The Open Archives Initiative Protocol for Metadata Harvesting: an Opportunity for Sharing Content in a Distributed Environment

Muriel Foulonneau (mfoulonn@uiuc.edu)

Grainger Engineering Library
University of Illinois at Urbana-Champaign

UIUC
June 2006
Outlines

- Improving resource discoverability
  - Hidden Web, portals and distributed digital libraries
- Interoperability
  - Metadata and protocols
- The Open Archives Protocol for Metadata Harvesting
  - The protocol, examples of services and repositories
- Issues for digital libraries of distributed objects
Improving resource discoverability
Sharing content

- New services, new representations of the content, new audiences

- Bring your content to attention of new users outside your immediate community
  - 37% of visits to images of the State Library of New South Wales came from the PictureAustralia portal in 2002/3
Integrated Access to CIC Metadata

The CIC metadata portal provides access to a broad range of noteworthy and often singular information resources held by Midwestern research universities belonging to the CIC consortium (Big Ten Athletic Conference Universities, the University of Illinois at Chicago, and the University of Chicago).

The portal allows you to search across many types of content: text, videos, sheet music, thesis and dissertations, websites, software. Where available, links are provided directly to the information resource on the hosting CIC institution Website.

http://cicharvest.grainger.uiuc.edu/
Thematic access to resources

Which African State are you interested in?

Notice: The area covered by the resources have been automatically interpreted; this may lead to discrepancies.

<table>
<thead>
<tr>
<th>Country</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sudan</td>
<td>166</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>title</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Library Association Archives: Executive Board and Executive Director Executive Directors Salary File</td>
<td>The materials described in this descriptive record are archives (official records of an organization).</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>Type</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Fine Arts and Applied Arts Urban and Regional Planning Department: Scott A. Cray &amp; Gay's Papers</td>
<td>The materials described in this descriptive record are manuscripts (personal papers of an individual).</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>Type</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>University of Illinois at Urbana-Champaign archives finding aid - University of Illinois at Urbana-Champaign</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>Type</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Sugar Plantation in Central, Blue Nile Province</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>Type</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Africa Food - Sights and Sounds of a continent - University of Wisconsin Madison</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>Type</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Peanut Cultivation in the Tariq Area of the Nile Valley</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>Type</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Africa Food - Sights and Sounds of a continent - University of Wisconsin Madison</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>Type</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Young Girl in Field of Long Staple Cotton in Blue Nile Province</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>Type</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Harvesting from a Date Palm</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>title</td>
<td>Type</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Aerial View of Desert Development Project in the Gabor</td>
<td></td>
</tr>
</tbody>
</table>

Website maintained by the University of Illinois at Urbana-Champaign

In partnership with CIC institutions.

For information and comments, please contact Maria K. Edmunds.
Russian Publics collection at UIUC
On the CIC metadata portal
Search on Google

Web

Tip: Search for English results only. You can specify your search language in Preferences.

4 results stored on your computer - Hide - About

Gospodin i Izvoshchik. **Un Seigneur et un fiacre.** Ein Herr und ein ...

**Un Seigneur et un fiacre.** Ein Herr und ein Miethkutscher. Description, (unknown). Creator, (unknown). Subject. Source, 'Volshebnoi fonal' ...

Madame Bovary : la scène du fiacre - [ Translate this page ]
Et, à droite, ce gentilhomme tout barbu de fer, sur un cheval qui sa cabre, est son petit-fils Louis de Brézé, **seigneur** de Breval et de Montchauvet, ...

Guenog - Tout sur l'église - [ Translate this page ]
Saint Michel présentant **seigneur** et une dame. Le **seigneur** est vêtu d'une cotte ... et un moine tonsuré, robe blanche et capucule bleu (sa **fiacre**) ? ...

**ATHENA: MAUPASSANT, Un Lâche; Thierry Selva** - [ Translate this page ]
Il prit enfin le marquis de la Tour Noire et le colonel Bourdin, **un grand seigneur** et **un soldat**, c'était fort bien. Leurs noms portaient dans les journaux ...

**Amazon.fr : Un amant de rêve ; Livres** - [ Translate this page ]
Achetez ces 2 articles : **Un amant de rêve** et **Le seigneur** arrogant ... elle vient à peine de s'engager dans Baker Street qu'**un fiacre** s'arrête à sa hauteur. ...

mfoulonn@uiuc.edu
University of Illinois at Urbana-Champaign

June 15th, 2006
Multiple services use different features
Interoperability
Content and services

- Building services

=> New services need content with similar features
What is interoperability

- Interoperability is the capacity for different systems to talk to each other

- “01-04-04”
  - this is a month
  - 01=“Jan”

- I need
  - A standard language
  - An interpreter
Various types of interoperability

- **Technical**
  - Protocols, hardware, … Mac/PC, Netscape/IE …

- **Organizational**

- **Content – related = metadata**
  - What do you talk about? The “item” = Granularity and nature of the object
  - Semantic: date…. Created? Published?
  - Syntactical: 04 January 2004
  - Linguistic: 04 Enero 2004
Metadata

- Are used to
  - Manage
  - Provide information
  - Retrieve
  - Preserve
  - Define rights and conditions of use
  - Describe structure

⇒ Descriptive
⇒ Administrative
⇒ Structural
A metadata format

- Is a set of elements or information, mandatory or not, to apply together in order to reach one of the above mentioned objectives

- Standard
  - As a text
  - As a DTD in SGML
  - As a Xschema in XML

=> MARC, EAD, MODS, Dublin Core, LOM, MPEG7, MyHomeCookedSchema …
The Dublin Core Metadata Element Set

- 15 elements

<table>
<thead>
<tr>
<th>Content</th>
<th>Intellectual property</th>
<th>Instantiation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>Rights</td>
<td>Language</td>
</tr>
<tr>
<td>Description</td>
<td>Contributor</td>
<td>Identifier</td>
</tr>
<tr>
<td>Relation</td>
<td>Publisher</td>
<td>Format</td>
</tr>
<tr>
<td>Type</td>
<td>Creator</td>
<td>Date</td>
</tr>
<tr>
<td>Source</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Where metadata lay

- "Internal"
  - Webpage
- Embedded
  - TEI, EAD
- External
  - Catalogs
  - XML records …
    - Includes a link to the resource

=> Third party metadata

Library of Congress home page

```
<HTML>
<HEAD>
<TITLE>The Library of Congress</TITLE>
<META NAME="description" CONTENT="Home page of the Library of Congress, Washington, D.C. The Library of Congress is the nation's oldest federal cultural institution, and it serves as the research arm of Congress. [...]"> 
<META NAME="keywords" CONTENT="library of congress, home page, catalog, copyright office, [...]">
```
Sharing metadata: Federated search

- My user wants “mills”…. Whatever that comes from
  - Federated search

```
<title>My resource</title>
<date>04
```

Eg. Z39.50, SRU/SRW, WAIS
Sharing metadata: Data aggregation

- The portal gathers metadata (and resources?)

Eg. Search engines, union catalogs, OAI
OAI divides the world between data providers and service providers
The OAI framework

Service provider

Harvester

Repository
Data provider

Repository
Data provider

Repository

Aggregator

Data provider

Data provider
OAI repositories can be organized in sets

What do sets represent?

- Journals: issues
- Institutional repositories: Departments, research centers, etc.
- EPrint Archives: Subject, Publication Status
- Cultural Heritage Repositories: Collections with Intent

Set representations may be constrained by the software package used.

April, 2006
Multiple representations of an object

MARC Record

In XML

Dublin Core Record

In XML

Qualified Dublin Core Record

In XML

MODS record

In XML

Honoré Daumier Lithograph (Brandeis University)
OAI is based on standards

- HTTP protocol
- XML
- XML Schemas
- Dublin Core
OAI supports 6 verbs

- **Identify**
  
  http://aerialphotos.grainger.uiuc.edu/oai.asp?verb=Identify

- **ListSets**
  
  http://aerialphotos.grainger.uiuc.edu/oai.asp?verb=ListSets

- **ListRecords**
  

- **ListMetadataFormats**
  
  http://aerialphotos.grainger.uiuc.edu/oai.asp?verb=ListMetadataFormats

- **ListIdentifiers**
  
  http://aerialphotos.grainger.uiuc.edu/oai.asp?verb=ListIdentifiers&metadataPrefix=oai_dc

- **GetRecord**
  
An OAI response

<record>
  <header>
    <identifier;oai:images.library.uiuc.edu:emblems/324</identifier>
    <datestamp>2003-10-22</datestamp>
    <setSpec>emblems</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:creator>Müller, Johann Heinrich Traugott, 1631-1675</dc:creator>
      <dc:identifier>http://images.library.uiuc.edu:8081/u/?/emblems,324</dc:identifier>
    </oai_dc:dc>
  </metadata>
</record>
Examples of repositories

- Library of Congress
  [http://memory.loc.gov/cgi-bin/oai2_0](http://memory.loc.gov/cgi-bin/oai2_0)

- ContentDM at UIUC
  [http://images.library.uiuc.edu:8081/cgi-bin/oai.exe](http://images.library.uiuc.edu:8081/cgi-bin/oai.exe)

- Ohio State Knowledge Bank
  [https://kb.osu.edu/dspace-oai/request](https://kb.osu.edu/dspace-oai/request)
Examples of services

- OAster: http://oaister.umdl.umich.edu
- CIC Metadata Portal: http://cicharvest.grainger.uiuc.edu/
- American South: http://www.americansouth.org/
- NSDL: http://nsdl.org/
- Picture Australia: http://www.pictureaustralia.org/
- Language Archives: http://www.language-archives.org/
Turn key systems and modules

- CWIS: http://scout.wisc.edu/Projects/CWIS/
- ContentDM: http://contentdm.com/
- Digitool: http://www.exlibrisgroup.com/digitool.htm
- DSpace: http://www.dspace.org/
- EPrints: http://software.eprints.org/
- DLXS: http://www.dlxs.org/
- OAICat: http://www.oclc.org/research/software/oai/cat.htm
- DLESE OAI software: http://dlese.org/oai/index.jsp
Useful tools

- UIUC OAI registry
  http://gita.grainger.uiuc.edu/registry/
- OAI repository explorer
  http://re.cs.uct.ac.za/
- Errol
  http://errol.oclc.org/
Digital libraries of distributed objects
Metadata shareability issues

- Granularity
- Loss of context
- Completeness

Record 2 of 8

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Author/Creator</td>
<td>Richard Lewis, c.1700-1734</td>
</tr>
<tr>
<td>Publisher</td>
<td>Printed by Hall and Sellers</td>
</tr>
<tr>
<td>Type</td>
<td>text</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://name.umdl.umich.edu/AM1315">http://name.umdl.umich.edu/AM1315</a></td>
</tr>
<tr>
<td>Collection</td>
<td>American Poetry Database</td>
</tr>
</tbody>
</table>

DLF-NSDL Best practices on shareable metadata

http://oai-best.comm.nsdlo.org/cgi-bin/wiki.pl?TableOfContents
What is behind URLs
Conveying actionable URLs

http://rama.grainger.uiuc.edu/assetactions/

View

Resize

Select

Annotate

Share

Battery Park New York City. A Battery Burns surveys this scene.

mfoulonn@uiuc.edu
University of Illinois at UC

June 15th, 2006
Conclusions

- Interoperability: technical, content-related and organizational, well OAI is the easy part

- Works even better for particular communities with similar organizational structures and metadata formats

- Extensions of the protocol for:
  - Objects
  - Actionable URLs
References and useful material

- The Open Archives Website
  http://www.openarchives.org/OAI/2.0/guidelines.htm
- DLF/NSDL best practices for OAI and shareable metadata
  http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl?TableOfContents
- OAForum Tutorial
  http://www.oaforum.org/tutorial/
- Getting a Leg Up on OAI