Research in indexing: more needed?

K. G. B. Bakewell

Some research into indexing methods is analysed, but it is suggested that much remains to be done. In particular, more needs to be known about user-reactions to indexes, including such matters as alphabetization and presentation of indexes. The British Standard on indexing (BS 3700:1988) makes a number of assumptions and recommendations about users of indexes, but these need to be tested. One objective of the Society of Indexers is to conduct and promote research into indexing, and the Society should consider applying for funding to enable it to investigate user approaches and user needs.

In June 1990 a number of officers and Council members of the Society of Indexers met in London and carried out a SWOTC analysis, examining the Society's strengths and weaknesses, the opportunities which the Society has, and the threats and constraints. Four groups examined priorities for marketing, training, research and administration.

One of the Society's objectives, as laid down in its constitution (objective 2.4), is to conduct and promote research into indexing and related matters, including technological developments. In spite of this, a search through the indexes to The Indexer revealed only two entries under 'research', both in Vol. 13 (1982-1983). Both referred to book reviews: New technology and developments in the communication of research during the 1980s by A. J. Meadows (Leicester: Primary Communications Research Centre, 1980), a book which had no index and was reviewed in 13 (1) April 1982, p. 64; and Research in document classification and indexing (Canada) 1971-80 by Timothy C. Craven (Frankfurt: INDEX-Verlag 1981), reviewed 13 (2) October 1982, pp. 130-1.

Hans Wellisch's excellent book has no section dealing exclusively with research, though many do involve or report the results of research—for example, exhaustivity (pp. 121-4), filing (pp. 130-44), natural and controlled languages (pp. 245-8), personal names (pp. 279-99), romanization (pp. 337-46), string indexing (pp. 367-72), thesauri (pp. 381-8) and typography (pp. 392-405).

Most issues of The Indexer include at least one article—usually more—involving research. Several of these have been by Hans Wellisch: one thinks of his studies of early multilingual and multiscritp indexes in herbaria, the treatment of ‘indexes’ and ‘indexing’ in encyclopaedias, the meaning of the word ‘index’ and the oldest printed indexes. Among many other research-based articles are David Crystal's analysis of the relationship between linguistics and indexing, Ming-yueh Tsay's bibliometric study of indexing and abstracting, Cecelia Wittmann's evaluation of subheadings in award-winning book indexes, Elizabeth Moys's examination of the subject indexing of law books, and Hazel Bell's consideration of the indexability of fiction.

The computerized software which many of us use to prepare our indexes would not have been possible without research. Campbell Purton's investigation of the feasibility of microcomputers for indexing in 1982 led to the research which resulted in the now widely used MACREX and CINDEX programs.

The members of the BSI Committee which produced BS 3700 also carried out a great deal of research to enable the standard to be published.

There has, then, been much research activity on the part of individual indexers and individual members of the Society, but whether the Society itself has carried out objective 2.4 is another matter.

Evaluation of indexes

The best-known attempt to evaluate index performance is what is now known as the Cranfield experiments, which emphasized four 'bench marks' to be used in assessing indexes: exhaustivity (the extent to which the subject content of a document is analysed by the indexer); precision (a measure of the relative efficiency of a system by a comparison of the number of relevant documents retrieved with the total number of documents, relevant or irrelevant, produced by a search); recall (the number of documents produced in answer to an enquiry which are relevant to a user's requirements); and relevance (the extent to which the system allows the retrieval of relevant documents). The College of Librarianship Wales (as it then was) carried out two important investigations of indexing systems, one concerned with documents on information science and the other, the EPSILON Project, concerned with the performance of nine printed subject index types.

As Regazzi has pointed out, the Cranfield studies opened up important research areas in evaluating indexing systems and information retrieval systems because they drew attention to the problems involved in measuring the quality of the output of such systems. They and the Aberystwyth studies were, however, concerned primarily with the evaluation of indexes to collections of documents and not the indexes to individual documents. The Wheatley Medal selection panel have their criteria, as does the
Society of Indexers in assessing those applying for admission to its Register, and Geoffrey Hamilton has provided very useful guidance on how to recognize a good index.17 The Society's current President, David Crystal, has suggested that the science of linguistics could provide a useful method of evaluating indexes: 'Investigating the relationship', wrote Professor Crystal, 'will be a challenging task; but at the end of the road is a linguistic theory of indexing language, and that is a goal worth pursuing.'18 As far as I am aware, nobody has yet pursued it.

The Subject Access Project at the University of Syracuse's School of Information Studies, led by Pauline Atherton, examined the feasibility of using book indexes for subject retrieval in the humanities and social sciences.19 There were serious problems, not least being that only 55% of the 113 books examined possessed indexes. Other difficulties included absence of an introductory note to explain the scope of the index; inconsistency; too many entries with ten or more undifferentiated locators; inadequate cross-references; and absence of multiple access points.

In 1974, eight indexers gave their views in a symposium on the inadequacies of book indexes.20 One contributor deplored the absence of indexes in many books on management. Other shortcomings (in addition to those identified in the Subject Access Project) included insufficient space for the index resulting in inadequate detail or coverage; revision of texts without revising the index; lack of specificity in the index; alphabetical confusion; errors in location references; lack of harmonization between indexes in similar subject fields; in contrast to the previous point, failure to tailor an index to fit the specific book; and failure to index such matter as illustrations, tables, bibliographies, prefaces, introductions and appendices. Mrs H. B. King quoted a sixth former studying A level history: indexes? I don't trust them. I use them, but more and more I study the table of contents and am forced to read great chunks to find out what I want.'21

Clearly there is scope for research into the shortcomings of indexes if they are to be used as effective means of retrieval.

Length of indexes

What is the optimum length of an index, or is this like asking 'how long is a piece of string'? BS 3700 simply states, 'The index should be sufficiently detailed to meet the expected needs of the user, given the amount of detail provided in the document indexed', continuing 'Factors affecting the length of an index include the character and purpose of the document indexed: a technical or academic treatise is likely to require a more detailed index than a popular biography; an abstract may need more index entries than a popular journal article.'22

Twenty-seven years ago Margaret Anderson carried out a small survey of the percentage of space occupied by indexes in books on various subjects.23 She found that indexes were nearly always set in type smaller than that used for the main text, and the increase in the number of lines in a page of index over that in a page of text could be as little as seven per cent or as much as 120 per cent. Most indexes of some books on history examined by Mrs Anderson fell into the range of 58 per cent of the text. Biographies tended to be more lightly indexed than history and more likely to have no index: of 22 biographies published between 1899 and 1965, 16 had indexes in the range 1–4 per cent, with a tendency towards increasing length in the newer books. Of the remaining six, four had indexes between five and six per cent. Two had unusually long indexes (9.7 per cent and 9.5 per cent). The length of indexes in geographical books was found to be very variable, and the length of indexes in books on science was more clearly related to the standard of difficulty of the text.

Wellisch has just over five pages on the length of an index.24 He begins by pointing out that a long index is not necessarily also a good one and that 'many "long" indexes are so big because they contain unnecessary entries—the result of overindexing—while at the same time they may lack some rather obvious and necessary entries, especially for synonyms and related terms which do not actually appear in the text but may well be sought by users'.

Mrs Anderson concluded her short article: 'To pursue the subject in a worth-while way would require much more extensive sampling, and better mathematics. One thing that has not been touched on is the relation between the length and the usefulness of an index.'25 There is scope for research here.

Alphabetization

Considerable controversy was caused by the comment in the first draft of BS3700:1988 that letter-by-letter arrangement was deprecated. The final version contained a compromise; word-by-word arrangement was preferred, but letter-by-letter arrangement 'may be required for the continuation of an existing index'.26

Neil Fisk felt that the trend towards word-by-word indexing was disastrous for technical and scientific work because 'the choice between one word or hyphenation or two words is optional for very many hundreds—nay, thousands—of technical word-compounds, and for quite a number of scientific names and expressions too'.27 While preferring word-by-word, J. Arthur Greenwood conceded that the system requires cross-references when 'a catchword has plausible solid, hyphenated, and two-word forms'.28 Pilot studies by Hartley, Davies and Burnhill indicated that students use a variety of methods when asked to put words in order for an index. University students tended to use sense and meaning of words as prime considerations but schoolchildren were more likely to use the letter-by-letter method of alphabetization.29

Another controversial issue is whether to include prepositions and other 'unimportant' words in the alphabetization. Some authorities suggest that they should not be recognized,30,31 while Wellisch32 feels that they should. BS 3700 states that, no matter which system of alphabetization is used, 'care should be taken to group in one alphabetical
sequence all the subheadings that qualify a single heading" and the examples given clearly do not ignore prepositions in the arrangement.

As Wellisch states, "No other issue in librarianship and indexing has evoked more heated debates, has been beset by more arcane and silly rules, has been the target of more well-deserved satire and ridicule, or has frustrated users more than the arrangement of headings in catalogs and indexes." It would be useful to have some research to establish what users really think about such matters.

**Presentation of indexes**

S. I. Wicklen considered the typography of indexes in the second issue of *The Indexer*. He concluded that "typography does have its contribution to make in the arrangement and disposal of printing material, thus aiding the indexer and the reader", and that "Each book should be treated individually as a problem in design". Robin Kinross, Kenneth Day and Burnhill, Hartley and Davies have also considered the typography of indexes. The most significant piece of research in this area has come from Nan Ridhalgh. She concludes that the typescript presented by the indexer is a most important part of the eventual printed design; that indexes should be kept as simple as possible so that the information is easily retrieved by the reader; that the designer, indexer and author should agree a layout at the outset; and that an indexer who has a full understanding of the design features of an index will be able to submit the typewritten index in the required style so that the translation to the printed version is a smooth and efficient process.

**PRECIS**

**PRECIS** (Preserved Context Index System), the subject of a prolific literature, was itself the product of a great deal of research. A small evaluation study at Liverpool suggested that users (students of librarianship and information studies) were generally happy with **PRECIS**. One aspect of **PRECIS** which was never really investigated was its potential for book indexes, though it was used effectively for the index to the **PRECIS** manual. An investigation suggested that users were generally happy with the **PRECIS** system, but it has now been replaced in the British National Bibliography by **COMPASS** (Computer Aided Subject System), a simpler (and less effective?) system which is claimed to be "a subject authority system that contains the best of the basic qualities of the **PRECIS** system without its complexities, and also builds on **PRECIS** to produce a system that can be created, searched and displayed in an automated interactive environment." It would be interesting to see the results of an investigation of user reactions to this system.

**User reactions**

An important aspect of index evaluation is the assessment of user reactions. What do users think about alphabetization and presentation? Do people use indexes at all? Some years ago the Chairman of the Research Committee at Liverpool Polytechnic (as it then was), attempting to establish the research record of a particular department, spent some time browsing through the Research Report rather than turning first to the index that I had so carefully compiled. We know from Norman Hillyer’s excellent contributions to *The Indexer* that many books are still published without indexes. Do the readers of these books care about the absence of an index? Does the lack of an index affect sales?

The index to BS 3700:1988 contains four subheadings under the entry 'user needs', relating to assigned keywords, entry construction, entry layout and index construction. 'The choice of concepts to be named in the index depends on the expected needs of the user . . .'. 'Entries should allow for the different approaches likely to be made by the user . . .'. 'The index should be constructed according to a logical, balanced and consistent pattern, easily recognizable by its potential users'. The quality and consistency of an index will be enhanced if the indexer . . . is fully briefed as to the likely requirements of the potential users of the index'. What are these expected needs, likely approaches and likely requirements?

Holmstrom has pointed out that, however well versed in the subject matter of any book its indexer may happen to be, he (or she?) cannot foresee what items may at some future time chance to be of interest to some individual reader and that there is a case for a scientifically controlled investigation aiming to correlate different styles and details of indexing with the opinions formed by index users.

Christina Beale and Jean Stirk both feel that more could be done to teach people how to use indexes effectively, and this, as pointed out by Michael Marland and Cecilia Gordon, can usefully be done at school. Helen Chandler, essentially a user rather than an indexer, has shown that user needs are not paramount in the indexes to European Communities documents.

David Crystal has suggested experimenting with slightly different variations of index entries to test user reactions, and Christina Beale stated in 1980: 'Much more work, however, still needs to be done on how, for example, the indexing needs of library users could be studied or how the needs of book-index users might be assessed'.
Conclusion

Thirteen years after Christina Beale's paper, and nine years after David Crystal's, there is still much to be done and, who knows, there might be an opportunity to do it. The British Library Research and Development Department held an Open Day on 10 November 1992, during which John Burchell gave a very useful seminar on research grant priorities. Stressing that he was presenting his own views, not stating official policy, Burchell suggested various areas for research priorities in the future, including: the use and non-use of information services; user groups and their needs; education for information use; and the economics of information. There seems no reason why the use and economics of indexes to individual documents should not be investigated, as well as the use and economics of information services, or why the Society of Indexers should not submit a research proposal to the British Library.

References

43. For example: Austin, Derek and Dykstra, Mary. PRECIS: A
Got that? or, Too many cooks?

Belatedly having discovered a thirty-year journal cumulative index that was published twenty years ago, I was entranced to find the following apologia in the preface (all names have been concealed to protect the innocent and guilty alike):

'An index for the first 30 volumes of (X) has been planned since 1965 or earlier. Several starts, in fact, were made during these years, but none reached a satisfactory end... The index has six major parts although there are nine different headings. There are more headings than parts because reviews have been listed separately under two headings in addition to their listing in the Author Index, while Notes and News and Current Research have been indexed separately to preserve their separate status in the journal even though their subjects are the same.

'The Author Index lists all contributions to the journal (except Memoirs) in alphabetical order according to author... The Title Index lists titles of all contributions other than Reviews, Notes and News, Current Research, and Memoirs. Listing is in alphabetical word order. Definite and indefinite articles, in any language, when they are the first word in a title, do not affect alphabetical order, but they are printed as the first word of the title.

'The Geographic Index lists research contributions but not reviews and notes... Some contributions are listed under two or more area headings. A topical index was planned but was not compiled in time to be included...[there is lots more in this vein].

Finally, 'Many people assisted in the development of this index. Z took entirely upon herself the task of extracting bibliographic information from the [139 indexed issues]. Every entry was checked twice against its journal entry by Y and her assistants. X typed the manuscript. Galley was read... by W. [signed] V'

On looking at the index itself one finds, of course, that the use of titles as the main index entry results in massive information loss. For instance, 'curved single-piece fish-hooks' is found only under C, not under F or H. Likewise there are several entries beginning 'Contributions' or 'Development' which are useless without knowing what is contributed or developed. What a grievous waste of well-intentioned effort and money such an index represents. One can only hope that some lessons have been taken on board for the next cumulative, which should be due soon after 1995.

C. L.

An example of endless-loop cross-references from the glossary of Ashton-Tate's Advanced Programmer's Guide for dBASE II and III Plus:

Endless Loop. See Loop, Endless.
Loop, Endless. See Endless Loop.