Summary of OAI Metadata Best Practices

This handout introduces (1) the range of metadata formats which may be used with OAI, (2) quality issues that currently limit the usability of OAI-shared metadata, and (3) "best practices" for increasing the quality of shareable metadata.

Successful institutional sharing of collection information depends on high-quality metadata. Simple or unqualified Dublin Core (DC) provides a low barrier for institutional sharing of metadata, but more descriptive metadata formats may be additionally shared via OAI to provide richer records. Initial implementation efforts have revealed quality issues that can limit the reliability and usability of records. Establishing "best practices" for shareable metadata among OAI data providers and service providers will alleviate these concerns.

SEE ALSO: **OAI "Cheat Sheet"** (best strategies for converting metadata formats into shareable OAI), **OAI Tools** (technology available for generating, converting, managing, and harvesting metadata), and **Summary of the DLF Aquifer MODS Profile** (recommendations for using the MODS metadata format to share metadata on digital collections of cultural heritage materials).

"Best Practices" for Shareable Metadata

Better communication between OAI data providers and service providers can help identify and eliminate problems with metadata. The OAI "Best Practices" project (a joint effort of the National Science Digital Library and the Digital Library Federation) is developing specific prescriptive recommendations in several areas of OAI implementation: http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl?MetadataContent. Following are some general guidelines for improving the quality of shareable metadata:

Data Providers

- Understand and use XML to convert richer metadata formats and to provide the most appropriate views of the metadata.
- If you are interested in inclusion in a specific OAI-based system, find out their preferred metadata formats and best practices.
- Provide the information most useful for discovery and selection. (For example, technical metadata is usually not necessary – users typically won't care what scanning equipment was used.)
- If exposing metadata about complex or multipart objects, consider whether it makes sense to expose all of the metadata associated with the objects. (For example, if a digitized book is made available, it usually doesn't make sense to include a metadata record for every page of the book unless there is something unique about each page that merits their inclusion. It makes more sense to include just a top level record describing the book as a whole.)
- Use the appropriate metadata format for the resource being described. (For example, don't try to use all of the elements of a finding aid using simple Dublin Core, use EAD. Transfer only the top level information to the DC record.)
- Provide context for metadata, so that a user from outside your institution would be able to identify the resource described.
- Be consistent in the content & format of metadata. It is much easier for service providers to work with 'imperfect' metadata if the imperfections are consistent. Best practice is to provide documentation of the decisions and standards used for exposed metadata.
- Keep elements clean and precise.

Service Providers

- Develop routines for normalizing and enhancing metadata.
- Establish "level of granularity" appropriate for object-level records.
- Provide feedback to metadata providers about metadata interoperability issues.

related DLF projects:

DLF Aquifer
Bringing collections to light through the Digital Library Federation

document authorship:

METASCHOLAR
An Emory University Digital Library Research Initiative

last revised: 29 November 2005
Milewicz & Shreeves - v. 13

Quality Issues

Some aspects of current OAI-shared metadata that limit usability:

- **Inconsistent:** Formatting inconsistencies within and between OAI data providers create uncertainty about how to interpret metadata and impediments to reliably harvesting it.
  EX: discrepancies in “Date” format (1901-07-01; July 7, 1901; ca. 1901, etc.)
  EX: failure to use controlled vocabularies (LCSH, ERIC Thesaurus, etc.) or to indicate that they are being used
  EX: broad range of “granularity” (from providing records for each page of a book to providing only one record for the book)
- **Insufficient:** Metadata offers partial or too little information for understanding resource.
  EX: contextual information lost in conversion to simple DC; also referred to as the “on a horse” problem, i.e., a photograph titled “On a horse” loses significance if no metadata indicate that it belongs to a “Teddy Roosevelt” image collection (Wendler, 2004)
- **Incompatible:** Metadata elements contain information not relevant to or compatible with OAI-based systems, or contain information that interferes with successful harvesting.
  EX: HTML tags in records – may interfere with validation process
  EX: Technical metadata included in descriptive metadata

Metadata Formats

While DC is the minimum requirement for sharing metadata using OAI, other metadata formats may be used, provided they have an XML Schema Definition Document (XSD). Data providers may greatly increase the usability of their metadata by providing metadata in its original format in addition to Dublin Core. Many of the metadata formats currently used by academic and research libraries may be converted into OAI-compliant XML files. See instructions for crosswalking formats at http://www.ukoln.ac.uk/metadata/interoperability/. Common formats are listed below, along with links to implementation guides:

- ***DC*** (simple or unqualified Dublin Core) **http://dublincore.org/documents/2001/04/12/usageguide**
  Lowest common denominator format, required for the OAI protocol. Increasingly viewed as insufficient for understanding resources, since much descriptive metadata must be condensed into fifteen elements.
- ***QDC*** (qualified Dublin Core) **http://dublincore.org/documents/dcq-rdf-xml/#DCQual**
  Specification of qualifiers for Dublin Core elements, to make DC more compatible with RDF & XML and to offer more flexibility in description of metadata elements. The above URL details how to express QDC in RDF and XML. For an older, more complete description of QDC, go to: http://dublincore.org/documents/2000/07/11/dcmes-qualifiers.
- ***MARC*** (MAchine-Readable Cataloging) **http://www.loc.gov/marc**
  Provides rich, consistent metadata for describing resources, and may be converted into an OAI-compliant XML format. MARC-21 represents the incorporation of USMARC and CAN/MARC into one format. Instructions for using MARC-21 in XML: **http://www.loc.gov/standards/marcxml**.
- ***EAD*** (Encoded Archival Description) **http://www.loc.gov/ead**
  Standard format for encoding machine-readable *finding aids* (i.e., abstracts and indexes), providing detail about collections that goes beyond the description available through MARC. Expressed in SGML. There is not currently a standard format for expressing EAD in XML.
- ***METS*** (Metadata Encoding & Transmission Standard) **http://www.loc.gov/standards/mets**
  Standard format for digital objects, providing access to *structural* as well as administrative and descriptive metadata. May be used with MARC (as source for descriptive metadata). See implementation instructions at **http://www.loc.gov/standards/mets/mets-schemadocs.html**.
- ***MODS*** (Metadata Object Description Schema) **http://www.loc.gov/standards/mods**
  A schema for bibliographic element sets, based in MARC-21 and expressed in XML, which can be used for transferring MARC-formatted data from existing records or creating new ones. The newest version of the schema is available at **http://www.loc.gov/standards/mods/v3/mods-3-0.xsd**.
- ***TEI (Text Encoding Initiative) Independent Headers*** **http://www.tei-c.org/**
  Guidelines for creating cross-platform electronic representations of texts. Not a complete format, but used to create shareable metadata records.

REFERENCES

Foulonneau, M., et al. (2004, November 15). *Getting a Leg Up on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)*. Presentation for the National Science Digital Library Conference, Chicago, IL. http://nsdl.comm.nsdl.org/meeting/session_docs/2004/2620_National_Science_Digital_Library_Conference.doc

Shreeves, S. L. (2004, October). *Report on DLF OAI and Shareable Metadata Best Practices Work*. http://comm.nsdl.org/download.php/473/ReportOAIBestPractices_ver2.doc

Tennant, R. (2004, May 14). Bitter harvest: Problems & suggested solutions for OAI-PMH data & service providers. *California Digital Library* http://www.cdlib.org/inside/projects/harvesting/bitter_harvest.html

Wendler, R. (2004). The eye of the beholder: Challenges of image description and access at Harvard. In D. I. Hillman and E. L. Westbrooks (eds.), *Metadata in Practice* (pp. 51-69). Chicago: ALA Editions.