Indexing the future of information

Glenda Browne

This article describes changes that have occurred recently in the work indexers do, and considers the implications of these changes for our future.

The Australian Society of Indexers has run five conferences over the past ten years. Two themes have emerged - the importance of partnerships, and the need to consider our role in the electronic publishing environment. Our first conference, run by the Victorian branch in Marysville in 1995, was titled ‘Partners in Publishing’, while the most recent, run by the ACT Region branch in 2001 in conjunction with the Canberra Society of Editors, was called ‘Partnerships in knowledge’.

The second conference, run by the ACT Region branch in wonderful old Ranelagh House in Robertson (NSW) in 1996, was titled ‘Electronic Age’, focusing on our role in aiding access to information in electronic documents on the web and on CD. The first NSW conference, held in Katoomba in 1997, was called ‘The FutureProof Indexer’, as we wondered whether we would be made redundant by computer search systems.

Fortunately we discovered that we were still needed – in fact, we were crucial for quality information access – and the Victorian branch felt confident enough to return to the erudite essence of indexing with ‘The August Indexer’ (as in ‘refined’ or ‘imposing’), held in Hobart in August 1999. Our current conference has the very positive title ‘Indexing the World of Information’ as we consider not only the continuance of our traditional roles, but expansion both to new opportunities and to work on information around the world.

I believe the future is already here, in that significant changes in information provision have occurred in the last few years. There have been great technological developments. As I remind my kids: ‘When I was your age, I had to walk to the TV to change the channel!’ On the other hand, some of the change has been overwhelming. As humorist Ashleigh Brilliant said: ‘There has been an alarming increase in the number of things I know nothing about’. For us these new concepts are mainly reflected in acronyms. Are you up to scratch with XML, XFML, XTM, taxonomies, ontologies, OWL, RSS, CSS, CMS, DTD, DRM, DC, Semantic Web, RDF, RTF, PDF, DOI, ROI, EAD, wikis, NIPs, and SEO?

The things I think are required of indexers and indexing now, and which will be more important in the future are:

- Single sourcing using XML
- Embedded indexing
- Teamwork/Distributed content indexing
- Innovative specialized indexing/Globalization
- User testing

And the most important thing in all the work we do will be networking with other indexers, and communicating with a range of professionals to ensure that the work we do fits into the environment we are working in.

Single sourcing/multipurposing

Single sourcing means creating a document once, in such a way that it can be output in a number of different ways as required. This often means creating the document in XML format and outputting it in different print and online formats. It can also mean outputting different levels of detail, for example, a briefer version for children. XML is more powerful than HTML for this purpose as it separates the content of a document from its display.
When fully implemented, single sourcing can involve the re-use of single paragraphs. You could, for example, write a definition of ‘maternity leave’ once, and reuse it in many contexts. This means when the definition changes it will only have to be changed in one place, and all other uses will be automatically updated. This level of single sourcing requires good organization, and metadata to describe each of the separate document elements.

**Single sourcing example**

The *NSW Public Health Bulletin* is published 12 times a year; the final issue includes a subject and author index to the year’s content. In addition, a cumulative index is updated on the web every year (http://www.health.nsw.gov.au/public-health/phb/Subject_Index_for_2002web.htm). From 2003 the *Bulletin* has also been included in PubMed, a bibliographic database accessible online (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi). For PubMed the *Bulletin* publisher provides citation details in XML format, but PubMed provides the subject headings. The PubMed requirements are slightly different from those required for the print index, requiring restructuring of data entry; for instance, PubMed requires each part of each name in a separate field, whereas in previous indexing they were joined in a string. PubMed also requires a URL so it can link directly to the full text using its ‘LinkOut’ facility.

The solution for this project has been to use an MS-Access database with all the fields required for each different index. The database reporting feature is then used to extract the exact fields needed, and to add the XML coding. An MS-Word macro is used for some final changes (eg, converting ‘NSW’ to ‘New South Wales’).

Figure 1 shows a typical entry in the print version. The cumulative version can be viewed on the web at www.health.nsw.gov.au/public-health/phb/Subject_Index_for_2002web.htm. Figure 2 shows a sample of XML output for the bibliographic database.

The use of single sourcing for the *Bulletin* index means that data can be entered once, and output in different formats monthly, yearly, and in a cumulative version. This approach requires technical skills in database construction, and makes the indexing job much fiddlier and less concentrated on pure indexing.

**Embedded indexing**

Embedded indexing involves the addition of tags in the text of the document to indicate required indexing terms, rather than the creation of a stand-alone index that refers to the document, but is not connected to it. The advantages of embedded indexing are that the indexing flows with the text it refers to, so if pages are added or deleted the indexing doesn’t have to change. Indexing can start before page numbers are finalized, and need not be re-done for new editions (although it will have to be edited where changes have been made to the content).

The disadvantages of embedded indexing are that it is more time-consuming to do, and harder to edit. It should be done on the ‘live’ copy of the document; that is, nobody should be making editorial changes to one copy of a document while the indexer is working on another copy. If this is not possible, indexing and editorial changes will have to be merged into the same document. Embedded indexing is commonly used in technical indexing of computer manuals and other documents for which prompt publication is essential, and where updates are made regularly. It is also used in HTML Indexer web indexing software to automatically take care of deletion and movement of pages.

Embedded indexing can also be used when one document is being split into a number of separate documents, or where separate documents are being combined. In these cases editing is needed to ensure that the index makes sense in its final context.

Embedded indexing is important for single sourcing. The indexer has to index for the most widely used format and accept that the index for the other formats may not be ideal. (If time permits you can, of course, tweak the index for each individual situation, but you have then lost the efficiency of single sourcing).
CUP XML

Cambridge University Press (CUP) has been asking indexers to embed their indexes in the text of the document (or provide them in a form that can be embedded) for the last year or so. This is a highly significant development as it is not a new approach for a new format, but impacts on traditional book indexing methods.

The rationale behind the approach is that embedded indexing will enable CUP to ask indexers to start indexing before copy editing has finished, thus cutting some time from the publication schedule. It also means they will be able to output their documents in different formats, and to issue later editions without totally reworking the index. (However, if copy editing results in significant changes to the text, then the index created with the earlier version will not be appropriate. And, even if changes are only minor, the person doing data entry will have to decide exactly where tags go when the indexer’s copy and the final copy are not identical).

For indexers, it means the job takes longer, and has more clerical aspects to it. Indexers can work in their usual way, creating an index in specialized indexing software without adding page numbers. Then they have to allocate a unique number to each index entry, and either embed these numbers in the MS-Word or Adobe Acrobat PDF version of the document, or write the numbers in the printed proofs for later entry by others. Some indexers transfer the index to MS-Excel in entry order and automatically add consecutive codes. Others use specialized indexing software that can automatically increment locator numbers. A simple example is shown in Figure 3.

Indexers report that the process takes extra time (15 percent extra working time, or double the time for data entry). Some have said the publisher pays them for the extra time spent (although most indexers haven’t claimed for the extra working time, or double the time for data entry) although others say the pay doesn’t cover the extra work.

As far as I can tell this approach may have practical and economic benefits for publishers, but offers nothing for indexers. It simply adds to the clerical load without adding anything to the quality of the work process or the final output. It may be that later refinements of the process can make the clerical load less daunting, for example by the provision of automatically numbered paragraphs. A correspondent on Index-L wrote that CUP mainly has authors indexing their own works. If authors find the embedding process daunting, it might end up creating more work for indexers who can manage both the intellectual and technical aspects of the job.

It will be interesting to see whether this approach does offer CUP the benefits they expect, and whether other publishers will then take it up. The adaptable indexer will be the one with the best future.

If you are interested in following this up, you can download the 47-page author instruction manual from the US CUP at https://authornet.cambridge.org/information/productionguide. I haven’t indexed using this process but have spoken to other indexers who have, and have followed the discussion on Index-L about the process.

Teamwork

Indexer involvement in larger projects, particularly intranet and website design, involves more work in teams. There has also been convergence in the jobs people do, so that a metadata and taxonomy/thesaurus supervisor might be an indexer, librarian, or information architect. Indexers working in large teams need to be aware of the special knowledge and contributions of other people in the team, and have an idea of how they fit together. Indexing is only one part of the whole information access plan. Work done on big projects depends to some extent on the capabilities of the software package being used. To this extent, the computer is another partner in the project.

Most projects don’t start from scratch, but bring data and metadata from pre-existing documents. One indexing challenge is to salvage the metadata that is useful, and build on this, and to discard the rest. Sometimes this involves the NEHTBGE principle – near enough has to be good enough, as there is not always time to give top priority to all the content. The 80:20 rule is relevant here – often 20 percent of the content gets 80 percent of the use, so it makes sense to give this the best treatment. Sometimes legacy metadata or simple full text search is used for the least used bits of content.

When starting a large project it is important to be aware of the client’s priorities and aims for the indexing you are doing. These sometimes range from pure information access to issues such as personalization (offering different content to different users), customisation (allowing users to adapt their view of the site for their needs) and branding (focusing on the image of the company that the site gives).

In practical terms it is important to be clear who the client is, and what the ultimate goals of the project are. Sometimes your direct client and the ultimate client are different. Good communication and written documentation are essential. Finally, when quoting for work involving large teams and a number of stakeholders, you need to allow for time spent in meetings and incorporating feedback. It would not be unusual for this to take 30 percent of an indexer’s time.

TEXT

A faceted classification is based on a controlled vocabulary, and implementation requires indexing of documents with metadata from the controlled vocabulary...

...Faceted classification often works well for ecommerce applications, where specific attributes are applicable to all products.

INDEX

<table>
<thead>
<tr>
<th>e-commerce applications</th>
<th>faceted classifications</th>
<th>websites, e-commerce</th>
</tr>
</thead>
</table>

Figure 3. CUP XML numbers in margin of text with highlighted locations in text. The index entries are shown at the bottom. The ‘b’ and ‘f’ are used to indicate the beginning and end of a range.
Teamwork/Distributed content indexing

One example of teamwork in indexing is what I have dubbed DCI – Distributed Content Indexing. Often content management systems used for intranet creation allow Distributed Content Authoring (DCA) – that is, their workflow enables a number of people and departments within the organization to contribute content, rather than just those within an intranet content group. Organizations with DCA will often offer DCI as well, that is, the people who author the content will also index it.

The advantages of DCI are that you can get prompt indexing (it can be done as soon as the content is created, with no delays waiting for a specialist indexer), and the terms used in indexing will be similar to those used in content. For DCI to work, it is essential that the authors are committed to indexing, that there are sound policies to ensure good, consistent metadata creation, that authors are trained in the general principles of indexing as well as specific company policies, and that there is a metadata editor to oversee the project and ensure consistency and completion of metadata.

Innovative specialized indexing

There is a huge potential for the creation of innovative indexing projects to satisfy user needs. The great challenge is to identify the ones that could be financially successful. There are likely to be niche markets for specialist users and products from specialist indexers.

Dragon index

Recently my son, a Dungeons and Dragons fan, decided to index his issues of Dragon magazine. While I thought ‘author’ and ‘title’, he thought ‘prestige classes’ and ‘feats’ – that is, he wanted to draw out specific aspects of the content, rather than just lead readers to each individual article. This is a reminder that it is crucial to determine user needs before going ahead with any project. A short sample is given in Figure 4. It is interesting that he used singular entries (elsewhere in the index plurals are used for entries such as ‘Games’), capitalized entries, and inverted entries – all of which may be avoided by professional indexers.

After creating the index my son emailed the editor of Dragon magazine to ask if he would be interested in seeing it. The editor replied that he would, but that there were already three free indexes to Dragon magazine on the web. This suggests to me that there is a demand for specialist indexes in this field, but that they are being provided free by hobbyists.

Globalization of specialist projects

When indexing is very specialized it is most likely to have global implications. There are now projects involving specialist indexers from many countries, working together through the internet to create large products. In a virtual company there are few limits on where the indexer lives, so long as time zones can be calculated to allow occasional phone contact, and the indexer is willing to pay a bit in bank fees every time they cash a foreign currency cheque.

The Australian IT industry is predicting large-scale overseas outsourcing of IT jobs. This is less significant in indexing, where language and context are so much more important, but it is already happening to some extent. There are advantages to us being able to offer our services in the global market; there are also risks that people in countries with lower wages will be able to significantly undercut our charges, resulting in less work. I’m thinking Fiji sounds nice...

User testing

User testing is crucial for large projects to ensure that the approach taken by indexers is one that suits the information seeking approaches of users. Unfortunately, the little research that has been done with users of book indexes has found that many of them have great difficulty with index use, and that features we take for granted are not clear to them. Users often:

- Don’t like cross-references
- Don’t know the alphabet
- Don’t understand indexes
- Want more entry points
- Search more broadly than indexers index
- Appreciate alternative information access mechanisms

Indexers can:

- Involve users in index planning
- Consider alternatives such as ‘best bets’ links to the most popular pages
- Remember the 80:20 rule – 20 percent of the content gets 80 percent of the use
- Add more internal guidance
- Integrate thesauri to lead users to appropriate terms
- Use font variation to guide users
- Obsess less over minor issues
- Offer training in index use and search techniques

Skills needed

To manage all these new requirements, I believe indexers need the following skills:

- Sound knowledge of basic indexing principles
- Knowledge of the principles of metadata creation and search engine operation
- Knowledge of the principles of taxonomy/thesaurus creation
- The ability to evaluate and tweak automated systems
- Database design skills
- Project management skills
- Knowledge of usability issues, and user testing skills

Skills needed

To manage all these new requirements, I believe indexers need the following skills:

- Sound knowledge of basic indexing principles
- Knowledge of the principles of metadata creation and search engine operation
- Knowledge of the principles of taxonomy/thesaurus creation
- The ability to evaluate and tweak automated systems
- Database design skills
- Project management skills
- Knowledge of usability issues, and user testing skills

Figure 4. Dragon magazine index excerpt.

Feat, General
Avoid Critical Hit
Dex 13+, Dodge, Skill Focus (Tumble), BAB +8 or more, ranks in Tumble, 292:114
Feat, Kaiju
Battle Roar, 289:70
Improved Trample, 289:70

The Indexer Vol. 24 No. 1 April 2004

35
Ongoing themes
So, our two major themes continue – to be ‘FutureProof’ we have to maintain and develop our skills, and find out what clients want and need, and how we can provide it. In doing this we must create and maintain partnerships with other information professionals – with indexers, editors, publishers, records managers, archivists, bibliographers, librarians, information architects, taxonomists, web managers, database managers, technical writers and authors.

Networking
Some sites with information useful for people wanting to expand the scope of their work are listed below. These sites are mostly Australian, but similar information will be available in other countries.

- AusSI/INDEX-L/alialINDEXERS – details at http://www.aussi.org
- ALIA – http://www.alia.org.au
- Editors – for example http://www.users.bigpond.com/socedsnw
- SIGIA-L – http://mail.asis.org/mailman/listinfo/sigia-l
- IA get-togethers – http://IAwiki.net/CocktailHours/Sydney

Conclusion
Indexing, I believe, has a great future, with opportunities not imagined ten years ago. But it is also a very challenging future, with not only new opportunities, but significant adaptation to be made to keep up with current jobs. To stay aware of developments, and remain relevant, indexers have to learn new skills and make contact with other indexers and related professionals. In my conference talk I used a cake image, with the thought that indexers are not the cake; we’re not even the icing on the cake, but we’re the hundreds and thousands on the icing on the cake (shaped like an ‘I’). To work at our best, we need to fit in with, and stick to, the icing and the cake.

This article is based on a talk given at the Australian Society of Indexers’ conference ‘Indexing the World of Information’, held at the Carlton Crest Hotel, Sydney, NSW, Australia, on 12–13 September 2003.

Glenda Browne trained as a scientist and librarian, and has been a freelance indexer of books, journals, online help and websites since 1988. She teaches occasional courses in web indexing and thesaurus construction. With Jonathan Jermey she has written Website indexing: enhancing access to information within websites, the second edition of which is due to be published by Auslib Press in March 2004. Glenda is Treasurer of the NSW Branch of AusSI, and edited the AusSI Newsletter for three years. Email: webindexing@optusnet.com.au

A fluffy index
Will Self is described in the blurb to Feeding frenzy, his latest collection of journalism and assorted writings (Penguin, 2002), as ‘the Jack the Ripper of English prose’. He says of himself, ‘I align myself with the utopian socialist libertarian tradition of English thought. I am fiercely anti-establishment.’ Nowhere is this more evident than in his index. In it he perversely manages to break just about every rule of professional indexing, making inconsistency into a virtue.

In his Introduction, Self explains that the pieces chosen for inclusion are ‘simply numbered rather than titled, the date and publication appearing at the end; and they are arranged in no particular order, nor are they categorised.’ Indeed, they are not even printed in order of original publication, but apparently totally at random.

He continues:

If this gives the text a certain agglutinative quality, I hope that it’s aerated by the provision of a more than usually fluffy index. This covers not only people, places and things, but also ideas, obsessions and my own irritating stylistic tics.

In fact, although many of the items reprinted here are selections from his brilliantly idiosyncratic restaurant reviews (originally published in The Observer), you will not find any of the names of the restaurants in question in the index. Instead there are bizarre entries such as

epicene hips
epicene museum curators
epicene sub-aristos
kidney-shaped expanses of water
shoebox-shaped rooms

Most of his entries are spare and succinct, but he has no fewer than 32 (undifferentiated) page references for ‘irony’, and 42 for ‘television’. He is clearly enamoured of cross-references, which bulk out his index to an inordinate extent:

delusions, extraordinary popular see extraordinary popular delusions
extraordinary popular delusions 270, 326
popular delusions, extraordinary see extraordinary popular delusions
Queen Mother, the see Bowes-Lyon, Elizabeth
Quixote, Don see Don Quixote
Victoria, Queen see Queen Victoria

He deploys see also references with equal deftness:
pullovers 66, 104, 163 see also sweaters; woollies
sweaters 30 see also pullovers; woollies
woollies 66, 229, 241, 315, 316 see also pullovers; sweaters

In short, almost a textbook example of how not to make an index, but at the same time what one reviewer (Steven Poole in The Guardian) described as the ‘crowning glory’ of the book.