If we're not there yet, how far do we have to go ?
A review of web metadata at the University of Melbourne

Eve Young[1] and Baden Hughes[2]

1. Metadata Coordinator,
Information Acquisition & Organisation Section
Information Division
The University of Melbourne, Victoria 3010, Australia
e.young@unimelb.edu.au

2. Research Fellow,
Department of Computer Science and Software Engineering,
The University of Melbourne, Victoria 3010, Australia
badenh@cs.mu.oz.au

1. Introduction

The University of Melbourne instituted a metadata standard for web content in 1999. After several ad hoc reviews, we describe a case study in which we conduct a systematic, broad coverage review of the use of the institutional metadata standard across all web content. First we review the development of the institutional metadata standard, building on Dublin Core. Next we conduct an in depth review of practice, finding that Dublin Core and University of Melbourne metadata is sparsely implemented for reasons including changing authoring habits and evolving web content types. Finally a number of issues for future investigation are identified within the institutional context.

2. Grounding, Problem Identification

The University of Melbourne's World Wide Web Publishing Policies and Guidelines (University of Melbourne, 1999-2001), first published in 1999 specified the inclusion of nine largely administrative meta tags, some relating to authorization and currency of the page's information.

Despite this explicit standardisation, it was long suspected that a substantial amount of content was out dated and inaccurate. The need for systematized administrative metadata was shown through 2 research projects carried out by The University of Melbourne's Web Centre in early 2002, to investigate and analyse the scope of the non-compliance, and identify the issues that contributed to it. This research was on a small percentage of the web pages as being manual it was found to be too time-consuming to be carried out on a broader scale.

The first item investigated was the currency of the expiry date element, a notation indicating the date at which a content review and renewed publishing authorisation was expected. The review found 72% non compliance with the expiry date element (Zajacek, 2002b). Such a low degree of compliance made it difficult to realise the administrative and resource discovery benefits if tags are not implemented or implemented correctly. Questions relating to the low compliance rate for the expiry date tag prompted additional research. Analysis of 608 files within one directory on the main web server revealed a number of interesting statistics: 167 pages, or 27% had a set value in the meta tag; 2 pages had 'not applicable' as the value in the

meta tag; 381 pages, or 63% had no date in the meta tag and 58 pages or 9% had a value that was empty or unreadable. Given that the expiry date tag was the primary mechanism for managing and reviewing the currency of web page content, the analysis revealed that only 27% of the pages audited had an appropriate value in the tag. Even if the automatic email notification functionality had been activated, the quality of the values in the expiry tag would have rendered this functionality worthless.

The second item investigated was the institutional A-Z index. After analysing 78 web pages (pages linked from the A, B and C entries in the A to Z index on the web site) the research revealed that the meta tag with highest degree of compliance was content type (a default value) with 84.6%, the lowest was expiry date with 11.5% compliance. The A-Z index audit found more unknown than known maintainers in a 78 page audit (Zajacek, 2002a). Other information has been hard to come by. It is proving difficult to identify and document present maintainers, as the expiry date tag analysis has shown

3. Towards Standardization

In May 2001, Eve Young submitted a paper identifying the growing importance of metadata, the emergence of standards such as Dublin Core, and recommended the university consider the application of a metadata standard to official web pages. The Information Strategy Advisory Committee (ISAC) endorsed Young's recommendations in July 2001, and requested the establishment of an expert committee (the "Metadata Working Group" (hereafter, MWG)) to advise on the implementation of a uniform approach to the creation of metadata on university web sites.

The MWG membership drew on expertise from across the university and included two ISAC representatives, a metadata expert, academics, IT managers, librarians and web managers, and the metadata coordinator. The MWG researched existing metadata standards (Dublin Core, IMS, EdNa), investigated the implementation of metadata standards in other large, information rich organizations such as universities and libraries, analysed the effectiveness of existing meta tags within the university web pages, determined the metadata requirements in terms of resource discovery and management of the web site, and explored the use of complimentary standards such as document type and audience.

The MWG recommended Dublin Core as the metadata standard using 12 of the then 15 elements and some of the date qualifiers on the basis that Dublin Core was an international standard; relatively easy to implement; used by many other universities and governments; suited current requirements and environment; and ensured future interoperability with other emerging metadata standards such as IMS and LOM, which may be used on campus in other contexts.

Thus the University of Melbourne metadata standard directly imports from Dublin Core the following elements: DC.Title, DC.Creator, DC.Subject and Keywords,

DC.Description and Description, DC.Publisher, DC.Contributor, DC. Rights, DC.Date, DC.Date.Modified, DC.Language, DC.Format, DC.Identifier.

The entire set of Dublin Core metadata elements were not adopted since only these core elements were viewed as being of interest to the institution (adopting a minimal barrier to entry was one related consideration, however, we see later that even the mandated minimum is met a very small proportion of the time.)

It is also worthwhile noting that the adoption process included no provision for the underlying Dublin Core standard changing and hence for the institutional adoption to be modified accordingly, a point to which we will return later.

Additionally the MWG identified the importance of metadata in the administration and management of the web pages, including content and publishers, and the integration with other projects such as the development of a web archiving strategy and the implementation of a content management system for web publishing. Motivated in part by the Dublin Core extensions for administrative metadata (Ianella and Campbell, 1999), some local administrative inclusions were added to the core Dublin Core elements: UM.Creator.Email, UM.Date.ReviewDue, UM.Authoriser.Name, UM.Authoriser. Title, UM.Maintainer.Name, UM.Maintainer.Email.

The MWG developed a draft metadata standard based on Dublin Core that was initially endorsed in July 2003. This was successfully trialled prior to implementation in the Metadata Pilot Project and was initially implemented on web pages required to use the university web page templates, as specified in the guidelines (University of Melbourne, 1999-2001).

4. Broad Scale Compliance Analysis

In order to provide data for analysis, a full crawl of the University of Melbourne web presence was conducted in March 2005. The software used was the Internet Archive's Heritrix suite (Internet Archive, n.d.), an open-source, extensible, web-scale, archival-quality web crawler project. In total 57Gb of data was retrieved from www.unimelb.edu.au and its associated sub-domains over a period of 146 hours. A total of 1.4 million documents were retrieved.

4.1 The nature of the University of Melbourne web environment

A number of caveats are in order in interpreting these statistics: 1) the statistics do not include sites blocked through robots.txt style exclusions; 2) the statistics do not include sites requiring any form of authentication; and 3) the statistics discussed below only include explicitly linked retrievable documents.

MIME Type	***# Documents***	***Size (Mb)***

MIME Type	*# Documents*	*Size (Mb)*
text/html	659,171	9,000
image/jpeg	246,153	8,712
image/gif	217,913	1,539
application/pdf	62,862	2,440
text/plain	8,877	659
application/msword	7,619	1,222
application/msexcel	2,639	2,562
application/mspowerpoint	1,551	2,224
application/postscript	983	714
All Others	213,887	28,221
Total	1,421,645	57,893

A number of interesting observations can be made with regard to the crawled data and its relation to earlier metadata curation efforts.

1) HTML is no longer the dominant format : The University of Melbourne's metadata creation processes have been primarily oriented at creating Dublin Core-extended metadata as simple HTML meta tags (University of Melbourne Web Centre, 2004). However, data gathered from the March 2005 crawl shows that pure HTML content in fact is no longer the largest constituent at either a numerical or size-wise rank. As such, a metadata standard which makes this assumption is increasingly outmoded either for reasons of expression or compatibility with content creation tools.

2) Web-accessibility of "non-native" document types : Many of which are not addressed by the UniMelb guidelines for metadata creation but which do offer some potential for restricted metadata inclusion eg MS Word, MS Excel. However, emerging document types such as XML and RDF do not easily allow for the embedding of metadata internal to the resource.

3) The emergence of dynamic documents: An analysis of the "All Other" categories shows that many (around 38%) of these documents are in fact dynamic, that is generated server side on demand by technologies such as PHP, ASP and JSP. No thought currently has been given to the inclusion of metadata in automatically generated documents of this type although it would be possible using simple templates.

4.2 Usage of University of Melbourne metadata

An analysis of the HTML documents collected from the March 2005 crawl was undertaken to determine the degree of compliance with the organisational requirements for metadata. The mechanics of this analysis were to write and execute scripts inspecting the HTML <HEAD> elements for relevant metadata tags as well as the HTML <BODY> elements.

Element	***% HTML Pages with Metadata in <HEAD> Element***	***% HTML Pages with Metadata in <BODY> Element***	***Total % HTML Pages containing Metadata in either <HEAD> or <BODY>***
DC.Title	23.9%	8.5%	32.4%
DC.Creator	18.0%	11.6%	29.6%
DC.Subject and Keywords	26.4%	12.5%	38.9%
DC.Description and Description	28.5%	8.0%	36.5%
DC.Publisher	18.4%	24.4%	42.8%
DC.Contributor	16.3%	13.4%	29.7%
DC.Rights	2.4%	14.0%	16.4%
DC.Date	18.7%	17.7%	36.4%
DC.Date.Modified	13.8%	25.1%	38.9%
DC.Language	6.2%	15.2%	21.4%

Element	*% HTML Pages with Metadata in <HEAD> Element*	*% HTML Pages with Metadata in <BODY> Element*	*Total % HTML Pages containing Metadata in either <HEAD> or <BODY>*
DC.Format	5.3%	13.3%	18.6%
DC.Identifier	4.9%	7.8%	12.7%
UM.Date.ReviewDue	17.2%	29.1%	46.3%
UM.Creator. Email	46.4%	15.7%	62.1%
UM.Authoriser.Title	32.3%	25.9%	58.2%
UM.Maintainer.Name	48.9%	12.5%	61.4%
UM.Maintainer.Email	42.6%	24.7%	67.3%

A number of observations can be drawn from the statistics above:

1) Differences between core Dublin Core and institutional metadata: institutional metadata is clearly more regularly contributed, despite the automatic creation of some of the Dublin Core metadata by authoring environments.

2) Alignment with broad Dublin Core norms: These figures are generally in line with the findings of broad scale Dublin Core-oriented metadata communities such as OAI (Ward, 2003) and OLAC (Hughes, 2004), at least in the ratio of the relative elements.

3) Correlation with manual inspection statistics : these broad scale experiments suggest that the trends detected in earlier focused studies such as Zajacek (2002a, 2002b) are valid. In fact, it appears that either owing to the passage of time or broader base data that the statistics from earlier surveys may in fact err on the side of optimism.

4) Differences between metadata included in <HEAD> vs <BODY> elements: for institutional metadata, there is a strong tendency to include metadata in the <BODY> elements where it is immediately visible on the page rather than in the <HEAD> elements. In part this reflects the emphasis of the training materials distributed to date.

5. Reflections and Challenges for the Future

While at first glance it may appear that the percentage based coverage of the HTML collection at the University of Melbourne is quite low, it is useful to remember the actual number of documents represented here – more than 650,000. It is true that many of these documents are non-compliant for quite easily identifiable reasons – exclusion of metadata in template based pages such as those within the learning management system as an example.

Large scale search engines such as Google are not using meta tag information any more but perform full text indexing (see Richardson, 2004). Hence the benefit to general web search of metadata creation according to a given standard within the institution is almost zero for external searchers, although it may still retain currency for other administrative purposes eg the authorization of web content publication. This leads to the need to distinguish between the institutions need for web content management, and how metadata facilitates this goal, and decoupling from web search experience in general.

The pending introduction of institution wide Content Management System is likely to have pervasive effects. Notably, the existing metadata standards failed to address distributed content creation (or underestimated the pervasive effect of "publish to web" type technologies to all staff), and as such we are eager to address this as early as possible with new generation tools and practices. Embedding metadata creation at the point of content creation, leveraging inbuilt capacity to create as much core metadata as possible is certainly desirable.

A number of items for future work have been identified from the analysis carried out in this paper.

In the first instance, a return to first principles with regard to the currency of the Dublin Core Metadata Set and the expression of University of Melbourne metadata is warranted. While the existing metadata standard was in part driven by resource discovery needs, these motivations have largely been surpassed with the arrival of fully featured web search engines, and as such, the role of institutional metadata appears largely to be administrative.

At a more practical level, there is the need to consider the impact of new content management systems at the institution and subsequently how web metadata should be created in this new environment. Given the largely administrative focus of University of Melbourne metadata, changes to work practice to encompass true publishing authorisation and embedding of non-repudiation needs to be considered. One possible way forward is to build a "compliance audit" service, perhaps integrated into the Content Management System deployment, for run time verification of metadata compliance, with a "watermarking" service which automatically imprimaturs compliant pages in the absence of manual inspection. However, this will require the formalisation of University of Melbourne metadata as a true Dublin Core application profile and an associated formal schema, and the creation of controlled vocabularies for extensions.

In retrospect, there are a large number of pages which will be updated only at an irregular interval, and as such substantially increasing the coverage of institutional metadata in the short to medium term may require the deployment of an automated metadata creation service such as DCdot (Powell, 2000) or an augmentation service

such as OLACdot (Hughes, 2005). Early experiments with DCdot show significant promise, but need to be more carefully evaluated in light of recent research in the area (Greenberg, 2005).

Finally, we see training as of critical importance. While significant effort was invested in training key personnel, and the propagation of the institutional standards and training notes online, only a small number of face to face classes have been held.

6. Conclusion

Despite being identified as one of the leading universities with regard to metadata implementation (Ivanova, 2004), on reflection we see that The University of Melbourne still faces significant challenges in the deployment of metadata across institutional web content. Indeed "getting there" may itself be an ill-conceived notion – even over a 2 year period we have found that compliance is in fact a moving target with the evolution of external standards, web content creation tools, and web content demography. While a strong basis for institutional metadata was formed by the adoption of Dublin Core, the disparate content creation environment and rapidly changing composition of web content has induced a less than satisfactory application of these standards. Automated metadata creation and assessment, forming a significant component of future work may address this problem in part, although only a longitudinal study, with adequately established baseline metrics will demonstrate if we are any closer to the holy grail.

References

American National Standards Institute, 2001. Dublin Core Metadata Element Set is now standardized as ANSI (American National Standards Institute)/NISO (National Information Standards Organization) Z39.85 (Oct. 2001) http://www.niso.org/standards/resources/Z39-85.pdf

Jane Greenberg, 2005. Final Report for the AmeGA (Automatic Metadata Generation Applications) Project. Accessed 25 April 2005 at http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf

Baden Hughes, 2004. Metadata Quality Evaluation: Experience from the Open Language Archives Community. Proceedings of the 7th International Conference on Asian Digital Libraries. LNCS 3130. Springer Verlag. pp. 320-329.

Baden Hughes, 2005 (to appear). Towards a Flexible Framework for OLAC Metadata Enrichment. Proceedings of the DELOS Digital Repositories: Interoperability and Common Services Workshop.

Internet Archive, n.d. Heritrix. Accessed 25 April 2005 at http://crawler.archive.org/

Renato Ianella and Douglas Campbell, 1999. The A-Core: Metadata about Content Metadata. Accessed 25 April 2005 at http://metadata.net/admin/draft-iannella-admin-01.txt

Nelly Ivanova, 2004. Metadata and Australian Universities: An Environmental Scan. Proceedings of AusWeb 2004: The Tenth Australian World Wide Web Conference. Southern Cross University.

Andy Powell, 2000. DCdot Dublin Core Metadata Editor. Accessed 25 April 2005 at http://www.ukoln.ac.uk/metadata/dcdot/

Joanna Richardson, 2004. Competing in a World Scooped by Google. Proceedings of AusWeb 2004: The Tenth Australian World Wide Web Conference. Southern Cross University.

Jewel Ward, 2003. A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative. Proceedings of the 2003 Joint Conference on Digital Libraries. IEEE Computer Society Press. pp. 315-317.

Eve Young. 2004. Metadata at the University of Melbourne. Accessed 25 April 2005 at http://buffy.lib.unimelb.edu.au/ird/metadata/index.htm

Eve Young. 2003. Metadata Working Group Papers and Reports. Accessed 25 April 2005 at http://buffy.lib.unimelb.edu.au/ird/metadata/index1.htm

Eve Young and Martine Booth. University of Melbourne Metadata Implementation. Presentation DC-ANZ Conference 2003, University House Canberra, Feb 2003. Accessed 25 April 2005 at http://www.dc-anz.org/conf2003/DC-ANZAgenda.html

Martine Zajacek, 2002a. Non-compliance and Implementation issues raised by the Metadata Working Group. University of Melbourne. Accessed 25 April 2005 at http://www.unimelb.edu.au/development/wag/2002-02/agenda.html

Martine Zajacek, 2002b. Expiry Date Tag: analysis of sample report. Mss to Web Advisory Group, University of Melbourne http://www.unimelb.edu.au/development/wag/2002-02/agenda.html

University of Melbourne, 1999-2001. World Wide Web Publishing Policies and Guidelines. _ Accessed 25 April 2005 at http://www.unimelb.edu.au/guidelines/pre2004/web-guidelines.pdf

University of Melbourne Web Centre, 2004. University metadata standard: what, why & how. Mss. University of Melbourne. Accessed 25 April 2005 at http://www.unimelb.edu.au/webcentre/training/trainingfiles/metadata_training_notes.pdf

Author/s:
Young, E.; Hughes, B.

Title:
If we're not there yet, how far do we have to go ? A review of web metadata at the University of Melbourne

Date:
2005-06

Citation:
Young, Eve and Hughes, Baden (2005) If we're not there yet, how far do we have to go ? A review of web metadata at the University of Melbourne.

Persistent Link:
http://hdl.handle.net/11343/33834

File Description:
If we're not there yet, how far do we have to go ? A review of web metadata at the University of Melbourne