used to determine matches. Some investigators considered a match to have occurred if there was agreement on the indexed concepts, while other investigators required agreement on terminology. This paper discusses indexing consistency as measured by the number of identical terms assigned to the same articles by two different indexing organizations using the same thesaurus as a source for the entry vocabulary.

Although many subject authority lists—for example, *Library of Congress Subject Headings* and *Thesaurus of ERIC Descriptors*—exist, the authors chose the *CAB Thesaurus* as a model for this study because its use lends itself particularly well to documentation. This study, however, presents general conclusions that may apply to other authority lists.

Phyllis Reich is Head of Reference and Database Coordinator and Erik J. Biever is Head of the Plant Pathology Library, both at the St. Paul Campus Libraries, University of Minnesota, St. Paul, Minnesota 55108.

THE NAL AND THE CAB DATABASES

AGRICOLA, produced by the National Agricultural Library (NAL), and CAB, by the Commonwealth Agricultural Bureaux, are two major agricultural databases. Both are available through commercial vendors and also appear in print and CD-ROM formats. In 1984, the CAB began indexing from a thesaurus that it published that year. In 1985, the NAL adopted the use of the CAB thesaurus, with some modifications, as the indexing subject authority list for AGRICOLA.[4] The modifications included Americanized spellings and additional terms for subject areas not covered well by the CAB thesaurus. The added terms are primarily in the fields of home economics, human ecology, and food science. Although lists containing these modifications are available to NAL indexers, they have not been distributed for use by searchers.

NAL's indexing policy permits the use of some enrichment terms that are neither in the CAB thesaurus nor in the modified thesaurus. However, indexers place these terms in the identifier field of the indexed document rather than the descriptor field. NAL also assigns each indexed document at least one broad subject category code. These codes, with their scope notes, appear in an NAL publication and do not form a part of the CAB thesaurus. The category codes do not affect the choice of index terms. The *NAL Notes to Indexers*, which NAL generously made available to the authors, states, "Vocabulary terms are assigned independently from the category codes."

At the time the sample for this study was taken, the 13 different indexing units that make up the CAB Documentary Service were located at or near specialist research centers and libraries in the United Kingdom. Currently, these units are housed in one location in England. Each of the units has a staff of information specialists, many of whom have experience in the disciplines for which they have indexing responsibility. A given document can be independently indexed by a number of these units. Some documents included in this study were indexed by as many as three separate units.

METHODOLOGY

If database users and producers accept subject authorities such as thesauri as indexing standards, it can be assumed that for an authority list providing 100 percent vocabulary control, an identical set of terms will be assigned to a document independent of the time, place, or person indexing the document. To measure the vocabulary control—or to put it differently, the indexing consistency—conferred by the use of the CAB thesaurus as a source of indexing terms, the authors examined the descriptors independently assigned to the same set of documents by CAB and NAL during 1986, when each was using the thesaurus as an indexing standard.

A DIALOG database search in the AGRICOLA and CAB files retrieved a set of journal articles indexed by both NAL and CAB. In each case, the authors searched the journal titles *Agronomy Journal* and *Journal of Animal Physiology and Animal Nutrition* for the publication year 1986. These publications are core journals covering different areas of the agricultural sciences and are indexed by both NAL and CAB. The sample consisted of 185 articles from *Agronomy Journal* (sample #1) and 51 articles from *Journal of Animal Physiology and Animal Nutrition* (sample #2), for a total of 236 articles. The authors entered the results of the DIALOG search on *Agronomy Journal* into a dBase III Plus file, with one article per record and with fields within each record for the titles and descriptors. Spelling errors that occurred in the search results were corrected, and British spellings were Americanized. Geographic terms assigned by CAB, including names of countries, states, and provinces, were removed because NAL does not place these terms in the descriptor field. The authors wrote programs to 1) count the number of NAL and CAB descriptors for each article, 2) compare NAL and CAB descriptors for each arti-

TABLE 1
SUMMARY OF DESCRIPTOR STATISTICS

	Sample #1	Sample #2	Average
NAL descriptors per title	8.8	5.9	8.2
CAB descriptors per title	9.8	5.7	8.9
NAL descriptors identical to terms in title	1.2	2.8	1.5
CAB descriptors identical to terms in title	1.9	2.6	2.1
Identical NAL and CAB descriptors	2.2	2.6	2.3
Identical NAL and CAB descriptors	24%	45%	27%

cle and record the number of matches, and 3) compare NAL and CAB descriptors to article titles and record the number of matches.

NAL and CAB assigned an average of eight to nine descriptors to articles in sample #1 and an average of five to six descriptors to articles in sample #2.

The results of the search on *Journal of Animal Physiology and Animal Nutrition* were downloaded from DIALOG and imported into a dBase III Plus file with a structure similar to the *Agronomy Journal* file. Because some article titles appeared in German, the authors created fields for the original title as well as the differing translations used by NAL and CAB. With a few small modifications, the same set of programs was used in analyzing the *Journal of Animal Physiology and Animal Nutrition* file as with the *Agronomy Journal* file. In the case of translated titles, the authors matched descriptors to the appropriate translations.

RESULTS AND DISCUSSION

NAL and CAB assigned an average of eight to nine descriptors to articles in sample #1 and an average of five to six descriptors to articles in sample #2. The observed difference in indexing depth between the two samples was consistent to both NAL and CAB, leading the authors to speculate that differences in indexing depth can be attributed to differences in article specificity in the journals from which the samples are taken or to differences in the indexing vocabulary for the subject areas.

The number of descriptors identical to the terms in the title of an indexed document—a measure of the indexer's use of natural language—ranged between 1.2 and 2.8. The average number of identical descriptors assigned by NAL and CAB was between 2.2 and 2.6 (see table 1). Although the average number of matching terms did not vary significantly between the two samples in this study, the percentage of matches—24 percent for sample #1 and 45 percent for sample #2—varies considerably. This variation exists because more terms were assigned to articles in sample #1 than in sample #2. The chance that two descriptions of the same contents will involve identical terms decreases as the number of terms used to describe the contents increases. To derive the indexing consistency figures, the authors employed the formula used by Middleton—$C = 2c \div (a + b)$—with C representing indexing consistency for a specific citation, a and b indicating the number of terms assigned by both indexing organizations, and c indicating the number of matching terms.[5]

INDEXING DEPTH

As Lancaster reports, one of the conclusions of the Cranfield Project, a major study of the performance of four different indexing systems, is that indexing depth and the specificity of the indexing language are the two principal factors affecting recall and precision in any retrieval system.[6] The Cranfield Project found that an inadequately specific vo-

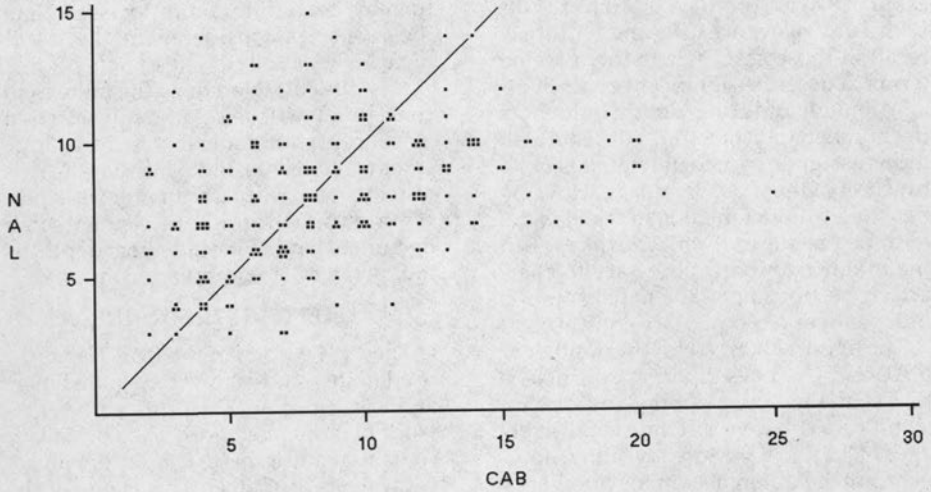

FIGURE 1

Each dot represents an article, the abcissa being the number of CAB terms and the ordinate being the number of NAL terms for that article. There are a number of cases in which more than one dot occupies the same position. For points on the diagonal, the number of CAB terms is equal to the number of NAL terms.

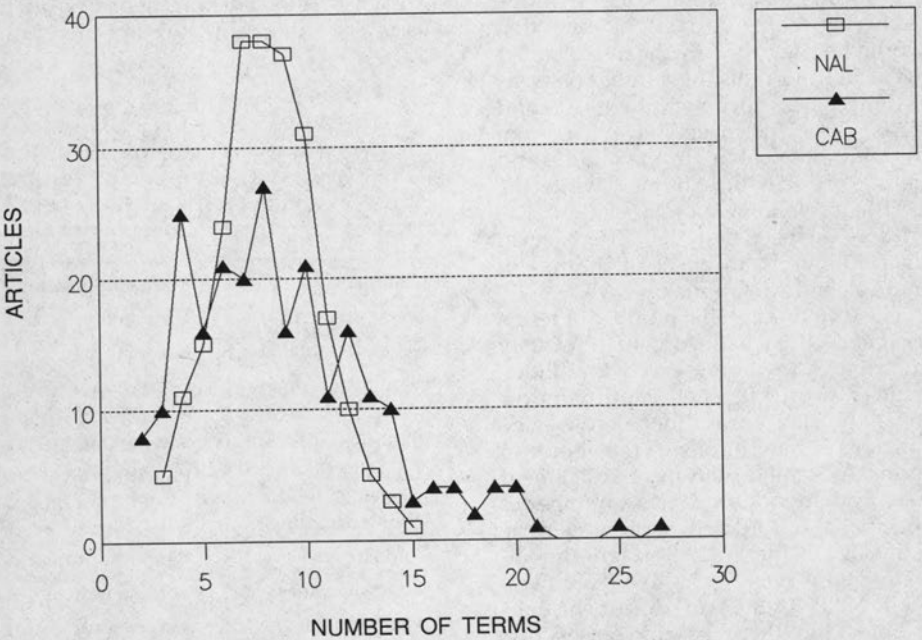

FIGURE 2

Terms assigned per article by NAL and CAB

cabulary will produce search results with a low relevance ratio and will affect recall in those cases where there are no terms to describe significant concepts.

Although indexing depth influences the number and sometimes the choice of terms assigned to a document, it is not a function of the index language. How exhaustively a document is indexed is determined by the indexer, who decides on the number of terms necessary to characterize a document adequately, and is independent of vocabulary control.

The figures obtained for the number of NAL and CAB descriptors per article—that is, the indexing depth—indicate no significant difference in the overall averages of 8.2 and 8.9. More revealing, however, are the differences in the number of assigned terms on a per-article basis. Figure 1, a scatter diagram of the number of NAL and CAB terms per document, shows relatively few instances of identical indexing depth. The diagram depicts a scattering distribution rather than the clustering along a straight line that would be suggested by the overall averages. Figure 2 plots the number of terms assigned per document by NAL and CAB. The NAL curve shows a normal distribution, peaking in the middle and falling off at both extremes, while the CAB curve shows more variation. The authors suspect that the difference in distributions is an effect of the use of multiple indexing units by CAB.

The difference in the number of terms assigned to each document obviously contributes to the observed low index match rate of 27 percent. Depth of indexing may play an additional role inasmuch as it may influence the choice of terms. A detailed analysis of a document may lead an indexer to use a number of specific descriptors in lieu of generic terms. For example, CAB assigns the collective term *rumen gases* to a document, while NAL, for the same document, enumerates the three specific gases discussed, ignoring the collective term. One record from CAB lists five species of grasses, while the corresponding NAL record gives one term, collecting the species under the botanical family name, thereby including more grasses than those species considered in the document.

It is difficult to analyze the failures in conceptual matches. Although selection of concepts for indexing is, to a large extent, related to indexing depth, it is also often a product of an indexer's perception of the significant elements in a document and, like indexing depth, is independent of vocabulary control.

THE *CAB THESAURUS*

The 1984 edition of the *CAB Thesaurus* used in this study is a very detailed hierarchical list of 48,000 terms with a network of cross-references. The vocabulary contains many terms that are approximate equivalents of another. This is one of the factors in the observed inter-indexer consistency failures. An interesting picture emerges from a detailed comparison of some of the assigned descriptors (see table 2).

Each of the indexing agencies often uses terms that, although not identical,

TABLE 2
EXAMPLES OF NAL AND CAB EQUIVALENCIES FOR IDENTICAL CONCEPTS IN THE SAME DOCUMENTS

NAL	CAB
Leaf analysis	Plant composition
Tissue analysis	Plant analysis
Developmental stages	Growth stage
Plant development	Growth rate Growth period
Planting date	Sowing date
Cold injury Survival Winter hardiness	Winter survival Cold resistance
Cold stress	Winter hardiness
Winter hardiness	Cold resistance
Weather data	Agricultural meteorology
Simulation models	Mathematical models
Crop density	Plant density
Residual effect	Residues

serve essentially the same function: to describe the same concept in a particular document. The thesaurus's hierarchical and cross-reference structure frequently leads the searcher to the alternate terms used by the different indexers. However, a thesaurus that includes related terms and term combinations with slight differences in meaning poses problems for its users. For example, a searcher wishing to express the concept of plant response to low temperatures would have to use all the relevant terms provided in the thesaurus because there is no indexer agreement on terminology, as shown in table 2. In this case, bringing all the related terms together, but indicating that only one will be a recognized descriptor, would have best served the input/output function of the thesaurus. A highly specific indexing language allows precise retrieval. If, however, database searchers cannot be confident that they will be able to predict indexer terminology, they will use the thesaurus's generic, specific, and related terms for a given concept in a Boolean OR statement, sacrificing precision for acceptable retrieval.

Although indexing depth influences the number and sometimes the choice of terms assigned to a document, it is not a function of the index language.

INDEXING POLICIES

To some extent, an indexing agency's policies and protocols influence term selection. A CAB publication outlining the Bureaux's indexing policies makes it apparent that term assignment at CAB is oriented toward production of the printed indexes.[7] Indexers assign descriptors for the CAB print indexes first. Additional terms may then be selected for online searching. Descriptors for the printed indexes are assigned and arranged hierarchically, while descriptors in the online database are independent, to be linked by the searcher when appropriate. CAB's practice of making its indexing policies available to the public is the exception rather than the rule. Although searchers have access to vocabulary control lists, they often do not have access to the policies driving term selection. For example, NAL's policy, contrary to the instructions in the CAB thesaurus, requires scientific names for crops before they are harvested and common names after harvest. If database searchers were aware of this policy, they could make use of the information to increase the precision of their searches.

THESAURUS DESIGN

Because many thesauri were designed for use by information specialists—that is, indexers and search analysts conducting client-mediated searches—their suitability for most end users is questionable. If thesauri are to function as output devices for end users searching files on CD-ROM and locally mounted databases from remote sites, they will have to be made available in computer-readable formats. While it is essential that both indexers and searchers have access to the same thesauri, the method of access need not be the same. Unless the thesauri are attractive and easy to use, many end users will avoid them. To best serve end-user needs, the thesauri should be integrated into the database search software. These thesauri would work well if organized, using hypertext programming techniques.

CONCLUSION

Indexing depth and the indexer's perception of the significant elements in a document—two factors in the observed low inter-indexer consistency matches—are variables outside the control of a thesaurus. The other major factors in the low consistency matches relate to term selection once decisions have been made about the important concepts in a document, and the number of terms necessary to describe the identified concepts. Term selection is a function of the entry vocabulary—that is, the thesaurus—and indexing policies. The highly specific index language of the CAB thesaurus

and the inclusion of terms and term combinations that are near equivalents contribute to indexing inconsistencies. Consistency—and, therefore, the ability to predict the indexer's terminology—appears to be more difficult to attain with increasing vocabulary specificity.

Users generally have not had access to the policies driving term selection. For a thesaurus to function effectively as an output device for the searcher, these policies should be linked to the thesaurus's cross-reference structure, directing users to the terms or term combinations dictated by policies.

Results indicate that broad search strategies, which, in effect, negate the precise retrieval capabilities of a highly specific indexing vocabulary, are often necessary for adequate document recall.

REFERENCES AND NOTES

1. F. Wilfrid Lancaster and E. G. Fayen, *Information Retrieval On-Line* (Los Angeles: Melville, 1973), p.244.
2. Michael R. Middleton and Aurora Di'Orio, "A Comparison of Indexing Consistency and Coverage in the AEI, ERIC and APAIS Databases," *Behavioral and Social Science Librarian* 3:33–43 (Summer 1984).
3. Lawrence E. Leonard, "Inter-Indexer Consistency Studies, 1954–1975," *University of Illinois Graduate School of Library Science Occasional Paper* No. 131 (1977).
4. Sarah E. Thomas, "Use of the CAB Thesaurus at the National Agricultural Library," *IAALD Quarterly Bulletin* 30:61–65 (1985).
5. Middleton and Di'Orio, "A Comparison," passim.
6. F. Wilfrid Lancaster and J. Mills, "Testing Indexes and Index Language Devices: The ASLIB Cranfield Project," *American Documentation* 15:4–13 (1964).
7. *CAB Database Production Manual* (Wallingford: CAB International, 1987).

CORRECTION

Appendix A of Delia Neuman's article "Designing Library Instruction for Undergraduates" (52:176, March 1991) is slightly misleading. The text of the article refers to the four categories of goals and objectives of MAJIK/1, but the appendix incorrectly numbers one of the objectives and elevates it to the status of a goal category, thereby creating five categories. The correct numbering of the MAJIK/1 goal categories and of the incorrectly numbered objective is as follows: "I. Introduction to periodical indexes," "II. Instruction in the use of periodical indexes," "II. 3. The User will match abbreviated journal titles to the full titles," "III. Instruction in the use of the UMCP Serials List," and "IV. Instruction in the arrangements of periodicals in the various UMCP Libraries."