Why postcoordination fails the searcher

Bella Hass Weinberg

Postcoordination, in which terms are combined at the searching stage rather than at the time of indexing, has been the main form of database access since the 1950s. Reasons for the failure of postcoordinate searches include the absence of specified relationships between terms, the complexity of formulating Boolean searches, and the high frequency of terms in large databases. Recent writers on indexing electronic text have called for precoordination to enhance the precision of retrieval. Among precoordinate indexing structures, a book index with coined modifications is the most precise. The time and cost associated with such customized analysis will, however, limit its application in the electronic environment.

In summarizing the history of information science, Robert Hayes notes that the concept of postcoordination underlies all computer-based information retrieval systems.[1] Postcoordination, the combination of terms at the searching stage rather than at the indexing stage, was viewed as a major advance in the field of information science; it constituted a rejection of traditional methods of organization of information in libraries: precoordinate subject heading systems and hierarchical classification.[2]

Among the advantages claimed for postcoordination is freedom from the restrictive word order of traditional indexing systems (deciding, for example, whether the preferred sequence is Topic–Place or Place–Topic), and from the possibility of placing a topic in only one position in a library classification scheme.

From the perspective of the maintenance of controlled vocabularies, postcoordination is advantageous because it allows for a smaller list of authorized terms: many compound terms need not be enumerated, as the components can be combined at the searching stage. For example, a thesaurus in the field of literature could separately enumerate national/linguistic/ethnic adjectives (*American, British, French*) and genre terms (*fiction, poetry, drama*). From these six descriptors, nine compound terms can be synthesized: *British poetry, French drama,* etc., but they need not be stored as such.

All these advantages are familiar to me; I even wrote some of them into the revised American National Standard for thesaurus construction[3] (in my capacity as chair of the committee that developed the standard). Yet postcoordination does not work for me. Two examples of the failure of postcoordinate searches done for me follow. One search was conducted for an academic purpose two decades ago, and the second for a personal information need very recently.

The dissertation search

In the mid-1970s, when I started working on my dissertation, I decided to have an online search done to make sure that all the documents relevant to my topic would be cited. At the time, the only bibliographic database covering library and information science was ERIC (Educational Resource Information Center); it indexes journal articles and unpublished reports.

Two decades ago, end-user searching was unheard of, but I did have input into the search statement in conjunction with the reference librarian. The focus of my doctoral research was *automatic indexing* from *full text*; since the ERIC thesaurus did not then have a descriptor for the latter concept ('Full Text Databases' was added in 1988),[4] the Boolean free-text combination that was searched was 'index? AND full()text'. (For those unfamiliar with DIALOG commands, 'index?' represents the truncated form of *index*, which will match *indexes* and *indexing*; 'full()text' signifies that the two words must be in the specified order and adjacent, i.e., no other words may come between them.)

In that prehistoric era, one did not get a printout of search results immediately; searches were processed in batch mode. The big batch of paper that I ultimately received included 400 false drops of the following type: any abstract which indicated that the *full text* of a document was available and that it had an *index* was retrieved. (It would have been unwise to restrict the search to titles, as this brief element would not necessarily reveal all the documents that dealt with indexing from full text; full-text journal articles were unavailable online at the time and hence abstracts were searched.) This search cost me $90.

The personal search

I have a substantial collection of large hats, which I wear mainly to synagogue on Saturday mornings. Invariably, I am running late when I need to find a hat, but it takes a long time to locate the one I need among the many hats stored individually in cardboard boxes. Transparent acrylic shoeboxes have simplified my life, and I decided to invest in acrylic hatboxes. I knew they existed because I had seen them pictured in a mail-order catalog of assorted containers—for the exorbitant price of $33 each. Instead of paying the middleman's markup, I decided to buy a substantial quantity directly from the manufacturer at wholesale prices.

The primary collection of and subject index to manufacturers' catalogs in the United States is *Thomas Register*.[5] The classification and index of this reference work warrant a thorough analysis, perhaps at another time. In the printed tool, I had the choice of looking through all the entries under the 'acrylic' headings (with numerous cross-references to 'plastic') or all those under the 'boxes' headings, both of which span many pages. Then I thought, What a perfect candidate for a postcoordinate search!

I proceeded to the computer-assisted reference service at my university, and together with the librarian formulated the search statement '(lucite OR acrylic OR plastic) AND hatbox?'. This free-text search yielded zero hits, as did a search with the variant spelling 'hat box?'. My search of the printed tool had revealed the controlled term 'Boxes: Hat' (the index has a cross-reference from 'Hat boxes' to this heading), and using this term with the adjacency operator in the online search yielded two hits, a perfectly acceptable number. (The print display under 'Boxes: Hat' was not lengthy, but the words *lucite*, *acrylic*, or *plastic* did not appear in the descriptive portion of any of the entries. The online records have numerous descriptors for each company.) We printed out the names and addresses of the two companies, and I went on my merry way.

I called both companies and was disappointed to learn that they produce *hatboxes* (from cardboard) as well as *plastic* boxes, but neither produces *plastic hatboxes*! I resolved to declare the failure of postcoordination.

In a previous article, in explaining why standard periodical and database indexing does not serve the researcher, I noted that scholars are generally not seeking a combination of concrete topics, but rather an aspect of a single topic.[6] Here I am claiming that even when a combination of topics is sought, postcoordination often fails.

Isolating the factors

In discussing indexing, it is important not to confuse its many distinct facets. A classic book on isolating factors in the design of indexes is by Jessica Milstead.[7] A recent article emphasizing the importance of not muddling the numerous factors involved in research on database indexing is by Dagobert Soergel.[8]

Many people associate postcoordination with free-text, and the latter with full-text, but all three are independent phenomena. Postcoordination can be, and often is, done on controlled vocabulary terms. Free-text searching can be done on titles and abstracts, not just full-text.

Within free-text searching, a major reason for false drops is *homography*: in natural language, many words have the same spelling but different meanings. My failed searches cannot be attributed to homography or even to the lack of controlled vocabulary, only to the fact that my search terms merely coexisted in the document record; they did not have the desired relationship.

Why postcoordination fails

Problems with postcoordination have been apparent since the technique was first implemented. These problems have been exacerbated, however, with the exponential increase in the size of bibliographic and textual databases, which yield unacceptable numbers of document records to scan and intolerable numbers of false drops.

In the recent literature of library-information science, one finds several statements that precoordination is necessary in the electronic environment and, more specifically, that book index structures would be best. In the following sections I review the reasons for the failure of postcoordination and posit that book index structures allow for the most precise retrieval.

1. *Absence of relationships*—The typical postcoordinate search using Boolean AND specifies only that two terms must co-occur in a document, perhaps in a certain range of proximity, but the nature of the relationship between the terms is not expressed. Both of the unsuccessful searches described above can be explained in this way. As Preschel aptly put it, '. . . the concatenation expresses the topic that the user *hopes* to find in the database, not necessarily a topic that *is* actually in the database.[9]

2. *Complexity*—End-users are not willing to devote the time to master the use of Boolean operators, let alone the subtleties of proximity searching and truncation. Many gateway software packages have been developed to convert a natural language query into a formal search statement, but each package must make the difficult choice of linking search terms automatically by either the AND or the OR operator.[10] Despite the publication of a Common Command Language,[11] it is still necessary to learn numerous search languages to interact with online catalogues, online bibliographic databases, and CD-ROMs. An incorrect command or

the absence of a feature such as truncation may fail to retrieve the records of interest, which remain hidden in the system.

Although Liddy and Jorgensen have shown that book index structure—especially if there are multiple sequences and complex cross-references—is misunderstood even by educated users,[12] the panorama of the printed page reveals the possibilities to the user, and after some fumbling, most do manage to zero in on the entry of interest.

3. *Frequency*—The number of machine-readable databases, the number of records within them, and the amount of machine-readable text have all grown exponentially over the past few decades. A postcoordinate search is therefore likely to yield an excessive number of postings. The user then employs artificial means to reduce the number—for example, limiting the search by publication date,[13] and possibly eliminating relevant documents in the process.

Frequency is a major factor in the development of book indexes. Subheadings are assigned (or retained at the editing stage) when main headings have numerous locators. A professionally compiled book index never overwhelms the user with an unmanageable number of postings.

A variety of techniques is being used to deal with the information overload resulting from postcoordinate searching; the one most in vogue at present is automatic relevance ranking.[14] This is essentially number crunching: documents in which a query term has the highest frequency are displayed first. Marchionini *et al.* have demonstrated that the results of implementing such algorithms are poor.[15]

Calls for precoordination in the electronic environment

In 1991, the Library of Congress held an invitational conference called 'The Future of Subdivisions'. Experts on subject analysis were asked to consider changes to the Library of Congress Subject Headings (LCSH) system. It had been pointed out often that LCSH was designed for card catalogues and that its structure should be changed in the electronic environment.

In the proceedings, Elaine Svenonius made a strong case for precoordination in computerized catalogues, noting the problems of frequency, complexity, and the significance of word order.[16] For example, 'Philosophy —History' (history of philosophy) is different from 'History—Philosophy' (philosophy of history). Combining the two words with Boolean AND eliminates this distinction, and hence many false drops will result upon retrieval.

Nancy Mulvany, in a paper on online help systems, argued that presenting a precoordinate book index structure to users is far more helpful than requiring them to formulate Boolean queries.[17] In her pioneering paper on indexing CD-ROM, Barbara Preschel pointed out that it is possible to have book-index-like structures in this medium, which, unlike continuously updated online databases, is fixed.[18] Yet CD-ROMs housing bibliographic databases generally have the same indexing as their online counterparts: separate descriptors designed for postcoordination. Moreover, CD-ROMs often have interfaces that are radically different from those of the major online vendors; complexity of search commands compounded with the problems of postcoordination portends poor results.

Variety of precoordinate structures

Precoordinate indexing systems are of many types. Separate descriptors may be linked without specification of the relationship between them, or complex coding systems may accomplish such specification. Relational coding is difficult for indexers to apply and for searchers to decipher. Such indexing systems have, over time, been discarded. PRECIS[19] is a prime example.

While slashes, asterisks, and alphanumeric codes are non-intuitive methods of indicating relationships between terms, natural language offers little words, mainly prepositions and conjunctions, that can be very effective for this purpose: *of, by, and, from, on*, etc. The use of such *function words* in string indexing systems is taboo, however, no matter how necessary these words are to clarify relationships. This is especially evident in the *Guide to indexing and cataloging with the Art & Architecture Thesaurus*, which features ambiguous precoordinate strings such as 'tables—wood—bonding—manufacturers'.[20] The intended meaning, 'bonding of wood onto tables by manufacturers', is by no means apparent. Book indexers debate about the filing of function words, but all agree that they should be applied judiciously for the clarification of heading-subheading relationships.

Pragmatic conclusions

In a recent encyclopedia article, Farrow stated that book indexers do not read the text they are analyzing because there is no time.[21] I beg to differ. When I index books, I most certainly do read them—and far more slowly than the typical reader. Each phrase must be considered for indexability; unstated concepts must be expressed as headings; and the best formulation of subheadings carefully considered.

Book index structures are ideal from a user perspective, but are also the most time-consuming to create, and hence are undesirable from an economic perspective. In a review of Fugmann's *Subject analysis and indexing*, Lancaster criticized the author's near total disregard for cost-effectiveness in arguing for carefully controlled indexing.[22] Jim Anderson has stated that not all publications merit detailed human indexing;

works that are not important deserve only automatic analysis.[23]

Machine-readable documents that receive only full-text word indexing are pretty much doomed to oblivion as users despair of wading through thousands of records to locate the specific information of interest. But many, perhaps most, publishers will not be willing to pay for the time-consuming analysis of book-like indexing in the electronic environment.

Precoordinate indexing of machine-readable text will require continuous revision in electronic databases that are constantly updated. In contrast with the common practice in the print environment of deleting subheadings for headings that have only five locators, it may make sense to retain all subheadings in the online environment because of the possibility that text will be added and further differentiation of entries required. Thus Wheeler's advice regarding the coining of modifications for all headings in the initial stages of book indexing becomes relevant to electronic indexes.[24]

In the context of access to library collections, I have presented a theory of relativity for cataloguers, arguing that neither name headings nor subject headings are permanent.[25] Reindexing of bibliographic databases is rarely done, except for the substitution of modern synonyms for obsolete or deprecated terms. As the amount of electronic text grows, however, we will have to refine the indexing to show how new documents or passages differ from older ones, in order to assist the user in making a selection from the vast quantity of machine-readable text. Coined modifications, not standard subdivisions, are required for this purpose.

Perhaps we can succeed in getting the message across to the information industry that precoordination is essential for access to the information superhighway. If so, the future of book indexers will be secure, regardless of the dominant medium of publication in the next millennium.

References

1. Hayes, Robert M. Information science and librarianship. *Encyclopedia of library history*, ed. Wayne A. Wiegand and Donald G. Davis, Jr. New York: Garland, 1994, 277.
2. Doyle, Lauren B. *Information retrieval and processing*. Los Angeles: Melville Publishing, 1975, 173.
3. National Information Standards Organization. *Guidelines for the construction, format, and management of monolingual thesauri: An American National Standard*. Bethesda, MD: NISO Press, 1994. (ANSI/NISO Z39.19–1993.)
4. *Thesaurus of ERIC descriptors*. 12th ed. Phoenix, AZ: Oryx Press, 1990, 106.
5. *Thomas register of American manufacturers and Thomas register catalog file*. 82nd ed. New York: Thomas Pub. Co., 1992.
6. Weinberg, Bella Hass. Why indexing fails the researcher. *The Indexer* **16**(1) April 1988, 3–6.
7. Milstead, Jessica. *Subject access systems: Alternatives in design*. Orlando: Academic Press, 1984.
8. Soergel, Dagobert. Indexing and retrieval performance: The logical evidence. *Journal of the American Society for Information Science* **45**(8), Sept. 1994, 589–99.
9. Preschel, Barbara M. Indexing for print, online, and CD-ROM. In *Indexing: The state of our knowledge and the state of our ignorance: Proceedings of the 20th Annual Meeting of the American Society of Indexers*, New York, 1988, ed. Bella Hass Weinberg. Medford, NJ: Learned Information, 1989, 55.
10. Benson, James A. and Weinberg, Bella Hass, eds. *Gateway software and natural language interfaces: Options for online searching*. Ann Arbor, MI: Pierian Press, 1988.
11. National Information Standards Organization. *Common command language for online interactive information retrieval*. Bethesda, MD: NISO Press, 1994. (ANSI/NISO Z39.58–1992).
12. Liddy, Elizabeth D. and Jorgensen, Corinne L. Reality check! Book index characteristics that facilitate information access. In *Indexing, providing access to information: Looking back, looking ahead: The Proceedings of the 25th Annual Meeting of the American Society of Indexers*. Nancy C. Mulvany, ed. Port Aransas, TX: ASI, 1993, 125–38.
13. Weinberg, Bella Hass and Cunningham, Julie A. Online search strategy and term frequency statistics. In *Productivity in the information age: Proceedings of the 46th ASIS Annual Meeting*, Washington, DC, **20**, 1983, 32–5.
14. Koll, Mathew B. Automatic relevance ranking: A searcher's complement to indexing. In *Indexing, providing access to information: Looking back, looking ahead: The Proceedings of the 25th Annual Meeting of the American Society of Indexers*. Nancy C. Mulvany, ed. Port Aransas, TX: ASI, 1993, 55–60.
15. Marchionini, Gary, Barlow, Diane and Hill, Linda. Extending retrieval strategies to networked environments: Old ways, new ways, and a critical look at WAIS. *Journal of the American Society for Information Science* **45**(8) Sept. 1994, 561–4.
16. Svenonius, Elaine. Proposal #2 [The expanded use of free-floating subdivisions in the Library of Congress Subject Headings system]: Arguments in favor. In *The Future of subdivisions in the Library of Congress Subject Headings System: Report from the Subject Subdivisions Conference, 1991*, ed. Martha O'Hara Conway. Washington, DC: Library of Congress, Cataloging Distribution Service, 1992, 36–8.
17. Mulvany, Nancy. Online help systems: A multimedia indexing opportunity. In *Challenges in indexing electronic text and images*, ed. Raya Fidel *et al.* Medford, NJ: Learned Information for the American Society for Information Science, 1994, 91–101.
18. Preschel, 58–9.
19. Austin, Derek. *PRECIS: A manual of concept analysis and subject indexing*. London: The British Library, 1984. In a personal communication, Hans Wellisch informed me that just about all the institutions that had adopted PRECIS have since dropped it. A recent article in this journal reports on the use of a simplified version: Jacobs, Christine and Arsenault, Clement. Words can't describe it: Streamlining PRECIS just for laughs! *The Indexer* **19**(2) Oct. 1994, 88–92.
20. *Guide to indexing and cataloging with the Art &*

Architecture Thesaurus, ed. Toni Petersen and Patricia J. Barnett. New York: Oxford University Press, 1994, 46. This point was made in my review of the *Art & Architecture Thesaurus* and its *Guide*, forthcoming in *Journal of the American Society for Information Science*, March 1995.
21. Farrow, John. Indexing as a cognitive process. *Encyclopedia of library and information science* **53**, supplement 16, 1994, 159.
22. Lancaster, F. W. Review of: Robert Fugmann. *Subject analysis and indexing: Theoretical foundation and practical advice. Journal of Documentation* **50**(2) June 1994, 150.
23. Anderson, James D. Standards for indexing: Revising the American National Standard guidelines Z39.4. *Journal of the American Society for Information Science* **45**(8) Sept. 1994, 632.
24. Wheeler, Martha Thorne. *Indexing: Principles, rules and examples*. 5th ed. Albany: The New York State Library, University of the State of New York, 1957, 18.
25. Weinberg, Bella Hass. A theory of relativity for catalogers. In *Cataloging heresy: Challenging the standard bibliographic product: Proceedings of the Congress for Librarians, 1991, St. John's University, New York*, ed. Bella Hass Weinberg. Medford, NJ: Learned Information, 1992, 7–11.

Dr Bella Hass Weinberg, a Past President of the American Society of Indexers, is a Professor in the Division of Library and Information Science, St John's University, Jamaica, New York.

More on the ISBN

An article in *The Indexer* in April 1992 described the use and importance of the International Standard Book Number (ISBN) in publishing.[1] Now Shane O'Neill, Head of Professional and Reference Publishing at Macmillan Press, in *LOGOS* discusses the bibliographical aspect of books, emphasising the increasing need to systematize information about the book trade and the importance of identifying individual titles.[2]

There have been bibliographies since at least 1777, but modern bibliography is based on the work of three publishers: the Whitakers, Frederick Leypoldt and Richard Roger Bowker. The person most prominent in Britain was Joseph Whitaker, who produced *The Bookseller* (1858) giving information about forthcoming books and also *The Reference Catalogue of Current Literature* (1874) which became *Books in Print* in 1948, and originally consisted of the catalogues of 135 publishers and an index.

In the USA Leypoldt was working on similar lines. *The Publishers Trade List Annual* came out in 1873 and the same year saw the introduction of *Publishers Weekly*. Bowker was involved with Leypoldt in the production of the latter and also produced an index to the former. The production of what is now (American) *Books in Print*, frequently known simply as *Bowker*, was in Bowker's mind, but its projected cost prevented its publication until after his death. In more specialized fields were the H. W. Wilson Co's *Cumulative Book Index* and *Humanities Index* and Ulrich's *Periodicals Directory*. Australia, New Zealand and Canada also produced their own versions of *Books in Print*.

The production of such lists and of national catalogues (a distinction between them being made in Britain, US and Germany, but not in most other countries) made it essential that some sort of identifier for each title be made. After various ideas had been put forward the 10-digit ISBN evolved. J. Whitaker and Sons Ltd became the British ISBN Agency and Bowker its American counterpart. The adoption of the ISBN, and the fact that it is encompassed with the 13-digit European Article Number and the American Article Number barcode system, allow it to be used for point-of-sale systems.

Although Whitaker and Bowker are the two original firms involved in book trade bibliography there are also some serious rivals. All aim at the same market, and the reception of data electronically means that much of the work can be done instantaneously.

PHILIP BRADLEY

References

1. Bradley, Philip. Book numbering: the importance of the ISBN. *The Indexer* **18** (1) April 1992, 25–6.
2. O'Neill, Shane. Bibliographical publishers: Historic pioneers, contemporary innovators. *LOGOS* **5** (2) 1994, 76–85.

An unusual review of Tristan Smith

Reviewing Peter Carey's novel, *The unusual life of Tristan Smith* (Faber, 1994), in the *Sunday Telegraph* (4 Sept. 1994), David Robinson concludes, 'A glossary at the end elucidates the more obscure jokes, but what sort of novel requires an index?'

The answer to that is, of course, any novel in which readers may wish to locate specific passages or collate dispersed references to the same theme. But the question itself raises a puzzle—this novel in fact has no index. The six-page glossary at the end of the 414-page text is followed by four blank pages. Does Mr Robinson mean that this novel *should* have had an index—in which case he has answered his own question? Or does he think 'index' just another term for 'glossary'?