Some indexing decisions in the Cambridge encyclopedia family

David Crystal

The publishing history of *The Cambridge encyclopedia*, and its associated projects, is reviewed, as a perspective for discussing some of the indexing issues involved. The paper explains the reasoning behind the extensive index in *The Cambridge factfinder*, describes the automatic indexing of the encyclopedia database, and discusses the relative merits of word-by-word and letter-by-letter indexing in this genre.

When in 1986 I was asked whether I would be interested in editing a new general encyclopedia, I enquired of the publishers what their expectations were about the job, for I was by no means sure that I was qualified. I never had a satisfactory answer, other than the observation, made in all seriousness, that it helped to have written an encyclopedia already (the allusion was to my *Cambridge encyclopedia of language*, then at proof stage). Nearly a decade later, I think I could now write a fairly detailed job specification—and I know that indexing awareness would be a critical element in it.

An excursion into publishing history

What began as a single enterprise has now grown, like the proverbial Topsy, into a whole family of encyclopedias. I find it difficult myself to keep track of the various editions and impressions, so perhaps a summary will be helpful, before proceeding to discuss some of the indexing issues I have encountered in the family's five-year history. But before I can give this summary, I need to take a short trip down a side road in the history of British publishing.

The first book, *The Cambridge encyclopedia*, was published in 1990, and the name itself demands a gloss. As its preface states, the project was conceived by Cambridge University Press, but the editorial development and production were in the hands of W & R Chambers. Why? Because in the mid-1980s there was a proposal to launch a new joint imprint—an interesting collaboration between two prestigious firms: Chambers, with a long early history of large encyclopedia publishing (but nothing from them in that genre since the fourth edition of 1935);[1] and CUP, with a long history of reference publishing, but no tradition of general encyclopedia production. Although several books were published under the joint imprint, the venture lasted only until 1989, when Chambers fell under the wing of Groupe de la Cité, and there was an amicable divorce. The various planned joint projects were shared between the two companies, and the encyclopedia went to Cambridge. That is why it is called *The Cambridge encyclopedia*, and not *Chambridge* or some other.

This story is not solely of historical interest. The encyclopedia was planned (as all such projects are nowadays) as an electronically stored database, and located in Holyhead, where the editorial office is. The database was compiled from a mixture of sources. For people and places I had available material used by the fourth edition of the *Chambers biographical dictionary* (with some of its ongoing revisions) and by the Chambers–Cambridge *World gazetteer*. One of my first jobs, accordingly, was to make a selection from these sources of the kind of people-and-places data which should form part of the envisaged general encyclopedia database. The new encyclopedia had its own criteria for coverage, such as its concern to deal with international issues properly. This meant, for example, that the heavily Scots bias of the original Chambers would need to be eliminated, and various lacunae (such as important personalities in South Africa and Australia) dealt with. Topics were written from scratch, bringing together a team of over a hundred contributors who submitted entries on their specialisms, and who also reviewed the accuracy and up-to-dateness of the biographical entries.

The idea was always that this database, as it grew, would be able to service the needs of a variety of reference books. An entry on a particular country, for example, would contain a great deal of compartmentalized information, including a complete list of its political leaders, and this could be edited for use in many different ways. (As an illustration, the unabridged *Cambridge encyclopedia* gives all 20th-century leaders, the *Concise* only those since World War 2, the *Factfinder* only those since 1990, and the *Biographical* everyone since the country began.) While the joint imprint was in operation, this notion presented no difficulty; however, once the divorce took place, certain complications emerged.

The database continued to be used by both firms in a variety of ways, but under certain constraints to avoid identical books eventually being produced using

the same material. Such an outcome seemed unlikely, given the different publishing directions in which the two firms were moving, but the safeguards had to be there. So, for example, Chambers used the database to help inform their *Book of facts* (1992) and several of their smaller-scale reference publications, such as the 'Making Sense of' series, as well as their *Encyclopedic dictionary* (1994). The exploitation of the database by CUP I review below.

As time passes, of course, the common material in the original database will become a smaller proportion of the whole. The way Chambers is now developing its copy of the original database will be very different from CUP's way. Both firms are adding new material and revising entries, but without joint consultation. As a result, the two databases have already begun to diverge, just as languages do when they leave a common source. However (and again as in the history of languages) many publications over the next few years will betray their common origins, and it will be important for readers to appreciate what the different books are trying to do, if they are not to be misled by the residual similarities. This is only likely in the biographical domain, where both publishers have produced major works—notably the *Chambers biographical dictionary* and the *Cambridge biographical encyclopedia*—two books which are as different as chalk from cheese.[2] A superficial response will note the similarities between many entries, and draw all kinds of erratic conclusions (see, for example, a letter in *The Bookseller*, 23 September 1994, and my reply, 28 October).

The Cambridge family

There are at present five members of the Cambridge family.

- *The Cambridge encyclopedia*. First published in 1990, this is a 1500-page book, containing some 30,000 entries, organized alphabetically, but with thematically-structured Ready Reference and Colour sections. It went through two corrected and updated impressions, in 1991 and 1992—but these were no cosmetic exercises, given the massive changes which were affecting international affairs during those years. A completely revised second edition, in a larger format, and containing some 36,000 entries, appeared in 1994.
- The *Cambridge concise encyclopedia*. First published in early 1992, this is a book of just under a thousand pages, containing some 19,000 entries. It went through a corrected and updated reprint later in 1992 and again in 1994. A completely revised second edition, derived from the second edition of the big encyclopedia, will appear in mid-1995.
- The *Cambridge paperback encyclopedia*. This was a paperback edition of the *Concise*, first published in 1993, and updated in 1994. Its next edition will be mid-1995.
- The *Cambridge factfinder*. This was essentially a massive amplification of the Ready Reference approach used in the big encyclopedia, containing a great deal of tabular and almanac-style information, in 843 pages, including a 125-page index. First published in 1993, the book was revised and updated in 1994, and a new edition is planned for 1995.
- The *Cambridge biographical encyclopedia*. First published in 1994, this 1300-page book covered some 25,000 people. It added an unabridged Ready Reference section and a thematic dimension to the alphabetical approach which characterizes this genre, capturing the new emphasis in a change of title from 'dictionary' to 'encyclopedia'. A *Concise biographical* is in preparation for 1996, and will contain an index of occupations and roles, organized thematically, which could also be used in relation to the larger book.

Each of these books is now subject to an annual update publication schedule—an exercise much facilitated by the way the database is organized and indexed (see below).

Indexing decisions: the Factfinder *case*

An operation of this kind is totally dependent for its success on efficient information retrieval—though efficiency is a somewhat distant notion, given the limited evidence about how and why people actually use encyclopedias. Most of the information in the above books is organized alphabetically, and presents us with standard questions of indexing procedure. However, in the case of the *Factfinder*, which is thematically organized, a more fundamental issue emerged.

The factbook genre is wide-ranging, including such compilations as almanacs, record books, year books, and gazetteers. (I exclude here those books which use a term such as *factbook* or *factfinder* in the title, but which are actually encyclopedias, such as the *Hutchinson factfinder concise encyclopedia*.) Alphabetical order is not usually helpful, except within small sections. Information is presented thematically, and the conceptual organization adopted is usually summarized in an introductory table of contents. This may be accompanied by some level of indexing, but often it is not. For example, the 410-page *Guinness Sport Yearbook* (1994) organizes its sports alphabetically, but the various competitions and competitors within each category are given no further guide. The 960-page 1993 *World Almanac* is totally thematic, with several sections organized alphabetically, and has only a preliminary 27-page index of a very general kind. The 718-page Chambers *Book of facts*, (1992) is also totally thematic, with many of its sections organized alphabetically: this has only a table of contents followed by a 7-page alphabetical listing of the general categories already given in the contents.

The problem in using this genre is that it presupposes a relatively high level of conceptual awareness.

It seems to assume that only people who already know something about football will wish to consult the section on football, and that they will probably want to read (or skim) the whole of that section anyway, so there is therefore no need to provide a more detailed breakdown of the information it contains. This is probably a fair assumption for a factbook which is thematically restricted, such as on sport or politics; but the wider the range of a book, the less sensible this assumption becomes; and in the case of a work of general reference, it is nonsensical. In short, the broader a work of reference, the more essential an index.

The kind of problem which arises can be seen from the following example (which happens to be a real-life instance, relating to a school homework query). Using the Chambers book, I needed to look up *seagull*. The contents pointed me to the section on birds (86-99), where I found lists of birds listed alphabetically within 10 types (flightless, birds of prey, songbirds, etc.). I flipped through these looking for the category within which seagulls would probably be located, and found *seabirds*. *Seagull* was not there. I worked through the other categories and eventually found it, but listed as *gull*, and under *shorebirds*. I had no idea that a *sea* gull was a *shore* bird. There are two problems, both of which would have been solved by an appropriately detailed index, including a cross-reference (from *seagull* to *gull*). It took me a few minutes to find the entry: it should have taken a few seconds.

At least in this case my general knowledge allowed me to focus immediately on the category of birds. If the enquiry had been about a particular writer, and I had been unsure whether he was a poet, a novelist, a playwright, or whatever, I would have had no alternative but to thumb through the various sections until I found my target. And as there is never any guarantee that a factbook will have included a particular fact, the pressure is on the user to continue looking through wider and wider sections of the work, until eventually throwing it aside in frustration.

There is no alternative, it seems to me, but to provide a factbook with an index which tries to anticipate all the enquiries its users are likely to make—in principle, every person, place, animal, natural history category, and so on. The length of the index should be governed by the thematic coverage of the book, and it is likely to take up a larger proportion of the whole work than in the case of a single-subject book. When the present project began, I expected the index to take up about 10%, and advised the publisher accordingly. In the event, the 125 pages of 4-column index which emerged took up 18%—after negotiation with the commissioning editor, who was slightly concerned that, if allowed my head, he would end up with too long an index tail wagging too short a maintext dog. For example, I decided not to index the maps, or the individual titles used as illustrations of an author's work. The result is a letter-by-letter index (apart from *St* and *Mac*), with most entries consisting of a single page reference, though in a number of instances I have included clusters of references, to help those who wish to browse. The index in no way replaces the table of contents, which presents a 10-page hierarchical classification and thus allows a top-down approach useful for those who want an overview of a subject area; rather, it acts as an essential complement.

Automatic indexing

The remainder of this paper deals with the Cambridge encyclopedia database as a whole. The first stage in the operation is a level of automatic indexing, an essential step provided by our database management system, INMAGIC.[3] This system consists of two parts: the database design (the *data structure*) and the information contained in the database. The data structure specifies each unit of information which is included in the database, using a list of *fields*, each of which holds a specific type of information. For example, the fields of the encyclopedia database include one for pronunciation, one for birth/death dates, one for the maintext of an entry, one for the headword specification, and so on. Each record in the database (each encyclopedia entry, in our case) is organized using the same set of fields—though of course not every field is needed for every entry (e.g. the birth/death field is needed for biographical entries, not for topics). Unlike some database management systems, INMAGIC does not limit the number of records in a database, the size of records or fields, or the number of entries in fields. This lack of constraint on the size of the database was the chief reason for choosing INMAGIC for the encyclopedia project.

Another important factor in choosing INMAGIC was its efficient indexing system. It was essential that we should have a rapid way of retrieving any phrase, word, or word element from any part of the database, as literally any item in a record might need correction or updating at any time. Also, the need for consistency checking throughout the database as a whole (e.g. ensuring that all instances of *mediaeval* are spelled in the same way, or enabling us to find all entries which talk about *lasers*) requires a total indexing operation. Basically, we want every word and word stem indexed. INMAGIC allows us to do this, taking a fair amount of time to do the indexing (for large numbers of new entries or changes within entries, we let it index after working hours), but afterwards offering extremely rapid search times. It takes less than a second to find all the entries containing any one word in our two-million-word database, and to make them ready for display on screen in sequence.

Up to fifty fields in a data structure can be indexed, and one makes a decision at the outset about which fields should have an index. (The chief reason for not

indexing everything is to save disk space. If a field is left unindexed, it can still be searched, but it takes much longer—in our case, some two minutes to go through the whole database looking for a word.) INMAGIC offers four options for indexing a field: it may be given a *term index* (which contains up to the first 59 significant characters of each entry in a field); it may be given a *keyword index* (which contains every word in all the field entries, regardless of the field's length); it may have both term and keyword indexes (which gives maximum flexibility, but uses most disk space); or it may have no index. (You can change your decision at a later stage, and re-index a field, though this is something which we have rarely had to do.) A word is defined as a series of consecutive, significant, alphanumeric characters separated by one or more non-significant characters. The significant characters are defined by the sort code (see below) for the field: numbers and letters are significant; spaces are not; and there is a choice available as to whether punctuation is indexed. The indexes can be displayed on screen, or written to a file and printed.

Sort codes are INMAGIC filing patterns which meet different search needs. There are nine such codes available, most of which ignore upper/lower case and punctuation. Sort codes can be changed at a later date, and we have done so (see below). The chief alternatives are word-by-word vs letter-by-letter filing, and numeric filing of numbers vs alphabetic filing. With alphabetic sorting, INMAGIC compares terms from left to right, one character at a time, with the shorter term sorted first. In alphabetic sorting of numbers, 110 sorts before 20 because 1 comes before 2. It is possible to exclude leading articles (the word which appears at the beginning of a field entry, typically *the* or *a*, up to a limit of 127 characters) and a certain number of stop words (words which you do not want the system to index, regardless of case or punctuation, typically such words as *and, the, of, to, der, gli*, up to a limit of 255 characters). Table 1 shows the characteristics of the nine sort codes, and Table 2 illustrates keyword and term indexes using sort code 1 for a small sample of data.

Table 1 Characteristics of sort codes

Characteristic	sort codes								
	1	2	3	4	5	6	7	8	9
Dates	-	-	-	Y	-	-	-	-	-
Word-by-word	-	-	-	-	Y	-	Y	-	Y
Numbers numeric	Y	Y	-	-	-	-	Y	-	Y
Case/punctuation significant	-	-	-	-	-	Y	-	-	-
Leading articles ignored	Y	-	Y	-	Y	-	-	-	Y
Stop words ignored	Y	-	Y	-	Y	-	-	-	Y
UDC	-	-	-	-	-	-	-	Y	-

- Sort code 1: sorts letter-by-letter, ignoring punctuation, case, and spaces; numbers sort in numerical order; leading articles and stop words are ignored.
- Sort code 2: same as 1, but leading articles and stop words are not ignored.
- Sort code 3: same as 1, but numbers are sorted in alphabetical order.
- Sort code 4: used for date fields, ensuring that they are indexed for chronological sorting, in ascending order.
- Sort code 5: sorts word-by-word; consecutive punctuation and other non-alphanumeric characters are treated as a single space; numbers are sorted alphabetically, and leading articles and stop words are ignored.
- Sort code 6: sorts according to the ASCII collating sequence: it is letter-by-letter, alphanumeric, with numbers sorted as characters, and all punctuation and case distinctions significant.
- Sort code 7: same as 5, but numbers are treated numerically, and leading articles and stop codes are not ignored.
- Sort code 8: sorts according to the Universal Decimal Classification System, as far as possible.
- Sort code 9: sorts word-by-word, but in all other respects is the same as 1.

Table 2 A sample of data organized using sort code 1

Publication 2	Numeric filing: 2 before 13
Publication 13	
Publication 13.5	Letter-by-letter: 13.5 before 13-6
Publication 13-6	
Publication #3	
PUBLICATION #500	
Public policy	Letter-by-letter: PUBLICA before PUBLICP
A public's right to know	Leading articles not signifcant

Term Index Listing

1	PUBLICATION 2
1	PUBLICATION 13
1	PUBLICATION 13.5
1	PUBLICATION 13-6
1	PUBLICATION3
1	PUBLICATION500
1	PULBICPOLICY
1	PUBLICSRIGHTTOKNOW

In choosing which sort code to use for the database, certain options could be ruled out straight away as largely irrelevant to our concerns (4 and 8). Distinguishing case and punctuation is not likely to be very important (6), nor are the different approaches to the sorting of numbers (2 vs 3, 5 vs 9). We are left with two major questions, very familiar to indexers: do we go for letter-by-letter (1/2) or word-by-word (5/7), and within these categories, do we index (2, 7) or ignore (1, 5) leading articles and stop codes?

We can deal with the latter point briefly. Indexing grammatical words is important if there are many entries of the type *A-1678* or abbreviations such as *IN* (Indiana) or *in* (inches). A general encyclopedia has a fair number of these, in such cases as *The Hague* and in various abbreviations (e.g. the states of the USA, used in every American biographical entry, or measurements in natural history). We have the disk space to include all these items. The choice therefore resolves into the time-honoured question of 2 vs 7—word-by-word or letter-by-letter.

Word-by-word?

I was initially in no doubt that word-by-word indexing would be the most suitable method for a general encyclopedia, and assigned sort code 7 to every field except pronunciation (for which I used 6, as it contained several special symbols) and the birth/death field (specifying years only, for which I used 1). An encyclopedia deals, first and foremost, with concepts, and it therefore seemed sensible to use an indexing system which is semantically grounded, i.e. recognizing the basic status of the word. In this way, I reasoned, *bee* and *bee dancing* would be located together (rather than being separated by, say, *beech trees*, *Thomas Beecham*, and the various members of the remarkable *Beecher* family); and the 40-odd entries which begin with *sea* would fall together (*sea bass*, *sea cucumber*, *sea pen*, *sea slug*, etc.), rather than be arbitrarily separated by such entries as *Sealyham terriers*, *Ronald Searle*, and *Seattle*. There are many such cases where it was certainly elegant and possibly functional to use semantic groupings. It seemed likely that readers would appreciate the convenience of seeing juxtaposed what semantically belongs together, and that the browsing factor (which is an important feature of a general encyclopedia) would be fostered more in this way.

For the second edition, I was equally in no doubt that this had been the wrong decision. The new edition opts throughout for sort code 2 (with just one exception, in a field where we need to check on our consistent use of punctuation, and where we therefore employ sort code 6). What motivated this fundamental change? The initial impetus came from reader reactions. The Preface to the first edition welcomed reader responses to the book, and we got a substantial postbag as a consequence. Comments about treatment are not relevant here, but those relating to coverage are. Many people drew our attention to a perceived omission, and in most cases this was fair comment. But in several cases the omission was misconceived: the reader was unable to find the entry, though it *was* present in the book, and in all these cases the problem was due to the word-by-word ordering. I had first-hand experience of this myself on a number of occasions, choosing the wrong location for my first point of call. Of course, any failure to find an entry which I *knew* was in the book would immediately remind me of what the issue was, and I would automatically go to the alternative location. Many general readers, evidently, do not do this (which means that they have not consulted or understood the indexing guidelines in the front matter).

Why is word-by-word ordering such a problem, in a work of this type? The issue is essentially a linguistic one, to do with the problems of identifying a word within compound forms—of which there are thousands, in a general encyclopedia. In written English, there are two, related difficulties: change in written language conventions, and vacillation over present-day usage. A compound lexical item may in theory take any of three forms: spaced, hyphenated, or solid. Thus we will encounter *flowerpot*, *flower-pot*, and *flower pot*; *masterclass*, *master-class*, and *master class*. There is no principled way of predicting which it will be, though there are certain tendencies.

- The more familiar a new compound form becomes, the greater the trend to write it solid: a hypothetical neologism written as *drivel pump* is likely to be a newer concept than if it is written as *drivel-pump*; and *drivelpump* (along with its possible derivatives, such as *drivelpumping* and *drivelpumper*) suggests the concept has achieved real familiarity.
- The existence of other forms of a certain type can also influence a neologism: thus when *market-maker* came into use in the late 1980s, it was immediately hyphenated, presumably on analogy with the many other forms which already existed (*home-maker*, etc.).
- Potential for grammatical ambiguity can also make one form preferable, and many publishers formalize this, such as by hyphenating attributive forms but not predicative ones (*wear a white collar*, but *a white-collar worker*).
- There may also be regional variations, notably between American and British English: American English uses far fewer hyphenated forms than British English, and will opt for solid forms in many cases where the latter prefers some degree of separation (*hardworking* vs *hard-working*).

In short, there is no guarantee that, if you encounter a compound word in what you read, and wish to look it up, you will find it printed in your reference work using the same spacing convention. This is presumably why one reader complained that he could not find *sea gull* (given by us as *seagull*) and another that he could not find *sea lion* (given by us as *sea-lion*).

But why did these readers not find these items? Because their working assumption is that A–Z reference entries are organized letter-by-letter. The readers went to the place in the alphabet where they expected the item to be, encountered one series of printed items, did not find the item they were looking for, and

promptly gave up. Notice that in both the above cases, the readers had gone to the list of spaced lexical items first, and not looked for another list of solidly printed items. There are also cases of failed look-up happening the other way round, with someone going to a solidly printed list, and not finding a spaced item. The person who did not find *Belloc, Hilaire* had presumably looked between *bell-flower* and *bell-ringing*, and failed to see the item ten entries (over a column) further on. *A-level* is separated from its other possible alphabetical location by some 26 pages.

The alphabetic principle is instilled at a very early age, becomes the mainstay of dictionaries, and is the default value for any A–Z work. People assume an encyclopedia is going to be organized like a dictionary. Actually, there is also evidence that many readers do not draw a clear distinction between an encyclopedia and a dictionary anyway. In the *Datasearch* service we offer (in which readers fill in an enquiry form that enables them to interrogate our database to find all entries in which a chosen word is located), we regularly get people asking us dictionary questions—enquiring about the etymology of a word, for example, or asking for a definition. And with the growth of encyclopedic dictionaries (most recently, the 1994 *Chambers encyclopedic dictionary*), I would expect this distinction to become increasingly blurred.[4]

There were several other arbitrarinesses imposed by the word-by-word convention which caused editorial irritation. *Mazo de La Roche* appears five pages away from *Paul Delaroche*; *Marie Madeleine La Fayette* three pages away from *Marie Joseph Lafayette*. The approach also assumes that readers will be able to recognize 'words' in a foreign language, and thus know to look at the very beginning of letter *T* for *Ta-t'ung*, and at the beginning of *I* for *I Ching*. It similarly assumes a knowledge of abbreviations, so that *V-1* is found at the beginning of letter *V*, and *i.t.a.* at the beginning of letter *I*. All of these points are familiar to indexers, and to any users of other word-by-word encyclopedias, such as the Britannica *Micropedia*. After three years of living with these problems, we made a decision to change to letter-by-letter for the second edition.

Was there any evidence of people using other organizational principles than letter-by-letter? In only two cases—again, both familiar to indexers. Perhaps because of the influence of telephone directories, people assume that all *Mac* (*Mc, M'*) names are the same. To separate them on a letter-by-letter basis causes immediate confusion. Similarly, *St* is treated as *Saint*. These points are of course unrecognized in the INMAGIC sort codes, and we have to take special steps to ensure that entries beginning with these items appear in the right place. In the revised edition, these are the two exceptions to the letter-by-letter rule.

How fundamental was the change? How much of the encyclopedia has been affected by the alteration? How many entries actually moved? Given the universal range of subject-matter included in a general encyclopedia, this figure might be of some interest. I arrived at an estimate using the following steps.

The work contains approximately 25,000 entries in the A–Z section, of which some 75% have single-word headings. Of the multi-word headings, most have such a lengthy initial element (as in *Californian Indians*) that the possibility of an alternative indexing position is unlikely to arise. Even in those cases where the first element is short, only a minority of the entries are affected: for example, only 20 of the 60 entries beginning with *sea* actually end up with different neighbours in the second edition; in the *black(-)* series (e.g. *black box, blackbird*), the figure is 27 out of 69; and in the *blue(-)* series (e.g. *blue book, bluebird*) it is 14 out of 29—the highest proportion in the book. But these clusters of entries are exceptional. In the vast majority of cases, the number of entries which come together in a word-by-word approach, preceding a sequence of solid items, is less than 5.

Table 3 shows the situation, for the A–B section of the book (15% of the whole). There were a total of 324 spaced items located ahead of solid items. Of these, 75 (23%) stay in the same place in both editions. Of the remainder, calculating how many move their location between editions depends on the criterion employed. For example, in the word-by-word sequence of *Ark of the Covenant, ark shell, Arkansas*, and *Arkwright*, we could say that two entries move (the first two items jumping forwards over the third) or that only one entry moves (the third item jumping backwards over the first two). If we use the stronger criterion, we find a total of 249 entries (77%) jumping at least one place as a result of the new indexing principle. Extrapolating to the A–Z section as a whole, this suggests that around 1700 items could be said to be affected—but this is still only 7%. Using the weaker criterion, the figure is likely to be about 5%.

I have not found another case of an encyclopedia changing its indexing principle between editions, and I would be very interested to learn of other instances. The two editions, side by side, might make an interesting corpus for indexing research. For my part, I await with interest the first crop of readers' letters about the new edition. I confidently expect there to be a drop in false coverage complaints, and if not, I shall have to think again. I shall keep readers of *The Indexer* posted.

Notes

1. For the early history of encyclopedia publishing, see Crystal, David. On editing a modern Cyclop(a)edia. *Proceedings of the Royal Institution* **64**, 1992, 245–70.
2. Not so the US edition of the *Chambers Biographical Dictionary*, which until 1993 was published there by CUP under the title of the *Cambridge Biographical Dictionary*.

Table 3 The number of spaced multi-word items (e.g. black box) occurring before solid items (e.g. blackbird) in the A–B section of The Cambridge Encyclopedia[a]

A: No. of entries in a cluster[b]	1	2	3	4	5	. . .	14	18	42	Total
B: No. of instances of each cluster	81	25	19	8	6		1 (*blood*)	1 (*blue*)	1 (*black*)	142
No. of entries affected (A × B)	81	50	57	32	30		14	18	42	324
No. of entries jumping at least one place[b]	51	33	51	27	22		9	17	39	249

[a] Excluded are regal titles and other bynames (e.g. *Alexander III*, *Abbas the Great*), which are by convention located first in an alphabetical sequence. Hyphenated prefixes (*anti-*, *etc.*) are also not taken into account, as these are treated as if the items were solid.

[b] For example, in the sequence *access course*, *access time*, *accessory*, we see a cluster containing two spaced items (line 1 of the Table), one of which moves (line 4 of the Table).

At least one US reviewer of the latter, in *Wilson Library Bulletin* (October 1994, p. 78), was misled into thinking that the latter was a revision of the former.

3. Strictly, the software model is a 1992 upgrade called INMAGIC Plus. Further information about INMAGIC is available from INMAGIC Inc, 2067 Massachusetts Avenue, Cambridge, MA 02140-1338, USA.
4. It was never a rigid distinction anyway, outside Britain. US dictionaries have long contained some encyclopedic information (typically, about people and places), and this is seen also in the Continental tradition represented by such families as Larousse and Düden.

David Crystal is Honorary Professor of Linguistics at the University of Wales, Bangor; editor of the journals Linguistics Abstracts *and* Child Language, Teaching and Therapy, *and of the Blackwell* Language Library *series as well as of the Cambridge family of encyclopedias. He is President of the Society of Indexers.*

Just what we always suspected

David Ellis *et al.*: 'On the creation of hypertext links in full-text documents: measurement of inter-linker consistency', in *Journal of Documentation*, **50**(2), June 1994, 67–98: this is an extremely difficult technical article which applies 'arithmetic coefficients and topological indices' to measure the degree of similarity between different linkers' work on the same texts. It transpires (surprise, surprise) that there is as much variation between linkers [sc. button-placers or 'indexers'] as is found with 'normal' indexing.

Indexers in fiction—another

Candida Crewe, in her novel, *Mad about bees* (Heinemann, 1991) adds to the number of rather low-key, perhaps old-fashioned, indexers we have noted in works of fiction:

Liza was working away in that scatty style she found so winning. Papers, manuscripts, dictionaries had grown up around her like a miniature building site. And she was settled in the middle of it all, unfazed by her self-made chaos. Between her lips a cigarette had been left to its own devices, while she concentrated on typing at her manual typewriter. One finger was going at the keys with the frenzy of a woodpecker's beak laying into a bit of bark. . . .

Liza was employed by a number of publishers in a freelance capacity. She compiled indexes, mainly of biographies and history books. It was the perfect job for her. She could do it at home, and take on as many or as few titles as she chose. In the school holidays (her children's and her husband's) she would cut down considerably. The pay was pretty paltry, but she was not poor. . . .

'This book seems to be taking longer than most,' Liza said in answer to her husband's question concerning her work. 'Every paragraph, practically, has at least ten new references, all of which have to be included.'

'Index'll be as long as the text if you're not careful,' Samuel remarked.

'Very likely.'

'You're too conscientious. Maybe you don't have to log absolutely everything.'

Liza smiled. 'Don't you worry. I'm very selective. But even so the whole thing takes for ever.'

(Reprinted by permission of William Heinemann Ltd.)