

Metadata in Digital Libraries and an Overview of the OAI-PMH

Jewel Ward <jewelw@lanl.gov>

Visiting Scholar, Keio University

National Institute of Informatics

16 June 2003

Acknowledgements

- Herbert Van de Sompel (LANL)
- Michael L. Nelson (ODU)
- Simeon Warner (Cornell University)
- Tim Cole (UIUC)
- William H. Mischo (UIUC)
- Thomas Habing (UIUC)
- Hussein Suleman (then at VaTech)
- Irma Holtkamp (LANL)
- Jane Greenberg (UNC-CH)
- John MacMullen (UNC-CH)
- Naomi Dushay (Cornell University)

What is Metadata?

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource. Metadata is often called data about data or information about information (Hodge, 2001).

Types of Metadata

The Library of Congress Making of America II project has identified a threefold division of metadata (Making of America II 1998):

- Descriptive metadata - primarily used for resource discovery. Formats currently used include the MARC formats, Encoded Archival Description (EAD) and Dublin Core.
- Structural metadata - data that a system can use to help present a particular digital object to a user.
- Administrative metadata - data that allows the management of a digital collection (Day, 1999).

Administrative Metadata

- “Policy” would be another term for administrative metadata.
 - Administrative metadata facilitates both short-term and long-term management and processing of digital collections
 - includes technical data on creation and quality control
 - includes rights management, access control and use requirements
 - preservation action information
- (Cornell University Library/Research Department, 2003).

Administrative Metadata

- Preservation is essentially about management. In this scheme, preservation metadata (as with rights metadata) is a specialized form of administrative metadata (Day, 1999).
- For further information on preservation metadata, refer to Michael Day's article and/or OAIS.

Administrative Metadata

Sample Elements:

Technical data such as scanner type and model, resolution, bit depth, color space, file format, compression, light source, owner, copyright date, copying and distribution limitations, license information, preservation activities (refreshing cycles, migration, etc.) (Cornell University Library/Research Department, 2003).

Administrative Metadata

Implementations:

- MOA2, Administrative Metadata Elements
<http://www.clir.org/pubs/abstract/pub87abst.html>
- National Library of Australia
Preservation Metadata for Digital Collections
<http://www.nla.gov.au/preserve/pmeta.html>
- CEDARS
<http://www.leeds.ac.uk/cedars/metadata.html>
(Cornell University Library/Research Department, 2003).

Descriptive Metadata

- To repeat the Library of Congress' definition:
Descriptive metadata - *primarily used for resource discovery. Formats currently used include the MARC formats, Encoded Archival Description (EAD) and Dublin Core.*
- These formats are used for resource discovery and resource description (example, the Open Archives Initiative Protocol for Metadata Harvesting).

Data & Service Providers

- Data Providers (DPs) – Repositories – refer to entities who possess resources and metadata and are willing to share metadata with others via well-defined OAI protocols
- Service Providers (SPs) – Harvesters – are entities who harvest metadata from DPs in order to provide high level services to users (such as search and discovery).
- Data equals server, Service equals client

OAI-PMH Metadata

- Repositories are required to expose their metadata as the Dublin Core Metadata Element Set (DCMES).
- Repositories are strongly encouraged to expose their metadata in more expressive formats.
- Examples of other formats in use:
 - MARC
 - RFC-1807
 - Open Languages Archives Community Metadata Set
 - Electronic Theses and Dissertation Metadata Set (Nelson, 2001)

How the OAI-PMH Works

OAI “VERBS”

Identify

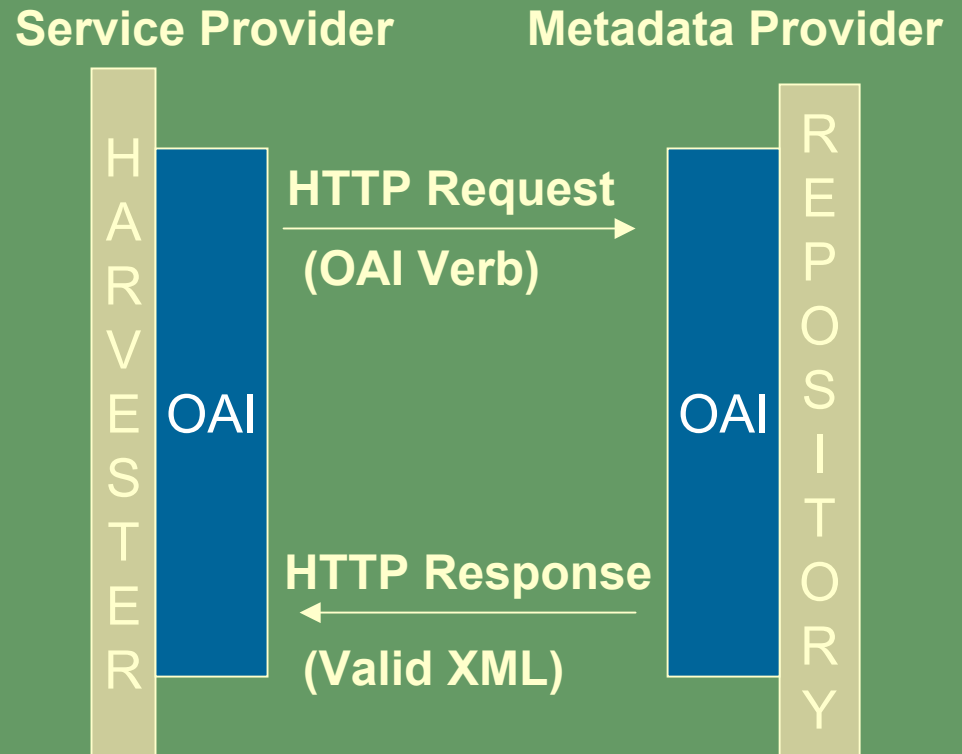
ListMetadataFormats

ListSets

ListIdentifiers

ListRecords

GetRecord



baseURL+verb

Examples

- <http://arXiv.org/oai2?verb=Identify>
- <http://arXiv.org/oai2?verb=ListSets>
- <http://arXiv.org/oai2?verb=ListMetadataFormats>
- http://arxiv.org/oai2?verb=ListIdentifiers&metadataPrefix=oai_dc
- http://arxiv.org/oai2?verb=GetRecord&identifier=<recordID>&metadataPrefix=oai_dc
- http://arXiv.org/oai2?verb=ListRecords&metadataPrefix=oai_dc

Example Response

```
<?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
  http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2003-06-03T20:13:50Z</responseDate>
  <request verb="GetRecord" metadataPrefix="oai_dc"
    identifier="oai:arXiv.org:acc-
    phys/9411001">http://arXiv.org/oai2</request>
- <GetRecord>
..
</GetRecord>
</OAI-PMH>
```

Example OAI_DC Record

```
- <record>
- <header>
  <identifier>oai:arXiv.org:acc-phys/9411001</identifier>
  <datestamp>2003-02-05</datestamp>
  <setSpec>physics:acc-phys</setSpec>
  <setSpec>physics:physics</setSpec>
</header>
- <metadata>
- <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
  xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
  instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
  http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:title>Symplectic Computation of Lyapunov Exponents</dc:title>
  <dc:creator>Habib, Salman</dc:creator>
  <dc:creator>Ryne, Robert D.</dc:creator>
  <dc:subject>Accelerator Physics</dc:subject>
  <dc:description>..</dc:description>
  <dc:description>Comment: 12 pages, uuencoded PostScript (figures included)</dc:description>
  <dc:date>1994-10-31</dc:date>
  <dc:type>text</dc:type>
  <dc:identifier>http://arXiv.org/abs/acc-phys/9411001</dc:identifier>
</oai_dc:dc>
</metadata>
</record>
```

Problems with Metadata

- Missing Data

- Data missing from the DC “format” and “type” elements

- Incorrect Data

- Content that should be in one kind of element was in a different kind of element. For example, “creator” content in the “language” content.
 - Default strings showing no content available: “unknown”, “—”, “...”, etc. (Dushay & Hillmann, 2003).

Problems with Metadata

- Confusing Data
 - Strings of names inconsistently ordered:
 - Smith, John, George Jackson, Humphrey Little and Stanley Black
 - A more correct format should read: Smith, John; Jackson, George; Little, Humphrey; Black, Stanley.
- Insufficient Data
 - Minimal DC requirements combined with inconsistent usage make it difficult to interpret the metadata for more refined services, like searching (Dushay & Hillman, 2003).

Problems with Metadata

- The top five elements are used 71% of the time while the bottom five elements are only used 6% of the time
 - Most- to least- used: creator, identifier, title, date, and type
 - Most- to least- used: language, format, relation, contributor, and source
- 54% of DPs use only two elements (creator, identifier) for approximately 50% of all element usage

Problems with Metadata

- The DCMES is not used to its fullest extent <understatement>
- Due to this underutilization of the DCMES, it will be difficult for the OAI community to build relevant cross-resource services based on it (Ward, 2003)

Domain-Specific Metadata

Bioinformatics

Multiple Metadata formats:

- MAML (Microarray Markup Language)
 - GeneXML
 - GEML
 - Genbank, PDB, GAME, SBML, CellML, DAS
- OLAC
 - Others: GIS, Genome Sequencing

Domain-Specific Metadata

Open Languages Archive Community

Metadata Elements:

Contributor, Coverage, Creator, Date, Description, Format, Format.cpu, Format.encoding, Format.markup, Format.os, Format.sourcecode, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Subject.language, Title, Type, Type.functionality, and Type.linguistic

<http://www.language-archives.org/OLAC/olacms.html>

Metadata Conversion

LaRC MARC to DC Mapping (excerpt)

LaRC MARC Metadata Set	Dublin Core
D245a, D245d, D245e, D245n, D245p, D245s	title
D513a, D513b	coverage
D520b	description
D072a, D072b(001), D650a, D659a	subject
D090a(000), D013a, D020a, D088a, D856q, 856w	identifier

Liu, et al. (2002)

Metadata Conversion

DC to Sandia Mapping

Dublin Core element	Sandia Metadata Field
identifier	report numbers
identifier -- URI	URL
subject	subject category codes
title	title
subject	keywords
creator	personal names
creator	corporate names
date	date
format -- extent	extent
description	notes
rights	classification & dissemination

Liu, et al. (2002)

Bibliography

- Cornell University Research Library/Research Department. (2003). Retrieved 5 June 2003 from <http://www.library.cornell.edu/preservation/tutorial/metadata/table5-1.html>.
- Day, M. (1999). Issues and Approaches to Preservation Metadata. Retrieved 5 June 2003 from <http://www.rlg.org/preserv/joint/day.html>.
- Dushay, N. & Hillmann, D. (2003). Analyzing Metadata for Effective Use and Re-use. Draft submitted to the 2003 Dublin Core Conference. Available http://www.cs.cornell.edu/naomi/DC2003/dushay_hillmann__draft.pdf.
- Hodge, G. (2001). Metadata Made Simpler. NISO Press. Retrieved 5 June 2003 from http://www.niso.org/news/Metadata_simpler.pdf.

Bibliography

- Liu, X., Maly, K., et al. (2002). Technical Report Interchange Through Synchronized OAI Caches. Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries..
- Nelson, M. (2001). Better Interoperability Through the Open Archives Initiative. The New Review of Information Networking. 7, 133-146.
- Ward, J. (2003). A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative. Proceedings of the Third ACM/IEEE Joint Conference on Digital Libraries, 315-317.

Further Information

- Open Archives Initiative
 - <http://www.openarchives.org>
- UKOLN Metadata Resources
 - <http://www.ukoln.ac.uk/metadata/>
- RLG – Issues and Approaches to Preservation Metadata
 - <http://www.rlg.org/preserv/joint/day.html>
- Open Archival Information System
 - Preservation Metadata and the OAIS Information Model
http://www.oclc.org/research/pmwg/pm_framework.pdf

Further Information

- Digital Libraries: Metadata Resources
 - <http://www.ifla.org/II/metadata.htm>
- MARC
 - <http://lcweb.loc.gov/marc/RFC-1807>
- Open Languages Archives Community Metadata Set
 - <http://www.language-archives.org/OLAC/olacms.html>
- Electronic Theses and Dissertation Metadata Set
 - <http://www.ndltd.org/standards/metadata/etd-ms-v1.01.html>

Questions?