

Paper presented at Electronic Age 1996 A conference for the information industry 20-21 April 1996 Ranelagh House, Robertson NSW. To be printed in the AusSI Newsletter and LASIE 27(3):45-49

## **Beyond free text searching**

**Garry Cousins**

© 1996 Garry Cousins.

In early November 1995 Macquarie Library Pty Ltd , publishers of the Macquarie Dictionary, asked me to consider taking on the twin tasks of proofreading the scanned text of Manning Clark's six-volume History of Australia , and indexing it, their plan being to issue the text of Clark's History as an indexed CD-ROM. Over the next two months I met with the publisher, editor and computer programmers at Macquarie to discuss specifications for the project, and also prepared some estimates. I began indexing in earnest in mid-January 1996 and at the time of writing (April 1996) I am just over half-way.

My contribution to this work-in-progress is one part of a team effort, which also includes the contribution of a publisher, an editor, production assistants, and computer programmers. Today I won't be talking about big questions like the overall presentation of CD-ROM indexes, or the design details of search engines for CD-ROMs, but rather of matters which relate to my brief for this project, which is restricted to the tasks of proofreading and coding the text with index entries. In particular, I want to tell you how the execution of these tasks for a CD-ROM have entailed some departures from normal book indexing practice. Of course, the very first departure from book indexing was the fact that I was doing all my work on-screen: the text was on-screen, not on page proofs.

In several preliminary meetings we discussed the logistics of proofreading and indexing the History . It was decided I would proofread the disks against the original hard copy and make corrections directly on-screen. The logistics of indexing were more complicated.

### **Free-text searching**

Many text-based CD-ROMs already exist in the market which use free-text searching, with varying sophistication, as the sole means of information retrieval. To the credit of the publisher at Macquarie, Richard Tardif, it was taken for granted in this project that free-text searching is not an efficient way to navigate one's way precisely around a text of any size, particularly when one is searching for conceptual information as well as simple names. Relying wholly on a text-search facility which can only locate literal strings of text has several serious drawbacks for searchers:

1. a concept which is not mentioned literally is overlooked, even though the subject may be discussed at length
2. the search criteria must match the text exactly. For example if you are searching a text online for Mozart's date of death a Boolean search combining the elements "Mozart" and "death" will miss a piece of text which runs: '... and so Mozart died in 1791', even though that passage contains the crucial information
3. a very successful search may oblige the searcher to scan dozens or even hundreds of entries, with no clue as to which aspect of the topic each refers.

Because of these shortcomings Macquarie did not want to rely solely on free-text searching, but wanted a comprehensive subject index, with a cross-reference structure which would take into account variations in vocabulary likely to be used by searchers. Searchers needed to be able to enter a topic discussed but not necessarily mentioned literally in the text, and have the search engine respond either with instances of text which discussed the topic, or a reference to related segments of text which did. However, we conceded that some types of information in the text could be retrieved quite efficiently using free-text searching, and did not need to be coded. When the programmers at Macquarie write the search engine for this

CD-ROM they will be writing what will actually be a hybrid of a text-search capability and a text-linked index.

### **Choice of terms**

It was simply not practical financially to code everything in the text: subjects, dates, personal names, corporate names and placenames. Some compromises had to be made. We decided it was essential that I make entries for:

1. all subjects (e.g. Aborigines, convicts, emancipation, gold rushes, transportation, etc.)
2. all decades (e.g. 1820s)

The terms chosen for the subjects, and the decades, would be typed in at the beginning of each paragraph in which the subject or decade was referred to. These terms and dates would be coded for the programmer by enclosing them in angle brackets, like a sort of pseudo SGML. For example, <1830s> or <emancipation>. The paragraph was our reference unit, although in many respects an arbitrary one.

We agreed I would not index:

1. personal names
2. placenames
3. corporate names
4. specific years (eg. 1854)

This decision has been slightly modified in the course of its application. We decided that most personal names could be searched very well using ordinary text searching; for example, if John O'Donohue is only mentioned three times in the whole work, text searching with good software on a fast machine can find the mentions efficiently. But in order for it to work there were two provisos: there had to be relatively few homographs in the text, and names being searched could only be mentioned once or a handful of times in the entire text.

A search of the printed indexes to the six hardcover volumes revealed that homographs were not numerous (e.g. 11 in the 390 pages of volume 1). Although it would mean the searcher would occasionally have to discriminate between different persons, or places and persons with the same name, we decided the risk of confusion was small, and that homographs were not a big obstacle to simple text searching for personal names.

Names which were mentioned often, however, called for different treatment. Text searches for frequently mentioned names are likely to be too successful: where does one start if a simple text search for William Wentworth returns 300 hits? Names such as these obviously had to be treated as they are in book indexes: they had to be subdivided into their various aspects.

This entailed the second major departure from book indexing: frequently mentioned persons were to be coded as index entries, but infrequently-mentioned persons were not. The latter were to be found by simple text-search, but the former were to be indexed/coded with a subheading. It will be the programmers' job to merge these two access points in the CD-ROM's search engine, so that when a searcher types in a name, the program will first check the differentiated, human-created list to see if it is a frequently-used name. If it is, the searcher will be shown the subtopics relating to that person; if not, the program will proceed with a simple text-search for the name.

To help me decide which names to include in the differentiated list, I consult the indexes to the hard-copy volumes. I decided to use the same rule-of-thumb that applies in book indexing, namely that a heading should be subdivided once more than 7 or 8 references accumulate. If a name in the hard-copy index had more than this number of references I provide subheadings as soon as I begin to code the name. I soon got used to consulting the

hard-copy index each time a new name appeared, but it was hard to get used to coding some names while letting others flow by untouched, so to speak.

Some compromises were needed in order to meet budget: it was decided that placenames and corporate names could also be located by free-text searching, and not subdivided. This is not a problem for the vast majority of placenames or corporate names mentioned only once or a few times, but makes searching for a handful of key places like Sydney and Van Diemen's Land difficult to do with precision. Rather the searcher needs to make their search specific. For example, to find information about the introduction of gas lighting in Sydney you would search first for gas-lighting, not Sydney. One good spin-off from this compromise is that it has forced me to index more specifically.

Budget constraints also forced us to accept some compromise with regard to indexing dates. Originally it had been hoped that dates could be coded down to the specificity of a day, but this proved far too expensive. We settled on coding decades and centuries, leaving individual dates to be found by free-text searching. This has drawbacks more apparent in some sections of the History than others. It is quite workable in , say, volume I which covers a period spanning from the 14th century to the 19th century, but in volume II, which covers only sixteen years, it is of limited use.

These compromises regarding names and dates notwithstanding, the primary purpose of the index was to provide subject access to the text. We decided to index all subjects, regardless of whether or not the actual topic or subject name appears in the text. Taking the paragraph as the basic unit, a keyword enclosed in angle brackets would be inserted, or embedded, at the beginning of each paragraph in which the topic was mentioned. If the treatment lasted for more than one paragraph, the coded keyword would be embedded at the beginning of each paragraph, until the discussion stopped. If a particular subject was mentioned often, the specific nature of the mention would be pinpointed with a subheading; for example, <transportation: abolition of>.

### **Materials and working methods**

Macquarie has provided the text of the History in hard-copy (some 2500 separate pages) and as 19 disks containing 94 Microsoft Word files, which have been produced using a scanner. The scanning job is very good and although the detail in footnotes suffered a little, the copy is quite clean.

The six already existing indexes to the individual hard-cover volumes are of little direct use: the references to page numbers mean little once the text is up on screen as one long scrollable document, 2500 pages or 5000 screens long. But they have proved very useful as a means of ascertaining in advance names which will require subheadings.

Now that I have been working for some time I have settled into a routine: I proofread a file, usually a chapter long, on-screen first, with the original hard copy by my side as the master copy. I soon learnt that the scanner made some regular mistakes like translating the letters "cl" as "d", or "in" as the letter "m", or replacing em dashes with hyphens or en dashes, and have devised a list of such errors which I look for as a matter of course each time I open a new file. I make corrections directly on-screen.

Then I start coding the text. After reading a paragraph I decide on appropriate dates and keywords and insert them in angle brackets at the beginning of the paragraph. These coded terms will, of course, be hidden from the reader in the final product, which is just as well, because they can occupy considerable bulk. A paragraph might have seven or eight keywords, often with subheadings, attached to it, so that the coding runs for several lines. The paragraph beginning might also carry one or more dates; if several dates were mentioned in the paragraph, several dates would be coded and inserted: <1820s><1830s><1840s><1850s> etc. I have included a sample of coded text at the end of this paper.

It is essential to keep a thesaurus or authority list to maintain consistency in the choice of keywords. I decided to use a dedicated indexing program, CINDEXTM, to compile my authority list. Because I am coding the text in Microsoft Word for Windows I can also have CINDEXTM open in a window simultaneously, so that after inserting a keyword in the text, I can copy it to the Clipboard, switch to CINDEXTM with one keystroke (Alt-Tab), and paste the entry into the authority list. I just have to be careful to duplicate every term when I am inserting five or six keywords in the text. Although they won't be in the final subject index, I include file numbers in the page field of the records in the authority list, so that when I have to go back and edit a heading I can locate the relevant file quickly. So a record might look like this:

>population

>in Van Diemen's Land (1850)

P>43

In this record, "43" refers not to a page, but to File 43 (which contains, say, pages 180-197 from chapter 9 of volume 3).

The authority list also includes see and see also references, made in exactly the same way as for book indexing. On the final product I imagine the searcher will click on the cross-reference in order to bring up the related or preferred topic.

Generally the 94 files correspond to chapters, but the match is not always perfect, so I also use CINDEXTM to compile a chapter list with corresponding file numbers and pagination. For example:

File 1 = vol 1, chapt 1 (pp 1-24)

File 2 = vol 1, chapt 2 (pp 25-29) ...

File 41 = vol 3, chapt 7 (pp 140-160)

File 42 = vol 3, chapt 8/9 (pp 161-179)

File 43 = vol 3, chapt 9 (pp 180-197)

File 44 = vol 3, chapt 10/11 (pp 198-216)

File 45 = vol 3, chapt 11 (pp 217-239) etc.

This has proved to be of considerable help in navigating my way around the files when editing.

### **Sample of coding in one paragraph from Manning Clark's History of Australia**

<1840s><1850s><moral campaigns><women: moral protection of><Grey, George: recommends resumption of immigration><immigration: to South Australia><labour: employment agencies>No other colony besides New South Wales produced a woman of the stature, single-mindedness or industry of Mrs Chisholm, but in all the others reception committees, immigration officers and philanthropists laboured for the protection of the immigrants' morals and encouraged the growth of those virtues of self-reliance, industry, purity and family affection so dear to the heart of Mrs Chisholm. In Melbourne there was a Ladies' Female Immigrant Society, presided over by the head of the Anglican Church, a fine example of benevolent usefulness, and a most necessary antidote to the rottenness, sloth and moral evils to which the migrants too often succumbed because of the monotony of their long journey. In South Australia, after Grey recommended its resumption in January 1843, assisted immigration resumed with such a bang that 35 per cent of all assisted migrants going to the Australian colonies were sent to South Australia. There a Benevolent and Strangers' Friend Society administered relief to the needy and promoted the moral and spiritual welfare of immigrants. Its secretary, a Mr Maguire, activated by much more exalted principles than gain, and a seasoner of all his work with true Christian humility, placed unmarried females in homes and found employment for the afflicted and disconsolate who

were far from the land of their birth. There was also a Colonial Labour Committee which assumed the responsibility of finding employment for members of both sexes and of influencing both employers and workers to uphold agreements, taking care not to interfere with the price of labour but requiring master and man to make their own terms. In that colony the zeal to protect the morals of female immigrants reached such a pitch by the beginning of 1851 that the first mate on the Joseph Soames had 5 pounds deducted from his gratuity for speaking to the female immigrants, notwithstanding the remonstrances of the surgeon on the ship not to do so, and despite testimony that in all other ways his conduct had been decorous and proper. In Hobart Town in Franklin's day the wife of the Lieutenant-Governor, the wife of the head of the Anglican Church and the wife of the Chief Justice appointed themselves guardians of the material and moral welfare of migrants.

### **Bio**

Garry Cousins qualified as a librarian but has been working as a freelance book indexer since 1988, mostly in the social sciences and law. Garry teaches indexing at the University of New South Wales and Macquarie University , and has also taught indexing to editors at LBC Information Services. He is the Australian editor of Brief Entry, a newsletter for law indexers. Garry was accepted as a registered member of the Australian Society of Indexers in 1990 and was the inaugural New South Wales branch president of the Society from 1990 to 1992.