Indexing in the Electronic Age

20-21 April 1996
Robertson, NSW

Table of contents

- Program

Papers

The full series of papers were published by *LASIE* in September and December 1996

- Automatic indexing and abstracting, by Glenda Browne
- Conceptual indexing for CD-ROMs: beyond free text searching, by Garry Cousins
- Future indexing developments in WORLD 1, by Sandra Henderson
- Indexing the Internet : pinning jelly to the wall?, by Roxanne Missingham
- Encyclopaedia of Aboriginal Australia: how is it indexed?, by Geraldine Triffitt

Electronic Age '96 A conference for the information industry

20-21 April 1996

Ranelagh House, Robertson NSW

Day One Saturday, 20 April 1996

Key Issues

- The electronic age: key issues for indexers - Lynn Farkas, Datascape Information
- Indexing the Internet : pinning jelly to the wall? - Roxanne Missingham, Div. of Wildlife and Ecology, CSIRO

Current Initiatives: A Multimedia Viewpoint

Case study presentations on indexing for multimedia and CD-ROMs, by representatives of organisations which have produced such products

- Helen Routh, CCH Australia Ltd
- Lindsay Parsons, Scantext;
- Richard Barber, ACEL
- Geraldine Triffitt for David Horton, Australian Institute of Aboriginal and Torres Strait Islanders Studies - The Encyclopaedia of Aboriginal Australia (CD-ROM)

Current Initiatives: Publishers' Viewpoints

Publishers' panel on the role of electronic indexing in their organisations, and the impact on publishers' use of indexers

- Paul Mullins, AGPS;
- Helen Routh, CCH Australia Ltd;
- Richard Barber, ACEL
- Evan Predavec, Butterworths

Current Initiatives: An Indexer's Viewpoint

- Conceptual indexing on CD-ROMs: beyond free-text searching - Garry Cousins
- Informal demonstrations and displays in Conference Exhibition area

Conference Dinner

Awarding the Australian Society of Indexers medal - ASI Medal Committee

Day Two Sunday, 21 April 1996

Internet and its Impact

- What contribution can indexing make to the Internet? - Tony Barry, Australian National University
- Indexing the Web: an exercise in hypertext navigation - Dwight Walker, Australian Society of Indexers
- Converting indexed information to the Internet: a case study of a dictionary conversion - David Nathan, Australian Institute of Aboriginal and Torres Strait Islanders Studies

Automating the Indexing Process

- Indexing registry and word processed documents - Peter Eden, RIMS Australasia

- Using fuzzy retrieval and relevance ranking in library catalogues - Rick Clark, Contec Data Systems
- Automatic indexing and abstracting - Glenda Browne
- Future indexing developments for World 1 - Sandra Henderson, National Library of Australia

Paper presented at Electronic Age '96 A conference for the information industry 20-21 April 1996 Ranelagh House, Robertson NSW. To be published in Online Currents, the AusSI Newsletter 20(6):4-9, July 1996 and LASIE 27(3):58-65

Automatic indexing

Glenda Browne

Introduction

This paper will examine developments in automatic indexing and abstracting in which the computer creates the index and abstract, with little or no human intervention. The emphasis is on practical applications, rather than theoretical studies. This paper does not cover computer- *aided* indexing, in which computers enhance the work of human indexers, or indexing of the Internet.

Research into automatic indexing and abstracting has been progressing since the late 1950's. Early reports claimed success, but practical applications have been limited. Computer indexing and abstracting are now being used commercially, with prospects for further use in the future. The history of automatic indexing and abstracting is well covered by Lancaster (1991).

Database indexing

Extraction indexing

The simplest method for indexing articles for bibliographic databases is **extraction indexing** , in which terms are extracted from the text of the article for inclusion in the index. The frequency of words in the article is determined, and the words which are found most often are included in the index. Alternatively, the words which occur most often in the article compared to their occurrence in the rest of the database, or in normal language, are included. This method can also take into account word stems (so that *run* and *running* are recognised as referring to the same concept), and can recognise phrases as well as single words.

Computer extraction indexing is more consistent than human extraction indexing. However, most human indexing is not simple extraction indexing, but is **assignment indexing** , in which the terms used in the index are not necessarily those found in the text.

Assignment indexing

For assignment indexing, the computer has a thesaurus , or controlled vocabulary, which lists all the subject headings which may be used in the index. For each of these subject headings it also has a list of **profile words** . These are words which, when found in the text of the article, indicate that the thesaurus term should be allocated.

For example, for the thesaurus term *childbirth* , the profile might include the words: *childbirth, birth, labor, labour, delivery, forceps, baby* , and *born* . As well as the profile, the computer also has **criteria for inclusion** -- instructions as to how often, and in what combination, the profile words must be present for that thesaurus term to be allocated.

The criteria might say, for example, that if the word *childbirth* is found ten times in an article, then the thesaurus term *childbirth* will be allocated. However if the word *delivery* is found ten times in an article, this in itself is not enough to warrant allocation of the term *childbirth* , as *delivery* could be referring to other subjects such as mail delivery. The criteria in this case would specify that the term delivery must occur a certain number of times, along with one or more of the other terms in the profile.

Computer database indexing in practice

In practice in database indexing, there is a continuum of use of computers, from no computer at all to fully automatic indexing.

- No computer.
- Computer clerical support, e.g. for data entry.
- Computer quality control, e.g. checking that all index terms are valid thesaurus terms.
- Computer intellectual assistance, e.g. helping with term choice and weighting.
- Automatic indexing (Hodge 1994).

Most database producers use computers at a number of different steps along this continuum. At the moment, however, automatic indexing is only ever used for a **part** of a database, for example, for a specific subject, access point, or document type.

Automatic indexing is used by the Defense Technology Information Center (DTIC) for the management-related literature in its database; it is used by FIZ Karlsruhe for indexing chemical names; it was used until 1992 by the Russian International Centre for Scientific and Technical Information (ICSTI) for its Russian language materials; and it was used by INSPEC for the re-indexing of its backfiles to new standards (Hodge 1994).

BIOSIS (Biological Abstracts) uses computers at all steps on the continuum, and uses automatic indexing in a number of areas. Title keywords are mapped by computer to the Semantic Vocabulary of 15,000 words; the terms from the Semantic Vocabulary are then mapped to one of 600 Concept Headings (that is, subject headings which describe the broad subject area of a document; Lancaster 1991).

The version of BIOSIS Previews available on the database host STN International uses automatic indexing to allocate Chemical Abstracts Service Registry Numbers to articles to describe the chemicals, drugs, enzymes and biosequences discussed in the article. The codes are allocated without human review, but a human operator spends five hours per month maintaining authority files and rules (Hodge 1994).

Retrieval and ranking tools

There are two sides to the information retrieval process: documents must be **indexed** (by humans or computers) to describe their subject content; and documents must be **retrieved** using retrieval software and appropriate search statements. Retrieval and ranking tools include those used with bibliographic databases, the 'indexes' used on the Internet, and personal computer software packages such as Personal Librarian (Koll 1993). Some programs, such as ISYS, are specialised for the fast retrieval of search words.

In theory these are complementary approaches, and both are needed for optimal retrieval. In practice, however, especially with documents in full-text databases, indexing is often omitted, and the retrieval software is relied on instead.

For these documents, which will not be indexed, it is important to ensure the best possible access. To accomplish this, the authors of the documents must be aware of the searching methods which will be used to retrieve them. Authors must use appropriate keywords throughout the text, and ensure that keywords are included in the title and section headings, as these are often given priority by retrieval and ranking tools (Sunter 1995).

The process whereby the creators of documents structure them to enhance retrieval is known as bottom-up indexing . A role for professional indexers in bottom-up indexing is as guides and trainers to document authors (Locke 1993).

One reason that automatic indexing may be unsuited to book indexing is that book indexes are not usually available electronically, and cannot be used in conjunction with powerful search software (Mulvany and Milstead 1994).

Document abstracting

Computers abstract documents (that is, condense their text) by searching for high frequency words in the text, and then selecting sentences in which clusters of these high frequency words occur. These sentences are then used in the order in which they appear in the text to make up the abstract. Flow can be improved by adding extra sentences (for example, if a sentence begins with 'Hence' or 'However' the previous sentence can be included as well) but the abstract remains an awkward collection of grammatically unrelated sentences.

To try and show the subject content, weighting can be given to sentences from certain locations in the document (e.g. the introduction) and to sentences containing cue words (e.g. 'finally', which suggests that a conclusion is starting). In addition, an organisation can give a weighting to words which are important to them: a footwear producer, for example, could require that every sentence containing the words *foot* or *shoe* should be included in the abstract.

Computer abstracting works best for documents which are written formally and consistently. It has been used with some success for generating case summaries from the text of legal decisions (Lancaster 1991).

After recent developments in natural language processing by computers, it is now possible for a computer to generate a grammatically correct abstract , in which sentences are modified without loss of meaning.

For example, from the following sentence:

"The need to generate enormous additional amounts of electric power while at the same time protecting the environment is one of the major social and technological problems that our society must solve in the next (sic!) future"

the computer generated the condensed sentence:

"The society must solve in the future the problem of the need to generate power while protecting the environment" (Lancaster 1991). Text summarisation experiments by British Telecom have resulted in useful, readable, abstracts (Farkas 1995).

Book indexing

There are a number of different types of microcomputer based software packages which are used for indexing.

The simplest are **concordance generators** , in which a list of the words found in the document, with the pages they are on, is generated. It is also possible to specify a list of words such that the concordance program will only include words from that list. This method was used to index drafts of the ISO999 indexing standard to help the committee members keep track of rules while the work was in progress (Shuter 1993).

Computer-aided indexing packages , such as Macrex and Cindex , are used by many professional indexers to enhance their work. They enable the indexer to view the index in alphabetical or page number order, can automatically produce various index styles, and save much typing.

Embedded indexing software is available with computer packages such as word processors, PageMaker , and Framemaker . With embedded indexing the document to be indexed is on disk, and the indexer inserts tags into the document to indicate which index terms should be allocated for that page. It does not matter if the document is then changed, as the index tags will move with the part of the document to which they refer. (So if twenty pages are added at the beginning of the document, all of the other text, including the index tags, will move 20 pages further on).

Disadvantages of embedded indexing are that it is time-consuming to do and awkward to edit (Mulvany 1994). Indexers who use embedded indexing often also use a program such as Macrex or Cindex to overcome these problems.

Embedded indexing is commonly used for documents such as computer software manuals which are published in many versions, and which allow very little time for the index to be created after the text has been finalised. With embedded indexing, indexing can start before the final page proofs are ready.

Embedded indexing will probably be used more in the future: for indexing works which are published in a number of formats; for indexing textbooks which are printed on request using only portions of the original textbook or using a combination of sources; and for indexing electronically published works which are continually adapted. In some of these applications the same person may do the work of the editor and indexer.

The most recent development in microcomputer book indexing software is Indexicon (Version 2), an **automatic indexing package** .

Indexicon

Indexicon -- How it works

Indexicon is published by Iconovex , and is available as an add-on program for MS-Word and WordPerfect on IBM-compatible computers, and for MS-Word on the Macintosh. All versions cost US$129. Indexicon 2.0 for MS-Word requires MS-Word for Windows 6.0 or above; a 386 or better CPU (486 recommended); Windows 3.1 and 8 MB RAM (Indexicon Spec Sheet 1996).

To use Indexicon, the book to be indexed must be available electronically in a word processing format. The user chooses from six levels of detail, and Indexicon creates an embedded index at that level using the indexing facility available with MS-Word or WordPerfect. The user can then edit the tagged entries in the original document. Indexicon indexes are subject to all the problems of embedded indexes, including the time-consuming editing process.

Indexicon comes with a primary lexicon containing about 55,000 words, and it allows the user to create specialised lexicons. The lexicons include index values for terms, which indicate how likely it is that a given term should be tagged as an index term. In general this depends on the degree of specialisation of the term, so that a commonly-used term is less likely to be indexed than a more specialised term.

The primary lexicon also includes compound terms, so that pairs of words such as *control tower* , or *remote control* , are always indexed as a pair.

All words in the lexicon also show what part of speech they are. Thus the word *lead* would have two parts indicating that it can mean the metal *lead* or the verb *lead* .

Indexicon will invert proper names if it recognises them, however if the surname carries another meaning (as with the names Brown, Miller, and Young) then the name is not recognised as such. Indexicon can also be set to omit proper names and geographic names if it can recognise them (Iconovex 1996).

Indexicon -- Reviews

Indexicon has been reviewed a number of times. The Indexicon Spec Sheet (1996) on the Internet says: "Indexicon is a tool capable of handling everyone's indexing needs". PC Magazine (13/9/94; quoted on the Indexicon Spec Sheet 1996) says: "With Indexicon, creating an index is as quick and easy as spellchecking".

However, a review of Indexicon Version 1.00b by Mulvany and Milstead (1994) found that it did not live up to the promises on the packaging that it was the "Standard for Indexing" and

could produce "professional quality indexes". In a response to this review, Steven Waldron, President of Iconovex, acknowledged many of the points raised, and stated "The purpose of INDEXICON ... is NOT to replace professional indexers" (Waldron 1994).

The Macintosh version of Indexicon was reviewed by Erfert Fenton (1996). He says: "Before I wrote this review I was skeptical of computer-based indexing programs. Having written it, I'm even more skeptical."

Fenton found that Indexicon missed many terms. When tested on a chapter on Macintosh fonts it missed the terms *pica, em dash* , and *leading* (pronounced *ledding*). It started many terms with adjectives (e.g. *slushy winter roads*) and it included many inappropriate entries (e.g. *Uncle Steve Yahoo*). The reviewer found much evidence of the fact that the computer did not understand what it was reading, and was therefore unable to make valid judgments.

In a test using Indexicon Version 2.0 to index a short article on the use of in vitro fertilization to save tigers from extinction, we identified the following problems:

Indexicon did not recognise and invert any of the names in the article. In one case this was because the person's surname had another meaning (*Ann Miller*); in the other two cases the name appeared in a string with other capitalised words and the whole string was indexed (e.g. *Leslie Johnston of National Zoo*).

Indexicon included some inappropriate entries (e.g. *Biologist's hopes*) and some strange constructions (e.g. *Reproductive tract, Nicole's* -- Nicole is a tiger).

Bengal tiger cubs was indexed in direct order, but *Tigress, Siberian* was inverted. Presumably this is because *Bengal tigers* is included in the lexicon as a compound word.

Indexicon does not generate cross-references so these must be identified and added by the indexer at the editing stage.

Finally, Indexicon did not group terms, so that *Tiger* and *Tigers* were given as separate entries.

In this exercise Indexicon set to the highest level of indexing indexed all important terms; in other experiments which we did many important terms were omitted, while non-significant terms were included.

Indexicon -- Potential uses

Iconovex states that Indexicon is suitable for use with documents which would not otherwise be indexed, and as a first step for professional indexers.

It is currently used to index manuals (e.g. corporate policy and procedure manuals), large contracts and large quantities of e-mail. Technical writers who index their own work have been using it as a first step in indexing.

Among indexers, Indexicon is most likely to be useful for specialists, who are more likely to take the time to create specialised lexicons, and to work with the program to enhance its efficacy in their special field. For journal indexing, where the same indexer works with similar material, in a consistent format, year after year, it might be worth taking the trouble to set up a specialised lexicon, and use Indexicon as a first step. But Indexicon is not good enough at picking key concepts and leaving out worthless ones, to be useful, in general, as an aid to indexing books.

If Indexicon improves, and if the embedded indexing software used in word processing programs improves, it may become more cost-effective to start indexing with Indexicon, and then enhance the index by editing.

As the ability of computer software to recognise personal names develops, it may also become useful as a tool for automatically generating name indexes (Feldman, Lawrence e-mail 15/03/96).

Effect of automatic methods on professionals

As computer programs become more sophisticated, and more information appears in electronic form, there will eventually be less 'traditional' indexing work available. This loss may be balanced in the short-term by an increase in the number of databases and an increase in the number of indexing and abstracting projects attempted. The proportion of freelance versus in-house work may also change.

Humans should still be used for important works, which perhaps can be identified by studying usage and citation patterns (Anderson 1993). Indexers and abstracters will have to become more selective, and decide on the quality of the works they might index and abstract, as well as the subject content.

If we remain better than computers we must show this, and indicate that there are economic returns (to the publisher) and academic returns (to the index or abstract user) from a quality index or abstract.

On the positive side, indexing and abstracting skills will be needed in the development of computer systems, and to check the output from computers. Indexers will be needed to set up and maintain thesauruses, and to train writers as 'bottom-up indexers' so that their work is readily retrievable.

Indexers will have to become entrepreneurial and computer literate. Indexers with skills in the related areas of computing, editing, librarianship and bibliography may be best suited to take advantage of new opportunities. We will have to be able to identify gaps in the organisation of knowledge and to fill those gaps in a commercially effective way. To do this we will have to be computer literate. Not only will we have to know how to use various computer tools for indexing; we will also have to know how information is organised and used electronically, so that we can best understand the needs and make our own contributions.

Acknowledgments

I would like to thank Terry Maguire, language director of Iconovex, the publisher of Indexicon, for a trial copy of the software, and prompt answers to all of my questions. I would also like to thank Jonathan Jermey and Bill Browne for their support and patience while I prepared this talk and paper.

References

Anderson, James D. 1993 ,

Indexing standards: Are they possible? What good are they? Why bother? In: *Indexing, Providing Access to Information: Looking Back, Looking Ahead, Proceedings of the 25th Annual Meeting of the American Society of Indexers,* Alexandria, VA, May 20-22. Port Aransas, Texas: American Society of Indexers.

Farkas, Lynn 1995 ,

Economics and the future of database indexing. In: *Indexers – Partners in Publishing, Proceedings from the First International Conference* , Marysville, Vic, March 31 to April 2. [Melbourne]: Australian Society of Indexers.

Fenton, Erfert 1996 .

Indexicon 1.0: Indexing program for Word 6. Macworld Communications.

Hodge, Gail M. 1994 .

Computer-assisted database indexing: the state-of-the-art. *The Indexer.* Vol. 19, No. 1, pp. 23-27.

Iconovex 1996. *Indexicon 2.0: Automated Indexing for Microsoft Word: User's Guide.* [Bloomington, MN]: Iconovex. Indexicon Spec Sheet 1996.

Koll, Matthew B. 1993.

Automatic relevance ranking: A searcher's complement to indexing. In: *Indexing, Providing Access to Information: Looking Back, Looking Ahead, Proceedings of the 25th Annual Meeting of the American Society of Indexers,* Alexandria, VA, May 20-22. Port Aransas, Texas: American Society of Indexers.

Lancaster F.W. 1991.

Indexing and abstracting in theory and practice. London: Library Association.

Locke, Christopher 1993.

Weaving the Social Fabric: Illuminating Manuscripts. In: *Indexing, Providing Access to Information: Looking Back, Looking Ahead, Proceedings of the 25th Annual Meeting of the American Society of Indexers* , Alexandria, VA, May 20-22. Port Aransas, Texas: American Society of Indexers.

Mulvany, Nancy C. 1994.

Embedded Indexing Software: Users Speak Out. In: *The Changing Landscapes of Indexing, Proceedings of the 26th Annual Meeting of the American Society of Indexers,* San Diego, California, May 13-14. Port Aransas, Texas: American Society of Indexers.

Mulvany, Nancy And Milstead, Jessica 1994.

Indexicon, The Only Fully Automatic Indexer: A Review. *Key Words* , Vol. 2, No. 5, pp. 1, 17-23.

Shuter, Janet 1993.

Standards for indexes: Where do they come from and what use are they? In: *Indexing, Providing Access to Information: Looking Back, Looking Ahead, Proceedings of the 25th Annual Meeting of the American Society of Indexers* , Alexandria, VA, May 20-22. Port Aransas, Texas: American Society of Indexers.

Sunter, Steve 1995.

Humans and computers: partners in indexing. In: *Indexers -- Partners in Publishing, Proceedings from the First International Conference* , Marysville, Vic, March 31 to April 2. [Melbourne]: Australian Society of Indexers.

Waldron, Steven 1994.

Message to INDEX-L@BINGVMB.BITNET on 31/10/94.

Bio

Glenda Browne, PO Box 307 Blaxland NSW Australia 2774, email glendabrowne@optusnet.com.au

Glenda is a freelance indexer with a background in biotechnology and information management. After a stint in sole charge of a small hospital library she moved into library teaching at Mt Druitt TAFE, which she now combines with indexing (in partnership with her husband Jonathan Jermey) and caring for their children. She has indexed articles for the CSIRO Index database, and has created book indexes on subjects ranging from Communication to Chemistry. Glenda's index to *Pharmacology and Drug Information for Nurses* was Highly Recommended in the 1995 AUSSI medal awards.

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

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 seaching; 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 is 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, CINDEX, to compile my authority list. Because I am coding the text in Microsoft Word for Windows I can also have CINDEX open in a window simultaneously, so that after inserting a keyword in the text, I can copy it to the Clipboard, switch to CINDEX 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 CINDEX 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.

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

Future indexing developments in World 1

Sandra Henderson

Manager, National Bibliographic Publications

National Library of Australia

What is WORLD 1?

WORLD 1 is the Australian business name for a new system being developed by the national libraries of Australia and New Zealand, to replace the existing ABN, Ozline, NZBN and Kiwinet services. The development project is known by the name NDIS (National Document and Information System). The National Library of New Zealand is expected to announce in the near future the name under which the service will operate in New Zealand. The project is based in Canberra, with computer company CSC undertaking the development.

WORLD 1 will provide access to information both within and external to the national databases, and to documents of all kinds, not just cataloguing and indexing records. In some cases full text documents and images of documents will be included in WORLD 1, and gateways provided to other services.

Features of WORLD 1

Unlike the current ABN and Ozline services, which offer only single-database searching, WORLD 1 will offer users a choice of single database or cross-database searching, including combinations of cataloguing and indexing records in a single logical view of the database. It will be possible, for example, to search an education topic in APAIS, plus the Australian Education Index, or in a "database" consisting of the relevant index databases as well as cataloguing records in that subject area. A single result set will be provided, containing records from all sources searched.

Gateways will be provided to other information sources, allowing users to access overseas databases and retrieve records and information. In addition, it is planned to mount some large overseas databases in the WORLD 1 system during phase two of the project.

Searching in WORLD 1

The search facilities of WORLD 1 will offer powerful and friendly searching, with a single search system for all users. The users will be able to choose a search level appropriate to their level of expertise, or opt for command language searching. Much attention is being given to the development of appropriate online assistance.

The index databases

At the National Library of Australia we produce two index databases. In each case the process of preparing the indexes includes selection of articles from journals, newspapers and monographs, analysis of content, assignment of subject headings from appropriate thesauri, and the assignment of other access points (such as name headings or additional subject terms). In one case the indexer is also required to write an abstract for the item if there is no abstract in the original item.

APAIS (Australian Public Affairs Information Service) is one of Australia's oldest indexing services, having been in operation since 1945. Items for indexing cover the broad areas of social sciences and humanities, and a total of 11,500 items per annum are indexed. The Library maintains the APAIS Thesaurus, for use with this index. APAIS appears in print (monthly issues January-November, and an annual cumulation), online (the online database

covers 1978 to present) and as one of the databases on the CD-ROM AUSTROM, published by RMIT in cooperation with the Library.

AMI (Australasian Medical Index) was started in 1985 and indexes the Australian journal and conference literature in health and medicine. Approximately 5,000 items per year are indexed, and appear in the online database and on the CD-ROM HealthROM, from the Commonwealth Department of Health and Family Services. The Medical Subject Headings (MeSH) thesaurus from the US National Library of Medicine is used as the thesaurus for this product.

At the present time both indexes are prepared using very old systems, without the capacity to load data other than via direct online entry. In the case of APAIS, there is no capacity for data validation or data correction after loading.

Indexing in WORLD 1

Although work has yet to begin on development of phase two of NDIS, which incorporates indexing, planning is well advanced, and the required functions were specified in the original Request for Tender. WORLD 1 will allow indexers to either work online to add new records immediately to the database, or offline, with upline loading or submission of files of record on disk. For those choosing to work online, as we will at the National Library, there will be workforms with Windows features, pull down menus and pick lists, the ability to cut and paste, and the setting up of links between indexing records and between index and catalogue records. Global changes of data, and online validation of records will be available to aid in rapid loading and maintenance of data.

Authority control will be enhanced, with both MeSH and the APAIS Thesaurus online to be used in data validation and selection of terms. Other authority files, such as lists of journals and names, will also be available for online use. In time it is hoped thesauri for other Ozline databases could also be made available online, as an aid to both indexing and searching.

Data Links

It will be possible to set up a number of links between records. It is advantageous to link journal article records to the national bibliographic database serial records to allow users to locate holdings of desired items. Links to authority records will make it possible to use the various authority files to enhance search retrieval and allow for automatic updating of records as thesaurus terms are updated or amended over time.

Since the start of 1995, most articles indexed for APAIS have been captured as images, and these will be made available online, linked to the corresponding citation. Negotiations with the Copyright Agency Limited have been underway for some time to resolve issues of copyright fees and payment of royalties. An independent company, SilverStream, is making those images available as a set on CD-ROM. A similar project has commenced to image articles indexed for the Australasian Medical Index.

WORLD 1 will also provide for the first time an ability to make modifications to the publication process, to improve the appearance of the printed publication, and offer users a facility for on-demand tailored bibliographic products. Users may wish to opt to produce their own regular bibliographies using a variety of indexing and cataloguing records in a format and output medium which is suited to their requirements.

When will WORLD 1 be available?

At the time of writing the implementation of Phase 1 had been delayed by some difficulties with a third-party software component. Work on other functionality within WORLD 1 has continued, and the system is expected to be launched in early 1997. Phase 2, with indexing and cataloguing functions, is to be available in late 1997.

Further information on WORLD 1

There is an internet WORLD 1 discussion list, which is open to any interested persons.

To join the list, send a message to

listproc@nla.gov.au

with no subject line and 'subscribe world-l firstname lastname' (without the quote marks) in the body of the message.

A newsletter, WORLD 1 connections , is available on application to

The Editor, WORLD 1 Connections,

Services to Libraries Division (Box 9),

National Library of Australia, Canberra ACT 2600.

Enquiries about WORLD 1 can be directed to the Help Desk at the National Library (toll free 1800 026 155) or email market@nla.gov.au

Web Bibliography

ABN (Australian Bibliographic Network)

Paper presented at Electronic Age '96 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):32-42.

Internet Indexing : pinning jelly to the wall?

Roxanne Missingham

"Don't panic" Hitchhikers Guide to the Galaxy (1)

"Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?" (2)

Introduction

We are, without doubt, in the information age. Information technology pervades our work, our homes and our recreation. Each day we use this technology to turn on our heating, cook our meals, produce reports, communicate with others, for entertainment and to obtain information.

Traditional publishing is changing in response to the new technology. Conferences convened by the Australian National Scholarly Communications Forum (3) in the past 3 years have described the move to electronic publishing, particularly on the Internet, by the academic/research and government sectors. This move has had an impact on all the links in the information chain - from the writing of ideas (now done on word processors), to editing and peer review (through word processing and electronic mail), publication (desktop publishing for print and electronic out put) and in locating or finding publications (electronic indexes).

The changes in information technology have an impact on the steps or strategies used to find information for a research inquiry. Research tools such as abstracts and indexes are now online, on CD ROM and on the Internet. Automation of information sources has occurred in two ways. Firstly existing abstracting and indexing services have made their information available electronically. This has enabled easier, quicker searching. Secondly new tools are being developed, such as Internet indexes, which automatically generate a database of information. Retrieval of information from the Internet is much more complex than for publications in print form. This complexity has stimulated the production of automated indexes. Attempts to index Internet information using standard library cataloguing or database indexing methods, seen in Cyberstacks and Netfirst (an OCLC service) are yet to be as successful as automated indexes, partly because of the sheer volume of material to be indexed, and also because of the newness of the medium. The challenges to researchers and information specialists to find information are being addressed through innovative solutions, both automated and manual.

This paper demonstrates that Internet indexing or searching is not yet fully solved. Search tools are critical to effectively accessing information on the Internet, but need to be used with care. A comparison of 7 major indexes (commonly called search engines) finds great differences in search results and effective access to information. Overall trends and issues for the future development of Internet search engines are then discussed.

Why is the Internet important?

The Internet is the most rapidly growing medium of communication and publication. The Internet Domain Survey, of January 1996 found almost 9.5 millions hosts were connected to the Internet, an increase of approximately 142% on the figure for July 1995. Academic, research and government sectors now publish on the Internet.

The survey also found that:

- the net domain is over twice as large as it was six months ago

- about 76000 systems are now named www (i.e are World Wide Web sites), up from only 17000 six months ago
- No one has any clue how many users there are, but most people would agree that there is at least one user per host.

(Network Wizards, 1996)

Estimates of those with access to the Internet range up to 65 million people (Four Corners, February 1995 estimated 35 million, Compuserve and America Online Figures, were quoted as 65 million in In, around and about, September 1995). The original Internet community comprised the university and major research sectors. (AARNET, the Australian Internet segment now owned by Telstra was originally a consortium of the Australian Vice Chancellors Committee and CSIRO). Schools, libraries, corporations and to a increasing extent the general public have joined the Internet. All these groups are using the Internet to publish and to search for information (or "surfing" the "sea" of information on the Internet)

Publishing via the Internet, whether by electronic mail or making information available through gopher, ftp or world wide web has the advantage of being:

- quick (you can communicate or publish apparently "instantly");
- widely accessible (to the very broad Internet community);
- cheap to produce material (many Web mark up tools are free or come with desktop publishing/word processing packages);
- relatively simple to produce (taking much less time than more complex desktop publishing); and
- pretty much in the total control of those publishing (as opposed to the significant controls of publishing houses and traditional editorial control).

Many of the traditional publishing and distribution processes we have relied upon for consistency and control are becoming less important as electronic publishing increases. Internet publishing of scholarly and non-scholarly material is changing forever the way we organise and seek material.

A critical issue in this new publishing medium remains over how, with such a vast number of Internet documents, relevant material can be found. Those publishing on the Internet wish to communicate with an audience and be read, as they did in print. The location of information is thus an important issue to those publishing as well as information seekers.

What is an Internet index?

The Internet consists of many files across the world. There are over 22 million World Wide Web pages, millions of gopher files, applications and databases. To provide a finding service "Internet indexes" or search engines, as they are often known, have developed. These play much the same role as traditional indexes used to find books in the library, or journal articles, conference papers and other publications.

Internet indexes all offer different methods of access (use of search terms) and different information about the files or "publications. The same process is used for the collection and distribution of information as a traditional printed index. Major differences exist in the production of the database used in the index, and its access.

The process for construction of traditional and Internet indexes is:

Traditional Print/database index Selection of sources Creation of Citations Addition to a database /abstracts/indexes End product/access - Print - CD ROM/online

Internet index Sources selected (usually by self nomination) Robots collect information (sometimes called 'harvesting') Database created Online access using a "search engine"

The terminology used by Internet indexes/search engines can be confusing. Common usage of some key terms is:

Search engine: used to describe 1) the actual software which searches the database of Internet files/resources or selected information on the files and also 2) the actual Internet index or complete searching service offered online.

Search: the matching undertaken by the search engine of the terms entered by the user against the terms existing in the database. There are a wide variety of algorithms used for this process.

Index: used to describe the service provided through the Internet search site comprising the information collected and made available through the search software or search engine for queries.

Robot: a program which automatically (usually on a regular basis) searches the Internet for a defined set of sites and collects information about the files at each site. Sometimes it may collect copies of the whole files, generally called "harvesting" after the software, Harvest. Others collect selective information, such as the title, location and first paragraphs.

In this paper the term Internet Indexes is used to describe the search service available, rather than search engines. Index indicates a process of selection, storage and retrieval using the search engine.

Internet indexing challenges

The Internet cannot be considered to be just another step in the history of indexing. It offers huge challenges and needs a very different approach to indexing to enable effective information retrieval. The need for quality access to information has increased, because of the number of searchers. There is also a need for tailored search tools for specific subjects. These tools save searchers time and provide access to reliable, quality information including evaluative comments from professional indexers, or what Internet Indexing services now called "professional web reviewers".

Indexing the Internet offers many challenges:

- the Internet contains millions of documents/files;
- the location of Internet documents/files changes frequently;
- there is no quality control for Internet information, no consistency in use of terminology, or even in the use of titles;
- keeping up with new sources is very difficult; and
- indexes are complicated by the fact that most will only index sites which nominate themselves (somewhat similar to the current cataloguing in publication process).

Internet documents have only two absolute features. The first is a location (where they are stored and can can be found) and the second is content. There are no standards that require authors or title to be used, nor a requirement that heading information include title or sub title information. Internet indexing is quite different to journal article indexing where this information is clear.

Evaluation of indexes

Why do we need to evaluate indexes? There is a multiplicity of indexes for print and electronic information which have been used by researchers and librarians for many years. They have been used because they provide an effective and efficient mechanism for finding information on specific subjects. In commencing to research a topic or inquiry the best search strategies are selected (see Basch). Indexes are used to obtain the best citations (or information lists) as efficiently as possible. Information found needs to be relevant and

accurate to enable futher refinment of the search or to enable the researcher to obtain the information sources listed in the results.

Indexes are finding aids for publications. Different indexes are used for different subjects. Different search strategies are also used. By evaluating indexes it is possible to find which are best suited for different types of research, as well as the best way to use them. These decisions have to be made for the use of traditional printed and automated indexes, and apply equally to Internet indexes.

Traditional indexes, whether in print, CD ROM or online, have been evaluated in terms of:

1. geographical coverage country of publication of material included, for scientific indexes this is usually US/UK/European orientation
2. language English/French etc
3. dates of coverage normally indexes cover material published within specific dates eg within 2 years of the date of the volume
4. discipline (subject area) such as zoological material, biological material
5. scope
6. inclusiveness (breadth) a wide range of material within subjects or selected material based on quality of research or sometimes accessibility of material
7. comprehensiveness (depth)for instance, does it include all material on zoology, including grey literature or only citations of major journals in the field
8. level (e.g. for the general public, for researchers)
9. format (CDROM, Online, print etc)

(from Lane 1989)

This is quite a comprehensive list, but on a day to day basis probably the most important criteria used in deciding which index to use are:

- subject/discipline
- scope
- date
- inclusiveness
- comprehensiveness.

All of these criteria apply to Internet indexes, together with additional factors.

Debates rage across the Internet over indexing. There is a very wide range of opinion on what should be done to ensure that information can be found through Internet indexes. At one extreme there are those who propose a completely new approach to information discovery or "mining" through a completely automated system. They argue that the only human components of information discovery will be the humans creating information on the Internet. In order to fully automate this process creators will be required to provide good quality information on title, authorship, publication date and subjects of their Internet publications. Without this information the ability of any Index to provide an excellent search result is limited. With high quality information personal robots could be used to find information on a regular (perhaps daily) basis according to personal profiles. This approach is based on a philosophy that "everyone can get what they want" and that information gathering is a long term activity gradually building each individual's knowledge. It presumes that some form of quality control can be built into the Internet to enable an easier detection of relevant information.

At the other end of the spectrum are the indexing and searching professionals who see their role is changing, but becoming more important in the "information mine" of the Internet. An example of the application of traditional indexing procedures (a very time consuming approach) to the Internet is Cyberstacks starting with a selection of subject areas, such as agriculture. This approach has its difficulties in application too, mostly due to the volatile nature of the web where indexing a resource can really be like pinning jelly to the wall, as it may be there today and gone, or completely changed tomorrow. Not only can the resource's name, contents and location change regularly, but its accessibility and format change easily as well.

If the Internet is growing, becoming an increasingly important research publishing format and more accessible, how can we determine what to use for each inquiry? A set of criteria to evaluate Internet indexes is proposed below, based on Internet index use and feedback from research skills students at the University of Canberra.

1. contents:
    1. subjects areas covered
    2. inclusiveness (breadth)
    3. comprehensiveness (depth)
2. (ftp, gopher, www) (see note 4)
3. search techniques available
4. interface design
5. accessibility (speed of access, convenience of form, regularity of updating of index)

    When accessibility is very poor, however (eg response time is very slow) this may be the first factor to eliminate an Internet Index from use.
6. format of results (useful information includes title, address, extracts of text or a summary)

To test the proposed criteria, it has been applied to some actual search engines. In order to evaluate the usefulness of the Index three specific searches were also undertaken:

1. an actual research topic: a library client has asked for research on the effect of forest disturbance on fauna;
2. a general question: find "newjour" - the list of new Internet journals; and
3. a general search for the Australian Society of Indexers.

All searches were undertaken in the last week of March 1996 and results are valid only for this time.

The Evaluation

Alta Vista Excite Infoseek Inktomi Lycos Magellan Open web Yahoo

Location www.altavista.digital.com www.excite.com www.infoseek.com inktomi.berkeley.edu www.lycos.com www.mckinley.com www.opentext.com www.yahoo.com

Contents:* Access to 11 billion words found in 22 million Web pages. A full-text index of over 13,000 news groups updates in real-time. Fully automated index Access to more than 11.5 million web documents. Reviews database (over 50,000 web site reviews). Usernet more than 1 million articles from 10,000 newsgroups Classified advertisements (past two weeks). One of the very large web indexes. Has main database, subject lists, world news. Web sites, gopher sites, (4) Reputedly over 5 million web 'a2z' directory to browse by category or find by keyword. Includes short descriptions of each site and links to related sites. Point Top 5% list by professional web reviews Contains a wide range of sites with reviews by professional web

reviewers who award 1 to 10 points in three areas: Depth (Is it comprehensive and up-to-date?), Ease of exploration & Net appeal Large index Well established, one of the most popular with end users.

Subjects areas covered All- sites added by nomination ditto ditto ditto ditto ditto ditto

Inclusiveness Very large number of sites ditto again very inclusive large number of sites very inclusive high up on this scale

Comprehensiveness Down to document text level ditto also down to document level ditto ditto listing more of sites

Scope of coverage ftp, www, gopher, newsgroups all plus review database and classified database web, ftp, gopher, listserv, other web, gopher main database covering ftp, www, gopher, etc, 'a2z' database, top 5% database sites and reviews sies and reviews/ subject lists

Interface design Drop down menus where appropriate, simple & clear Uses frames, coxes, radio buttons. You may find the form does not display in Mosaic 2. Uses tables for mock columns, subject categories are at this level. Simple input box. Uses drop down menus & single input box Drop down menus where appropriate, simple & clear, subject categories are at this level List of subejct categories and input box, uses frames and tables Simple entry form with room for 4 words or phrases, fields selection drop boxes & Boolean for each term Input box, list of subject categories, no frames, table

Search techniques available

Simple queries (can search phrases, use "+", truncation,) specify fields(eg .gov). Advanced searches (Selection Criteria: AND, OR, NOT and NEAR, phrase Results Ranking Criteria: words for sorting), Start date, End date

Advanced

searching also

available Simple queries (terms are +'d), Advanced queries (+ in front of a search word for ALL of the documents, to Excluded Words e.g Jaguar -car , full Boolean operators and syntax. All searching - capitalisation may be used, "" for a phrase, [] for near searches (within 100 words), +word must be in results, -word must change

not be in results Simple queries (terms are +'d), Advanced queries (+ and - to include and exclude words). Simple, enhanced (can "and" or "or") and formless searching. Refine search on bottom of page. Assumes "and"for terms, no capitalisation, "-" (minus) symbol between words, the search will find records containing one word but not the other. Drop downmenus allow forsearching in specific fields, can search by phrase, &offers full Boolean searches. Can"customize"

restrict search to titles, URLs, and/or comments; search case-sensitive or case insensitive; "or"; search; change result limit from 100.

Accessibility Spped of access comparable with most US sites, not overly heavy use of graphics Speed of access comparable with most US sites, frames slower, than no frames frames option is noticeably slower than no frames quite quick, better than most index sites quite quick quite quick

Format Contains title, basic information location relevance ranking, title, summary title, location,relevance, score, size, abstract, Similar page link relevance ranking, title, occurrences of search words, location title, relevance, ranking, outline, abstract, location, size title, contents, summary, location title, relevance score, size, location & find similar pages title (hyperlinked) and single sentence

Search #1

Comment

28 pages of material was found 2 very useful citation

Comment

Again hit rate in the 400 range. This site had 2 useful citations

Comment

Many hits, limited userfulness

Comment

Many hits WCMC material scored first, not as much relevant material as on excite

Comment

Again many hits, useful material located (2)

Comment

29 pages of material, no useful citations

Comment

only 2 citations found - neither useful

Comment

Search produced no useful results - navigated to society and culture, environment, then forests

Search #2 references to individual files in the site come up before the main site individual files on the site come up before the home page pointed to newjour page #3 in list directly pointed to correct page in the top 2 hits pointed directly to correct page #1 on list #3 on the list, #s 1 & 2 are postings to ADFA epub list not found

Search #3 ASI came up quickly ranking algorithm has American society higher than Australian! ASI came up quickly at the top of the list. ASI came up quickly as #1 on list ASI came up quickly as #1 on list 20 pages of results, #1 was Australian Computer Society Gopher, ASI was #6 on list #5 on list not found

Overall Excellent (****) Very Good (***) searching techniques limited Very Good (***) unusual use of search language (near means within 100 words) Good (**) appears not to have as much in the index, search techniques limited Excellent (****) search techniques limited, excellent format Good /Very Good (***) ranking hid some useful material, limited searching techniques Good /Very Good (***) ranking hid some useful material, does not retrieve as many citations as Lycos Very popular and interesting in approach to subject descriptions

* see note 4

Internet myths

The Internet is clearly becoming an increasing important information source. Using the Internet productively for information can only be achieved through finding the relevant information. Indexes are the key to this process. The indexes have not yet solved the problem of how to find quality information. Many factors indicate that finding Internet information is at a very early stage of development. Some are outlined below.

1) Anyone can do it

While the basic Internet software is becoming more sophisticated it can still be difficult to use. There can be problems in accessing sites on the Internet, and to search well takes a great deal of time. Searching can be complex and require the use of a number of Internet

indexes, as well as a degree of serendipity. Libraries can expect an increasing role as intermediaries in the search for information from the Internet.

2) All the information you need is on the 'net

Many major organisations connected to the Internet provide a wealth of publications and information. Many others however have only selected information available at present. Any inquiry for "published" information needs to be evaluated to see what format is the most appropriate. Care needs to be taken to make sure that essential information, which may be in print, is not missed in the search.

3) You can find resources on the 'net

As is clear from the comparison of the 8 Internet Indexes, there is a need to determine which index is the most appropriate for a search. Criteria of contents, scope, search techniques, interface design, accessibility and format of results are the most relevant. Finding resources is complicated but sites being unavailable or moving.

4) Researchers will use the 'net all the time

Many researchers are more than fully occupied undertaking research and policy work. There is a need for good searchers to build up skills in Internet searching. While many who have the Internet at their desktop will do their own searching, an Information Centre or specialist searcher can provide a cost effective result (especially in terms of searching time).

5) Indexers will be obsolete

While Internet Indexes are predominantly automated, the last 12 months has seen a trend to using people to evaluate sites. Excite, Lycos, Magellan and Yahoo now all provide searchable site descriptions or keywords allocated by "professional web reviewers". While these are limited by short descriptions, or American use of terms, they are a growing service.

Conclusions

Automated Internet indexes are:

- limited by the lack of a controlled vocabulary for searching and the lack of consistency found in good quality citation indexes;
- limited by the sites/files on the Internet themselves;
- generally quite comprehensive (ie index whole documents) and general rather than specialised by subject;
- essential to begin any research using the Internet; and
- moving to more quality, subject information.

The sheer volume of data in a relatively uncontrolled form is like providing the front door key to the whole of the Library of Congress to our clients without a catalogue or floor map. Adding the current search tools is in some ways like giving them a torch and saying "go for it".

Limitations of Internet indexes:

- do not list sites which exclude access to robots;
- do not list sites which are down at the time the robots are working and therefore are missed;
- have great trouble finding any title composed primarily of stop words, such as "I can do it";
- may be confused by "spamming" by creators who include huge lists of terms (which may not be related to the content of the file) in the top of their files.

There are personal preferences in using Internet Indexes, just as there are in using online and printed indexes.

An ideal Internet index should:

- contain an indication of the sites covered by the index;
- allow for full Boolean searches (AND, OR, NOT);
- allow for phrase searching;
- be able to be used through a wide range of software (Netscape, Mosaic, Internet Explorer);
- index all files to document level;
- contain a minimum of graphics;
- weight site home pages higher that individual files;
- rank above 50% only those using all words in the Boolean search
- use a thesaurus which can be easily accessed by searchers;
- provide the following information in the results screen:
- title
- URL
- short summary
- relevance ranking
- file size.

The Internet resembles jelly in its constantly moving contents. Indexes for the Internet are also changing quickly. The information structure of traditional indexes, including clear statements about authorship, title and publication details do not exist in Internet files. Some have suggested the usd of metadata at larger "pins" to hold the jelly to the wall, but while publishing on the Internet remains uncontrolled it is unlikely that standards such as metadata can be made mandatory. New solutions for indexing have involved a mix of automated and human effort. The flexibility of the new indexes is stretched by the volume and range of files, including some creative attempts by authors to ensure they are indexed under terms not necessarily related to their files.

For Internet Indexes to be even more useful more specific searching needs to be available, i.e. Boolean operators, phrase searching and an indication of contents of the individual index. Connections between professional indexes and searchers and those creating Internet indexes will ensure that Internet indexes develop to become more useful than ever before.

Note: In the time taken to finalise this paper Inktomi has become "Hotbot" with an increased coverage of sites. Keep watching this space!

(1) Adams, Douglas (1979) The Hitchhikers Guide to the Galaxy, London, Pan Books p 24

"...he also had a device which looked like a largish electronic calculator. This has about a hundred tiny flat press buttons and a screen about four inches square on which any one of a million 'pages' could be summoned at a moment's notice. It looked insanely complicated...had the words DON'T PANIC printed on it in large friendly letters."

(2) From "Choruses from 'The Rock'" Eliot, T S (1936) Collected poems 1909-1936, Faber & Faber, London p 157

(3) The proceedings of the first forum were published as Mulvaney, J. and Steele, C (eds) (1993) Changes in scholarly communication patterns : Australia and the electronic library, Australian Academy of the Humanities, Canberra.

(4) * From Inktomi (http://inktomi.berkeley.edu/counting.html) (1 April 1996)

"There are three ways to count URLs for an index:

We count the number of documents actually retrieved from the Web and added to our index. This is a fraction of the total number of distinct URLs contained in those documents. We feel this is the most honest measure of coverage.

Lycos counts unique URLs whether their documents have been retrieved or not. If the lycos crawler comes across a link pointing to a document it gets counted as being in their index. Most of the documents in Lycos's index have never been retrieved. Here is a quote from their news page that provides the breakdown; by our metric they have 1.178 million documents:

July 15th, 1995 Latest catalog: The newest big catalog contains

- 5.07 million URLs, including 1,177,750 files downloaded.
- 5,077,834 unique URLs, including:
- 1,177,750 documents fetched totaling 8,703,484,067 bytes
- 3,900,084 unexplored URLs with descriptions
- 1,834,323,446 bytes of Lycos summaries
- 1,078,127,917 bytes of inverted index.

A third way to count URLs, used by Open Text, is to count the total number of non-distinct URLs in the collection. Thus, if a link to a popular document (such as Yahoo) appears 100,000 times, then it counts as a 100,000 documents."

Bibliography

Adams, Douglas (1979) The Hitchhikers Guide to the Galaxy, Pan Books, London

Basch, Reva (1994) Secrets of the super searchers, Eight Bit Books, Wilton, CT

BELNET (1996) A selection of Internet search tools, http://www.mtm.kuleuven.ac.be/Services/search.html 1 April 1996

Courtois, Martin, William Baer, Marcella Stark (Nov 1995). "Cool tools for searching the Web - a performance evaluation" Online Magazine, 19(6) : pp.14-32

Eliot, T S (1936) Collected poems 1909-1936, Faber & Faber, London

Lane, Nancy (1989) Techniques for student research: a practical guide, Longman Cheshire, Melbourne

Lebedev, Alexander (1996). Best search engines for finding scientific information in the Net, http://www.chem.msu.su/eng/comparison.html 1 April 1996

Mulvaney, J and Steele, C. (eds) (1993) Changes in scholarly communication patterns : Australia and the electronic library, Canberra, Australian Academy of the Humanities

Network Wizards (1996) Internet domain survey: January 1996, http://www.nw.com/zone/WWW/report.html 26 March 1996

Seidelman, R. In, Around and about, September 1995

Stanley, Tracey (1996). "Alta Vista vs. Lycos", Ariadne on the Web, Issue 2, http://ukoln.bath.ac.uk/ariadne/issue2/engines/ 1 April 1996

University of Michigan School of Information and Library Studies (1996) Matrix of WWW Indices. A comparison of Internet indexing tools, 1 April 1996

Winship, Ian (1995) World Wide Web searching tools - an evaluation, 1 April 1996

Paper presented at Electronic Age '96 A conference for the information industry 20-21 April 1996 Ranelagh House, Robertson NSW. Printed in the AusSI Newsletter 21(9):4-7, October 1997 and LASIE.

How is it indexed?

Geraldine Triffitt

The *Encyclopaedia of Aboriginal Australia* was published as a two volume work and a Macintosh version CD-ROM in 1994 by Aboriginal Studies Press, the publishing arm of the Australian Institute of Aboriginal and Torres Strait Islander Studies. This was a ground-breaking work, which has been recognised in the number of awards it has won for both formats. Reviewers comment on its attractiveness, its colourful illustrations, the quality of its sound.

It was the inspiration of David Horton who edited the encyclopaedia and together with Ian Howie-Willis wrote most of the 2000 entries.

Horton expressed his vision for the *Encyclopaedia* in the introduction.

> 'Among my key guiding principles in constructing this work have been an emphasis on people and a lack of objectification. Just as this is not an encyclopaedia of material culture or archaeology, or anthropology, nor is it an encyclopaedia of Aboriginal studies. It is an encyclopaedia of Aboriginal Australia, or an encyclopaedia of the Aboriginal people.
>
> I have deliberately set out to create an encyclopaedia of Aboriginal society in its own right as a complete system, as distinct from its appearance in first chapters or introductions to books about the 'real' Australia.'... 'I intend this to be not only an encyclopaedia of Aboriginal Australia, but an encyclopaedia for Aboriginal Australia.'

There was considerable input from Aboriginal editors, contributors and advisors in the preparation of the encyclopaedia. The contents reflect the interests and priorities of Aboriginal people, with the emphasis on biographical information (450 of the 2000 entries are about people), individual language groups and communities, which account for well over half the encyclopaedia entries.

There are 18 main topics, Art, Economy, Education, Food, Health, History, Land ownership, Language, Law, Literature, Media, Music, Politics, Prehistory, Religion, Social organisation, Sport and Technology. Each topic has a major essay written by the editor for that section. Within the topic are shorter essays on major subjects and short entries of up to 250 words. 'See also' references at the beginning of each entry link the expected and unusual, for example prominent people associated with a language group or subject. These were chosen with care to gradually add to the reader's knowledge and to avoid dead ends. At the end of the entry, one or two bibliographic references allow the reader to further pursue the topic.

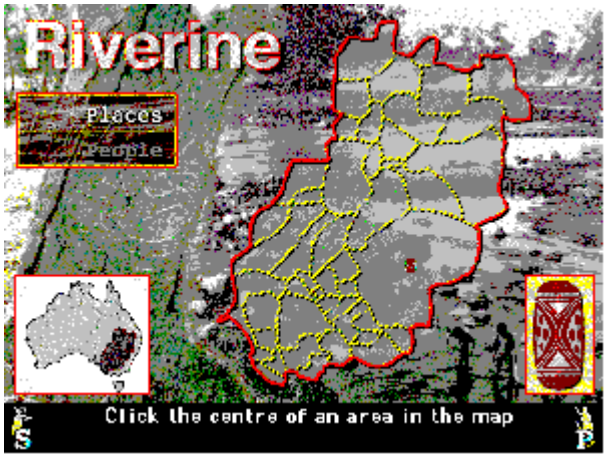
In the second volume, after an extensive bibliography, appendices give statistical, legislative and financial information, lists of winners of awards, bilingual schools, and the location of Christian missions.

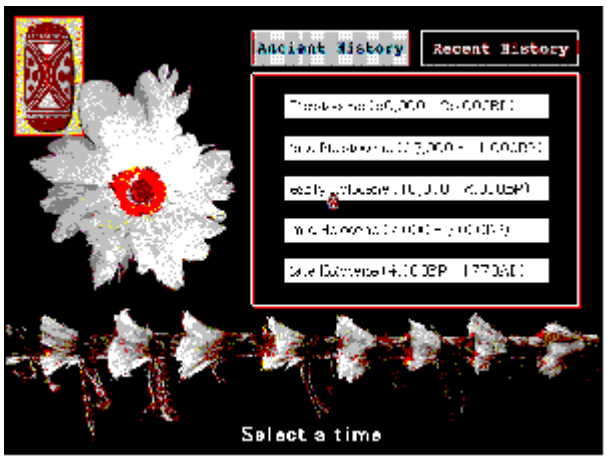
David Horton and film-maker, Kim McKenzie, designed and developed the CD-ROM - a colourful combination of text, pictures, sound, film and video. This has the same corpus of information and the 'See also' reference structure, enhanced by hypertext links on text and graphics. It is a particularly effective educational tool, which attracts the interest of all who view it.

The structure of the CD is as follows: over an image of the plateau at Kakadu, accompanied by the ululations of Central Australian women, the beginning screen gives the credits, customizes the sound and typeface, provides a help facility, and the start button. Pressing

the start button brings up the options of map, timeline, list of main topics, or typed in individual search topic. (Figure 1)

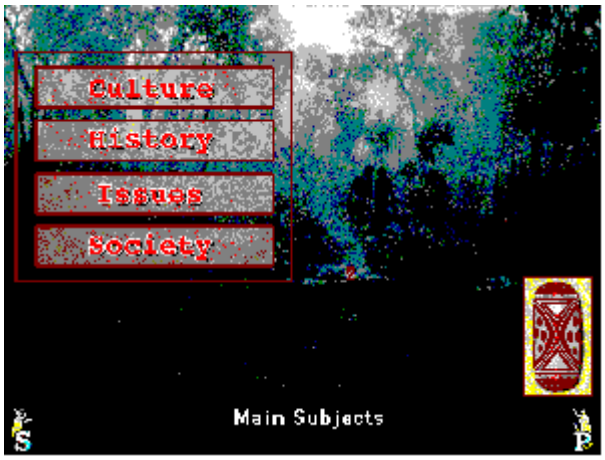
The map is divided into regions. Clicking on a portion of the map will show the relevant language group (Figure 2). The entry gives text, a reading list and possibly a relevant photograph, sound grab or video clip. 'See also' references and hypertext links refer to people, community, or a main topic. An alternative interface on the map lets the user access information on particular places within that region.



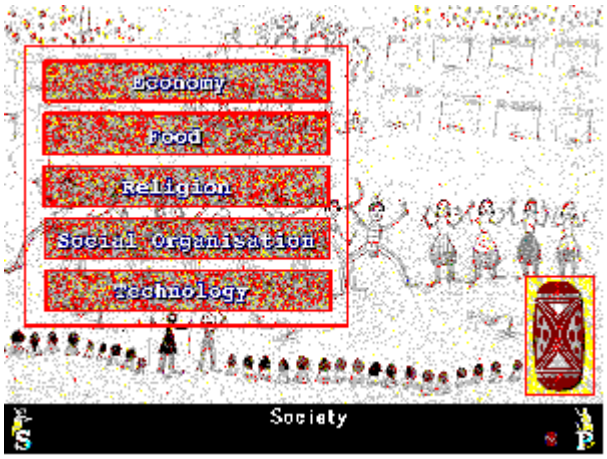


The timeline is divided into Ancient History and Recent History. Ancient History, which has 5 options, is the period from Pleistocene, 60,000-26000BP to Late Holocene, 4000BP -

1770AD. (Figure 3) Recent History has two levels of entry: the first is a division into 50 year periods, then each of these periods is further divided into ten year periods and the entries are arranged chronologically by date. (Figure 4)



The 18 main topics, are divided into four sections, Culture, History, Issues and Society. (Figure 5) Each of the main topics has an alphabetical list of headings for entries which can be searched. (See Figure 6 for the divisions for Society and Figure 7 for the plants listed under the Plant Food subdivision of Food)



The Type and Search button allows searches in the following categories: person, place or event or a typed in topic search (Figure 8). The tribe/language option refers back to the divisions on the map.

An ingenious navigation tool is the shield, the symbol of the Australian Institute of Aboriginal and Torres Strait Islander Studies. This provides an easy 4 way option. clicking at the top goes to homebase, the left side goes back, the right side searches, the bottom returns to the beginning screen and the quit button. Every navigation is recorded, allowing a quick return to a previous search.

How is such a mammoth work indexed (besides having entries ordered alphabetically)? There is no single highly detailed index to the whole work if an index is defined as

> a detailed alphabetical key to names, places, and topics in a book with reference to their page number, etc. in the book' (Macquarie Dictionary,1981)

This does not seem an appropriate definition for a CD-ROM.

The printed version has an Entry Guide comprising lists of entries under subject, region, state, tribal groups, and author. A time line describes major events, and an 'index' gives references to preferred terms, many of which are variant spellings for language groups This means that indexing is only at the level of entry heading. These same lists underlie the CD-ROM navigation tools. For example, the variant spelling list operates for the search screen when a term is typed in, the region list underlies the map, and the Recent History time zones list events in chronological order (Figure 4)

Does the *Encyclopaedia* need an index? It was a deliberate decision on David Horton's part to choose a sophisticated 'See also' reference system and ordered lists of entries and not an index. This was not a decision made in isolation, for example, he consulted teachers who suggested regional or state lists of entries would be useful for teaching purposes.

I have some personal reservations. As a reference librarian, I use an *Encyclopaedia* as a ready-reference tool to provide information quickly on topics accessed through an index. Consequently I find the *Encyclopaedia of Aboriginal Australia* useful for enquiries on people, language groups (although many are not represented), places and events, but can be frustrating when seeking specific information on other topics. For example when seeking to find out when Aboriginal people gained the vote there were no entries under Elections, Voting, Suffrage or other likely terms.

However the *Encyclopaedia* was not designed for someone like me who has worked in an Aboriginal studies reference environment for ten years and has built up my own network of finding aids. As quoted above, Horton's vision was of 'an encyclopaedia for Aboriginal Australia'. It is equally an encyclopaedia for anyone wanting to find out about Aboriginal culture and history. It is designed to be read, to lure the reader on to further discoveries. The CD-ROM begs to be played with, to test its audio-visual capabilities. The *Encyclopaedia* is not just a ready-reference tool.

Horton avoided the all-embracing key article on a topic used in some other encyclopaedia. These demand access to topics through a linear index. His *Encyclopaedia* challenges the

need for a printed index, in which disparate items follow in alphabetical order, not because of any relationship between preceding or succeeding terms.

The philosophy of the *Encyclopaedia i* s one of exploration, of finding information within a context, of navigating a web of linked pathways and references. In many ways this corresponds to the Greek notion of an encyclopaedia as 'a circle of knowledge which a reader could enter at any point and follow around'. (Barnes, 1996).

Kim McKenzie, sees the way in which the *Encyclopaedia* has been designed for CD-ROM as being an index. As well as the 'See Also' reference system of the hard copy, hypertext links lead the user from maps, to topics, people, events and back again using as search terms the entry headings of the hard copy version, not free text. By pressing the appropriate button, the user may see Ernie Dingo in a video, hear language spoken, or the songs of an area, or read text against a background of a photo of a particular event. This would seem to embody the essence of an index, which is to guide the user from a specific term to the body of information to which the term refers.

Both versions have the advantage of integrating information so that the searcher can use different pathways to reach a piece of information. The permutations of these pathways are vast, even for a few search terms. Searchers learn the context of the information by navigating the linkages. The advantage of a CD-ROM is the speed that this can be achieved.

Does the Encyclopaedia work in the way it is constructed? Ask the Aboriginal people who visit the Institute Library, who, having found a map of their area on the CD-ROM, find information on their people, hear the sound of their songs or watch videos of events in their history. Teachers complain that they cannot get their students away from the CD-ROM to go to other classes. Students, new to Aboriginal studies, find the Encyclopaedia opens up a new world to them and refers them on to more specific areas of study. The final word should go to one of the Encyclopaedia's reviewers

> 'When the set arrived, I sat down to give the pages a preliminary flick-through to see what sort of task I'd undertaken and when I looked up, four hours later, I was mortified to discover how the time had just flown by. I use reference books in my work: they're a tool to inform and refresh the memory, and people turn to them for scraps of specific information, not a "read". So it's almost a confession then, that later the same night I was burning the midnight oil - unable to put down a volume of an encyclopedia. (Sykes, 1994)

REFERENCES

Barnes, R. (1996)

> Keeping the encyclopaedia: an academic dilemma in hard times. *ANU Reporter* , 14 August: 4.

The Macquarie Dictionary (1981).

> Sydney: Macquarie Library

Sykes, R. (1994).

> [Review of] the Encyclopaedia of Aboriginal Australia. Sydney Morning Herald 9 July.

I am grateful for the comments and help from David Horton and Kim McKenzie in preparing this paper, and the editing care of my partner, John.

Bio

Geraldine Triffitt, National Library, Canberra ACT 2601 Australia

President of the ACT Region Branch of the Australian Society of Indexers and Subject Bibliographer (Linguistics), Library, Australian Institute of Aboriginal and Torres Strait

Islander Studies, Canberra, Australia.
Email: gtriffitt@interact.net.au