An introduction to metadata requirements for an e-print repository

Circulation: PUBLIC
Gareth Knight
Arts & Humanities Data Service

Summary
This paper describes three conceptual forms of metadata (descriptive, technical and administrative) and advocates the establishment of quality checks to ensure metadata consistency.

Defining metadata

A common definition is to describe metadata as 'data about data', or any data associated with a resource that describes that particular resource. Library catalogues are a familiar example of metadata. An entry in a library catalogue is made up of metadata elements defined by a formal standard, such as the name of the author and the date of publication. Library catalogue metadata is primarily designed to help readers find books in a library, but there are many other potential uses for metadata in different settings. For e-print repositories, metadata can be divided into three conceptual types that have different purposes (although there are some crossovers):

- **Descriptive metadata:** used for the indexing, discovery, and identification of a digital resource.
- **Technical Metadata:** stores technical details of the stored resource necessary to identify and preserve the content. Technical metadata might include information such as the structural divisions of a resource (i.e., chapters in a document).
- **Administrative metadata:** represents the management life-cycle for the object, which may include information the user may need to access and display the resource, including its preservation, processing history and rights management information.

There is no 'one size fits all' solution for metadata. Instead, a wide variety of metadata formats have been created for different purposes, and with varying degrees of depth and richness. Repository software (ARNO, CDSware, DSpace, Eprints, Fedora, i-Tor, and MyCoRe) ensures interoperability through support for the OAI (Open Archives Initiative) Protocol for Metadata Harvesting (OAI-PMH).

The OAI-PMH distinguishes between two different forms of providers - 'Data Providers' (archives which expose metadata to harvesters) and Service Providers (which harvest the metadata and create services with it). To provide basic communication for archives with different underlying structures, OAI-PMH specifies unqualified Dublin Core (a simple 15 element metadata schema) as a basic requirement (Weibel, 1998). This can be used in many different ways (federated searching, bibliographic extraction, citation analysis, etc.) to harvest metadata, search the results and present them to the user.

In comparison to Internet search engines, which index keywords, the use of OAI & Dublin Core has significant benefits. By searching human-created metadata rather

than machine-extracted keywords, a user will be able to search relatively small collections in significant detail. For example, they may restrict their search to works by a specific author or with a particular publication date. It is simple to find resources that contain the phrase "Tony Blair" using a search engine that indexes keywords, but it is not as easy to find only articles written *by* Tony Blair.

Descriptive metadata

The primary goal of e-print repositories is to provide efficient, global access to scholarly output. This aim can be supported by the creation of descriptive metadata, which is often provided by the depositor when they deposit an e-print in a repository. Descriptive metadata collected for an e-print usually consists of the more common elements found in the unqualified Dublin Core metadata schema - title, creator, subject, description, publisher, and type (DCMI, 2003). Dublin Core is widely used, and offers a simple metadata schema suitable for resource discovery.
However, although unqualified Dublin Core is acceptable for many tasks and will allow basic interoperability between repositories, there are a number of reasons why it may be useful to have additional supplementary metadata:

- The unqualified Dublin Core element set may not include all of the elements that you require. The Theses Alive! project (http://www.thesesalive.ac.uk/), for example, utilizes the more specialised Electronic Theses and Dissertations (ETD) schema which includes elements that are not present in unqualified Dublin Core (display title, department, document type) or have less clear meanings (DC's Date vs. ETD Defence Date). Other community projects provide additional subject-specific elements to cover chemical classification (Chemistry Preprint Server); refereed (CogPrints); jurisdiction, audience, postcode and mandate (Australian Government Locator Service)
- In some circumstances the Dublin Core elements may be too vague. Although it is a widely accepted standard, there are many differences in the type of data that is entered into the elements. For example, a media organization (such as ITN) and an article author will have a different understanding of the creator element. As an unqualified metadata schema, the Dublin Core elements may not be sufficiently precise for your records. In these circumstances a repository that supports qualified Dublin Core should be used.
- The community in which you operate may specify a different, more complex, metadata format. For example, IMS/IEEE LOM for e-learning (Carpenter, 2003). Specialized metadata formats provide subject or material specific elements (e.g. MARC records for books, ISAD (G) records for archival material, or TEI headers for electronic texts). Various projects are currently investigating the possibility of refining the Dublin Core set - the DCMI (Dublin Core Metadata Initiative), for example, are currently working upon a series of specialised elements and vocabulary suitable for different communities (W3C, 2003).

Many more complex metadata schemas can be mapped to Dublin Core, that is, at the loss of some information, an unqualified Dublin Core record can be extracted from the original record. Keep this in mind, as it may be worth storing sophisticated metadata internally, and then converting it to Dublin Core so that other repositories and services such as portals can access information about the e-prints in the repository using the default OAI-PMH.

Quality control

By design, most e-print repositories rely on their depositors to provide the metadata for e-prints. Quality control is thus both very important and potentially difficult to enforce. Although a Dublin Core metadata record is short and simple, many authors find it difficult to complete fields in a consistent manner that aids interoperability and retrieval. (James et-al, 2003). Common mistakes include incorrect references, inconsistent use of keywords, spelling, date formats, extraneous punctuation, acronym use and different subject descriptors (James et-al, 2003; Pinfield**,** Gardner & MacColl, 2002**;** Boyce, 2000, p. 414; Colorado Digitization Program, 1999). There is also the possibility of "overloading the core" by placing a large amount of unstructured information into some elements (Miller, 1997).

While it can be time consuming and costly to significantly improve the metadata provided by depositors, it is wise to perform regular checks to ensure quality and consistency. Contemporary e-print software packages, such as eprints.org (http://www.eprints.org) include an approval process that requires the system administrator to review submitted data before it is made available, providing an opportunity to check and improve metadata if necessary. Controlled vocabulary and other restrictions can also be placed upon the type and amount of information the depositor can enter into a element, and depositors can be notified if essential elements have been left empty (Colorado Digitization Program, 1999). Overall, it is usually cheaper to correct metadata problems as they appear, rather than conducting retrospective work later (Hillmann, 2003)

Digital Preservation

Digital preservation is necessary to ensure digital information that has continued value remains accessible and usable in the long-term (Hedstrom, 1999, p. 189). Preserving digital information is difficult because of the relatively short media lifetimes of media, software and hardware (Chen, 2001, p. 24). Preserving information, that is metadata, about the software, file formats, and hardware needed to access a digital object, such as an e-print, is widely regarded as an important first step to ensuring its long-term accessibility. The wide range of functions that preservation metadata is meant to fulfill means that most of the currently published schemas are complex, yet share many common elements. Preservation schemas (for example, Cedars, OCLC, NLNZ) typically contain structural metadata that describes how a digital object is put together and encoded, administrative data that records the intellectual rights and provenance of the object, as well as providing an audit trail of recorded changes to the resource (e.g. migration from one file format to another), and fixity information (such as checksums) that can be used to authenticate the data.

In contrast to support for resource discovery metadata, e-print repository software packages (e.g. DSpace's "bitstream format registry") capture a minimal amount of technical metadata by default (file format, MD5 checksum, date of submission). This provides basic information useful for distinguishing between different file formats, but is not necessarily sufficient to validate the data format (is the HTML well-formed?), distinguish between different versions (i.e. the difference between Adobe Acrobat 3-5), or recognize specific requirements (a specific Internet browser, external style sheets) that contribute to the decoding and display of e-prints.

Although there is little incentive to create technical metadata in the short-term, repositories should begin to consider how they can store this information to ensure long-term access. Some repositories have compensated for software limitations by establishing their own internal file naming conventions that carry specific meaning (e.g., a unique identifier, depositor's code, etc.) (James et-al, 2003).

Repositories should also consider recording administrative metadata. Popular repository software such as DSpace & eprints.org currently record checksum information necessary to authenticate the object. This can be enhanced through the storage of process metadata to record when specific actions were performed on an e-print, such as withdrawal and migration to another format.
The Dublin Core 'rights' element (dc:rights) has been identified as a universal method of storing copyright information (although it cannot be used to restrict content to certain individuals or groups). Significant effort has also been made by the Library of Congress, the National Library of New Zealand and other third-parties to develop bespoke tools able to handle the complexities of current file formats and extract specific elements. It is possible these will become useful in supplementing existing work practices for the creation of e-print metadata in sufficient detail.

Conclusion

E-Print Repositories typically rely on the OAI-PMH and the Dublin Core schema to define their metadata requirements and make the metadata interoperable. However, it should not be automatically assumed that Dublin Core will meet all of a repository's requirements. To create an e-print repository, staff must ensure these solutions fulfill their own specialized needs, in terms of describing the resource, and consider techniques for creating metadata records that meet pre-determined quality threshold. If the e-print repository intends to remain in operation for several years they should also begin to consider the long-term implications of storage by investigating methods of storing appropriate preservation metadata.

References

Australian Government Locator Service (AGLS)
http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html

Boyce, P. (2000). For better or worse: preprint servers are here to stay. *College and Research Libraries News, 61(5*), pp. 404-407

Carr, L. 2004 EPrints Handbook
http://software.eprints.org/handbook/

Carpenter, L. 2003. Implementing OAI-PMH
http://www.oaforum.org/tutorial/english/intro.htm

CEDARS (n.d.). Metadata for digital preservation: the Cedars Project outline specification.
Retrieved on May 3, 2003, from
http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html

Chemistry Preprint Server, 2004
http://www.chemweb.com/preprint

Chen, S.S. (2001). The paradox of digital preservation. *Computer*, 34 (3), March, 24-28.

Day, M. (2003). Integrating Metadata Schema Registries with Digital Preservation Systems to Support Interoperability: a Proposal

DCMI (2003). Dublin Core Metadata Element Set, Version 1.1: Reference Description
http://www.dublincore.org/documents/dces

Harnad, S (2001). For Whom the Gate Tolls? Retrieved on January 29, 2003 from:
http://www.ecs.soton.ac.uk/~harnad/Tp/resolution.htm#1.Preservation

Hillmann, D. 2003,NSDL Metadata Primer
http://metamanagement.comm.nsdlib.org/outline.htm

James, et-al, 2003
http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf

Lupovici, C. & Masanès, J. (2000). *NEDLIB Metadata for Long-term Preservatio*n. NedLib
Report Series, no. 2. Retrieved on May 3, 2003, from
http://www.kb.nl/coop/nedlib/results/NEDLIBmetadata.pdf

Miller, 1997
http://ahds.ac.uk/public/metadata/disc_04.html#pads2

National Library of Australia (1999). *Preservation Metadata for Digital Collection*s. Retrieved on May 3, 2003, from http://www.nla.gov.au/preserve/pmeta.html

National Library of New Zealand (2002). *Metadata Standards Framework – Preservation Metadat*a. Retrieved on May 3, 2003, from
http://www.natlib.govt.nz/files/4initiatives_metaschema.pdf

Open Archives Initiative [OAI] (2003). *The Open Archives Initiative Protocol for Metadata Harvestin*g. Retrieved on May 3, 2003, from
http://www.openarchives.org/OAI/openarchivesprotocol.htm

OCLC [Online Computer Library Center] (2002). *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects.*
OCLC/RLG Working Group on Preservation Metadata. Retrieved on May 3, 2003, from
http://www.oclc.org/research/pmwg/pm_framework.pdf

Pinfield, S. Gardner, M. and MacColl, J. 2002. Setting up an institutional e-print archive

http://www.ariadne.ac.uk/issue31/eprint-archives/

Smith et-al, 2003
http://www.dlib.org/dlib/january03/smith/01smith.html

Weibel, S. 1998. Using Web Metadata
http://www.w3.org/People/EM/talks/www7/tutorial/part1/sld042.htm

SHERPA Project Document

An introduction to metadata requirements for an e-print repository

Gareth Knight
Arts & Humanities Data Service

03/08/2004