

Subject cataloguing : faster, better, cheaper

Wan Wong & Alison Dellit
National Library of Australia
wwong@nla.gov.au
adellit@nla.gov.au

Abstract

As the volume of born-digital materials increases, and the workforce changes, there is a growing need to find ways to allow faster and better ways to create or collect catalogue data. In this environment, many libraries are looking for ways to catalogue more efficiently, and some are even asking fundamental questions like whether subject analysis is necessary at all, particularly for full-text electronic materials. The National Library of Australia is actively investigating ways to make cataloguing more efficient, while preserving the consistency and integrity of the data. Some of the projects that the National Library has investigated include using online forms to generate basic MARC records, and improving our documentation. In this paper, we will discuss one of those projects, the development of a new tool designed to assist subject cataloguing.

Subject cataloguing is often a time consuming process. The new tool currently being developed will pull some of these existing tools together to facilitate the cataloguer's search for the most appropriate and up-to-date subject heading. Cataloguers will be able to check the currency of bibliographic records using a particular heading, and its scope notes if available. The tool is intended to fit in with, and offer improvements, to the way cataloguers perform subject cataloguing.

Future technological developments in automatic metadata extraction and automatic classification may fundamentally change the way subject cataloguing is done. Similarly Web 2.0 technologies have the potential to influence the way subject information is provided and accessed. The paper will look at the implications of these developments.

Introduction

The explosive deluge of information we face now, especially online information, has created a lot of opportunities and challenges for libraries. Within libraries cataloguers also find themselves faced with the challenge of needing to keep up with print as well as online publications, not to mention various other carriers of information not traditionally collected by libraries, including web sites, blogs, chat rooms, etc.

As far as subject cataloguing goes, the Library of Congress Subject Headings (LCSH) is still the predominant standard thesaurus used in libraries in the English-speaking world. Our cataloguing processes and online catalogues are still built around it. Libraries invest a lot of resources in training their cataloguers to understand how subject headings should be constructed and assigned, and in providing various tools

for cataloguers to use. Although often found wanting in guiding users through reference structures in subject authority records, online catalogues do provide users with a structured way of identifying relevant library materials and collections on a particular subject.

The subject cataloguing process has been studied in the National Library, and in 2006 a project was formed to investigate the feasibility of developing an in-house subject suggester tool. The tool will facilitate the cataloguers' efforts to search and construct the most appropriate subject headings for the item in hand. By functioning as a one-stop-shop, the tool will bring together various data the cataloguer needs to perform subject cataloguing.

In the meantime however we cannot stop crystal ball gazing with exciting Web 2.0 and Library 2.0 developments happening to date. We are witnessing the increasing use of social tagging in library environments and advances in automated keyword assignation. Are they friends or foes for the widely adopted structured controlled vocabularies in the world's libraries? This paper will look at some of the issues and implications surrounding those new technological developments.

The subject cataloguing process

The subject cataloguing process is a time consuming process. To assign subject headings as specific as possible to an item, cataloguers typically need to spend some time reading information from the item, such as introduction, foreword, table of contents, etc. After analysing the subject, the next step is to construct appropriate subject headings to best reflect the subject matter of the item.

In the National Library cataloguers have various tools at their disposal to assist them to construct subject headings. They include Classification Web, Library of Congress Authorities online, the Libraries Australia authorities database, the Subject Cataloguing Manual in printed form or online via Catalogers' Desktop, the Australian National Bibliographic Database (ANBD) as well as our own local database.

As preliminary investigation into workflow improvements in the subject cataloguing process, we have observed a variety of cataloguers performing subject cataloguing. It was extremely interesting to see that different people use different systems at different stages of the process. Their search strategies in some of the online tools such as Classification Web are not always the most effective either.

To develop a useful subject suggester tool, we are not planning to replace the cataloguers' judgement and subject analysis with automated classification or assignation of subject descriptors. Instead we are looking at whether we can pull most of the existing tools together to create a one-stop-shop so that cataloguers do not have to think what to look up next, but can complete most of the subject cataloguing process in one place.

A one-stop-shop

In developing a useful subject suggester tool, we are not planning to replace the cataloguers' judgement and subject analysis with automated classification or assignation of subject descriptors. Instead we are looking at whether we can pull most of the existing tools together to create a one-stop-shop so that cataloguers do

not have to think what to look up next, but can complete most of the subject cataloguing process in one place.

The various sources of data the tool needs to bring together include:

1. the whole LCSH file

The complete LCSH file needs to be the major search target for the tool, and the tool should allow cataloguers to search the whole authority record, from the authorised heading to all the references as well as all notes fields. While the broader and narrower terms present in any subject authority records can be searched through hyperlinks, the tool should also lead cataloguers into all possible related searches. These possible related searches may be other subject headings that are found to be in the same bibliographic records as the heading being searched for. For instance, if you are interested in the subject heading "Personal coaching", the tool can present you with a list of subject headings used in bibliographic records that have "Personal coaching" as their first subject heading, such as "Self-actualization (Psychology)", "Mentoring" or "Executive coaching". The tool may also provide a related search on narrower terms that are not listed in the authority record of a broad subject heading, like the various kinds of birds not listed as narrower terms in the authority record for the heading "Birds" but all with "Birds" as the broader term in their own authority records.

2. subdivisions listed in the Subject Cataloguing Manual (the "H lists")

To construct the most specific subject heading strings for an item using LCSH, cataloguers very often need to add subdivisions to main headings. We are looking into ways to incorporate information on free-floating subdivisions into the tool. One way to do this may be to search for the information in real time by linking into the *Subject Cataloging Manual: Subject Headings* which for National Library cataloguers is available via *Catalogers' Desktop*. The other way may be to present and maintain the information as static information but allow searching and display in relation to a particular subject heading. For instance, a cataloguer looking at using the subject heading "Chinese Australians" will be able to search for the pattern headings for Ethnic groups listed in the instruction sheet H 1103 and also consult the pattern headings for Classes of persons with ethnic qualifiers in the instruction sheet H 1100. Naturally the general free-floating subdivisions would be available for all headings.

3. the Australian extension to LCSH

Sometimes Australians do think differently and use different terms from our American colleagues. For instance "Education, Elementary" is the valid LCSH heading but "Education, Primary" makes much more sense to Australian users. We need to incorporate the Australian extension to LCSH in the tool and indicate very clearly on those occasions when we do deviate from standard LCSH.

4. National Library institution specific subject headings and thesaurus

Likewise in the National Library we have institution specific subject terms that we have developed and maintained in-house. We apply them to certain categories of collection materials such as music and original materials. To make it a one-stop-shop the tool will provide cataloguers with ready access to the information.

5. bibliographic data in the ANBD

There are a lot of times when cataloguers need to refer to bibliographic records and see how particular subject headings have been used. This is especially true when a cataloguer is trying to differentiate two or more similar subject headings from each other, or to define a concept contained in a heading. The tool links to Libraries Australia Search and tells the cataloguer from the outset how many records have included a particular subject heading, and among them, how many are from the National Library and how many are from the Library of Congress. This should prove to be a more efficient way to search bibliographic records instead of cataloguers needing to do separate searches in the local database or the LC online catalogue.

By interacting with Libraries Australia Search the tool can also do additional searches within the search results on a particular subject heading. These may include using the information in bibliographic records to suggest other possibly related subject headings and related DDC numbers.

6. correlation search of DDC

From our internal study, looking up a DDC number based on the first subject heading assigned is typically the last step a cataloguer takes before completing the cataloguing process. To provide a look-up facility for the DDC number is therefore a logical functionality to have in this new tool. We are investigating how to give cataloguers access to DDC numbers editorially or statistically mapped to a particular subject heading. Ideally the suggested DDC numbers will be provided in context by including the captions as well as the hierarchy from which the number is derived. It will also be interesting to see if we can extract data from the DDC tables for cataloguers to use.

Then to complement this information the tool can tap into records on Libraries Australia, search on bibliographic records that use a particular subject heading as their first subject heading and return a relevance ranked result of all the DDC numbers in those records. The relevance ranking mechanism will need to rank recent NLA records over records from other libraries, as well as ranking records indicating the use of DDC Edition 22 over records using other editions.

Relevance ranking

Indeed relevance ranking of result sets will be one of the major challenges in the development of this tool. The initial search of the tool will be a simple keyword stemmed search that searches the whole authority record including the authorised heading, the references and the notes fields. For instance, a search for "Teachers" will retrieve "Teaching" as well as "Early childhood teachers". At the same time it also searches the ANBD to retrieve records that have included "Teaching" as a subject heading, possibly with a multitude of different subdivisions.

This means that the relevance ranking of the results needs to take into account matching of the searched term with terms found in the authority record and matching of the searched term with terms found in bibliographic records. When a cataloguer searches on the term "Overfishing", the tool can look at all the See references in the authority record, such as in this case "Fish populations -

Overfishing". If a bibliographic record contains "fish populations" as a phrase anywhere in the record, the subject headings contained in that record can be treated as possible related searches and their relevance may be boosted.

The currency of the heading and the frequency of use in the ANBD should also be part of the relevance ranking mechanism. Newly created headings will get a high score in currency but a low score in frequency and it will require some extensive testing to make sure the balance is right.

Interactivity with ANBD and ILMS

As mentioned above, the tool will need to interact with Libraries Australia Search so that cataloguers can refer to bibliographic data and can limit them to records originating from the National Library or from the Library of Congress if they so desire.

More significantly, for the tool to give us the efficiency it needs to work seamlessly with our local library management system. Cataloguers working in the Cataloguing module should be able to bring up the tool very easily. When the cataloguer decides to use a particular subject heading or a heading string they can copy and paste part of it or the whole thing into Voyager. Or better still, they can export part of or the whole heading or a string back into a formatted 650 field with the correct subfields in the record. Likewise it would be ideal for the suggested DDC number to be dropped into a formatted 082 field by just one click.

Web 2.0 subject searching

Ontologies like LCSH have emerged as a way of enabling users to search for items by subjects. There are now several catalogues and catalogue prototypes, including WorldCat, the Pennsylvania State University Catalogue¹, and the new VU find² and Primo³ applications, which effectively leverage LCSH as it is now to assist end users to find items by topic. Using a controlled vocabulary is a way of ensuring that all the items on a particular topic are brought together - so books on the Murray Darling Basin and the Murray Watershed are recognised as being similar. It also disambiguates - so material on Dewey the person is separated from material on Dewey the classification system, and the fence in your backyard is separated from the art thief's best friend. Until now, our OPACs have been slow to leverage the information inherent in a Library of Congress subject heading, but we are starting to see some innovative applications.

But is this enough? Will subject assignment by trained professionals continue to be the most effective way to meet these needs? This section of the paper will look at some of the pressures that will influence the future of subject cataloguing, and which might fundamentally change our jobs in the future. The first important development is digitisation of material.

¹ <http://citeseer.ist.psu.edu/763420.html>

² <http://www.vufind.org/>

³

http://alphasearch.library.vanderbilt.edu/primo_library/libweb/action/search.do?vid=VANDERBILT&reset_config=true

We are already dealing with new material in digital form, whether through electronic galleys or online scholarship. On top of this, many libraries have digitisation programs, and Google is taking this up a notch. According to the March 22 *Economist*, Google is now digitising 10 million books a year, and we can assume this will fundamentally change the balance of what is available. Even if it is not publicly available, having the full text can assist in search and categorisation.

Digitisation

Full-text items already contain a number of words relating to the subject, in almost all cases. These words will generally occur more frequently, and good relevance ranking can distinguish between books dealing principally with Australian Aboriginal Health, and health books that casually mention it. It is a mistake, however, to assume that therefore relevance ranking of un-linked full-text items would end the need for subject headings. Firstly, because there will always be a wealth of material that is not full-text even if digitised - pictures, objects and even some manuscripts. Secondly, because this has the same limitations as a thesaurus added without structure - you are reliant on the author using the same words as the searcher. There are a couple of techniques being worked on at the moment, however, that seek to overcome this latter problem.

Automated analysis

Automated Metadata Extraction is a term that refers to the process of building computer software that can look at a webpage, or a full-text document, and guess what the metadata is. In the last few years there has been significant progress in making this work to extract information such as creator, date published, publisher and so on. So far, however, it has proved much more difficult to use it to apply subject descriptors to the system. The Library of Congress did a study that reported in late 2005⁴, concluding that this technology was a long way away. Research is continuing, and is a development we should keep an eye on.

There is a form of automated classification, however, which is very different in principle. Unlike the above approach, this approach does not attempt to teach the computer program how to understand language, or meaning. Instead, it regards the relationship between metadata and the original document as a series of patterns, and tries to analyse those patterns.

In Recommind's patented system⁵, for example, the computer program is "trained" by analysing a few hundred or a few thousand full-text documents and the subject headings that have been allocated to those headings by trained cataloguers. It looks for patterns - for example, if the term 'HMS' appears frequently in the document, then 'Navy' tends to be one of the subject headings allocated.

It will then attempt to allocate headings to full-text items that have not yet been classified, and those will be corrected, further training the system. Eventually, the program has enough understanding of patterns that it can successfully predict subject headings and allocate them.

⁴ Greenberg, J., K. Spurgen, and A. Crystal, *Final report for the AMeGA (Automatic Metadata Generation Applications) Project*. 2005, Library of Congress: Chapel Hill, North Carolina

⁵ http://www.recommind.com/2007/mindserver_categorization.html

This kind of technology is being widely used for different purposes - it lies behind a lot of the recommendations offered to you by Amazon, or Netflix, for example. It is most successful, however, with short full-text documents such as news articles, or with documents that deal with a similar topic, where the variance isn't huge. It is obviously also only applicable to text items. There are several programs which claim to be able to analyse the patterns in sound or pictures, but none that are being consistently used for subject analysis.

Tagging

This brings us to quite possibly the most interesting development in the library world - user-generated tags. There is nothing really new about tags. They are simply free keywords entered to describe an item - not necessarily even the subject. It is an approach that is not used by cataloguers because without a thesaurus, each cataloguer will use terms that are specific to him or her, and there is no way to link related concepts together. Even if three cataloguers were using free text words with each item, the result wouldn't be the wide variety of terms used by people searching for items.

But what if 100 cataloguers typed in free text terms? Well, for a start, the results wouldn't fit on a catalogue card. But the bits and bytes are small enough to store in a digital world. There is no reason that a metadata record cannot be longer than the original item. Obviously, there is no contemporary institution that can afford to have 100 cataloguers working on each item, which is why the essential element to tagging is that it is not carried out by trained professionals, taught to put things into categories, but by users, who are simply typing in terms they think they would want to retrieve that item through.

The big tag success story in the bibliographic world is Library Thing⁶. Library Thing 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 assign to books.

LibraryThing offered people tags as a way to find items within their own collections, but discovered that tags are a perfect way for people to find things in others collections - to browse to items that they 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. But 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

⁶ <http://www.librarything.com/>

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 assignation - tagging, automated analysis - we could potentially improve access to items in the collection, providing more points of entry for a search.

One way that we could use tags would be to cluster the synonyms around a Library of Congress Subject Headings, thus utilising both the interrelated structure of LCSH that lends itself to faceted browsing, and the diversity of tags. This would be easy enough by plotting terms that tend to co-occur. 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 cyberpunk, 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? Can we use tags as a way to generate new proposals for LCSH headings?

Perhaps even further into the future, could we create an ontology through mapping tag correlations with each other, bypassing the step of authority records altogether? Another site that uses tags is the photo uploading website Flickr. Flickr has the capacity to analyse the sets of tags assigned to an item, and to create clusters. So 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. And 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 a representation of what most people are tagging with "holiday". Other broad searches are 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. [

Conclusion

We are a very long way from this sort of technology replacing subject classification, or even from the point when we could say that it will do so. At the moment, the only significant body of tags for books is at LibraryThing, and it is hard to imagine libraries building up the mass necessary, given the low incentives for people to tag books they do not own, and are not trying to organise for themselves. The number of books not covered by LibraryThing dwarves those that are, particularly when you are considering newly released books, and automated analysis still doesn't work well for long documents. Even Flickr's tag clusters only work for one term at a time, not for searches of more than one term.

But it is important that the cataloguing profession is aware of trends in technology and social life that might change the fundamentals of what we think our role is. It is clear that in the future, where we increasingly incorporate user-generated data or

automatically-generated data into our data repositories, our cataloguing processes and resource discovery systems will be more focussed on how to make the various kinds of data work together to achieve the best results for users.

Making information accessible, now and into the future, is at the heart of our profession. Our input into the best ways to do that is invaluable for a society groaning under the weight of information overload. If we do not think about how to embrace and use new technologies, we will cede the space to others with less training, history and often, less public interest. To quote Ranganathan, "Every book its reader" remains the reason why cataloguing is needed. By providing good access to information using all available data intelligently and efficiently, we can then "Save the time of the reader", which is the librarians' role in this world of information overload.