Harvesting and Resolution Methods for Building OAI-based Services

Jeffrey A. Young
jyoung@oclc.org

CERN OAI3 Workshop# 4
Geneva, Switzerland
14 February 2004
Introductions

- Name
- Affiliation
- Plans
- Needs
- Technical experience
Review OAI-PMH Protocol

- Identify
- ListSets
- ListMetadataFormats
- ListRecords
- ListIdentifiers
- GetRecord
Find Repositories to Harvest

- [http://www.openarchives.org/Register/BrowseSites.pl](http://www.openarchives.org/Register/BrowseSites.pl)
- [http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai](http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai)
- [http://oai.grainger.uiuc.edu/registry/](http://oai.grainger.uiuc.edu/registry/)
- Friends lists
- Communities (e.g. [www.ndltd.org](http://www.ndltd.org))
Exercise: Getting Started

- What are your data sources?
- How will you add value?
- Who will design the system?
- Who will create/operate the software?
- Who will create/maintain the data?
- Who will advocate for it politically?
- Who will benefit?
- Who will pay?
Metadata

- Metadata is data about data
- Metadata formats: two extremes
  - Dublin Core
  - MARC
- Metadata can be relative
  - Who created this document?
  - Who created the metadata about this document?
- Keep in mind, though, that OAI works just as well for sharing XML content
XML/DTD/XSD/XSL

- XML - eXtensible Markup Language
- DTD - Document Type Definition
- XSD - XML Schema Definition
- XSL - eXtensible Stylesheet Language
eXtensible Markup Language

- Meta-markup language
- HTML – Hypertext markup language
- XHTML – eXtensible hypertext markup language
XML Overview

- Well-formed XML
- XML Namespaces
- Valid XML
  - DTDs
  - XML Schemas
- OAI Items vs. Records
  - Item identifiers
  - Multiple metadata record representations
XML Namespaces

- Ambiguous XML Elements
  - `<wind>NNE</wind>`
  - `<wind>Clockwise</wind>`

- Prefixes help identify and differentiate elements
  - `<weather:wind>SE</weather:wind>`
  - `<toy:wind>Widdershins</toy:wind>`

- But, prefixes are arbitrary and potentially ambiguous, so what we really need is a URI (ie. prefixes are a local shorthand for the URI)
  - `<weather:wind xmlns:weather="someURI">NW</weather:wind>`
XML Schema Definition

- Defines what an XML document contains
  - XHTML
  - oai_dc
  - MARC21 XML
What is our “item”? 

- Work – a distinct intellectual or artistic creation
  - J.S. Bach’s *The art of the fugue*

- Expression – the intellectual or artistic realization of a work
  - The composer’s score for organ
  - An arrangement for chamber orchestra by Anthony Lewis

- Manifestation – The physical embodiment of an expression of a work
  - CD, printed score, multimedia kit, etc.

- Item – A single exemplar of a manifestation
Exercise: Data Definition

- Design a metadata format for items in your project
  - List the elements you need
  - Consider the encoding rules
  - Consider using controlled vocabularies

- Assign an XML namespace

- Map a crosswalk to Dublin Core

- Create a sample item with both formats
  - Consider assigning OAI sets

- Report issues, problems, and concerns
Exercise: A Simple Harvester

- XOAIHarvester – a simple harvester written in XSLT


- The purpose of the Perl script is to manage incremental harvesting

- Caveat! OAI is merely the first step. Once data is harvested, OAI provides absolutely no guidance for doing something useful with it.
Concerns

- Data quality
- Duplicates
- Intellectual Property Rights (IPR)
- The appropriate copy problem
- Persistence
Repository Variables

- **MetadataPrefix**
  - oai_dc – the lowest common denominator

- **Set**
  - Hierarchical
  - Allows selective harvesting
  - Work best with community agreement
  - Client warrant
Exercise: Select/Create Tools

- http://www.oaforum.org/oaf_db/list_db/list_software.php
- http://www.cs.cornell.edu/people/simeon/software/utf8conditioner/
- http://harvest.physik.uni-oldenburg.de/dc/index.html
An Alternative Service Model

- ERRoLs are URLs to content and services related to repositories in the OAI Registry at UIUC

- [http://errol.oclc.org/](http://errol.oclc.org/)
Discussion

- Issues, Problems, Concerns?
Music Services

- Organizational issues
- Cultural issues
- Collection policies
- Best practices
- Consensus-building
- Controlled vocabularies
  - http://alcme.oclc.org/gsafd/
- Do items represent digital and/or physical entities?
- Authority control
Repository Descriptors

- Repository-level “description” elements
  - oai-identifier description – identifier layout
  - eprints description – content & policies
  - friends description – discover repositories
  - branding description – branding information
  - olac-archive description – archive info

- Record-level “about” elements
  - Rights statements
  - Provenance statements
XSLT Overview

XML Document

XSLT Stylesheet

XSLT Processor

XML Document
Validate Repositories

- [http://www.openarchives.org/data/registerasprovider.html](http://www.openarchives.org/data/registerasprovider.html)
- [http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai](http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai)
- [http://www.w3.org/2001/03/webdata/xsv](http://www.w3.org/2001/03/webdata/xsv)
Example Service Providers

- **ARC - A Cross Archive Search Service** (experimental research service)
  http://arc.cs.odu.edu/

- **Dokumenten- und Publikationsserver der Humboldt-Universität zu Berlin** (search service, German language user interface)
  http://edoc.hu-berlin.de/oaisearch/

- **iCite** (citation index)
  http://icite.sissa.it/

- **NCSTRL**—Networked Computer Science Technical Reference Library (search engine)
  http://www.ncstrl.org/

- **my.OAI** (value-added search interface to a selected list of metadata databases)
Resources

- http://www.openarchives.org/
- http://www.oaforum.org/
Everything you need to know

http://www.oaforum.org/otherfiles/oaf_d23_technical2.pdf