Cataloguing? Metadata? or Something Else?: Providing Access to Electronic Information [1]

Niall O'Driscoll

National Cataloguer, Australian Securities & Investments Commission (ASIC), Sydney

INTRODUCTION

Traditionally libraries have used cataloguing to organise and give structure to the information they have in their collections. Despite a history of more than twenty years using computer-based cataloguing tools, this information has been primarily restricted to books and journals on the shelf. However, a few short years have been enough to change the situation significantly. Now a library can expect to source a large proportion of the information needed by patrons from electronic sources outside the library, at the time of need. This paper attempts to look at some of the effects this change has had on cataloguing, and how cataloguers must transform themselves to ensure that their skills are still valid in this new world.

This paper contains more questions than answers. I don't think at this stage, in the transition from paper-based to electronic information, that any librarian knows all the answers. I hope the elaboration of some of the issues (and the pitfalls) of trying to apply traditional cataloguing practices in the rapidly changing environment of government libraries will be of help to librarians working in other arenas, and will help illuminate some of the questions they may need to ask about their own environment.

THE PAST

Very few professionals working in the library and information environment would need reminding of how a 'traditional' library is run. The changes brought about by the explosion of the Internet and other electronic forms of

[1] This is an edited version of a paper entitled 'New Tools and Techniques for Next Generation Cataloguers' that was presented at the *Digital Libraries Technology '99* Conference, Sydney,

information have occurred in the space of a very few years. Certainly m not all of us, are still working in institutions which reflect the long establ framework of traditional libraries, but with a growing reliance on r technologies and information sources. Rusbridge calls these 'hybrid' libr which 'reflect the transitional state of the library, which today can neith fully print nor fully digital'.[2]

In a traditional library information exists predominantly in print format. S sort of decision-making process (probably by a library manager with i from the reference staff) determines which resources are acquired. C acquired, or perhaps during the ordering process, materials are entered c catalogue so that they may be found and accessed by library staff and patr After being catalogued the materials are prepared for the shelves where t are usually placed in a classified sequence as a further means of aid access. Materials may then remain on the shelf indefinitely when not in by patrons, unless removed by some human intervention process of weed or culling, or lost, at which point the catalogue record would be amended removed.

Some points which may be learned from this are:

- the major cost of the information was generally that of purchasing Anderson calls this the 'scarcity mentality' – 'The paper base manifestations of information had a very real cost – libraries have oft been measured by the number of dollars spent on print material';[3]
- once purchased, the library 'owned' the actual physical item as 'chattel', even though copyright still rested with the author. A librar could expect to retain this chattel as long as it wanted, and could use i (within reason and the Copyright Act) in any way it chose.

---

[2] Rusbridge, Chris 1998, 'Towards the hybrid library', *D-Lib Magazine*, July-August. Also available online http://www.dlib.org/dlib/july98/rusbridge/07rusbridge.html

[3] Anderson, Craig D. 1998, 'Where do you put the genre sticker in cyberspace?: an investigation into the organization of Internet-based resources' in *Robots to Knowbots: the wider automation agenda: proceedings. Victorian Association for Library Automation 9th biennial conference. 28-30 January 1998 Melbourne*. Available online at: http://home.vicnet.net.au/~vala/valaweb/vala1.htm

In this context, good cataloguing is important, not only for the generally accepted reason of providing access to information, but also because of a library's need to manage (and obtain value from) resources it has already spent money on

Recent changes to the world of information have changed this situation markedly, even if the model spelled out above is still relevant in many cases. Some examples of this change are:

- with the advent of the Internet much more information is available 'free';
- the Internet is a very volatile environment – information that is available in one specific place in a specific format one day, may move somewhere else, change format, or simply disappear, tomorrow;[4]
- with all forms of electronic information copyright issues become much more complex as all forms of access to electronic information rely on a copying process of some sort (from storage medium to screen, etc) Access to electronic information will almost always involve some sort of licence (or implied licence) and the fact that a library may have paid for the information doesn't mean that it will have permanent use of it;
- electronic information, by nature, lends itself to methods of automated discovery and retrieval of varying levels of usefulness.

These factors all have an influence on cataloguing. Those attempting to apply traditional cataloguing approaches to this new environment need to consider the impact of such an approach to their services. However, even in this changing environment, models from the traditional world of librarianship may still be useful, even if only by analogy.

WHAT IS CATALOGUING?

At its simplest level, cataloguing may be considered as the tools and techniques used by librarians to structure information in a way that makes it more readily accessible. For the purposes of this paper I will often be

---

[4] This fact is mentioned by a large number of authors. For example, Wendy Smith 1998, 'Lost in cyberspace: preservation challenges of Australian Internet resources', *LASIE,* vol. 29, no. 2, June.

including many aspects of what might be considered by many to be inc practices because in the electronic environment the division between inc and cataloguing are becoming blurred. Subject gateway pages on l websites hardly fall into the realms of full cataloguing, but do atten provide some structure to information for the purposes of inform discovery

Many people have attempted to describe the functions of the catal conceptually. Two of the best known are Cutter's rules (1876) and the Principles (1961). Implicit in most of these formulae is the idea that t creating the catalogue are dealing with works which are already included library's collection. In other words, the process of cataloguing is ent removed from that of selection of works for inclusion in the collection (w would be covered by the separate library discipline of collec development). In a world where the library is no longer restricted to mate housed within four walls, but instead provides access to information w may reside anywhere, this distinction needs to be re-evaluated. The between the collection development librarian and the cataloguer has becc blurred, that is, in libraries large enough for these functions to be underta by different people. It is particularly true in cases where the library wishes provide access to information that is 'free'. The cost of providing informat has been transferred from the cost involved in purchasing the information, the cost involved in cataloguing, describing and giving access to it.

CATALOGUING - THE COST

A recent study in Western Australian university libraries costed the creatio a catalogue record at around $20.[5] When a $50 monograph, for example catalogued then the extra $20 cost may be viewed as value-adding – catalogue record gives added worth to the information by helping ensure th having been purchased, it may be found and thus used. Where, however, library wishes to provide access by cataloguing 'free' resources on i

[5] Wade, Rona and Williamson Vicki. 1998, 'Cataloguing costed and restructured at Curtin University of Technology', http://www.unilinc.edu.au/curtin.htm

Internet this $20 changes from being only one of a range of costs that are involved in providing the information to being the *major* cost. In this context, cataloguers need to change the way they work to include other rôles – the concept of collection development needs to be considered at the point of cataloguing, rather than as something that happens independently of the cataloguer.

To further complicate this issue, the putative $20 is traditionally a one-off cost. Admittedly serials and loose-leaf services do change, sometimes often, and most libraries will weed, discard or lose items from their collections, with the resultant need to amend their catalogues. Nevertheless, in a large proportion of cases, once a work has been catalogued, processed and placed on the shelf the cataloguer need never look at the record again. Also, in the context of a national cataloguing system like Kinetica or the old ABN, and in an environment where State and National libraries attempt to maintain legal deposit collections in perpetuity, it is to be hoped that at least one copy of any hard copy work will remain in at least one collection forever – and thus the catalogue record ought to be relevant to at least one library forever.

The online environment, on the other hand, is notoriously volatile – electronic information is very easy to modify, so most electronic information *is* constantly being modified. Some of this change is beneficial – information is kept up to date and information sources may be allowed to grow as extra information is added. Conversely, information may disappear or may change location without warning – or the content or subject focus may undergo major alteration. Anyone attempting to provide access (through cataloguing) to these sort of electronic resources needs to plan for their work to be constantly monitored and updated – and needs to plan for the resultant ongoing cost of maintaining their cataloguing. Naturally, it is possible to use software to take over some parts of this maintenance process. Nevertheless, there is a cost involved even in applying these software solutions *in case* the information changes. There is even a cost involved in a cataloguer viewing a source and

determining that changes identified by electronic means do not requi alteration to the catalogue record after all.

Another well-documented problem with the cataloguing of electi information sources is their sheer number – a number that is itself ha verify OCLC estimates the number of web sites at over 2 million, of w more than 1 4 million are publicly accessible [6] These websites contain million web pages – and between 1997 and 1998 the number of web increased by 65%. Meanwhile, Lawrence and Giles estimate the numbe web pages at 800 million.[7] Admittedly, OCLC's figure of 2 million web s might be compared with the 30 and 39 million records in RLIN and OC respectively, which could be used to demonstrate that the task of catalogi them is not beyond the attempt.[8]

The fluctuation between the relative importance of website to web page itself a problem Some websites are simple homogenous affairs which wo require only the one cataloguing record. Other web sites are larger and m heterogenous, their constituent web pages functioning as works in their o right, each in need of separate cataloguing The number of entities on the w requiring individual cataloguing, therefore, could lie somewhere between t figures of 1.4 and 800 million – with the figures currently growing at 65%!

METADATA

Many people have dismissed the possibility of human analysis a cataloguing of the Internet – certainly it is beyond the capabilities of a single library. Much effort is consequently being put into means of improvi access through enhanced web search engines, crawlers and the like, and al through metadata. That is, metadata that has been created by the author

[6] Statistics (based on June 1998 WEB sample), OCLC Research. Web characterisation project, http://www.oclc.org/oclc/research/projects/webstats/statistics.htm

[7] Lawrence, Steve and Giles, Lee 'Accessibility and distribution of information on the web'. Summary available at http://www.wwwmetrics.com

[8] Hass Weinberg, Bella 1999, 'Improved Internet access: guidance from research on indexing and classification', *Keywords, the newsletter of the American Society of Indexers* vol 7, no 2, March-April, pp 5-10

originator of an electronic document and embedded within it, to serve as a descriptor or tag and to aid in the recovery of the document.

By its very nature, electronic information suits itself to automated manipulation, by a range of increasingly sophisticated computerised services However, it is obvious to anyone in the information field that none of these services is entirely effective. In the absence of true artificial intelligence at this date, we see that search engines, web crawlers and the like are able to find the instance of words and strings of characters, and to use statistical functions to rank retrieval in some sort of 'relevance' order – we all know, however, that current computers cannot replace a human when it comes to actually analysing the meaning and content of a document

These search engines are also hampered by the same volume and volatility of electronic information that I have already discussed as being an impediment to human cataloguing of the Internet Lawrence and Giles have found that the best web crawler (Hotbot) only indexes 16% of the web; that most index less than 10%; and that altogether the crawlers only index 43% of the publicly available web sites Furthermore coverage is decreasing and can be months behind.[9]

Experienced professional librarians searching commercial, full text databases using sophisticated library search engines often experience poor retrieval and false drops – this is one of the things we are taught in library school! This occurs even though a commercial database would be limited in terms of its scope or coverage. Little wonder then that using less precise tools in the sprawling chaos of the Internet often produces less than perfect results Hence the increasing interest in author supplied metadata, as a means of improving access.

Metadata supplied by the author or publisher has many advantages:

- it is presumably in the author or publisher's interest that the reading audience can locate their documents, or else why create them at all? 'The

[9] Lawrence and Giles

publisher's documents need to be organised for items to be found. was the publisher's catalogue becomes a library of the publi products';[10]

- an author knows the subject of their work intimately, and ought to be placed to supply relevant subject keywords; and
- the use of metadata lessens the burden on the library commu obviously the prime advantage of metadata. If a search engine has bee up to give weight to metadata terms, more accurate results wil provided without the need for analysis and cataloguing by librarian indexers. The metadata is supplied as part of the process of creating document by the person or organisation creating it – hopefully a onerous task for the individual author of a single document than that cataloguer faced with a multitude of such documents which he or she to make some sense of

There are, however, a number of disadvantages to this approach which li its effectiveness as a total solution to the problem of organising electro information:

- despite the fact that the such schemas as Dublin Core and RDF act 'standards' it is unlikely that they will really produce 'standardisatic One of the major thrusts of these metadata schemes is the perceived n for flexibility and extensibility as a means of satisfying the differing u communities that are likely to use them. While this flexibility l advantages, it also has obvious disadvantages;
- an author may lack perspective on their work, and may not be the b person to describe its content; and
- authors and publishers have different aims to librarians, which will reflected in the way they create their metadata. A librarian, wh cataloguing a resource, is likely to be neutral in describing a particul work. Authors and publishers have the obvious aim of promoting the productions, and are more apt to be 'selling' it to a potential audienc rather than describing it objectively

[10] Barry, Tony 1999, 'The next waves of change: the future as seen from January 1999' *Information Online & On Disc 99*. Sydney, ALIA. p. 301

Perhaps the biggest failing of metadata is the fact that it has so far not made sufficient impact on the people it is aimed at: the authors of electronic documents. For instance the Metadata Workshop held in Luxembourg during June 1998

> identified that the current takeup of Dublin Core is slow, and that there is a lack of critical mass This seems to be a classical chicken-and-egg situation: authors and publishers do not invest in providing Dublin Core metadata if the Internet indexing services (the 'harvesters') do not utilise it, and harvesters do not collect Dublin Core and use it for selective indexing if there is not enough data available If this situation cannot be changed, Dublin Core might not turn into reality [11]

Lawrence and Giles found that only 0 3% of web sites use Dublin Core.[12]

OCLC has done some research on the use of meta tags in web documents as part of its Web Characterization Project which found the initially promising fact that 'nearly three-quarters of the sampled sites had at least one insubstantiation of the meta tags. On average, three meta tags are used on a public web site home page.'[13]

When we look at the figures more closely we see that, of the four meta tags that occurred in more than 10% of pages, two tags, 'Generator' and 'Content Type', are machine generated by the two most common html editors. These tags give information related to the way the document is created: that it is in html and has been created using a specific editing package. This is next to useless for document discovery and retrieval purposes.

The other two tags which occurred in statistically significant numbers were 'Keywords' and 'Description', which look a little more promising from the

---

[11] Report of the Metadata Workshop held in Luxembourg, 26 June 1999, http://www2.echo.lu/libraries/en/metadata2.html

[12] Lawence and Giles

[13] O'Neill, Edward T Lavoie, Brian F and McClain, Patrick D. 1998, 'Web characterisation project: an analysis of metadata usage on the web', http://www.oclc.org/oclc/research/publications/review98/metadata.htm

point of view of discovery and retrieval. OCLC's analysis of the terms in these tags showed that 'the most notable feature was that the keyw while usually pertinent in some way to the site's content, nonetheless often extremely broad.'[14]

The study also points out that 'Author', the fifth most common META occurring in 6% of the sample, is also often 'not particularly informativ many cases it is machine-generated from the name attached to the com used to create the document. Such a name may be something uninformative as the computer's IP address.

While it is to be hoped that the significant international effort that is b put into metadata standards, like Dublin Core, RDF, etc., will eventually some rewards, we can see that metadata is still of only limited use as a me of completely replacing cataloguing of net resources.

PRACTICAL SOLUTIONS

We have seen (and certainly experienced) that using automated harvest with or without some inclusion of author generated metadata, is less t entirely effectual as a single means of providing access to electro information. On the other hand, the amount of information becomi available in electronic form, combined with its extreme volatility mean t full human cataloguing of this material is beyond the resources of even t best funded library – especially as libraries still have to deal with their ha copy acquisitions, when 'more books are being published now than ev before, and the number of journals is increasing slightly.'[15]

A compromise solution has to be reached by librarians trying to mainta some sort of order in this increasingly chaotic environment. The following a a few pragmatic strategies which are worth keeping in mind when trying deal with this situation.

[14] O'Neill, et al.
[15] Haas Weinberg, p. 6.

FOCUS

To make a sweeping generalisation, librarians, especially cataloguers, are naturally tidy-minded people who want to create order everywhere Our professional training should make us realise, however, that we need to focus our efforts on activities that will have the most benefit for our users

> There is an enormous range of available information in the world The rôle of the library is to select, acquire, organise and make available an *appropriate subset* of these resources the goal of the library is to match the institution's needs and budget against the available information and its costs [16] *(my emphasis)*

No library would attempt to hold every single hard copy resource. Similarly, we should not try to give access to every electronic resource

We need to acknowledge the costs involved in cataloguing 'free' electronic resources in the way that we have always acknowledged the costs involved in acquiring hard copy resources. We should only try and catalogue those resources which are directly relevant and useful to our constituencies, with the explicit understanding that our catalogue records need to be constantly monitored and updated to ensure continued usefulness.

TOOLS OUTSIDE THE LIBRARY, OR OTHER PEOPLE'S HARD WORK

A very quick search of the literature shows that many libraries and other organisations have already embarked on projects to catalogue, index, organise or otherwise provide access to online resources. One would have to be very sure that their content is not relevant to one's users before replicating them in any sort of in-house system – Haas Weinberg points out the 'redundancy' of many library web sites.[17] A library has only finite resources which should be directed to where they can most benefit the library's constituency If a

[16] Rusbridge
[17] Haas Weinberg, p 10.

decision is made to catalogue publicly available information like web this should only be done where other tools do not adequately cov library's needs.

COLLABORATION

Similarly, there is much scope for collaborative efforts between librar provide access to online information. Our own history here in Australia ABN and its successor Kinetica show that these sort of collaborative e can have worthwhile results for all involved. The same difficulties pres by electronic resources are obviously still prevalent in the collabor environment, some of them more so. Somebody needs to take responsi for procedures for dealing with changes to sources, a responsibility that c easily fall through the cracks in a co-operative environment. Still, pro like OCLC's Cooperative Online Resource Catalog (CORC) may be worth keeping an eye on.[18]

PRIVATE ENTERPRISE

Haas Weinberg states that

> the primary problem I see is economic. Cataloguers have traditionally created records for materials that their libraries own. Book indexers are hired to add value to publications that will be sold, while journal indexers are paid to create thesauri and analyze documents for a database producer who charges for access to the resulting product. Nobody owns the Internet.[19]

Still, as people have paid for valuable databases in the past, people will doubt pay to use databases that add value to Internet based materials, and I sure there will be private companies offering such a database in the futu Libraries might make more constructive use of their resources by subscribi

[18] Hickey, Thomas 1998, 'Cooperative Online Resource Catalog explores uses for catalog o Internet resources' *OCLC Newsletter*, no. 235, September-October, http://www.oclc.org/oclc/new/n235/cooperative_online_resource_catalog.htm; and 'CORC: Cooperative Online Resource Catalog' http://purl.oclc.org/corc

[19] Haas Weinberg, p. 10.

to commercial databases of online resources rather than trying to organise them in-house in the same way that we now subscribe to journal indexes as well as the serials they index.

CONCLUSION

'Cataloguing has been described as the central mystery of libraries, although I have also heard it described as the international conspiracy of cataloguers.'[20] Many changes are taking place in the world of information – but the truth is still out there, and cataloguing still has its part to play in helping us understand and deal with this truth.

NOTES TO CONTRIBUTORS

Australian Law Librarian welcomes the contribution of articles, notes and letters to the editor. Articles should be 1500 to 3000 words

Refer to the *Style Guide* on the ALLG web site for details:

www.allg.asn.au/allgall/

In addition check the web site for:

- copy deadlines
- notes for advertisers,
- subscription rates
- contents of recent issues
- purchase of back issues

For further information contact the Editor:
Helen Wallace, Law Librarian, The University of Western Australia, Nedlands, W.A. 6907. email: hwallace@library.uwa.edu.au

[20] Barry, Tony, p 305