Optimising Synergy between Metadata, Database Platform and Business Needs: the Case of SIM at RMIT

Rhys Williams
Learning Technology Services
RMIT University
rhys.williams@rmit.edu.au

Troy Boulton
Learning Technology Services
RMIT University
troy.boulton@rmit.edu.au

Phil Anderson
Multi Media Database Systems
RMIT University
phil@mds.rmit.edu.au

Cherryl Schauder
Metadata Project Officer
RMIT University Library
cherryl.schauder@rmit.edu.au

Abstract:
In 2000 RMIT University launched a Web Refurbishment Project using RMIT Multi Media Database Systems' SIM (Structured Information Manager) software. The vision of the project was to create an enterprise-wide information environment which encompasses functions from communication and learning, to knowledge and information management, and record keeping. To achieve these goals the SIM database platform was configured to make extensive use of a mix of technical and business metadata to underpin the diverse functions required of the system. This paper outlines how metadata operates at different levels of the Web publishing system, from the perspectives of system development and programming, project design and management, and metadata standards development.

Introduction

In 2000 RMIT University launched a Web Refurbishment Project using RMIT Multi Media Database Systems' SIM (Structured Information Manager) software. The vision of the project was to create an enterprise-wide information environment which encompasses functions from communication and learning, to knowledge and information management, and record keeping. To achieve these goals the SIM database platform was configured to make extensive use of a mix of technical and business metadata to underpin the diverse functions required of the system. This paper outlines how metadata operates at different levels of the Web publishing system, from the perspectives of system development and programming, project design and management, and metadata standards development.

Defining metadata

There are many attempts in the literature at defining 'metadata'. A common definition is 'data about data'. Many librarians think of metadata as cataloguing of Web-based content. The National Library of Australia's Meta Matters Web site refers to it as 'A summary of information about the form and content of a resource' (National Library of Australia 1999). A division of the American Library Association describes metadata as aiding 'in the identification, discovery, assessment and management of information-bearing entities' (ALCTS 1999).

Essentially metadata consists of categorising and labelling to achieve desired outcomes. International metadata standards attempt to offer sets of categories and labels that will apply to the content of documents in a wide range of organisations and contexts. Document mark-up languages such as HTML and XML are also a form of metadata, and SIM makes use of XML in relation to the structure, look and feel of the pages across the site as will be explained later.

Marco, a US consultant in the area of data warehousing, identifies two types of metadata – technical and business (Marco 2000, p. 49). In the context of the RMIT project this is a helpful distinction. The technical metadata relates to the ways in which the SIM databases are structured and managed; the business metadata relates to the diverse functions required of the new Web site from the viewpoint of the users – staff and students.

Objectives and functions of the new Web site

RMIT is a multi-campus university with approx. 6000 staff and over 50,000 postgraduate, undergraduate and TAFE students. The aim of the Web refurbishment project was to achieve a consistent approach to publishing on the RMIT University Web platform, ensuring that all content is

- derived from an authoritative source,
- authorised and validated,
- presented in a consistent way across the RMIT Web presence, and
- separated from its form of presentation to facilitate updating.

Content would be created and managed by its direct business owners, academic and general staff, using common processes deployed on a decentralised basis, but supported by central application and hardware infrastructure.

The SIM portal would be the gateway into the university's Web presence, and the central navigation point to all information and services delivered by RMIT via the Web. Its main functions can be summarised as follows:

1. To provide information and services for staff, students and other audiences, contact information for staff and departments, information about programs and courses (closely aligning with the functionality of the academic management system), and general information about departments and research activities.

2. To make accessible to users all multimedia objects created by RMIT staff as part of their Web pages and a selection of media assets created for course delivery – i.e. establishing a media repository as an integrated part of the Web system, available to content creators from both the SIM Web platform and other RMIT Web systems.

3. To house prototype 'learning objects'/'online courseware', that are deemed to be repurposable or reusable by others, and facilitate ease of access to repurpose, – i.e. to integrate the Web system with the learning management system.

4. To hold comprehensive information about university policy, procedures and governance, as well as providing access to an increasing amount of intranet-only materials and resources.

5. To provide access to other key RMIT enterprise systems such as the Library, the academic management system, the online learning management system, human resources systems, and staff and student email systems. This will be achieved by applying the same visual design, branding elements and navigational conventions to all the portals, and by providing a single sign-on authentication system so that seamless access can be granted to all integrated applications.

How SIM uses metadata in achieving required functionality

The underlying SIM product used to develop the RMIT Web system was itself designed with metadata indexing and retrieval as well as full text retrieval in mind. In part, this comes from SIM's library-related heritage, as it uses the Z39.50 information retrieval protocol as its fundamental interface. Z39.50 itself forces a system to make a distinction between the way information is stored, and the manner in which it is accessed. It therefore allows users to interact with a potentially complex collection of documents and metadata via a simplified schema – effectively a metadata based view of the data which can be presented consistently, even when the underlying data is inconsistent.

Business metadata

Metadata facilitates the core business of the Web site by supporting (1) information retrieval and document management (2) content authorisation and validation and (3) consistent presentation of content.

1. Metadata for information retrieval and document management

A combination of appropriate business practices and the application to each page of a range of labels are crucial elements in achieving the desired functions of the project. Metadata plays a role with respect to both information retrieval and document management, including

1. obtaining precise results when undertaking basic or advanced searches across all the pages and media assets on the system
2. displaying selected elements such as titles and abstracts in search results
3. placing pages in logical positions within the hierarchy of the Web site
4. enabling pages and media assets to be filtered into subsets by document type, access/visibility level, audience type, interest area and other characteristics of the page
5. controlling review dates to keep pages and media assets up to date
6. recording and displaying vital ownership, copyright and editorial information relating to each page.

While every word on an RMIT Web page is searchable via the SIM search engine, the metadata makes it possible to facilitate precise retrieval across many thousands of documents and media assets by highlighting content-rich words and key features of pages. The search engine assigns special weightings to the words and phrases that have been entered in the abstract, keywords and title metadata elements, so that relevant pages are more likely to obtain higher rankings in search result outputs.

In preparing content for publication it is useful to search across the entire site for related or similar content to discover any possible duplicated, inaccurate or out of date material. Precise searching is thus not only for the end-user – but is also a vital information management tool for those who are publishing on the Web.

2. Metadata for content authorisation and validation

SIM requires all Web page publishers to belong to one or more editorial groups, each group member being assigned one or more roles – as a document creator, editor or approver. A document cannot be made visible to end users on the Web unless it has been approved by someone other than its creator. The document is labelled with its status at each stage of the publishing process – whether draft, proposed, pending, approved, suspended or deleted. ‘Meta-metadata’ records the name of the editor and approver and the date of the modification. This information is displayed in the footer of each page, and utilised in the editing interface to facilitate management of content throughout the content creation process. Each of these features or units of information is recorded as metadata, which in turn supports the desired outcome of authoritative and validated content.

3. Consistent presentation of content

The consistent presentation of content across the SIM portal is facilitated by means of :

- common design features,
- common approaches to branding,
- assigning a document type to each page, and
- pre-defined navigation pathways.

These are discussed briefly in turn.

Common design features

Named styles from a style template are applied in MS Word 2000 to components of each page by document creators. The document is then converted into XML by SIM for storage, and is then converted back into HTML or MS Word on demand when required for presentation or re-editing. This makes it possible to make global and consistent formatting changes, eg colour, font, etc to parts of pages across the entire site. SIM can also extract appropriately styled sections of text and automatically record them as metadata and/or display them in certain ways. Once the logical meaning of some text is identified by SIM, such as the text that makes up a title of a News document, then the system has complete flexibility in how that title is presented when the document is served up as HTML, as the content and its presentation have been separated.

Common approaches to branding

There are several factors which affect how a page is 'branded' when presented over the Web. Branding means selection of background colours, selection of page decoration graphics, overall HTML structure including number of columns on the page and location of elements such as the table of contents, and finally colour and font of text as specified by a cascading stylesheet. All of these branding factors can be dynamically adjusted for any page in the system based on the metadata associated with the document, including:

- its document type (such as News, Staff Profile etc)
- the editorial group that owns that document
- the identity of the user viewing that document
- the browser that the user is running, and even
- the navigation path within the site that the user followed in order to view that page.

Assigning a document type to each page

Information audits of Faculty and Department information categories and focus group discussions led to the devising of a limited set of categories relating to the main types of information conveyed on the Web. Document types include Generic, News, Staff profiles, and Academic programs and Media assets. Some document types have sub-types, for example News documents may consist of Media releases, Events and Awards. This document type information interacts with other metadata such as audience type, visibility and owner groups to determine how and where pages are presented in the site's navigational structure and search results.

Pre-defined navigation pathways

These are based on organisational structure and document types, so that there is a familiar feel to where documents are located in the table of contents that is displayed next to each page. These operate on a default system, but additions/changes to Path and Sub-Path in the metadata give document editors flexibility as to how the table of contents (and therefore the implied relationship between different items of content) should display.

Hypertext linkages between documents are achieved by deep-linking to the unique ID's of each document. A change to the site's structure or navigation model will not break such links within the site, or links from other sites referencing a document via its unique ID.

In addition, 'short' URLs can be set up for any page so that users can remember and revisit a page easily. For example, the Faculty of the Constructed Environment might normally have a URL of Faculties/Constructed Environment, but they also have a URL of '/CE'.

Technical metadata

The SIM system is data driven, starting particularly from the types of documents that were identified by the information architecture. In configuration files MDS staff then describe how the site is to be structured, by specifying the locations in which each document of each document type should appear.

Configuration files also specify the sorts of metadata that must be entered for each document type, and the type of validation to be done, be it a database lookup, a drop-down box, or a text input field. By changing these configuration files, new document types with new metadata validation rules can easily be added to a Web site. Because the structure of the site is dynamically enforced by the application of certain rules to the metadata stored with each document, it is in fact possible to completely change all of the URL pathways within the site by changing a single configuration file.

Automated navigation to client-focused content can be achieved by embedding searches of document metadata (for example, searches by audience types and interest areas) into either the built-in table of contents, or into the contents of a page. This is a particularly valuable feature of the system, which enables the business aims of presenting information from across the university to its different audiences based on a 'service' model, whilst still managing content in the system using production processes based on the university's organisational structure.

The dynamic, metadata-driven delivery of content from the SIM repository also facilitates advanced functionality such as dynamic highlighting of matched content in search results, and the specification of the location of search matches within the table of contents, as well as the hiding of documents, and selected content within documents, from non-authorised users in both the on-screen navigation and in search results.

Compliance with international standards

It was decided that for the sake of possible future Internet interoperability of publicly available pages of the Web site, the elements (or field names) in the RMIT metadata standard would be mapped to the Dublin Core elements (Dublin Core Metadata Initiative Web site, http://purl.oclc.org.dc). In addition, RMIT-specific metadata elements were added to cater for the management of the Web pages, the particular RMIT University context, and the characteristics of document types on the system. An example of an uploaded metadata screen is attached at the end of this paper.

RMIT's multimedia objects and learning objects stored in a separate repository would comply to some extent with the IMS metadata standard for learning objects, *IEEE learning Object Metadata base document working draft 6.1* (IMS Global Learning Consortium, Inc. Web site, http://www.imsproject.org). The wording of some elements was modified for the given context, and again some RMIT-specific metadata elements were added. In a recent communique, Neil McLean, Director of IMS Australia has reported on a program of action being undertaken by DCMI, IMS and IEEE, LTSC in relation to 'matters of mutual interest' (McLean 2001).

RMIT's metadata standard, as 'an application profile' is in good company with other projects in Australia and overseas such as EdNA, GEM in the U.S., the UK's Resource Discovery Network and the European Renardus project which have used

> 'general standards such as Dublin Core … alongside domain- or sector-specific standards' …. 'and with new elements … for local needs not covered by any of the existing standards … Application profiles are schemas that combine elements from multiple standards, perhaps with application-specific constraints such as the use of specific controlled vocabulary' (SCHEMAS 2000).

Research is being undertaken into the concept of registries 'where application profiles can be published and found by others'….'in various places, such as the Dublin Core Metadata Initiative, the Indecs project, XML.org, and indeed, the SCHEMAS project itself' (SCHEMAS 2000).

The present apparent lack of interest by Internet search engines to use available internationally-compliant metadata (see, for example Henshaw 2001), and questions relating to the future role of international standards are challenging issues that cannot be addressed here owing to time constraints.

The Subject metadata element and controlled vocabulary

The Dublin Core and IMS standards both advocate as best or common practice, the use of a controlled vocabulary in the subject field. Some hard questions on the issue were asked such as:

- Is there a ready-made thesaurus that would suit the particular content of the Web site?
- How long would it take to develop an in-house thesaurus, and how will it be maintained?
- Who will assign the vocabulary terms, and how will cross-references be integrated into the indexing and searching stages?
- Are users likely to make use of the vocabulary?
- What are the implications for the search interface?

With cost-benefit implications in mind, it was decided to give away the notion of using a controlled vocabulary. Instead, subject keywords – words and phrases assigned by the page uploaders/creators/editors would have to suffice. Title words and words in an abstract would provide further rich content descriptions for improving search results. The issue of controlled vocabulary has been discussed in more detail in a recent conference paper (Schauder 2001).

At the time of writing this paper many pages have been published to the prototype version of the new Web site. Performance testing of the search function suggests that the weightings given to keywords, titles, abstracts and full text words are resulting in satisfactory subject search results. Even where the metadata has been of a very poor quality, the system has appeared to compensate via the full text search.

It is interesting that the page uploaders often include valuable terms in their abstracts which they omit from the keywords field. As a result it was decided to increase the weighting given to the abstracts, and further testing will reveal the impact of this on search performance.

From a theoretical perspective it is not surprising that titles, abstracts and keywords, in conjunction with full text words, achieve satisfactory results. The human intellectual process of summarising the content achieves some of the benefits that would have been obtained via a controlled vocabulary, and, if the metadata for some pages is poor or inadequate, the full text words are there to support the search and ensure adequate retrieval rates.

While there is no attempt to index every page with a single thesaurus, there is, however, some application of small controlled vocabularies. Each page is initially categorised by a controlled list of document types; there are drop-down menus in certain common fields, eg ‘audience types’ and ‘interest’ areas; and particular document types have controlled subject indexing elements, eg ‘Services’ documents have a pull-down listing of services categories which are searchable from the Services index page. These all have the potential of being included as pull-down listings at the advanced search interface or at the index page of the relevant section of the Web site.

The challenges of implementing metadata

Viewed from a management perspective, this project may be seen as a significant change management initiative. The system’s deployment provides for an extensively distributed network of authors/content creators. But there is more. Authors/content creators and publishing groups are required to also perform metadata cataloguing, design, publishing, editing, and usability testing roles! To date, staff have typically provided a raw electronic copy of content to the Webmaster who ‘does the rest’. The new system does not rely heavily on technical Webmasters; rather it provides new roles for content managers who can assist authors with their extended role. Let us now examine metadata as a part of the publishing process.

The staff who upload pages to the Web site are general and academic personnel. They select an appropriate ‘starter document’ or metadata template, create their document in Word and then save it to the Web. As mentioned above, it was decided that for practical reasons, the uploaders and creators of the documents would enter the metadata themselves, rather than relying on specialist indexing staff. This takes place either in Word, or during the uploading process. When changes are made to the content of a page, the metadata is changed accordingly. An example of

an uploaded metadata screen is attached at the end of this paper. The process of filling in the metadata is generously supported with system-defaults, and the system refuses to upload pages if certain key metadata fields have not been completed. To gain acceptance the time taken must be kept to a minimum – certainly not more than a few minutes per page if possible. To date uploaders are choosing to fill in only the mandatory fields.

Challenging issues revolve around acceptance and training issues. They include:

- how to monitor the quality of the metadata – at least for high value documents
- how much time to spend on metadata training, and how to explain what it is and why it is important without frightening staff
- how to strike an appropriate balance with respect to the length and complexity provided in online metadata guidelines, manuals and help screens
- how important it is to control forms of personal and corporate names.

It will be essential to monitor the performance of the metadata as it relates to the different functions it supports, and to make changes to the metadata elements and processes as necessary.

The ingredients for a successful metadata implementation include:

- support of the project and the metadata component of it by senior management at a strategic level, and by middle management at a support and work flow level
- a sound information architecture which is appealing and logical to uploading personnel
- a good working relationship between system developers and project management, and between uploaders and the project team
- regular evaluations of metadata functionality and feedback from users
- sufficient flexibility to make changes and improvements to the architecture, templates and drop-down menus over time
- responsiveness by the system to relevant developments in Web technologies and concepts as they emerge.

Since there are few well-trodden paths, much of this development is ‘action learning’. As the Web site grows in size, and extends into new and ever more complex roles, the importance of synergy between metadata, the SIM database platform and the RMIT University business needs will only increase.

References

ALCTS 1999, Association for Library Collections and Technical Services, Cataloging and Classification Section, Committee on Cataloging: Description and Access, Task Force on Metadata *Summary report June 1999, Charge#3*, http://www.ala.org/alcts/organization/ccs/ccda/tf-meta3.html. Accessed 1 October 2001.

Dublin Core Metadata Initiative, Homepage, http://purl.oclc.org.dc. Accessed 1 October 2001.

Henshaw, Robin 2001, "What next for Internet journals? Implications of the trend towards paid placement in search engines", *First Monday*, 6 (9) Sept., http://www.firstmonday.dk/issues6_9/henshaw/index.html. Accessed 1 October 2001.

IMS Global Learning Consortium, Inc. Homepage, http://www.imsproject.org. Accessed 1 October 2001.

Marco, David 2000, *Building and managing the meta data repository*, Wiley, New York.

McLean, Neil 2001, *Harmonisation of metadata for Education and Training Communities*: Ottawa Communique 30 August 2001.

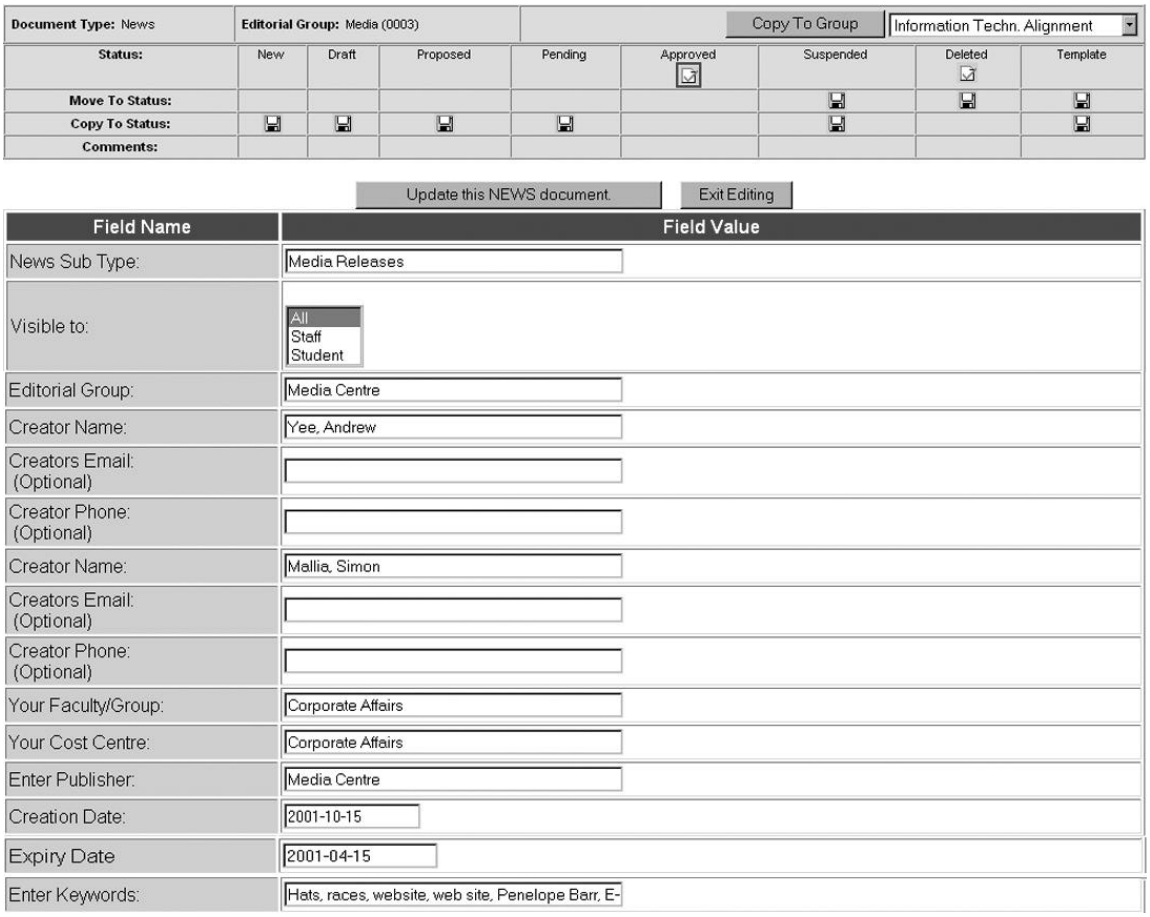
National Library of Australia 1999, *Meta Matters*, Introduction, http://www.nla.gov.au/meta/intro.html, Accessed 1 October 2001.

Schauder, Cherryl 2001, 'Metadata, subject retrieval and controlled vocabularies: practical issues' in *Seachange: cataloguing in a dot.com world, 14th National [ALIA] Cataloguing Conference preprints*, Organizing Committee of the 14th National Cataloguing Conference, Geelong, Vic. In press.

SCHEMAS:Forum for Metadata Schema Implementers 2000, *Metadata watch report #3*, http://www.schemas-forum.org/metadata-watch/third. Accessed 1 October 2001.

Attachment

Example (on this page and the next) of the metadata screen uploaded with a News (Media Release) Web page in the RMIT University Web Refurbishment Prototype Project. The title of the page is *RMIT researcher creates website for women to buy the perfect hat for the races.*



Document Type: News	Editorial Group: Media (0003)				Copy To Group	Information Techn. Alignment		
Status:	New	Draft	Proposed	Pending	Approved	Suspended	Deleted	Template
Move To Status:								
Copy To Status:								
Comments:								

Update this NEWS document. Exit Editing

Field Name	Field Value
News Sub Type:	Media Releases
Visible to:	All Staff Student
Editorial Group:	Media Centre
Creator Name:	Yee, Andrew
Creators Email: (Optional)	
Creator Phone: (Optional)	
Creator Name:	Mallia, Simon
Creators Email: (Optional)	
Creator Phone: (Optional)	
Your Faculty/Group:	Corporate Affairs
Your Cost Centre:	Corporate Affairs
Enter Publisher:	Media Centre
Creation Date:	2001-10-15
Expiry Date	2001-04-15
Enter Keywords:	Hats, races, website, web site, Penelope Barr, E-

Continues next page.

Example continued…

Select Audience:	All Current Students Current Students/Undergraduate Current Students/International Current Students/Postgraduate
Choose Interest Area:	Aerospace and Aviation Apprenticeships and Traineeships Architecture Building and Planning Art and Design Business
Rights: (Copyright Holder)	RMIT University
Image Alt Text:	Media Release
Thumbnail Image:	Media Release Defer ☐ Media Release Icon (ug81dumw6ddq.gif)
Image Alignment None Border Width 0 Vertical Space 0 Horizontal Space 0	
Choose Layout:	3 Column
Enter Alternate URL: (Optional)	Owner group prefix [/News/]
Coverage, Location: (Optional)	
Coverage, Year: (Optional)	
Contributor, Name: (Optional)	First Last Organisation
Contributor, Email: (Optional)	
Contributor, Phone: (Optional)	
News Source:	First Last Organisation
Related Link: (Optional)	Values for this image link field need to be entered from Word
Notes: (Optional)	
Title:	RMIT researcher creates website for women to b
Enter Abstract:	Finding the perfect hat for the races will be as easy as clicking on the web - thanks to HatExchange.com.au, a new web
Link Label	HatExchange.com.au
URL Target:	http://HatExchange.com.au