Correlative Indexes. X. Subject-Index Qualities

By CHARLES L. BERNIER
River House B-907, 1600 S. Joyce St., Arlington 2, Virginia
Received December 2, 1963

The two most important subject-index qualities are completeness and consistent organization. These qualities are subtle, variable, and measurable. Other index qualities are: guidance, subject-indexing,[1] format, typography, and price.

Omission of Index Entries.—Entries are usually left out of subject indexes to save the cost of putting them in. The index publisher may set a maximum or an average indexing density (number of entries or descriptors per document). Also, the indexer may fail to index novel subjects through inadvertence, lack of knowledge of the field indexed, or insufficient training in indexing. For the best indexes, there is no limit set on indexing density, within the bounds of recognized, good indexing practice.[2]

Because documents are indexed to enable their sure selection from the enormous number produced, and because information in most scientific and technical documents is an exceedingly precious commodity as measured by man-hours of high skill as well as dollars invested, the omission of a few dollars worth of valid technical-index entries per document on the basis of economics is unsound. The cost of work behind scientific papers and reports averages in the ten-thousand-dollar range.[3] The cost of good subject-indexing that can help to avoid duplication is less than 0.1% of this amount.[4]

Incomplete indexing (for reasons other than economics and indexer subject knowledge) can be corrected, in part, by training the indexer in indexing, and corrected nearly completely by thorough checking of all subject-index entries. If subject-indexing accuracy greater than about 80% is needed, then checking of all index entries is necessary.[5] If the omissions occur because of the indexer's lack of knowledge of the subject-matter indexed, there is little that can be done by the indexing organization to prevent them. It is difficult for an indexer who lacks formal education or equivalent experience in the scientific or technical subject matter indexed to understand why many of the entries he omits are needed. The indexer, well-educated or experienced in the field indexed, readily understands the need for entries that he inadvertently missed, when these are pointed out to him. Showing the indexer entries omitted will improve future indexing if he is knowledgeable in the field indexed; otherwise improvement will be marginal. Checking is usually carried out by another trained indexer experienced in the field indexed.[5] Discussion by the two indexers of the changes produced during checking gives maximum benefit from the checking operation. The indexer learns, as often does the checker, and quality of the indexing improves from the discussion.

Since there is usually some question as to what index entries are actually needed as keys to all properly indexable subjects in a document, there are normal, small uncertainties in the number and nature of entries or descriptors independently chosen to represent the same document. For experienced indexers knowledgeable in the subject field, these uncertainties range from 5 to 10% of the number of entries for all documents indexed.[6] Subject-indexing is recognized as not being so precise as author-, contract-number-, patent-number-, molecular-formula-indexing, and the like.[7] Checking and consultation between indexer and checker is needed to reduce the 5 to 10% error.

Depth of Indexing.—Before continuing the discussion of index qualities and their measurement, a few words about the term "depth of indexing" are in order. This term has been used variously to mean (1) indexing density (number of index entries or descriptors per document),[8] (2) relative specificity of the index entries chosen to represent subjects,[9] and (3) a measure of value- and use-orientation as opposed to word orientation.[10] In this last sense, "deep-indexing" is about equivalent to "subject-indexing" as used in this series of papers and "shallow-indexing" is roughly the same as "word-indexing."[1]

"Depth of indexing" has been considered an independent variable and under the control of indexer, employer, or budget. The number of subjects reported in a document is usually under control of the author. Also, the number of index entries needed to lead effectively to all subjects is an accident of language. If there are x subjects to be indexed in a document and $x - n$ subjects are actually indexed, the indexing is better described as "inadequate" or "incomplete" rather than as "not so deep." The effect is loss of access to the n subjects through the use of the index. It is as though the author had failed to report the n subjects or as though they had been edited from the report, so far as the index-user is concerned.

If the indexer generalizes rather than indexes to the maximum specificity, as is in accord with accepted practice,[11] the index is sometimes said to be "not so deep." More accurately, the indexing in this case can again be characterized as inadequate. Generalization by the indexer usually indexes a scope of subject far beyond what the author actually studied and reported and so may lead the index-user to subjects that do not exist in the reports retrieved, and worse, may cause the user to miss reports that he should see because he may fail to think of, and to search under, the more general terms as well as under the more specific terms. For example, if the author reports a study on carrots and the indexer generalizes by choosing instead "Vegetables" as the index heading, then the searcher who wants information about beets, potatoes, tomatoes, eggplant, spinach, turnips, squash, etc., and not about carrots will be misled by this entry under "Vegetables" to a study on carrots that will be irrelevant to his purpose. Also, the searcher may fail to look under "Vegetables" in the index to find studies on carrots. There will usually be no indication in the "Vegetable" entries to show that "carrots" are studied.

In a correlative-trope indexing system[12] in which "Vegetables" is the most specific indexing term in the vocabulary available to the indexer and to the searcher that covers studies related to "Carrots," there will be no difficulty in the recovery of all information about carrots in the selection of documents because both indexer and searcher have no other choice, in indexing to the maximum specificity, than to choose the term "Vegetables." All documents on carrots will be retrieved although they may be mixed with documents on beets, potatoes, turnips, etc.

Measurement of Completeness of Indexing.—The measurement of completeness of indexing is relatively straight forward, although somewhat tedious. The best way to measure this subtle quality is to have an experienced indexer who has adequate educational background in the subject-area indexed check the entries made against the documents indexed. Differences between checking and indexing are discussed with the original indexer to resolve as many points of issue as possible and to come up with fewer errors and mistakes than had independent double indexing been done. The object is to count the number of valid changes. Indexing agreed upon by two or more indexers is usually more valid than the indexing of one indexer alone so that the indexing agreed upon becomes a better standard from which to calculate the percentage of entries omitted.

Scattering of Like Information.—The index-user, -purchaser, and -publisher should realize that a primary objective of the good index and indexer is avoidance of scattering of like information.[13] Unnecessary entries are much more obvious to the user and may lead to complaints and corrective action. Entries not leading to new subjects have much less serious consequences than do missing and scattered entries.

The principal cause of scattering is synonymy. The same subject matter is indexed under two or more different terms or modifications (modifying phrases). Scattering because of synonymy is relatively easy for the alert indexer, checker, and index-editor to avoid. They simply have to know the meaning of everything they index and to recall instances of the same subject indexed before. The index-editor often has singleton entries to guide him to scattering from synonymy since singletons are often incorrect.[14]

Scattering can occur in modifications (modifying phrases) under headings as well as among headings. Paraphrasing in modifications and titles (when used) causes the scattering. Indexers are especially appreciative of the many ways in which the same thing can be said and strive for consistent organization. Rules for choosing headings and writing modifications help to avoid scattering from paraphrasing. Indexing rules have been discussed in an earlier paper of this series.[15]

In correlative-indexing systems, scattering is avoided by precise use of a carefully edited thesaurus (manual or mechanical), vocabulary, or glossary of indexing terms or descriptors.[16] The manual thesaurus, that shows relationships among indexing and cross-reference terms, and provides displays of related terms, is especially useful in avoiding scattering. Synonymy is eliminated by editing the thesaurus terms to be used in indexing.

The subtle fault of scattering is measured by systematically searching for all synonyms in the published index or among the thesaurus indexing terms. This searching in an index is tedious and not too certain of completeness. Scattering from paraphrasing is measured by reading all modifications under a sample of headings to detect like ones that have not been combined or brought near to each other.

Scattering from failure to index to the maximum specificity authorized by the author is measured by examination of the indexing of a sample of documents. The percentage is calculated of generic terms or descriptors chosen where more specific ones could have been used to represent what the author reported. These measurements, while not especially tedious, require the services of a trained indexer who knows the field indexed.

Cross References.—Good guidance to relevant entries is an added index quality. For published nonmanipulative indexes (usually in book form) cross references and notes guide the user to relevant entries. "See" cross references lead from synonyms or otherwise related terms to index headings and entries. "See also" cross references and notes lead to related entries.[17]

In manipulative indexes, *e.g.*, computer systems, there are various devices for guiding the searcher to the appropriate index terms or descriptors. Among the most effective of these devices is the thesaurus, which groups like terms so that all terms related to a given subject can be found mechanically or displayed together or linked by cross references.[17]

Calculation of the number of cross references in indexes is straight forward and is done by sampling and counting. Determination of the effectiveness of guidance devices is more difficult and requires the services of an expert. The need for a "See-also" cross reference depends upon the education and experience of the index-user. For example, an organic chemist would not need to be led from "2-propanone" to "acetone" or *vice versa*. He might appreciate being guided from "thiamine" to the systematic name, or the reverse. The user of a correlative trope index might not need any cross references from synonyms if all vocabulary terms are displayed on one chart of a relatively few terms properly arranged.

Choice of Popular Terms.—The indexer choses terms that the user will most likely look for first. Frequency of use of a term seems to be a reliable guide to its use as an index heading or as a descriptor. Counting of the number of little-used synonyms requires the services of one who knows the field indexed. The percentage of popular terms can then be calculated. In some fields, use of unpopular, systematic nomenclature in indexes may require special notice and mention.

Unnecessary Entries.—Index entries or descriptors that do not lead to subjects reported by the author waste the time of the indexer, of index-editor, and most important, of index-users.[1] The users may be led to subjects that are not new to them.

In manipulative systems, secondary descriptors can be used to increase specificity in searching.[17] Secondary descriptors are those not so useful as primary descriptors in leading to subjects studied. Searches with secondary descriptors alone may lead to abstracts or reports that are irrelevant. For example, the descriptors "Aircraft, Tire, Cord, Reliability, and Measurement" might be chosen to index a document dealing with "The measurement of the reliability of cord that will be used in tires for aircraft." Miscorrelations will occur if, for example, the primary descriptor "Cord" is omitted in searching. Many correlations of the above descriptors could lead to irrelevant documents. Examples of some guides to irrelevant subjects are: (1) aircraft tires, (2) cords used in aircraft (but not cord in tires), (3) aircraft reliability (other than that related to tires), (4) aircraft measurements, (5) tire reliability (other than related to cords), (6) tire measurements (unrelated to cords), (7) reliability measurement (not related to tire cords), (8) reliability of measurements, (9) aircraft-tire reliability (not affected by cords, *e.g.*, wear), (10) aircraft-tire measurement (of other than that of cords), (11) cords in tires, (12) tire-reliability measurement (not affected by cords), (13) aircraft-tire reliability measurement (not affected by cords), (14) measurements from aircraft, and (15) reliability of measurements from aircraft. These guides to unwanted subjects can lead to false drops or "noise" if reports on these subjects are in the collection.

The fact that very large information-retrieval systems actually function effectively[18] with only primary and secondary descriptors and without use of links or roles is due to the fact that: (1) the number of documents, in even very large collections, is tiny when compared with the population of all subjects that could be written upon, and indexed by, the same vocabulary of descriptors; (2) some combinations of descriptors are absurd; (3) precorrelation of descriptors is used to increase selectivity in subject areas of the collection where clustering of very similar documents has occurred; (4) scope notes have limited and defined the descriptors so as to make them more precise and selective; and (5) 50% of false drops is not disconcerting so long as the total number of references selected in a search is small—say fewer than 100. If, on the average, every other reference seen is relevant, the searcher has ample encouragement to continue searching.

In manual subject indexes, unwanted entries are avoided simply by not putting them in. Useful entries for the above example could be "**Cord**, reliability of aircraft-tire," and "**Reliability**, of cord for aircraft tires." These entries would represent a report of results of tests on tire cord before it is used in tires that gives no information about new methods of measurements of reliability. Had new methods of measurement been reported then the modifications above would contain the word "measurement" or an equivalent term, and a third entry reading "**Measurement**, of reliability of cord for aircraft tires," or preferably, a cross reference "**Measurement** (See also headings for the specific properties measured, such as **Reliability**)" would be used.[19]

Measurement of the number of unproductive entries (also a subtle fault) of published indexes is done by taking a random sample of entries and looking up the documents to see if the entries lead to new, indexable information or data. The services of a person with a recent educational background in the field measured is necessary in separating the new from the old. In the case of a mechanized information-retrieval system, a sample of questions is asked and the number of irrelevant references (or drops) counted. The irrelevancies can further be separated into those caused by the indexer choosing old information and the noise caused by correlations that do not lead to subjects indexed.

Format.—An obvious quality factor of indexes is the arrangement of typewriting or print on a page.

Indentures are helpful in emphasizing headings and subheadings. Spacing between lines improves readability. Column width has an effect upon readability and number of index pages required. Justified right margins make the page appear neater. Headings on each page indicating the entries starting the page speed searching. Pagination of subject indexes is seldom needed by users except those studying the index itself. Unless the organization of print on the pages is very poor, the effect of format on usefulness of the index may be negligible.

Typography.—Type fonts affect legibility, and consequently the usability of the index. If the type size is below 6 points, legibility drops sharply. Over 12 points, the words become long enough to affect rapid comprehension because of the limitation of visual span. Since number of the index pages varies directly as the square of the ratio of point sizes, the smaller type sizes are expecially attractive economically. The fact that indexes are searched and not read as text has been used to justify type sizes smaller than those used in text. There is evidence to show that type with serifs is more legible than that without serifs.[20] The use of different fonts to indicate different functions and parts of the index is excellent and common. Bold-face type is useful in indicating headings and primary (or asterisked) descriptors. Italics is useful in distinguishing cross references from reference entries and in emphasizing genus-species names. Good printing of clean impressions with sharp edges to the letters and even, black inking increase legibility.

The effect of type characteristics upon quality of an index is second-order except for very small or very large type, bizarre styles of type, and slovenly or odd-colored printing. Poor typography and format may be associated in the minds of users with lack of the other more-subtle qualities.

Evaluation of Indexes.—The use of one number to represent over-all quality seems unnecessary. The individual quality factors examined separately are adequate in evaluating and monitoring indexes. Cost and return can at least sometimes be determined before changing indexing procedures and form.

The potential return from good indexing can be appreciated from an estimate of duplicating the work that the document reports and from an estimate of the amount of duplication. If duplicating the work, and writing and publishing a report would amount to $40.000.00, then the 10 to 20 dollars that might be spent on indexing it adequately and publishing the index represents 0.025 to 0.05% of the larger number. Just how much duplication a given index prevents is unknown. If there were no indexes, classification, and primary or secondary distribution of documents, then the amount of unwanted duplication would be greater than it now is, but not 100% because the worker and his colleagues would not usually repeat the work. Actual estimates of loss of scientists' time from inadequate use of recorded knowledge have ranged from 30 to 80%.[21] If excellent indexes prevent as little as 1% of unwanted duplication, then their construction and use is easily justified on the basis of economics alone. If other factors are considered, such as the irreplaceable loss of time of scientists and engineers, then even greater increase, when needed, in the budget for indexes are probably justified.

REFERENCES

(1) C. L. Bernier and E. J. Crane, *J. Chem. Doc.*, **2**, 117 (1962).

(2) E. J. Crane, Ed., "CA Today—The Production of Chemical Abstracts," American Chemical Society, Washington, D. C., 1958, p. 38.

(3) U. S. Senate Committee on Government Operations, Subcommittee on Reorganization and International Organizations, "Coordination of Information on Current Research and Development Supported by the U. S. Government," Report 263 of the 87th Congress, 1st Session, May 18, 1961, p. 229, Appendix 1; "Cost of Research per Technical Article Describing Research Results," 1961, U. S. Government Printing Office, Washington, D. C.

(4) Cost data from the Defense Documentation Center and from *Chemical Abstracts*.

(5) Ref. 2, p. 61.

(6) Indexers with ten or more years of experience at *Chemical Abstracts* regularly found this range of uncertainty in number of kinds of headings of nonorganic subject entries chosen per abstract.

(7) Ref. 2, pp. 47–48.

(8) L. A. Schultheiss, D. S. Culbertson, and E. M. Heiliger, "Data Processing in the Library," Scarecrow Press, Inc., 1962, p. 36.

(9) Ref. 8, p. 176.

(10) J. C. Costello, Jr., *J. Chem. Doc.*, **3**, 165 (1963).

(11) Ref. 1, p. 120.

(12) C. L. Bernier, *Am. Doc.*, **8**, 47 (1957).

(13) Ref. 2, p. 49.

(14) Ref. 2, p. 64.

(15) Ref. 1, pp. 120–121.

(16) "ASTIA Thesaurus of Descriptors," Second Ed., The Office of Technical Services, Washington, D. C.

(17) C. L. Bernier, "Correlative Indexes. IX. Vocabulary Control," accepted for publication by *J. Chem. Doc.*

(18) The DDC information-retrieval system using correlation of descriptors by computer for 10^6 reports is not incapacitated by irrelevant retrieval

(19) Ref. 1, pp. 118–119.

(20) C. Burt, "A Psychological Study of Typography," University Press, England, 1959, pp. 8–9.

(21) I. Hirsch, W. Millwitt, and W. J. Oakes, "Increasing the Productivity of Scientists," *Harvard Business Review*, **36**, No. 2, 66 (1958).

Automatic Preparation of Selected Title Lists for Current Awareness Services and as Annual Summaries*

By ROBERT R. FREEMAN**
The Chemical Abstracts Service, Columbus 10, Ohio

JOHN T. GODFREY
E. R. Squibb and Sons Division, Olin Mathieson Chemical Corporation, New Brunswick, New Jersey

ROBERT E. MAIZELL
Olin Mathieson Chemical Corporation, New Haven, Connecticut

CHARLES N. RICE and WILLIAM H. SHEPHERD
Eli Lilly and Company, Indianapolis, Indiana

Received October 17, 1963

This paper constitutes a progress report in which our experience in the use of computers for searching *Chemical Titles* over a period of more than a year is reviewed. The limitations of *Chemical Titles*, and of titles in general, are well recognized by the authors.[1] In the present experiments we have accepted these limitations.

In 1961, Chemical Abstracts Service (CAS) began issuing *Chemical Titles* (*CT*) as a current awareness service for chemists and chemical engineers. The development of *CT* has been described elsewhere.[2] The by-product punched card and magnetic tape records of references to thousands of articles from chemical journals remained unexploited during the first year of publication.

A significant proportion of the users of *CT* are information scientists who search each issue for a group of chemists.[3] In a typical operation, references to papers of interest are copied, pasted, or typed on file cards, and individuals are notified. While effective, this method ties up many hours of the searcher's time.

Early in 1962, however, we realized that the machine records might serve as a useful basis for automatically retrieving references of interest to company research efforts. Time required for conventional searching of *CT* could be reduced or eliminated altogether, and computer output could be easily disseminated to individuals with minimal clerical work.

* Presented before the Division of Chemical Literature, 144th ACS National Meeting, Los Angeles, Calif., April 1, 1963.

** Now with the American Meteorological Society, P. O. Box 1736, Washington 13, D. C.