Subject-Index Production

CHARLES L. BERNIER

PRODUCTION OF SUBJECT INDEXES has gradually, and quite unintentionally, become complex, principally because of job simplification, quality standards and control, the large number of languages, and the size and complexity of the present literature. Processes used in the production of subject indexes may not be obvious from examination of the final product.

The forms in which indexes appear to the user vary greatly. However, principles of production, up to the time of storage, are similar. Subject indexes may appear in any of the following forms: books; unpunched cards with one index entry per card; punched cards for optical coincidence with one indexing term per card; edge-notched cards with an abstract, extract, or reference and with all index terms coded into the periphery; computer tape with document codes following the term codes; computer tape with strings of term codes for each document; and associative memories that are content-addressable and approached by parallel rather than by serial searching. Associative retrieval is largely outside of the scope of this article since it is a technique used to avoid indexing.

Since a subject index provides guides to subjects reported by authors, the first function of indexers is to select subjects. Documents, published articles, books, motion pictures, phonograph records, letters, telegrams, computer tapes, conversations, conferences, debates, dialogs, microfilms, stage productions, and other media of rational communication have subjects that can be indexed.

Not all subjects may be selected for indexing. For example, it may be the established policy of the indexing organization that only novel, emphasized, or extensively reviewed subjects be indexed because it is assumed that the index user does not need or want subjects brought to his attention on which no new work is being reported by the

Charles L. Bernier is Senior Research Information Associate, The Squibb Institute for Medical Research, New Brunswick, New Jersey.

author. The user presumably does not want to be referred by the index to a paper, only to find that he must turn to another to read, in adequate detail and more accurately presented, material which he may have known all along, i.e., material which is "old" to him. Subjects in which authors have themselves invested effort and study may be indexed, and not subjects derived from others. Subjects outside of the scope of the index may be omitted.

Subjects may be complex, and there may be several in the same document. The number of subjects and their complexity are under control of the author and not of the indexer, index publisher, or index user. The indexer, nevertheless, can separate compound subjects into simple ones in order to provide more effective guides to them.

Titles of technical documents can sometimes be excellent guides to subjects; however, many titles generalize. Differentiation between generalization for the sake of brevity and generalization proposed by the author is an important function of the indexer.

It has been found relatively easy for indexers to drift into indexing words used by the author rather than subjects upon which the latter is reporting. The difference between word indexes and subject indexes needs to be understand and remembered. Of course, subjects reported by the author are indexed by means of words; however, this does not create a word index. Words are windows through which subjects are seen. In general, the more words used, the more specifically a subject is seen. The more precisely words are used, the more accurately it is seen. Word indexing leads to omissions, and unnecessary and incorrect entries. Subject indexing is also different from extraction of data, whether numerical or verbal.

Once a subject has been selected, the next step is to paraphrase it, either in the words of the author or in standardized index terminology. The paraphrase becomes the embodiment of the subject. Experienced indexers may carry paraphrases in their heads. Paraphrases often consist of two to ten words; they do not have verbs, avoid repetition, use prepositions, and accurately express the subject. The title of a technical document may be suitable as the paraphrase of one subject. Usually more than one paraphrase is required to cover all of the subjects in a technical paper. Sometimes compound subjects can be combined into one paraphrase suitable for indexing. The complexity of the subject is usually fully expressed in the paraphrase.

After the subject has been paraphrased, the indexer chooses guides to the paraphrase. The guides consist of several words and a reference

number or code. The first word or term is equivalent to a subject heading. Following these are usually modifying phrases or subheadings. After the words in a guide are translated into standard index terminology, the guide becomes an index entry. Terms for the guide are as specific as are warranted by the author and are often the words of the paraphrase. If the author generalizes, so does the guide. Not all terms in a paraphrase are suitable as lead terms in a guide because they may be too general or may lead the user to entries under a subject heading that would be too heterogeneous. The subject heading "Review" would not be suitable as a guide to the paraphrase, "Review of Steel Manufacture in Toledo." However, "Review" would be suitable as a lead term for a guide to the paraphrase, "The Art of Writing a Review." In the first instance, the author is writing about "steel manufacture"; in the second, about "review writing." Most paraphrases require more than one guide. The exact number depends upon the complexity of the paraphrase. Indexing policy may eliminate some guides. The modifying phrases following the lead term usually contain other words of the paraphrase. Usually the paraphrase can be reconstructed from the lead term and the modifying phrase. It is not possible to write rules that prevent use of certain words as lead terms in guides because authors can and do study words. These words are necessary as lead terms. The reference following the modifying phrase guides users to the document indexed.

After a guide to a paraphrase has been chosen, the next step is to translate it into standard index terminology. Standard terminology eliminates scattering among synonyms or among different generic headings in the index. Scattering of related entries is a serious fault of poor indexes. If there is no systematic nomenclature available, then "See" cross-references are used to guide users from synonyms to the subject heading chosen. For organic compounds, for example, a standard nomenclature does exist; many cross-references can thus be avoided. The synonym chosen to be the subject heading is usually the term in most common use. In this way indexes track usage. The standard terminology used in an index may appear in the form of a dictionary, word list, standard list of subject headings, subject-authority list, or a thesaurus. Indexing under a "more general" term occurs when the vocabulary is limited in size so that specific terms must be indexed under "more generic" terms. An example is the indexing of "Toyomycin" under "Antibiotics." It is apparent that if the biomedical literature has upwards of 60,000 terms in it and if the subject-heading list

has only 6,000 terms in it, then indexing under "more general" headings must occur. All indexing of a "more specific" term under a "more general" one confuses generalization reported by the author with the generalization that meets the needs of the indexing system, unless special provision is made to separate these two kinds of generalization. Some indexing organizations permit "posting-up," or indexing also under a "more general" heading as well as under the "more specific." It seems more judicious to carry the relationships among general and specific terms in a thesaurus (manual or computerized) as a way of facilitating generic searches. Relationships shown in a thesaurus are relatively permanent in fields of science and engineering, and are similar to those found in *Roget's Thesaurus.*

Once the index entry has been created, it is recorded on an index card or slip of paper—one entry to a slip or card—to enable alphabetization.

In all kinds of indexing, errors are made. Most errors are made in subject indexing. In the field of chemistry, for example, 15 to 20 percent of an experienced indexer's entries are regularly changed. Of these changes, perhaps under 5 percent represent serious errors, such as omissions or incorrect headings. Index entries are checked by experienced indexers to reduce these percentages of error. Changed cards are then reviewed by indexer and checker to improve the indexing by eliminating errors introduced by the checker. Discussion of the changes is a form of corrective feedback. The incorrect reference numbers are corrected by a special check.

Card indexes or those published in the form of books or journals require alphabetization. Subject indexes are often alphabetized letter-by-letter rather than word-by-word. The complicated rules used for ordering of numbers and special symbols used in certain disciplines are often recorded in the introduction to the index. Commas are used to interrupt alphabetization for the purpose of bring similar entries together.

During indexing, the indexer cannot predict accurately the number of entries that will occur under a given subject heading, nor the content of related modifying phrases in the completed index. Because of this, index editors are provided. Editing brings similar entries together, eliminates dangling and circular cross-references, picks up errors that have slipped by both indexer and checker, speeds indexing by permitting more leeway in coining modifying phrases, and enables informed splitting or combination of headings. Unwanted subject head-

ings are crossed off and double indentions are made. For published indexes, indexers can function as index editors. Index editing is checked by other editors, preferably by those with more experience. Cross-references are selected during editing, often by means of an inverted cross-reference system. The index editor sees an entry and indicates on the inverted cross-reference that the corresponding cross-reference is to be used. An inverted cross-reference is simply a cross-reference alphabetized at the heading to which reference is made. For example, the cross-reference, "Iron. (See also *Steel.*)" has the corresponding inverted cross-reference, "Steel. Iron. (See also —.)" Inverted cross-references give indexers control of the index at all times. After cross-references have been justified during editing, they are inserted into the index and unwanted headings are crossed off. Subject-index cards for a published index can be shipped to the printer sooner than otherwise if the editing is preceded by a "first survey" during which the cross-reference and inverted-cross-reference systems are applied. The first survey helps avoid transfers from headings later in the alphabet to earlier positions on galley proof. During the first survey, small headings can be edited and special problems identified and labeled, and large headings bundled for editing later after the cross-reference system has been applied.

Type set directly from index slips eliminates errors introduced by copying the cards onto sheets. Indexing organizations photocopy the cards on high-speed microfilm cameras in order to protect them in the event of loss. Monotype composition is often used for technical indexes because it enables correction of individual type sorts, insertion of nonstandard type sorts, and use of type sorts most familiar to the user. Printing introduces errors that are eliminated by reading galley and page proof. The index publisher checks the galley proof against the index cards to ensure their correctness. Changes made on galleys are checked on page proof. Schedules are useful in speeding index production since sources of delay often then become visible, allowing improvements to be made.

Indexers should have formal education or experience in the subject field in which they index in order to save the time that they would require to learn the field bit-by-bit. An indexer who does not know the field tends to make more serious errors than one who knows it well. Also, one who knows the field can work much faster than one who does not. It is not necessary or desirable that the indexer be as highly specialized in the field of science, for example, as is the scientist who

carries on laboratory research or development. Since the new indexer is a subject specialist, he usually comes to the indexing organization without training in indexing. Such training is often obtained through coaching, during which the indexing is checked by an experienced coach who discusses changes with the new indexer.

Indexers earn salaries about equivalent to those of their colleagues in the subject field itself. The chance to keep informed about new developments has proven an especially attractive inducement to scientists.

Indexers have found that dictation of the index entries is more efficient than typing index cards or writing them longhand. Magnetic recorders are used to enable immediate correction to be made by erasure so that the transcriber is not overburdened with corrections. Typewriters are now available with chemical and other special keyboards to facilitate typing such things as numerical subscripts. Other keyboards designed for use with chemical literature enable typists to learn touch-typing of structural formulas for organic compounds.

Files are required for index cards and fire protection is provided in the form of vaults or fire-proof files. Special forms for indexing and entry into computers are available. The descriptive cataloging needed on some of these forms can be done by those who do not know the subject field.

It has not been found worthwhile to key-punch subject-index cards just for the purpose of alphabetizing them since the index slips or cards can be alphabetized by hand more rapidly than by machine.

Indexers use thesauri, dictionaries, or other standard word lists, including the indexes of earlier years. Standard reference works in the fields of alloys, bacteriology, chemistry, physics, and the like are invaluable in standardizing terminology and training the indexer in ancillary fields.

Sloping desk boards are used for draping galley proof while it is read or checked. Card holders for proof checkers attach to the proof boards. Index slips may be color-coded to represent different years of an abstract journal or index to other serial publications. Line composers are used in printing indexes to facilitate cumulations. Computer-operated composing machines offer great promise. Such machines have many fonts, a wide range of point sizes, automatic justification of right margins, automatic hyphenation, computer-controlled editing, and fine quality. With such machines, users can have a large number of fonts and can avoid the queer circumlocutions used to overcome deficiency

of type sorts. Many type faces (some more legible than others), formats, and papers are available for published indexes. Proper indentation of modifying phrases, double indentures, and subheadings also make it easier for the user to find what he wants, while running heads in larger type speed page location. Durable paper can now be purchased.

Although this article is primarily concerned with subject indexes, many of the observations apply to other kinds. Since authors tend to specialize, author indexes can serve as guides to remembered documents and subjects. Indexes to names of corporate authors or to sources of revenue for the work documented are also useful. Indexes to patent numbers are invaluable if one has only a patent number at hand. Taxonomic indexes aid in generic searches for classes of organisms. Molecular-formula indexes help the organic chemist who has not had the time to learn systematic organic nomenclature. Although classified indexes aid in generic searches, a better approach is often through the use of an automated thesaurus that enables selection of all species of a genus, parts of a whole, etc. Classified indexes may require subject guides for users who do not know the classification system and do not have time to learn it. Indexes to ring structures of organic compounds, which can also serve as generic indexes, have been produced. Citation indexes, long produced for legal use, are now available for science. Chemical-group indexes have been produced to aid searches for all compounds within the same groups. Indexes of notations, ciphers, matrices and sets have been proposed and developed. Their use should aid the chemist interested in synthesis of compounds and in correlation of physical and biological properties. Correlative indexes enable correlation of subject headings (descriptors) to increase selectivity lost by elimination of modifying phrases; Cartesian coordination of holes in cards provides one form of correlative index, the coordinate index.

Subject indexes have qualities that are independent of users, although these characteristics do affect use. Most of the qualities are subtle and invisible to the user. Completeness of indexes is one desirable quality. A policy, therefore, that limits the number of entries may eliminate guides to subjects. Or indexers may omit entries inadvertently and checkers may fail to pick them up; similarly abstracts may omit subjects that would normally be indexed and the indexers may not detect the omissions. Freedom from scattering of similar entries is a prominent quality of excellent indexes. Scattering, often among

synonyms or among modifying phrases, is avoided by provision of needed cross-references and by rules for writing modifying phrases. Thesauri are used for guidance, as are notes published in the index or its introduction; nevertheless, guidance external to the index burdens the user. Freedom from error in indexes should be sought during production, and correctness applies to the subjects selected as well as to the reference numbers. Indexing of what is novel, emphasized, or extensively reviewed produces indexes with desirable characteristics. Technical quality of production should also be high, for poor typography and printing may lead to the inference that more important and less visible qualities are also inferior. Format, too, can affect use, while price must of course be considered. An index priced too low may not be taken seriously, while too high a price may reduce use by limiting availability, even though loss of information or delay in obtaining it may often be far more costly than even the most costly of indexes; the cost of "not knowing" needs measurement.

The cost of index production has usually been unfavorable to prompt, efficient communication. It is as difficult to "prove" the value of indexes to the technical literature as it is to prove the value of proposed research and development. In demonstrating value of indexes, one must consider alternate routes to information and the cost of delays in these routes. The value of research and development to the society in which we live is now widely accepted and may be several orders of magnitude greater than its cost. The cost of research and development may be three orders of magnitude greater than is the price of bibliographic control, including production and use of indexes.

There are a number of unsolved problems in the production of subject indexes. Support, which is dependent on appreciation of the value of indexes, is such a problem. Appreciation may depend on a keen awareness of the value-cost-price relation just discussed. Those who authorize resources for indexes need written justification for such support and it has been exceedingly difficult to write adequate justification for excellent indexes.

Quality control in production of subject indexes is another continuing problem that depends, in part, on the appreciation of the value of such control. Since it is not superficially apparent that subject indexes may differ greatly in quality and since it may take users years to discover, for example, that more than half of the valid index entries have been omitted from an index, it is very difficult to gain this appreciation.

Another problem, that of delays in production of indexes, is often amenable to the therapy of dollars and scheduling. An estimate of the cost of such delays would be useful as additional justification—over and above the complaints of users—for their elimination.

Effective generic searching continues to be a problem, especially in published indexes built to the maximum specificity. Searching among a multitude of specific headings is a chore. Theoretically, an automated thesaurus with all genus-species and other relations built in, and with controllable generic searches available, should solve this problem. Another possible solution is a small, generic thesaurus. There is no doubt that thesaurus construction, updating, and use, need improvement and greater understanding. Indexes using these external controls should have more accurate and comprehensive relations, shown more promptly.

Computer-produced indexes of as high a quality as those produced by a human indexer are not yet available, although associative retrieval, especially from abstracts in English, appears to be a promising method. Automatic indexing depends upon the ability of machines to select subjects or their surrogates, to paraphrase them, and to provide useful guides to them. Frequency or infrequency of use of a term in a document is not a suitable criterion for selection as an index term, because authors do not deliberately repeat terms (or introduce them once only) in order to make the indexer or computer choose them for the index. The principal purpose of research on automation of indexing is to find means of saving the time of human subject-authority indexers and of improving the economic viability of much indexing.

When the full contribution of subject indexes to our civilization has been calculated, it will probably turn out that the price of the finest and most expensive indexes is extremely low when compared with the total production cost of the material indexed (including the cost of development and research) and with the value of this material to our civilization.

General References

Atherton, P., and Weaver, V. "A Method for Producing Journal Indexes." New York, American Institute of Physics and Vance Weaver Composition, Inc., 1964. (Processed.)

Bernier, Charles L. "Correlative Indexes. X. Subject-Index Qualities," *Journal of Chemical Documentation*, 4:104-107, April 1964.

Bernier, Charles L., and Langham, Cecil C. "Dictating Machines as Indexing Aids," *American Documentation*, 6:237-238, Oct. 1955.

Bernier, Charles L., and Crane, E. J. "Correlative Indexes. VIII. Subject-Indexing *vs.* Word-Indexing," *Journal of Chemical Documentation*, 2:117-122, Jan. 1962.

Binford, R. L. "A Comparison of Keyword-in-Context (KWIC) Indexing to Manual Indexing." An unpublished M.S. Thesis prepared for the schools of Engineering and Mines, University of Pittsburgh, 1965.

Crane, E. J., and Bernier, Charles L. "Indexing and Index Searching." *In* Robert S. Casey, *et al.*, eds., *Punched Cards. Their Applications to Science and Industry.* 2d ed. New York, Reinhold Pub. Corp., 1958, pp. 510-527.

Davis, Donald D. "Subject Indexing for Nuclear Science Abstracts." In *The Literature of Nuclear Science: Its Management and Use.* (Proceedings of a Conference held at Division of Technical Information Extension. Oak Ridge, Tenn., Sept. 1962. TID 7647.) Oak Ridge, Tenn., AEC, 1962.

Dopkowski, Philip L., ed. *Selected Bibliography on Indexing in Science and Technology: Theory, Application and Techniques.* Washington, D.C., American University, Center for Technology and Administration, School of Government and Public Administration, 1963.

Hillman, D. J. "The Structure of Document Relations." (Study of Theories and Models of Information Storage and Retrieval. Report No. 8.) Bethlehem, Pa., Lehigh University, Center for the Information Sciences, 1964. (Mimeographed.)

International Federation for Documentation (FID). *Abstracts.* (1965 Congress, 10-15 October 1965, Washington, D.C.) Washington, D. C., Secretariat, 1965 FID Congress, 1965. *See especially* the following abstracts: M. Bloomfield and E. Schafer, "Discussion of Some Indexing Problems;" B. Carter, *et al.*, "A Computer-Generated Index Publishing System;" W. R. Foster and D. F. Hersey, "Indexer Requirements for the Recognition of Scientific Content and Context;" and R. S. Hooper, "Indexer Consistency Tests—Origin, Measurements, Results and Utilization."

Jonker, Frederick. *Indexing Theory, Indexing Methods, and Search Devices.* New York, Scarecrow Press, 1964.

King, D. W. "Evaluation of Coordinate Index Systems During File Development," *Journal of Chemical Documentation*, 5:96-99, May 1965.

Kochen, Manfred, *et al. Adaptive Man-Machine Concept-Processing.* Yorktown Heights, N.Y., IBM Watson Research Center, 1962.

Metcalfe, John. *Information Indexing and Subject Cataloging: Alphabetical, Classified, Coordinate, Mechanical.* New York, Scarecrow Press, 1957.

Salton, G., and Lesk, M. E. "The SMART Automatic Document Retrieval System —An Illustration," *Communications of the Association of Computing Machinery*, 8:391-398, 1965.

Stevens, M. E. *Automatic Indexing: A State-of-the-Art Report.* (National Bureau of Standards Monograph 91.) Washington, D.C., U.S.G.P.O., 1965.

Wayne, Jean M., comp. *Indexing, with Emphasis on its Technique—An Annotated Bibliography, 1939-1954.* New York, Special Libraries Association, 1955.

Wetsel, F. R. "Time Studies in Producing Subject Indexes for Chemical Abstracts." Paper presented at the American Chemical Society, Division of Chemical Literature, Detroit, April 6-8, 1965. (Abstracted in *Chemical Literature*, 17:11, Spring 1965.)