

# Collecting the best data: improving cataloguing systems

Alison Dellit  
Metadata Librarian  
National Library of Australia  
[adellit@nla.gov.au](mailto:adellit@nla.gov.au)

## **Abstract**

*As libraries come to grips with enormous changes in information-seeking behaviour, many institutions are seeking to build “new generation” catalogues, which make resource discovery simple and fun. To fully take advantage of these changes, libraries also need to re-think what data we should be recording about our collections; and how we are recording it. The National Library of Australia is developing a new tool to streamline the process of selecting a correct subject heading. In the future, an even more radical approach to subject analysis and classification may be required to efficiently catalogue the increasing amount of born-digital information.*

## **Introduction**

As libraries come to grips with the enormous changes in the information landscape that the proliferation of the internet has engendered, many institutions are seeking to build "new generation" catalogues, which make resource discovery as simple and fun as it can be through Google or Amazon. To fully take advantage of these changes, however, libraries also need to re-think what data we should be recording about our collections; and how we are recording it. At the moment, really good metadata provides the opportunity to cluster results, provide connections between related searches and effectively relevance-rank results. However, pressures on technical services departments mean that the resources available for cataloguing are not increasing, while the flow of material is. At the same time, new technologies offer up the possibility of alternative ways of classifying material, or of simply providing better search results. This paper will discuss some approaches to improve and simplify subject cataloguing, and will discuss other ways of sourcing data to improve our resource discovery services.

## **Identifying what is important – why do we catalogue?**

In an era of challenges to cataloguing, it is particularly important that we are clear on why our profession exists. We do not categorise for the sake of categorisation, but because classification of material enables the preservation, maintenance and discovery of resources. Cataloguers have developed sophisticated systems to enable people to discover items in library collections. In the days of card catalogues, standards were developed that enabled searchers to not only find an item by title or creator, but also to browse subject indexes. A common vocabulary for subjects, such as the Library of Congress Subject Headings (LCSH), was a way of ensuring that all items on a common topic would have records next to each other. Although dozens of words can describe any one concept, standards provided ways to make sure that a consistent term was used by different cataloguers, and even across different libraries. Complex guides to LCSH were developed, which enabled cataloguers to construct the correct heading. Authority records, set up for each LC heading, contained information about which term to use for a concept, with "see also" references clarifying what related terms a cataloguer should use the term for. As much as possible, detailed guides to the use of authorities try to impose consistency in the choice of which subjects to list, although there is inevitably a level of subjectivity in subject analysis.

Card catalogues did not make it easy for users to find items by subject, but the standards developed by cataloguers meant that it was at least possible. As catalogues moved online, it became easier for people to use keyword searches to identify material by topic. Online catalogues enabled people to search the subjects – if one of the keywords they used for the search matched one of the words in the approved subject heading, they would return a result and be taken to all of the material about an item. Users could also search by keyword, returning relevant items by title, and then use the subject heading assigned to a relevant item to find more material, which may not have relevant terms in the heading. Online catalogues were also set up to exploit non-preferred and related terms in subject authority records although this was not always intuitive to users.

Throughout this time, cataloguers have carefully developed and maintained an array of standards to ensure that material was ordered and categorised in a way that facilitated access to that material, and effective management of it. Workflows have been developed around those standards, and both the standards and their workflows have become an essential part of our profession. At the heart of our profession, however, remains the purpose of our work – our workflows and standards help us to allow users to find what they are looking for. Just like reference librarians, cataloguers fundamentally work to assist resource discovery.

## **Why do we need change?**

The information landscape is changing rapidly and fundamentally. In the last decade, the proliferation of information in digital form – most notably on the World Wide Web – has led to profound changes in the information-seeking habits of Australians, and people in most developed countries. It has never been this easy to find basic facts quickly – from S.R. Ranganathan's five laws of library science (Ranganathan 1931), to recipes for Anzac biscuits (Australian War Memorial 2007), resources are available quickly to anyone with an internet connection. This proliferation of information is changing expectations for searching – people now expect that a simple keyword search will return relevant results, without having to mediate the search through finding the correct categorisation. They bring into the library forum the search habits learned by using the easy and intuitive systems of Google and Yahoo.

### **Increasing material**

This proliferation of information is spawning more information, and new original items. This is placing considerable pressure on technical services departments in libraries worldwide. Every day at the National Library of Australia, we face the challenge of tackling the ever-increasing intake of printed materials, as well as identifying and archiving online publications, and other born-digital publications. Self-publishing is becoming easier with print-on-demand services. Recent years have also seen the growth in online journals, and open source software publications. These pressures are acute, because cataloguing is such a time-consuming and expensive process. Careful categorisation and description takes time to complete, and using strict controlled vocabularies requires a lot of training and experience.

### **Workforce changes**

On top of this, there are pressures on the cataloguing workforce. The NeXus survey, conducted by Dr Gillian Hallam in late 2006, illustrated many of the problems. According to Hallam, librarians are an ageing section of the population, and a high proportion of cataloguers is planning to retire in the next few years. To date, the indication is that this cadre of skilled technical services librarians will not be easily replaced from the ranks of new graduates. Firstly, the number of university courses including cataloguing in Australia is declining, so fewer graduates emerge with thorough training in how to allocate Dewey numbers, or an understanding of Library of Congress Subject Headings. Secondly, there are strong indications that younger employees change jobs more frequently, and are unlikely to stay in a job for long periods. Many of the graduates entering cataloguing are older, and are in the middle, or even towards the end, of their working life. Being highly skilled, requiring extensive training and on-the-job experience, cataloguing is especially at risk.

Precisely at the time when new resource discovery services rely heavily on the consistency and accuracy of bibliographic and authority data, the future of having very experienced cataloguers in adequate numbers is uncertain (Hallam 2007) .

### **New resource discovery systems**

It would be easy if we could respond to these pressures by abandoning cataloguing altogether, or at least abandoning the subject categorisation part of cataloguing. However, it is only now that we are seeing the real benefit of careful subject categorisation, through resource discovery systems that effectively leverage the power of subject categorisation in faceted browsing, and search broadening options. In the last 12 months, a rash of new OPACs, and other library search systems have been released that provide a much easier and friendlier search experience. They include the excellent TALIS system at the State Library of Tasmania (Talis 2007), the open-source, freely available VU Find (VU Find 2007) ; Ex Libris's new Primo system (Primo 2007), the North Carolina State Library (NCSU 2007), and WorldCat (OCLC 2007). All of these systems rely on the classification or categorisation of items in Dewey, Conspectus and/or the Library of Congress Subject Headings.

Clustering, often called faceted browsing, is not a new concept. In fact, S.R. Ranganathan's Colon classification system (Ranganathan 1971), developed early last century, was an attempt to develop a series of clusters, well before the technology allowed searches as flexible as now. Faceted browsing systems deliver immediate results while offering an easy way for users to increase the precision of their search through a range of refining options.

This development is not surprising. As David Weinberger's *Everything is Miscellaneous* states: "The solution to too much information is more information" (Weinberger 2007). What Weinberger means is that to sort through a deluge of information, more metadata – information about that information – is what is needed. The more results that a search query produces, the more likely it is that someone will want to narrow his or her search. Faceted browsing based on topics helps users to define what they want – a search on civil war will produce a result set, but also options to restrict the results to those dealing with the American Civil War, the Spanish Civil War, or the English Civil War, for example. Users might then want to refine a result set further, to limit it to a format – books, journal articles, audio, pictures or websites, for example. While there are fewer systems offering them as yet, there are other technologies that classification can enable, to make it easier to find relevant material. Offering options to broaden a search, for example (like clustering in reverse), or linking directly to similar items, or items by the same author. The more information we have in the metadata, the more options we can offer to users to get to what they want in a large result set. Current faceted browsing library systems rely on consistent and accurate subject data in both bibliographic and authority records. For the immediate future, that means working out how to catalogue a growing body of material without substantial new resources.

The National Library of Australia has already taken many steps to meet these challenges. A few years ago, the Library took the step of realigning acquisitions and cataloguing functions to create multi-skilled teams, realigning the workflow to most efficiently use our library management system, and automating where we can. In 2007, a new position was created to undertake innovation projects in technical services, and one of the first projects was to examine how we could streamline the

most expensive, and arguably most important, part of cataloguing: subject classification.

## **Developing the subject suggester**

The project began with an observation study of cataloguers, which showed that assigning the correct subject headings and Dewey numbers to items is a time-consuming aspect of cataloguing. As a national deposit library, the NLA carries out a great deal of original cataloguing, so this is a significant expense for the library. A costing study undertaken by the library in 2006-07 also indicated that there was a great deal of difference in cataloguing speed and efficiency between very experienced staff and newer staff – so much so that original cataloguing was performed more cheaply by more experienced staff, even though they were generally at higher pay points. The aim of the project was thus to see if there were improvements in our cataloguing systems that might make cataloguing simpler to learn, and save time.

The initial stages of the project involved an observation study of a dozen cataloguers working on original material. This study indicated that most of the time of original cataloguing was spent using different systems to identify the correct heading: not time spent determining the "aboutness" of the work. In most cases, cataloguers determined one or two "aboutnesses" very quickly, and then settled in to work out the correct heading. To do this accurately, cataloguers needed to check both the authorities and catalogue records of similar material, to ensure the heading was authorised, and that it was being used in an appropriate way. Sometimes they needed to check the rules for assigning subdivisions as well. The process of moving in and out of systems was complex, and harder for those new to cataloguing and less confident with the structure and norms of LCSH.

The tool that we are working on is aiming to be a one-stop desktop. It should allow cataloguers to complete most of the steps in the subject cataloguing process in one place. This will certainly make the subject cataloguing process more efficient. In addition, because it is based on searching of subject authority records, it will ensure the resultant subject headings assigned by cataloguers are accurate and consistent. It also shows existing bibliographic data, so that any inconsistencies can be highlighted and rectified by cataloguers. If it can significantly speed up the process of finding correct headings, it might be possible to assign more headings to material, providing more points of entry for a search. The aim is for faster, more efficient and higher quality work.

### **Data in the tool**

To allow the tool to be a one-stop-shop, it needs to pull together data from different sources. These include:

- Full Library of Congress subject headings (LCSH) authority file: as it is still the standard taxonomy used in libraries, LCSH is at the centre of the tool.
- Information on subdivisions: to construct subject strings to make the subject heading as specific as possible, we want to include information on subdivisions in the tool. This includes both general free-floating subdivisions as well as free-floating subdivisions under pattern headings such as Industries and Classes of Persons.

- Australian extensions to LCSH
- National Library of Australia institution specific subject terms
- Bibliographic data based on Libraries Australia search
- Dewey Decimal Classification (DDC) data.

### **Relevance –ranked search**

The subject suggester tool will offer keyword searching of the full LCSH file. All parts of the subject authority record will be searchable: main heading, scope notes and references. Results will be relevance-ranked, with a match in the main heading or see reference, as well as the currency of the heading taken into account. Subject headings in the Australian extension to LCSH will be weighted favourably over conflicting LCSH headings and will be clearly flagged as Australian. A search of subject headings will be a federated search of the Australian National Bibliographic Database (ANBD) by linking the tool to Libraries Australia Search. This makes it possible to exploit related searches in the ANBD. One example may be that the tool can tell us what subject headings are often used together with a particular subject heading. For instance, the subject heading "History – Study and teaching (Primary)" may be often associated with "Education, Primary – Curricula".

We are hoping to integrate comprehensive data on the Dewey Decimal Classification system into the tool, providing links to statistically and editorially mapped numbers to a subject heading. Another possibility may be to pull out DDC numbers used in bibliographic records with a particular subject heading as their main headings, and do our own statistical mapping. All of these search functions are possible given current search technology, and the availability of up to date data. We have built a preliminary prototype using a Lucene search engine based on some preliminary data, although we have not yet integrated a Dewey search.

### **Extra functionality**

The subject suggester tool will need to work seamlessly with the National Library's integrated library management system (ILMS). The cataloguing module of the ILMS is naturally the work space for cataloguers. Therefore the tool will need to be activated from this module, and we will build functions to export the suggested subject headings and DDC numbers back into formatted fields in the module. This simple step will make it much easier to prevent spelling or transcription errors made by cataloguers re-typing headings or DDC numbers into the module, as well as save time. We are also hoping to include a tool to automatically add a Cutter mark to the Dewey, saving cataloguers from having to look up the mark manually and add it.

The tool is not designed to change the fundamental task of our cataloguers in classifying: to determine what an item is about, and ensure that it is accurately described in a way that facilitates discovery. If implemented effectively, it will decrease the need for cataloguers to develop an in-depth knowledge of the structure and vocabulary of Dewey and LC in order to achieve proficiency. By placing more of the burden for determining the correct heading for any given concept onto the computer, it focuses cataloguers on the difficult intellectual work: finding out what an item is really about, how it would best be described, and ensuring consistency of subject terms. It could also allow more library technicians to move into original cataloguing work. It should also free up more time to think through how future developments might change subject cataloguing, and how we can build more effective systems to leverage the metadata we create.

## **Technology is changing how we do our work**

### **Digitisation & full-text indexing**

Current cataloguing workflows are based on providing metadata about an item not easily viewed or searched online. It relies on a human looking at an object – generally picking it up, and often reading or listening to content, and then deciding how to describe the attributes of that object within certain constraints. By contrast, digital objects can be viewed and experienced simultaneously by thousands of people, and analysed by computers. The amount of historical research that is available in full-text form now is dwarfed by that which is not, but this will change in the next few years – possibly very quickly. The March 10, 2007 *Economist* estimated that Google is digitising 10 million books a year. The availability of content in digital form will fundamentally change how we approach our work, and it is very important that we acknowledge this.

Google's highly effective web search algorithm is proof that full-text searching (or at least searching of full-text interconnected web pages) can provide useful relevance-ranked result sets in response to subject queries, without a classification system. The success of Google book search indicates that it is possible to do this with the full-text of books. In running some of the most frequent searches carried out on the National Library's catalogue through Google Books, most produce a range of relevant results. The downside, of course, is that the search still relies on the words typed in by the searcher matching the concept. So, someone searching for information on "indigenous people" will not find works that talk about Aboriginal or native people, even though they are likely to be relevant, unless they also contain the terms "indigenous" and "people". So surely we can do better than this?

However, just because we accept that there is a need to be able to group like-minded material together in order to facilitate discovery, does not inescapably lead to the conclusion that we will always need humans to assign subject headings. Cataloguers need to be aware of and actively follow trends in technology that might meet these needs in different ways.

### **Automated analysis & classification**

There are several technologies that seek to add categorised headings to full-text documents by using some form of automated classification, and this technology is moving fast. In 2005, the Library of Congress released a report that concluded that semantic analysis – a process through which a computer analyses the relationships between words in order to be able to draw conclusions about meaning – was a long way from being useful for classification (Greenberg, Spurgen & Chrystal 2005). Since then, however, several projects have emerged that offer more hope for such a system.

These systems work by taking text and analysing the relationships of the words within the text to draw some conclusions. Based on this analysis, a search engine can do one of several things. It can, as the Associative Search engine of WebCAT Plus (Webcat Plus 2007) does, present the user with a list of related words and allow the search to be broadened to include these words as well, thus enabling more relevant items to be found. This process can be reversed for categorisation – documents are scanned for the related terms, and then allocated a classification based on their contents. According to Mike Keller in a talk at the National Library on September 12, 2007, this kind of technology is being used by Stanford University to

automatically categorise books. Like many of these systems, however, Stanford relies on humans – subject experts – to check the lists of related words created by the computer.

Other systems will use this analysis to provide clusters within the results. A particularly good example is the Carrot search engine (Carrot 2007), which will form “meaningful” clusters out of the results and return those clusters to the users. This is not attaching a categorisation to the material, but it is facilitating faceted browsing, without any need for an underlying taxonomy. Like the WebCAT Plus engine, the result set is a little clunky, and the clusters are not often perfect, but they do offer a way forward for navigation. It is worth pointing out that while we talking about applying these technologies to full-text documents, there are implementations for using them on metadata as well. A recent D-Lib paper discusses the experiences of applying clustering technology to the OAIster corpus, and concludes that there is some potential in this technology (Hageborn, Chapman & Newman 2007). This is an important development, and the paper is well worth reading.

### **Human intervention**

The most effective of these systems, it is important to note, still rely on human intervention to check and refine the lists of “associated words” that a computer comes up with, and to check and refine the clusters. In several of the systems discussed in this paper, the same theme will emerge – computer-based categorisation operates more broadly than humans, enabling greater discovery options, but is generally much less accurate. It suggests that the future may be for computer suggested categorisation, mediated by humans to ensure accuracy. This would suggest that we utilise the recall of computer systems, with the precision that humans can provide.

Other technology, such as the system patented by Reconnind (Reconnind 2007), can “learn” how to categorise material after being trained on already categorised material. So, for example, the system might analyse 2000 news articles that had had LC headings attached. It would find patterns between those with the same term allocated – such as the fact that articles with the phrase HMAS tend to be categorised with Navy. It is obviously a limitation that the system needs a reasonably large sample set of already categorised data to proceed, but the results can have a relatively high level of accuracy. These systems are proving reasonably effective with short full-text items, or with material dealing with a limited range of topics. The techniques are very similar to the technology used to provide recommendations by sites such as Amazon or Netflix, and are developing quickly.

While none of these technologies have become part of mainstream library use, it is likely that in the future they will be. Whether the system would be used to do the entire categorisation, or, as is more likely, to produce suggestions which are checked by a human cataloguer, will remain to be seen. There are also dozens of decisions to be made in applying them. For example, do we need to apply an ontology, a classification scheme, or subject headings, to a work in order to assist discovery, or can we apply the technology at the point of searching, to create clusters? How do we integrate the legacy data that we have into new systems? What depth should classification hierarchies have? Are different classification schemes needed for different types of collections, and/or for different audiences? Should proper nouns,

dates and named entities be specifically identified and classified? We need to start thinking about these questions and providing answers.

### **Our users: a new source of data**

There is another, significant and exciting source of data to assist categorisation, which is user-added tagging of material. There is nothing inherently new about tags – they are just keywords, not even necessarily subject keywords, attached to a record about an item, generally in online systems. These kind of descriptions are sometimes referred to as “folksonomies”; and the growth of Web 2.0 – where users are more involved in content creation – has spurred their development. One of the first services to popularise tags was Flickr, an online photo-sharing site. People attached tags to their photos as a way of enabling keyword searches to find their images. The tags could be anything – they did not have to be selected from a list, nor did they follow rules (such as singulars or plurals). This is an approach which gives most librarians heartburn – with no structure, surely material tagged "cat" will be forever separated from that tagged "feline"?

The reason tagging is powerful, however, is because it is done in such large numbers, which is a way of providing data on such a large scale that it can be used to draw some conclusions about the way people use language. If an individual cataloguer just added simple keywords to an item as descriptor, then these items would almost certainly be unfindable, as the terms would reflect the narrow world view that every individual has. But what would happen if 100 cataloguers typed in free text words to each item? Obviously, the MARC record would become very large, but it is likely that all the different ways of representing one concept would be reflected, so whether a user typed in “adolescents” or “teenagers”, they would find relevant items. Tagging works in the same way. As hundreds of users tag the same items, over and over again, patterns start to emerge in the tags.

So, the obvious question is: why would users want to tag items? Well, for a start, the most successful implementations of tags have come where users have a real stake in being able to find the resources that they are tagging, or allowing others to find them. In Flickr, people tag their own images so that more people will come and look at their work. Then people tag others’ images that they really like, in order to make those images more discoverable.

### **LibraryThing & bibliographic tagging**

The big tag success story in the bibliographic world is LibraryThing (LibraryThing 2007). LibraryThing is a social cataloguing site – it allows people to easily create a catalogue of items that they own. One of the features is tags – a field that allows people to just type in simple descriptors for their books. The system stores what at last count was up to 22 million tags, which users have assigned to books.

LibraryThing offered people tags as a way to find items within their own collections. Tagging has really taken off in this system because people want a way to sort and organise their own collections, and identify items within their collections. Interestingly, LibraryThing also makes it easy for people to import LCSH headings for their items. So why do they tag instead? In a seminal article in 2005, Clay Shirky (Shirky 2005) argues that tagging is easier for people than categorisation, because there is no sense that you have to find a definitive place for an item, but rather that items have associations with different terms, much like human memory creates associations between memories. There is less hesitation before tagging than before categorisation: put simply, we find it easier to do.

Tags are a perfect way for people to find things in others collections at LibraryThing—to browse to items that you might like, and to find items on the same or similar topics. I have less than 80 books entered into LibraryThing, and nearly 1000 tags have been assigned to those books. Natural groups of language and concepts occur in tags assigned to the same item. For example, some of the different terms that people have allocated to Hunter S. Thompson's *Hell's Angels* are just synonyms – one person's bikers is another's motorcycle gangs. However, many represent different concepts: sociology, drugs, gangs, journalism, violence, culture, and autobiography, for example. By using a mass of people, tagging allows more "aboutnesses" to be identified for any given book. Tagging is not alone in this – one of Recommind's findings in its automated classification is that the system will, on average, assign twice as many relevant subject headings to an item as an individual cataloguer would. This means that by using other ways of subject assignment – tagging, automated analysis – we could potentially improve access to items in the collection, providing more points of entry for a search.

### **A societal mind-map**

Not all tags relate to the subject of a work. People can tag by the colour and size of the item, characteristics of the author (African-American, or Muslim) and even by the value of the work (*Valley of the Dolls* is frequently tagged "trash" on LibraryThing, other titles are tagged "holidayreading"). These all provide points of entry for a search that are valid, and will increase usefulness of our search systems beyond the title/author/subject access points we currently offer. It may become useful, however, to define what attribute a tag is defining, and its relationship to the work. Or is this trying to impose order in the wrong way? It is likely that we would need some user management – sorting and arranging - of tags. LibraryThing uses crowdsourcing to sort its tags – allowing users to manually create associations between synonyms, so myth and mythology are linked together, as are chic lit, chick lit and chik lit.

One way that we could use tags would be to cluster the synonyms around Library of Congress Subject Headings (LCSH), thus utilising both the interrelated structure of LCSH that lends itself to faceted browsing, and the diversity of tags. This would make LCSH a much more user-friendly system - a bit like adding all the possible terms to a "see also" field in the authority record. But not all topic tags have a clear LCSH heading. It can take a while for a new heading to be approved - terms like steampunk, for example, can gain currency much faster than LCSH accepts. Do we only accept subjects that will correlate to an LCSH heading, or can we accommodate others? We could, for example, use tags to help generate proposals for new LCSH headings.

Even further into the future, could we create an ontology through mapping tag correlations with each other, bypassing the step of authority records altogether? Flickr has the capacity to analyse the sets of tags assigned to an item, and to create clusters. Thus, a search on Jaguars will sort the images that also are tagged with zoo/animal terms from those who also have tags about cars. The end result is a disambiguation between the two meanings of the word, which offers the user the choice of navigating to the correct topic – all without using a structured ontology at all.

The search works very well with straightforward topics, like Jaguar, Apple (the computers separated from the fruit), or mouse. It also works quite well with more ambiguous topics, from the point of view of offering options. A search on "holiday",

for example, produces clusters on vacations, Christmas, Easter, beaches and Paris. Not a comprehensive list, but representative of what most people are tagging with “holiday”. Other clusters offer options to narrow the search: a search on environment offers clusters based on green activism; nature shots or urban pollution, for example. The system also allows you to find broader terms that co-occur in tags, so a search on “Niagara Falls” will offer the option of searching on waterfalls, or Canada, for example.

Tags have many advantages. Shirky points out that the “signal loss” is much less than in any organised thesaurus. By this, he means that while the popular tags will dominate results, the less well-used tags are still there; they are not abandoned in order to come up with the best uniform solution. Tags will change over time, and move to reflect the terms people use, as well as the concepts that they represent. Not only are they easy to assign, but they more accurately reflect the way that people think and see the world than any one heading, no matter how democratically decided, ever could. This trend is part of “crowdsourcing”, a term used to describe the phenomenon of services based on the participation in a small way of thousands, even millions of people, which produces something much greater than the sum of its parts. The internet facilitates such easy participation that crowdsourcing has taken off in recent years. It represents a new sort of democracy – one based on participation in the development of a service.

Users will not simply tag library books, however, just because we want them to. LibraryThing is the only website to have millions of bibliographic tags, and it is clearly because there is a built-in incentive to tag in the service. If we are going to develop tags, we have to think about incentives to tag. Google image search, for example, has created a simple game out of tagging. This is the sort of approach libraries should think about as a way to develop plain language descriptors. Making it easy to use our site to create a bibliography is essential, enabling a “push to del.icio.us” feature and enabling compatibility with Zotero, for example. We need to develop collaborative arrangements to share tags between libraries, thereby generating a critical mass of tags that could not be created from one library alone.

## **Conclusion**

In the future, there will be a variety of different ways to assign subject headings, and indeed to facilitate discovery. While at the moment, there are no easy ways to assign subject headings automatically, it is likely that in the future there will be some combination of automatic analysis, and human checking, refinement and alteration. As well as this, the aggregation of large amounts of “crowdsourced” data could increase the wealth of our metadata immeasurably, and enable much more effective systems. These systems will not only save time, but will also probably increase the breadth of description, and preserve more, not less, information about our collections.

Those who bemoan technological developments because they mean the end of our profession have missed the point, and badly: our profession is more needed now than ever, and our usefulness is not our memory for Dewey numbers. We may stand on the brink of a brave new world, where computers can read texts, and mass amounts of metadata are available to be mined and exploited in presenting search results. The possibilities are wide and varied for how we organise this data, and plotting the right path, offering the right choices, working out what standards we

need, and what we don't, exploiting the right elements of the metadata, and clustering and grouping the right concepts are all challenges that lie ahead. Cataloguers have a much greater experience and theoretical understanding of the difficulties of organising data to provide access than any other profession. This is a time when our expertise in classification is sorely needed, and we need to dust off our basic principles, and think about the needs we are trying to meet. If we are not prepared to engage with technology with an open mind to changes, while focused on the outcomes that we need, then we will do the future of public access to information a grave disservice.

### **Acknowledgements**

While all opinions and any errors are my own, I am indebted to Wan Wong, Kent Fitch, Judith Pearce, Debbie Campbell and Deirdre Kiorgaard for their support, advice and help in refining my ideas, and the writing of this paper.

## References

- Australian War Memorial 2007, *ANZAC biscuits recipe*, Available from: <http://www.awm.gov.au/encylopcedia/anzac/biscuit/recipe.htm> [September 15, 2007].
- Carrot 2007, *Search demo*, Available from: <http://demo.carrot2.org/demo-stable/main>. [September 15, 2007].
- Ex Libris 2007, *Introducing Primo*, Available from: [http://alphasearch.library.vanderbilt.edu/primo\\_library/libweb](http://alphasearch.library.vanderbilt.edu/primo_library/libweb) [September 15, 2007].
- Greenberg, J, Spurgen, K & Chrystal, A 2005, *Final report for the AMeGA (Automatic Metadata Generation Applications) Project*, Chapel Hill, Library of Congress.
- Hagedorn, K, Chapman, S & Newman, D 2007, 'Enhancing search and browse using automated clustering of subject metadata'. *D-Lib*, 13, 7/8, Available from: <http://www.dlib.org/dlib/july07/hagedorn/07hagedorn.html>.
- Hallam, G 2007, *The future starts now: an analysis of the current library workforce*, Melbourne, Infonet.
- LibraryThing 2007, *What is LibraryThing?*, Available from: <http://www.librarything.com/> [December 5, 2007].
- North Carolina State University 2007, *Library catalogue*, Available from: <http://www.lib.ncsu.edu>. [September 15, 2007].
- OCLC 2007, *WorldCat*, Available from: <http://www.worldcat.org>. [September 15, 2007].
- Ranganathan, SR 1931, *The five laws of library science*, Madras, Madras Library Association.
- Ranganathan 1971, *Colon Classification, ed. 7 (1971): a preview*, Madras, Sarada Ranganathan Endowment for Library Science.
- Recommind 2007, *Mindserver Application*, Available from: [http://www.recommind.com/mindserver\\_categorization.html](http://www.recommind.com/mindserver_categorization.html) [September 14, 2007].
- Shirky, Clay 2005, *Ontology is Overrated*, Available from: [http://www.shirky.com/writings/ontology\\_overrated.html](http://www.shirky.com/writings/ontology_overrated.html) [September 15, 2007].
- State Library of Tasmania 2007, *Talis Online Catalogue*, Available from: <http://www.talis.tas.gov.au> [September 15, 2007].
- VU Find 2007, *VU Find*, Available from: <http://www.vufind.org> [September 15, 2007].
- WebCat Plus 2007, *Associative Search*, Available from: <http://webcatplus.nii.ac.jp/en/>. [September 15, 2007].
- Weinberger, D 2007, *Everything is Miscellaneous*, Chicago, Times Books.