METAINFORMATION INCORPORATION IN LIBRARY DIGITISATION PROJECTS

Michael Middleton

QUT School of Information Systems, Brisbane, Australia. m.middleton@qut.edu.au

This paper was accepted in Poster form and subsequently published as:

Middleton, M. (1999). Metainformation incorporation in library digitisation projects. In T. Aparac *et al* (Eds.), ***Digital Libraries: interdisciplinary concepts, challenges and opportunities; proceedings of the Third International Conference on the Conceptions of Library and Information Science, Dubrovnik, Croatia, May 23-26, 1999*** (pp. 334-338). Zagreb: Zavod za informacijske studije Odsjeka za informacijske znanosti, Filozofski fakultet; Lovkepp: Naklada Benja. [ISBN 953-6003-37-6].

The poster included illustrations and screenshots that are not present with this text.

Abstract: Approaches to dealing with metainformation in library digitisation projects are considered with respect to characteristics of defined elements. Some applications within projects being undertaken in Australia are described with reference to existing standards for resource description and vocabulary control, and the role of these within online cataloguing systems, and network interfaces.

Keywords: Digital libraries, metainformation, project management, online catalogues, information retrieval, Australia

Introduction

Much of the present access framework for digital libraries is being established by concurrent creation of metainformation for description of new objects or texts. The metainformation such as record description, or content indexing, is stored in structured databases for searching in an optimal manner. Retrieved surrogate records may then be used to link to the complete objects or texts if searchers desire.

Evolution of information retrieval systems by libraries, has increasingly included efforts to integrate finding aids for collections of both print and digital media. For example an Internet Web interface for a specific subject area may include a combination of links to catalogue records, to specific items in databases, to electronic full texts and to a digital bibliography of print material. An online system is expected to point seamlessly to both electronic and print media for user subject specialisations. Such an approach may be complemented by automatic current awareness or push technology profiles that monitor databases and return selected information to desktops.

Libraries are also undertaking retrospective conversion projects to transform print and image materials into electronic form. Many of these materials have previously been described within print or digital catalogues or finding aids, prior to commencement of the digitisation projects. Project management decisions include consideration of whether to convert this existing metainformation along with the source materials, and if so, whether to encapsulate the resulting digital form together with source information, or deal with it in a separate database.

However, this metainformation is often at different levels of granularity, or standardisation, or completeness, from that required for the digital interface. Sometimes it is not present at all. Therefore creation of metainformation for use with the source materials, rather than conversion, may be a more viable or cost-effective option.

In what follows, some characteristics of metainformation elements are reviewed in the context of international standardisation efforts. Then several case studies of projects in Australia are briefly considered.

Characteristics of metainformation

Metainformation (or metadata) may be characterised as information that describes: (a) the *agent* carrying the information, for example document description, or (b) the intellectual *content*, for example subject indexing. There are many resource description formats that attempt to standardise use of individual elements in particular contexts. A detailed compilation of these has been made by Dempsey *et al* (1997). The following examples of defined elements are given with respect to existing resource description standards.

Defined elements

Agent.

Document description

This has been most influentially developed for libraries per medium of MARC (MAchine Readable Cataloguing) format, for which there are many national instances within the ISO 2709 format for bibliographic information interchange.

Responsibility

This is information pertaining to the intellectual creation of the material. It has been adopted for example in the TEI (Text Encoding Initiative) using library cataloguing rules as a basis. It enables specification of elements such as *author, sponsor*, *funder*, *principal researcher*, and other contributions, although the form of the creator is not categorised into personal, meeting or corporate, as in library cataloguing.

Administrative

An example is embodied in the EAD (Encoding Archival Description) format. Here it is applied to elements like *access*, or *appraisal* for retention scheduling information.

Provenance
An example is *document source* that is provided for in the CIMI (Computer Interchange of Museum Information) format to track origin and ownership.

Configuration
This provides descriptions to assist with processing of data such as file format or record size, that are typically carried in header or label information of digital records. Alternatively it may be an element such as *type* used in SOIF (Summary Object Interchange Format) for specifying file types such as 'binary'.

Connections
An application is the *relation* element specified within Dublin Core that is intended to provide a means to express relationships between a discrete resource and other resources that may also be considered as discrete. These may be for example a periodical article and the periodical itself, an item in a collection and the collection itself, a file within a database and the database itself. Alternatively the relationship can be to resources that control the content of information in a particular field such as a thesaurus of descriptors, or an authority file of organisation names.

Conditions of use
This refers to elements that describe or link to availability statements, such the *rights management* element of Dublin Core that makes provision for links to a copyright notice, or a service for provision of information about terms of access, or rights management of the resource.

Content.

Topic
This may be expressed as keywords or descriptors from controlled vocabularies that are used to describe subject matter. Alternatively, a facet from a scheme embodying a standard notation may be used.

Coverage
This relates to extent of the intellectual content encompassed spatially or temporally.

Role
As used for example in museum description, this may be the context in which the subject matter may be used, for example game playing.

Most approaches to description embody a combination of these attributes, but not all of them. For example Dublin Core eschews administrative metainformation in order to keep descriptive elements intrinsic for an object.

Storage

The metainformation that is used for describing objects may be maintained separately from the objects themselves, most usefully in a structured database with links to the objects, or stored internally with the source information. External approaches include the specification

of specific MARC tags for reference to images from catalogue databases. The internal approach can vary as follows:

File naming
The naming of directories and files can be used as a mnemonic to provide a limited form of understanding of content of the source information within.

Markup
The tags defined in implementations of markup languages such as SGML are themselves metainformation. Together with the contents of fields specified for metainformation such as keywords, they provide a more powerful approach that may be utilised to structure databases.

Consolidated
An approach that combines descriptive markup information with incorporation into file structure, increases the utility of the metainformation. For example, the Blake Archive metainformation pertaining to configuration (Kirschenbaum, 1998), is combined with bibliographic description and provenance and contact information. This extensive identification is then inserted into that part of the JPEG image file reserved for textual metadata. It may subsequently be extracted with JPEGView.

Australian Applications

Management of digitisation projects must take into account many factors specific to metainformation. These include: (a) What, if any, resource description standards are to be used? (b) Is there an internal controlled vocabulary in use that requires conversion and maintenance? (c) Is there an existing external controlled vocabulary (e.g. AAT, NASA, APAIS) in use that can be maintained for the digital project? (d) Is a thesaurus to be created for the project? (e) Is there software support for authority file maintenance or thesaurus construction? (f) Is the metainformation to be encapsulated within the digital objects, or maintained as separate files or finding aids? (g) Is there to be a network interface? (h) What proportion of total cost may be devoted to conversion or creation of metainformation? (i) Can metainformation be created as part of existing processes (e.g. cataloguing department)?

Context

Digitisation projects in Australia are wide ranging. They illustrate convergence tenets by taking place in museum, archive, educational and library environments. At the national level they involve large scale projects for consolidating access to multiple collections of cultural institutions such as the Kinetica project of the National Library, and the Australian Museums On Line project. These are complemented by infrastructure research such as that supported by National Archives of Australia at Monash University, that aims to provide a framework for standardising sets of record-keeping metainformation.

At the level of specific sites, diverse projects include a Norwood public library project to capture local studies material including oral history sound recordings, and 3-D objects; the Powerhouse Museum's project to capture photographic images, and link them with the its collections information system; and digitisation of the Australian War Memorial collection of war photographs. Some Australian projects have been described by Iannella (1996), and a database of projects is maintained (Digitisation Forum Online, 1998).

Cases

IMAGES1. This is one of several projects underway at the National Library of Australia. It contains over 15,000 of the 40,000 paintings and 550,000 photographs of images relating to Australia, including all NLA's oil paintings, and selections from portraits, drawings, rare prints, objects and photographs such as those in Cazneaux Collection. The images have been scanned from transparencies or photographs and are displayed in medium and thumbnail resolutions. (National Library of Australia, 1998).

It is available through two interfaces. Many photographs appear in the consolidated OPAC with collection-level description. A URL embedded in MARC provides a link to the collective entry and thence to individual photographs. Alternatively the IMAGES1 interface provides direct links to discrete images. It enhances access by provision of searching by creator; other names associated with a work or collection, image number, and format (e.g. watercolour). MARC 950 tags are utilised to enhance format and subject access.

PICMAN. This is one of several projects at the State Library of New South Wales (1998). It consists of cataloguing of pictures and manuscripts collections, and includes images from six photographic collections that were earlier digitised for CDROM. Access is available via the Internet using AWAIRS software, though only non-copyright items may be displayed.

Search fields include creator, subject, publisher, title, contents and notes. However, these are complemented by a *date* range searching facility, a *format* searching pick list (posters, photographs, etc), a *persons* pick list (girls, adult females…), and a limited *area* pick list oriented towards the State of New South Wales' regions.

DIGILIB. This is a collaboration between architecture faculty and library (University of Queensland, 1998). Images converted from colour slides of Queensland historic buildings have been indexed for town, dwelling type, feature, structure, materials and context. A MARC-based form was developed to guide description. A controlled vocabulary is used so that a pick list may be presented for retrieval (e.g. structure='masonry', context='hills').

These three examples are representative of how digital metainformation is being put to effective use to access material in some of the many projects now underway.

References

Dempsey, L. *et al* (1997). A review of metadata: A survey of current resource description formats. Version 1.0. <http://www.ukoln.ac.uk/metadata/DESIRE/overview/>.

Digitisation Forum Online (1998). *Home*. <http://www.digitisation.net.au/index.html>.

Iannella, R. (1996). Australian digital library initiatives. *D-Lib Magazine*. http://www.dstc.edu.au/RDU/reports/DLIB-OZ/.

Kirschenbaum M. (1998). Documenting digital images - textual meta-data at the Blake archive. *Electronic Library*, **16(4)**, 239-241.

National Library of Australia. (1998). *IMAGES1* < http://www.nla.gov.au/images1/>.

State Library of New South Wales (1998). *PICMAN*. <http://www.slnsw.gov.au/picman/picman.htm>.

University of Queensland (1998). *Digilib*. <http://www.architect.uq.edu.au/digilibHOME.html>.