12

The future of metadata

Metadata in some form has been utilised in cultural heritage institutions since their inception. Metadata practices in these institutions were originally conceived for a specific set of functions (as described by Svenonius),[1] which were typically implemented in particular applications, for example, library catalogues. Yet, increasingly, attention is paid to the way in which metadata impacts usability, management and preservation of digital resources. These institutions are using metadata in ways that go beyond the single view of a resource that is provided by the exclusive use of a single metadata format.

Flexibility of metadata is gaining in importance. The expected functions of information applications in cultural heritage institutions change over time. The potential sources for metadata are diversifying, and harvesting metadata from end-users no longer seems a strange idea. Contextual information not previously considered important in many parts of the cultural heritage community is increasingly recognised for its value in explaining and interpreting a resource and guiding its users.

This diversification can be argued as leading to an *increase* rather than a decrease in the sheer volume of structured metadata. More and more pieces of structured information will likely exist for a given resource in the foreseeable future. Each will be related to that resource but each may be provided at different times, by different agents, or for different purposes. A major challenge facing the cultural heritage sector is to find efficient ways of gathering this data together (at least intellectually), to relating it to the same resource, and to use (and re-use) this metadata appropriately for new generations of applications as they emerge.

Automated metadata generation

As described in Chapter 4, automated metadata generation is already gaining a place in digital project workflows, and this trend will likely continue. Certain types of technical metadata are even best generated automatically rather than manually. Descriptive metadata presents a much greater challenge to automation, however.

Systems and algorithms to analyse documents and their structure are under constant development. The DC Dot utility,[2] for example, is a small tool, easy to test, that generates metadata from web pages mostly by suggesting keywords based on an analysis of the text. Tools like this are far from being useful for unattended subject analysis for most needs within the cultural heritage sector, but the state of the art is constantly improving. While automated classification methods were originally developed to work on full documents, recent work has attempted to apply the same principles to metadata records rather than full texts.[3] Automated content analysis for non-textual documents is an area of ongoing research as well, and these methods could potentially be used to generate metadata for storage together with human-generated metadata, or to perform searching on the fly. Research domains for content-based retrieval of still images and recorded music are particularly active.

Some middle ground between automated and manual methods is almost certainly the best choice for most digital library projects in the current environment. It is likely that many of the automated methods we do use will operate as 'supervised' activities, involving a human to make critical decisions, review output at key points, or develop 'training' material to adjust a system to a new type of content. Adopting automated creation methods does not have to require relinquishing control over the metadata creation process, and the use of these methods does not necessarily mean abandoning any core values deeply held by the cultural heritage community, such as the desire to provide effective collocation of like resources. Our challenge as new options for metadata creation evolve is to continually analyse how they can be used as tools to achieve our discovery and management goals for our collections. While our goals are likely to be influenced by new developments, we can still continue to demand metadata that meets our core values, no matter how it is created.

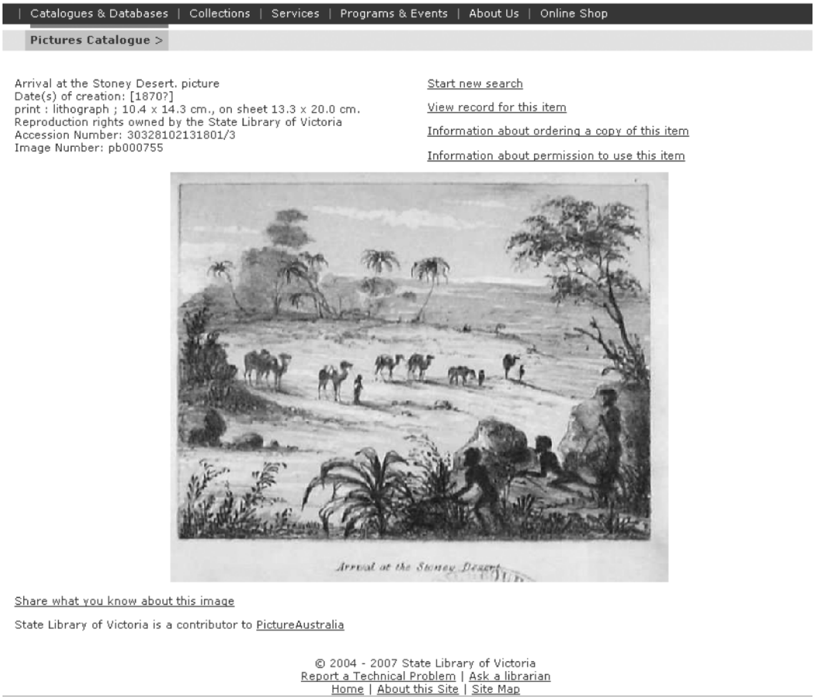
Web 2.0 ideas: participation and mashups

A recent surge in the number of applications and technologies based on the contribution of users has spawned a term to describe the phenomenon: Web 2.0. The value of these applications lies in the strength and breadth of the user community. Systems for user participation generally fall into four categories: recommendation/reviews, tagging, games, and coordinated efforts. All have potential for use in digital library applications.

Recommendation systems operate in two fundamental ways. The first is to solicit textual commentary and/or numeric ratings from users. This could simply be treated as descriptive metadata, its origin tracked, and this information displayed to a user. The second method is to use usage patterns to infer that a user ‘recommends’ a resource. This approach is taken in a wide variety of applications, from Google’s PageRank algorithm to Amazon’s ‘other customers who bought *x* also bought *y*’ feature. Options for ranking or generating recommendations based on circulation data have been analysed or implemented by North Carolina State University[4] and the California Digital Library. The latter concluded from a user study that circulation data as a source for recommendations is useful in only a limited fashion. ‘The preferred sources of recommendations cited by participants are faculty, bibliographies and footnotes’.[5] Analysis of which search results or browse options are used most often could also be a source of data for driving recommendation systems in digital libraries.

Tagging systems have emerged as online resources that allow users to manage items of interest to them, for example personal photos in the case of Flickr,[6] or Web bookmarks in the case of del.ic.io.us.[7] The *folksonomy* (the full set of terms applied by users to all objects in a given repository) that emerges from these activities is useful for study as well; it represents the vocabulary actual users use to describe resources. This vocabulary is often markedly different from the terminologies employed by controlled vocabularies in the cultural heritage sector. In the digital library realm, tagging systems could be of use both to solicit description from end-users (which could supplement or in some cases replace structured subject analysis) and as a source of new vocabulary terms. Experiments with tagging are happening in the cultural heritage sector, including PennTags[8] at the University of Pennsylvania, which allows users to tag catalogue records together with Web resources; Steve, a research project from the museum community studying user tagging behaviour for art museum resources;[9–11] and at the State Library of Victoria in Australia which invites users to ‘share what you know about this image’[12] (see Figure 12.1). The cultural heritage sector could also look to partnerships with commercial entities implementing tagging systems for cultural heritage resources, such as LibraryThing.[13]

Figure 12.1 Picture displayed by the State Library of Victoria (Australia) – note the link to 'share what you know about this image'.


| Catalogues & Databases | Collections | Services | Programs & Events | About Us | Online Shop
Pictures Catalogue >
Arrival at the Stoney Desert

Arrival at the Stoney Desert. picture
Date(s) of creation: [1870?]
print : lithograph ; 10.4 x 14.3 cm., on sheet 13.3 x 20.0 cm.
Reproduction rights owned by the State Library of Victoria
Accession Number: 30328102131801/3
Image Number: pb000755

Start new search

View record for this item

Information about ordering a copy of this item

Information about permission to use this item

Share what you know about this image

State Library of Victoria is a contributor to PictureAustralia

© 2004 - 2007 State Library of Victoria
Report a Technical Problem | Ask a librarian
Home | About this Site | Site Map

Tagging systems do not necessarily use completely unstructured data. Flickr has implemented a feature called 'machine tags' as part of a programming interface to Flickr content. Machine tags allow a user to place a tag in a particular namespace, i.e., put it in a particular category, such as 'medium:paint=oil'.[14] The del.icio.us system has added 'tag descriptions' to its service, allowing users to create titles and descriptions for tags used, that other users can view to more fully understand the meaning of a tag. If they prove to be user-friendly, structured tagging systems could be leveraged in digital library applications to more fully integrate metadata created by end-users with that created by specialised staff.

While most tagging systems operate by offering the user an easy way to perform some sort of resource management task he presumably already has a need for, other tagging systems use the concept of fun to entice users to tag, then use the user-created metadata to improve retrieval for a set of materials. Tagging games generally operate on the model of pairing two users and challenging them to use the same tag to describe a resource, or some variation of that model. Implementations of these games exist for still images, such as the ESP game[15] and the Google Image Labeler,[16] and for sound recordings, such as MajorMiner[17] and the Listen Game.[18] A particularly interesting variant on using users to create or enhance metadata is the reCAPTCHA service, which provides distorted text used to fight Web form spam while simultaneously collecting the correct text from an image that has been inaccurately interpreted by an OCR program.[19]

Large-scale collaborative efforts for metadata creation also exist. Wikipedia is likely the most well known, but several appear in the cultural heritage sector as well. In the archival community, users have been asked to contribute to the correction of OCR performed on manuscript documents, using an annotation platform.[20] Project Gutenberg is a long-standing cooperative project that operates on a distributed proofreading philosophy – asking any and all volunteers to proofread as small or large amount of text as they like. The current project evolved from an information sharing initiative begun in 1971, and the Project Gutenberg site today provides access to more than 20,000 electronic texts.[21]

User participation in creating digital resources doesn't have to stop with metadata creation. Both the Web 2.0 world driving user expectations and the increasing capabilities of distributed digital libraries are enabling the re-use of digital content in new and unanticipated environments. It is likely that the division between resources held by a cultural heritage institution and those owned by our users will continue to fade. This division has long been problematic in academic libraries, for example, where professors teach with both library resources and their own, but have traditionally needed to manage those resources in two different ways. Current sharing technologies are making this user-centred (rather than institution-centred) grouping of resources easier. The National Library of Australia's Picture Australia portal, for example, federates picture collections from libraries, archives and museums in Australia. They have gone a step further, however, and asked users to contribute their own digital images of Australia. Contribution by users is done not by asking them to visit a particular site or to go through a lengthy donation process, but rather in an environment in which they likely already operate: Flickr. On Flickr, a 'group' exists for PictureAustralia, and a user simply has to add the tag for that group. By adding a few simple extra pieces of data that would likely be provided by many users regardless, the image can be submitted for consideration for formal acquisition by the PictureAustralia collection.[22]

As part of the project Naming,[23] the community of an Inuit village was asked to help the work of curators in illustrating scenes and pictures of the Inuit community history. Without the output of this project, no trace would be left of the people and events of this place in the past, beyond oral history and human memory. Although not online, this experience illustrates how it is possible to request the contribution of those who are not information specialists to build a community around an archive of digital resources. Another example is the Kete archive in New Zealand,[24] which was created with the contribution of local associations to generate content from the community and illustrate the community memory.

The trend of many academic libraries towards setting up institutional repositories, which contain the cultural and scientific output of research institutions, is geared towards obtaining both digital content and metadata from users. These users are a carefully controlled bunch, however, defined as scholars who have established authority and credibility. These same researchers may also contribute their work to domain-specific portals such as arXiv.[25] The resources deposited in these repositories are not necessarily published and validated material, representing in many cases a category of material that research libraries have not previously collected. Researchers are often expected to deposit the articles themselves, although at some institutions librarians provide support for this activity. Because the metadata in this case is generally created by content specialists who are nonetheless novice metadata creators, it is necessary to provide user-friendly interfaces that allow researchers to quickly deposit their papers. The use of complex controlled vocabularies in systems such as this can be particularly problematic. Institutional repository applications must be designed with these realities in mind, providing metadata creation forms that are as short as possible, importing metadata from other applications whenever possible, capturing as much information as possible from the content itself, and so on.

To facilitate expanded use of digital content, in addition to participating in protocols such as OAI-PMH and soon OAI-ORE, cultural heritage institutions are increasingly looking to content sharing sites as means of reaching more users. Content sharing sites such as YouTube,[26] for example, provide mechanisms for embedding a resource (a video in their case) in a web page, either by the institution or by a user (see Box 12.1). This provides an easy way to allow third party sites, such as blogs, to cite digital objects.[27]

Box 12.1 **Example of a YouTube link to embed content in third party pages**

```
<object width=“425” height=“350”><param
name=“movie”
value=“http://www.youtube.com/v/tRpxKHlRQUc”>
</param><param name=“wmode”
value=“transparent”></param><embed
src=“http://www.youtube.com/v/tRpxKHlRQUc”
type=“application/x-shockwave-flash”
wmode=“transparent” width=“425”
height=“350”></embed></object>
```

Great benefit can be obtained from integrating user-contributed content and metadata into digital library systems. In order to embrace this benefit, applications and management strategies for digital resources must be expanded to take into account less structured as well as a wider range of metadata types and formats, while still providing high-quality services. User contributions can be part of an institution’s metadata management strategy.

Defining a strategy for metadata management

To meet the expanding need for robust metadata services, cultural heritage institutions must define a clear strategy for metadata management. This strategy should include adequate high-level descriptions of the various collections of content handled by the institution. It should identify the metadata that is needed about the acquisition process, the deposit process, the digitisation process, and all other processes that are part of the workflow for digital collections. The strategy should define the necessary metadata to manage the resource and to connect metadata with a series of services. As a result, the metadata management policy reflects the mission of the institution and the collections it holds. When institutional missions evolve, new tools are created, or new collections are obtained, an existing metadata management strategy should allow easier adaptation for these changes.

A sustainable metadata management strategy will also include a component devoted to metadata sharing. A metadata sharing strategy will outline the types of services in which the digital resources should be visible and how this visibility should be optimised to best serve the target audience. Collaborations in joint projects can be a good way of developing a sharing strategy. For example, the CIC Metadata Portal was a project jointly funded by Midwestern Universities of the CIC academic consortium.[28] Some Universities in the consortium already had OAI-PMH repositories, and most of them had previously shared their content previously through Z39.50 gateways and other initiatives, but few had formal strategies for how they would share their content and under what circumstances. Over the course of the project, many aspects of the OAI-PMH-based metadata sharing process were investigated. All of the participating Universities had implemented an OAI-PMH data provider by the end of the project. Each learned a little about how to better share their content. In addition, several of the participating Universities had implemented additional local services, providing guidance to other institutions, leading local networks, and developing their own services, for example, Wisconsin Heritage Online.[29]

When developing a metadata sharing strategy, many questions must be asked. How do I want my content to be viewed by other applications? How will this serve my objectives and my mission? How will I serve my existing audience? How will I look for new audiences? The metadata sharing strategy should include metadata design for resource discovery, but also the development of collection-level descriptions and the implementation of reliable identifiers and location mechanisms for digital resources.

It is also important to state any potential limitations that might be placed on the use of metadata records. While unrestricted sharing of metadata records best promotes the types of imaginative re-use described here, very real situations sometimes intervene and require the placement of restrictions on use. Be careful applying usage restrictions to metadata records; rarely does the loss of control that comes along with metadata sharing override the benefits to the institution of a looser usage model.

Providing the most liberal usage conditions for both metadata records and for content as possible promotes flexibility in the future use of these materials. New and revolutionary usage models are difficult to anticipate, and the effects of any restrictions given may be unnecessarily limiting. The Scientific Commons initiative, which proposes to assign machine-readable intellectual property rights to scientific digital objects, illustrates this principle by suggesting that rights be assigned such that agents would be 'legally allowed to use the next killer technologies at will'.[30]

Evolving institutional missions

Cultural heritage institutions are now heavily involved in the development of new services that often borrow technologies and practices from commercial applications. The evolution of these systems to provide ever better access to digital resources has expanded the missions of cultural heritage institutions. Metadata competencies are increasingly part of these missions. Dedicated metadata librarians have existed in cultural heritage institutions at least as far back as 1995.[31] This particular position, like many others, came into being as a diversification of a traditional cataloguing job. Since this time, many more metadata specialist positions have been created in cultural heritage institutions, with a wide variety of responsibilities.

Cooperation among institutions is also finding its way into mission statements. To support the increase of the creation and exchange of metadata within the cultural heritage community, useful tools are being developed to edit and transform metadata, for example the RDF tools from Simile,[32] free software packages for creating and manipulating MARC data such as MarcEdit,[33] utilities to edit embedded XML and ID3 metadata,[34] and higher-level tools for metadata management systems such as those from OCKHAM.[35]

Evolving sets of tools, competencies, and expertise are developing in cultural heritage institutions. Whereas it would be inappropriate to consider metadata as a *new* issue facing the cultural heritage sector, the continuing changes in information technology requires ongoing evolution of institutional missions. New types of services are often better built collaboratively rather than at individual institutions, and the role of the commercial sector in these developments can no longer be discounted. Including these services in institutional missions is increasingly important. Developing mission statements that cover the creation and distribution of digital content requires in-depth analysis of the way these institutions work and represent their content. This requires definition of the audiences and usage models that will be supported, with room for incorporating new models that have not been anticipated. Metadata models must change in response to changes in usage models. The purpose of a metadata manager in these institutions should be to oversee this evolution.

As demonstrated over the course of this book, all metadata is created in light of a specific purpose, intentionally defined or not. The cultural heritage institutions that are most effective in creating flexible, shareable, quality metadata will be those that explicitly tie metadata planning and creation practices directly into institutional missions that reflect the current environment, and can react to new developments. Only by positioning ourselves to fully participate in the information environment in which our users operate can cultural heritage institutions continue to fulfil our primary missions.

Notes

1. Svenonius, E. (2000) *The Intellectual Foundation of Information Organization*. Cambridge, MA: MIT Press; Chapter 2.
2. UKOLN. 'DCdot – Dublin Core metadata editor', available at *http://www.ukoln.ac.uk/metadata/dcdot/*.
3. Newman, D., Hagedorn, K., Chemudugunta, C. and Smyth, P. (2007) 'Subject metadata enrichment using statistical topic models', paper presented at the JCDL'07 Conference, Vancouver, Canada.
4. Antelman, K., Lynema, E. and Pace, A.K. (2006) 'Toward a 21st century library catalog', *Information Technology and Libraries*, 25(3): 128–39. Available at *http://www.lib.ncsu.edu/staff/kaantelm/antelman_lynema_pace.pdf*.
5. Whitney, C. and Schiff, L. (2006) 'The Melvyl Recommender Project developing library recommendation services', *D-Lib Magazine*, 12(12): available at *http://www.dlib.org/dlib/december06/whitney/12whitney.html*.
6. Flickr (*http://www.flickr.com/*).
7. Del.icio.us (*http://del.icio.us/*).
8. University of Pennsylvania, PennTags (*http://tags.library.upenn.edu/*).
9. Steve – The art Museum Social Tagging Project (*http://www.steve.museum/*).
10. Trant, J. (2006) 'Social classification and folksonomy in art museums: early data from the steve.museum tagger prototype', paper presented at the ASIST-CR Social Classification Workshop, 4 November 2006. Available at *http://www.archimuse.com/papers/asist-CR-steve-0611.pdf*.
11. Chun, S., Cherry, R., Hiwiller, D., Trant, J. and Wyman, B. (2006) 'steve.museum: an ongoing experiment in social tagging, folksonomy, and museums', paper presented at the Museums and the Web 2006 Conference. Available at *http://www.archimuse.com/mw2006/papers/wyman/wyman.html*.
12. See an example on the State Library of Victoria website: *http://www.slv.vic.gov.au/platebk/0/0/0/doc/pb000755.shtml*.
13. LibraryThing (*http://www.librarything.com/*).
14. Flickr. 'Flickr API/discuss', available at *http://www.flickr.com/groups/api/discuss/72157594497877875/*.
15. Carnegie Mellon University. 'The ESP game', available at *http://www.espgame.org/*.
16. Google Image Labeler (*http://images.google.com/imagelabeler/*).
17. Major Miner – Music Labelling Game (*http://game.majorminer.com/*).
18. University of California at San Diego Computer Audition Laboratory. 'Listen game', available at *http://www.listengame.org/*.
19. Carnegie Mellon University. 'Recaptcha', available at *http://recaptcha.net/*.

20. Couasnon, B., Camillerapp, J. and Leplumey, I. (2004) 'Making handwritten archives documents accessible to public with a generic system of document image analysis', paper presented at the First International Workshop on Document Image Analysis for Libraries DIAL'04, available at *http://doi.ieeecomputersociety.org/10.1109/DIAL.2004.1263255*.
21. Project Gutenberg (*http://www.gutenberg.org/*).
22. Picture Australia. 'PictureAustralia and Yahoo! invite you to contribute your photographs to PictureAustralia, using Yahoo!'s online image repository, Flickr', available at *http://www.pictureaustralia.org/Flickr.html*.
23. Library and Archives Canada, 'Project naming', available at *http://www.collectionscanada.ca/inuit/*.
24. *http://blog.kete.net.nz/*. See also Krajewski, P. (2006) 'La Culture au risque du 'Web 2.0': analyse à partir de la création d'une archive numérique communautaire open source néo-zélandaise, KETE', Dissertation DCB Ecole National Supérieure des Sciences de l'Information et des Bibliothèques, available at *http://halshs.archives-ouvertes.fr/halshs-00120016*.
25. Cornell University Library. 'arXiv.org', available at *http://www.arxiv.org/*.
26. YouTube (*http://www.youtube.com/*).
27. Google Books has implemented a similar mechanism to cite book paragraphs based on widgets. Tungare, M. (2007) 'Share and enjoy', *Inside Book Search* blog, Google: 6 September. Available at *http://booksearch.blogspot.com/2007/08/share-and-enjoy.html*.
28. University of Illinois at Urbana-Champaign. 'CIC metadata portal', available at *http://cicharvest.grainger.uiuc.edu/*.
29. University of Wisconsin-Milwaukee. 'Wisconsin heritage online', available at *http://www.wisconsinheritage.org*.
30. Wilbanks, J. (2007) 'Science commons – copyrights and experiences harvesting open content', presentation at the OAI5 conference, CERN, Switzerland, 20 April 2007.
31. McClellan, G. (1995) 'Job Posting: Metadata Specialist', *AUTOCAT: Library cataloging and authorities discussion group* (AUTOCAT@UBVM.CC.BUFFALO.EDU), 30 November 1995, 14:30:56–0600.
32. MIT Libraries. 'Semantic interoperability of metadata and information in unlike environments', available at *http://simile.mit.edu/*.
33. Oregon State University. 'MarcEdit', available at *http://oregonstate.edu/~reeset/marcedit/html/about.html*.
34. OptimaSC (*http://www.optimasc.com/products/dex/index.html*).
35. OCKHAM. 'Downloads/Services', available at *http://www.ockham.org/services.php*.