

CONVERGENCE: THE FUTURE OF INDEXERS AND OTHER PROFESSIONALS

Glenda Browne, www.webindexing.biz, webindexing@optusnet.com.au

Thursday, 1 February, 2007, 1415 to 1440 - 25 mins (20 talk, 5 questions)

The last decade has seen convergence in the work done by indexers, librarians, records managers and computer specialists, now working as information architects, taxonomists and digital librarians. Although some information professionals have found new and interesting work in these areas, others feel excluded by the focus on large digitisation projects and regret the loss of the intellectual pleasures of hands-on cataloguing and indexing.

This paper and presentation examine the current state of the industry and the opportunities and challenges for indexers and other information professionals in the future.

In some areas such as bibliographic database indexing it appears there will be a steady, unavoidable loss of traditional indexing opportunities. A similar fate may be on the way for the provision of indexes to individual periodicals, especially when they are included in large-scale database indexing projects such as Medline.

On the other hand, the huge increase in importance of the web and internets in the provision of corporate and government information has led to a widespread acknowledgement of the need for some vocabulary control, with traditional classifications, thesauri and subject heading lists morphing into taxonomies and ontologies. In addition, some of the indexing of individual databases and journals has been replaced by indexing done by aggregators of electronic content.

The proportion of documents that are indexed by humans (versus computers) may be decreasing, but as the total amount of information continues to increase there should be continued opportunities for us all.

This paper first considers the convergence of different information professionals, and then looks specifically at the future of bibliographic database indexing.

Convergence

From zoos, aquaria and herbaria to libraries, museums and archives, all sorts of collecting agencies:

- Acquire and accession
- Store
- Catalogue, classify, categorise, label, index
- Make findable and accessible
- Display and promote

And now intranets and websites do too!

The commonalities between these groups are recognised by the Collections Australia Network, which provides information about all of them from the one site (<http://www.collectionsaustralia.net>).

Intranets and websites

A librarian can work on an intranet, and an intranet indexer and thesaurus editor may well be a journalist, technical writer, content specialist, web manager or editor. Librarian and indexer roles in intranets and websites may be more consultative than hands-on, and while these sites may use thesauruses they may not **call** them that, and they may use alternative controlled vocabularies such as taxonomies (hierarchical), ontologies (structured for automated interrogation) or synonym rings (simple groups of synonyms for automatic search broadening).ⁱ

Records management

Records management agencies have traditionally used terms from thesauruses to describe items they are managing. Records management principles differ from those followed in libraries as they focus more on provenance (the source of the records) and index first with a term describing the functional area responsible for the activity being documented. Nonetheless, subject searches are the most common, so users have to be able to translate subject searches into functional searches. The Australian Society of Archivists has good examples of translating subject searches into functional searches (www.archivists.org.au/pubs/brochures/understanding.html), and StepTwo Designs has reported on a usability of a records management classification (http://www.steptwo.com.au/papers/kmc_caloundracouncil/index.html). Functional keywords are also supplemented by subject-based taxonomies/thesauruses, or by mapping from subject to function.

At the 2001 AusSI (now ANZSI) conference I spoke on 'Indexing the future of information'. One of my points was the need for greater communication between closely-related professionals, including librarians, technical writers and records managers. My paper was seen by a records manager in New Zealand, who asked permission to reprint it, and then by a records manager in Britain, who also asked to reprint it. I took this as an example of our shared interests.ⁱⁱ

User-generated indexing

Users are now doing their own indexing, with keyword tagging at social sharing sites such as de.li.cio.us, flickr, citeulike. The BBC and other organisations are harvesting user terms (folksonomy) for use in formal taxonomies, thus combining the best of both approaches.

Some convergence, but there are still separate silos

Despite the similar backgrounds and approaches needed by different information professionals, we still communicate very little. A paper by Philip Resnik and Gary Adams (1996) on the World Wide Web Consortium (W3C) website notes: "Conceptual" is something of a recent buzzword in the information retrieval business...for example, a search involving "agriculture" might do well to turn up documents about "farming".ⁱⁱⁱ Something librarians have known for a century, and should have been communicated not reinvented.

Convergence with computers

Convergence of the work of different information professionals has been added to by the convergence of the work done by computers with that of information professionals. While the creator of metadata for an intranet could be a librarian or an

editor, it could also be an expensive computer program ^{iv}
(<http://www.webindexing.biz/Articles/AutomaticCategorisation.htm>).

Bibliographic database indexing

This paper will examine bibliographic database indexing as an example of the convergence of computers and human indexers, and of the issues impacting on the future of indexing.

Bibliographic database indexing is under threat from a number of directions:

- There has been a decrease in human indexing and a rise in Machine-Aided Indexing (MAI), in fully automated indexing and in the absence of indexing (replaced by free-text, full-text search)
- There has been a decrease in the number of databases in many areas and in the number of databases that are indexed

These changes are due to improvements in machine indexing and the necessity to cut costs in database production. The need to cut costs stems from changes in government roles from funding databases ‘for the public good’ towards cost-recovery approaches. The decline is also related to changes to the availability of items for indexing (e.g., if a library which produces a database closes, or a clearinghouse does not receive the full range of donations it requires). User interest or lack thereof plays a role in the closure of databases, but can also influence their reinstatement. ^{v, vi, vii, viii, ix}

But maybe it's about time!

While indexers look with horror at the takeover of their intellectual roles by mere machines, the programmers of those machines believe they have a product of value that should have been used years ago.

Karen Sparck Jones (2004) wrote: ‘Operational bibliographic services were very reluctant to allow statistical methods any possible utility, especially given the tiny research experiments, and became substantially committed to the conventional boolean approach. The first Web engine builders had no such prior commitments and picked up the statistical idea... It thus took about twenty-five years for a simple, obvious, useful idea to reach the real world, even the fast-moving information technology one.’ ^x

How well do computers do?

The Center for Aerospace Information (CASI) at NASA uses machine-aided indexing with human review to map text to NASA thesaurus terms, apparently with comparable recall and better precision than human indexing. One problem with evaluating MAI is that you have to know how much human input was required in order to know how much the machine contributed. ^{xi}

While the NASA results were positive, anecdotal evidence from indexers suggests that MAI systems require a lot of editing, and that while they are good at picking out concrete nouns, they are not as good at identifying more complex topics such as behaviours. For an article titled ‘Behavioral and auditory evoked potential audiograms of a false killer whale (*Pseudorca crassidens*)’, the MAI descriptors were: *auditory evoked potentials; hearing; rubber; electrodes* and *gold*. The last three terms were extracted from the phrase ‘responses were received through gold disc electrodes in

rubber suction cups'. The more complex concepts in the abstract were not identified. The manual descriptors were: *auditory evoked potentials; sound spectrography; auditory thresholds; bioacoustics; and go/no-go discrimination learning*.

MAI techniques also do not work well when language is used creatively. For an article on endothelins (vasoconstrictive compounds) entitled 'ET: Phone home', the MAI suggested *Emergency Department* and a range of telecommunications terms.^{xii}

It is interesting that while search engines such as Google perform extremely well in the area they are most used in, they are not necessarily a good tool for all information retrieval needs. When Larry Page from Google was invited to submit the Google search engine to a standard evaluation of performance using measurements of recall and precision, Page replied that he considered those metrics irrelevant – his measure of performance was simply 'the time it takes our user to find what he is looking for' (Diakoff 2004).

Automatic indexing is now being used for multimedia retrieval as well as just text search. This is a more challenging area, as is apparent from the poor results sometimes returned. Nonetheless, the systems return some useful results, and will presumably continue to improve. The Automatic Linguistic Indexing of Pictures website^{xiii} shows examples of images with both automated and manually-generated results shown. An image of a skimobile with two people in was allocated the terms *ballet, doll, monument, indoor* and *plane*, for example.

Bibliographic database indexing – the human factor

If computers are taking over the traditional roles of human indexers of bibliographic databases, what is left for the people to do? Suggested roles include:

- Monitor and improve MAI and automated systems – computers, like children, can be trained
- Identify the bits that humans do better than machines – e.g., broad classification codes to complement free-text search on specific terms
- Identify new areas in which human indexing is the only way to go (high value, high sensitivity, specialised content analysis)
- Find ways to improve search, e.g., passage level indexing
- Train users to value quality search

8. The future

This is all discussed further in our book *The indexing companion*^{xiv}, which concludes:

The oyster defends itself against an intruder and produces a pearl. The information world is our oyster – whether it turns out to be a toxic heavy-metal-laden mass, or home to a pearl of great beauty, is yet to be seen. This book/talk points to the pearl.

ⁱ ANSI/NISO Z39.19 - 2005 *Guidelines for the construction, format, and management of monolingual controlled vocabularies*, www.niso.org/standards/index.html.

-
- ⁱⁱ Browne, Glenda. Indexing the future of information. *The Indexer* v. 24 n.1 pp. 32-36. This article was based on an address given at the AusSI conference 'Indexing the world of information' held in Sydney on 23 and 13 September 2003. It was later adapted and updated for *IQ* (Records Management Association of Australasia), and then reprinted in *Bulletin* (Records Management Society of Great Britain) issue 133 August 2006.
- ⁱⁱⁱ Resnik, Philip and Adams, Gary 1996. 'Multilingual Issues in WWW Indexing and Searching: Position paper for the W3C Distributed Indexing/Searching Workshop', www.w3.org/Search/9605-Indexing-Workshop/Papers/Resnik@Sun.html
- ^{iv} Browne, Glenda. Automatic Categorisation Parts 1-3. <http://www.webindexing.biz/Articles/AutomaticCategorisation.htm> First published in *Online Currents* Vol.18 Issues 1, 2 and 3, Jan-Apr 2003
- ^v Farkas, Lynn 1995 'Economics and the future of database indexing' in: *Indexers – Partners in Publishing, Proceedings from the First International Conference*, Marysville, Vic. [Melbourne]: Australian Society of Indexers.
- ^{vi} Manktelow, Nicole 2003. *Sydney Morning Herald* 22–23 February 2003.
- ^{vii} Drynan, Elizabeth 2002. 'Indexes in danger' *Online currents* v.17 i.2. Also archived at pandora.nla.gov.au.
- ^{viii} Drynan, Elizabeth 2005 'ACHLIS and ALISA bow out' *Online currents* v.20 i.9. Also archived at pandora.nla.gov.au.
- ^{ix} Johnstone, Pamela and Drynan, Elizabeth 2000. 'The value/future of bibliographic indexing and abstracting' *Online currents* v.15 i.6. Also archived online at pandora.nla.gov.au.
- ^x Sparck Jones, Karen 2004. 'IDF term weighting and IR research lessons' *Journal of Documentation* v.60 n5, www.soi.city.ac.uk/~ser/idfpapers/ksj_reply.pdf.
- ^{xi} Lancaster, F. Wilfred 2003. *Indexing and abstracting in theory and practice*. London: Facet Publishing, p.309.
- ^{xii} Greenhouse, Shelley, pers. comm. 12 May 2006
- ^{xiii} wang.ist.psu.edu/IMAGE/alip.html
- ^{xiv} Browne, Glenda and Jerney, Jonathan. *The indexing companion* Melbourne: Cambridge University Press, due March 2007