The Ah!-factor

Specifications for index compilation are usually given in terms of categories: 'Please include all names of people and places, book and periodical titles, and major concepts if there's room'. Warnings to the publisher that the index as composed according to the prescription is likely to exceed the allotted length result in the withdrawal of categories—'Okay, omit the book titles and plant species'.

Conscientious listing of all members of given categories must result in a non-negotiable total. Selection of entries to include in an index, though, usually has to reconcile two separate aims: to occupy the minimum necessary space, constrained by paper signatures and printing economics; and to include all those entries likely to be useful to the reader. This second criterion depends on the significance of the reference in the text, not on the category of entry (unless in a technical manual or dictionary). So rather than specifying classes for entry, like Noah's passenger list, we should postulate a grid through which our entries must pass; those of minor interest will be filtered out, greater ones will be retained, irrespective of subject. According to the length allotted for the index, the admission-qualification level will be greater or smaller. But the criterion is of textual importance, not category of entry. We recognize that 'he was as fat as Napoleon' or 'she wished she were the Princess of Wales' do not merit the inclusion of those names in the index, even though they fit within the specified category; a different principle rules. We seek to gauge calibre, not kind.

Thorough listing of all instances of specified types, indeed, can cause much extra work for indexers, first assembling the total mass whose length is unforeseeable, then excising entries to reduce the whole to the prescribed length. One hopes that such victims of indexing by category are paid by the hour, not according to the length of the visible product.

Selection by category is classification talk, not the textual analysis required for indexing. Librarians divide concepts into classes and deal with them accordingly; indexers rather study the individual texts and assess each element for the degree of importance it holds. A secondary instruction for indexing is to select only what is 'crucial, significant or pertinent. These adjectives are left undefined, though used repeatedly'.[1] The principle of choice here is subjective, depending on the indexer/reader's response to the text—and such response is difficult to quantify. It can only be sensitivity to the degree of our own reaction, awareness of relevance, of interest in its dictionary sense of immediate concern.

Ah . . . !

Perhaps we might rate these according to an 'Ah!'-factor. Thus, to more than a full page describing the role of Napoleon in the history of France, we would respond, 'Ah, yes!', and unhesitatingly enter his name, according it the full dignity of a main heading, perhaps to be later added to or subdivided. Half a page in Napoleon's biography describing his childhood, a recognizable and isolatable topic, we would greet, 'Ah!', as meriting a suitable separate subhead under his main entry. A series of minor references to a background figure who reappeared a few times during Napoleon's career provokes each time only an 'Mm' (Mm for mention) and warrants an unadorned page reference only. 'My uncle was as fat as Napoleon' we greet with a shrug and class as an aside, inadmissible in context. A really significant passage affecting several separate characters or themes could achieve an 'Ah-*ha*!' and merit several separate entries in the index under different headings.

Our index specification, then, becomes: 'Include all "Ah, yes!"es, of course, and the "Ah!"s; but you may have to leave out the "Mm"s, or give the "Ah!"s without subheadings'. Categories—places, books, plants, songs—become irrelevant; it is what is said that counts, not merely what is referred to (differentiated by Weinberg as 'comment' and 'topic').[2] Never mind the category, feel the grid. Variations in the space accorded the index—reductions from 8% to 5% of the text, or 5% to 3%—mean we adjust our interest admission level accordingly, screening out the least reaction-provocative remaining entries rather than ditching a further specific category.

A list of sub-Mms, not worth entering themselves, such as the names of the members of an orchestra or types of flowers in a garden, may together constitute one Ah! as a generalized subhead: 'Hertford Symphony Orchestra/members'.

It is the difference between a series of 'Mm's and a sustained 'Ah!' that determines the choice between a sequence of consecutive page numbers (9, 10, 11, 12, 13) and a continuous 9–13.

Computers, of course, are incapable of any sensitivity to the Ah!-factor, but can only adjudge the presence/absence of words.

In defence of strings

A corollary to the assessment of significance of an item rather than its category to determine entry in the index, is that the grade of entry then given it—main heading, sub, sub-sub, or undifferentiated page reference, and whether it stands alone or shares the reference with how many others—must also be consistent according to the Ah!-factor. If we have devised a subhead, 'at school', and then accumulate six page-references which fit exactly under that, not in themselves demanding further description or differentiation, then we let them stand as a run of

six. To hone the description more particularly would indicate a more important reference than any of them is; we are not according subheadings on the principle of subtraction from a total string of minor references, but on intrinsic virtue only. Recurrent characters need to be listed and retrievable; but if they never elicit an 'Ah!' to bestow a subheading, then a run of twelve page-references truly conveys a series of 'Mm' entries. They lie within that stratum of mentions that should not be excluded from an index, but are not important enough to waste space-consuming words on. We must acknowledge that this grade exists, and that thereby hangs many a string. We not only have a system for subjective analysis of weight of entry, but display it according to the same code.

Similarly, if three separate characters are merely present at a wedding, rating 'Mm' only, each receives only a page number for the occasion. If two of them have only two further page numbers each, making satisfactory complete entries of name and three page references, but one character has another eight 'Mm' references—then he must have a string of nine, not a falsely inflated 'at MC's wedding' to reduce the number at the expense of truth to significance, making it appear comparatively that he did more at the wedding than the two fellow-guests, instead of, in fact, doing as much more times in the book.

The main flaw in this system of grading the weight of references is the number of indexes that fail to follow it, where an undifferentiated string of numbers simply shows that an indexer has been lazy, or starved of page space, or relied on a computer flagging terms to merge references automatically, and left it at that. An undifferentiated string could—should—indicate a deliberate evaluation for each of 'Mm'.

If we compile our indexes according to the Ah!-factor, then, contemplating one whose entries have been assessed and admitted by significance rather than category, and where each is accorded a weight of entry in space and words that corresponds to its worth in the text, we may contentedly sigh, 'Ah—yes!'

References

1. Hyman, Richard Joseph. Indexes for analysis and diagnosis. *The Indexer* **13** (3) April 1983, 177–80.
2. Weinberg, Bella Hass. Why indexing fails the researcher. *The Indexer* **16** (1) April 1988, 3–6.

HAZEL K. BELL

Tiny, shiny and terrific! CD-ROM round-up

CD-ROMs (Compact Disks—Read Only Memory) and their future are much featured in the current trade press. The *NFAIS Newsletter* for November 1990 offers three articles on the subject: 'When is the price right? CD-ROM enters new territory'; 'CD-ROM and the next wave of technology'; 'CD-ROM—the information breakthrough'. The second, by Martin Brooks of Bowker Electronic Publishing, gives a clear basic description of the software:[1]

'Five years ago it was impossible to imagine placing a single 4.75" shiny disk in a computer and having access to over 550 megabytes of data... and that these data would be accessible using a relatively easy-to-use interface that did not require knowledge of cryptic commands and syntax.'

More and more databases are now being placed on CD-ROM; and established printed publications may be reissued in this alternative, enhanced form. Over 1,400 commercial titles have been released in CD-ROM versions, delivering large amounts of data conveniently at relatively low cost. CD-ROM developers and publishers offer unique schemes for indexing, sorting and providing access to data, claims Brooks.

'The amount of space available in the CD-ROM medium has enabled publishers to create products with almost unlimited indexing capabilities, allowing users to access data in ways never before imaginable or, at the very least, never before practical.' Details are given of the indexing of the CD version of Bowker's *Books in Print*: 'access by author, ISBN, keyword (any word of the title, author or subject type), Library of Congress catalog number, publisher, subject type, Ingram [a trade book distributor] title code, title, 4–4 author/title, series title, title key, audience, grade, illustration, language, price, and publication year. A Boolean search combining elements like "find all computer books that are illustrated for grades 6 and up that were published after 1988" can customize retrieval further. The CD-ROM software displays results in about 20 seconds.'

Priscilla Oakeshott gives us the statistics.[2] In February 1990 there were over 800 CD-ROMs on the market (each with a capacity of 10,000 pages), mainly conversions of existing databases or large full-text reference works. The emphasis is on business, medicine and science, though there are disks in all the 'online' subject areas: LISA, CAB, Sociological Abstracts, etc. Most are starting with the backfiles only and planning to offer regular updates. Although most of the CD-ROMs so far announced are new formats of existing online files, some are developing