Open Archives Initiatives Protocol for Metadata Harvesting Practices for the cultural heritage sector

Muriel Foulonneau
Relais Culture Europe
mfoulonneau@relais-culture-europe.org
Community report

- A community report on cultural heritage actors
  - Relais Culture Europe & Resource
- Heritage and cross domain issues :
  - Memory organisations

- Practices in :
  - Aggregated resources
  - OAI implementations

OAForum – September 2003
Muriel Foulonneau
OAI in the CH sector

⇒ A protocol based on metadata

- Reaching the hidden Web
- Cross-domain and cross-institution services
- Data re-usability
- Data exchange
- Downstream services
OAI Framework

- Shareable / Disclosed
- Asynchronous model
- Economic and organisational model
  - Digital heritage actors
  - "Contract" for data sharing
Schemas

- XML schemas
- Opportunity of OAI-DC?
- More complex schemas decided at service level

⇒ Need for proper schemas
⇒ Review metadata quality
⇒ Standardisation process
Records in a repository

Role of content providers to map their data to the common schema(s)
- Identify which data can be made available
- Necessity of standard crosswalks and metadata guidelines
- Managing relationships and EAD records

... and ensure high quality metadata
- Evaluate data update
- Identifiers
- How and when to define sets
American Memory OAI sets

Library of Congress collections for which records are available for harvesting through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

Last updated: March 10, 2003

Item-level metadata for the following Library of Congress collections of historical materials is available for harvesting using OAI-PMH version 2.0. Included are selected collections presented online through American Memory and selected collections accessible through the Prints & Photographs Division Online Catalog.

For more information on the Open Archives Metadata Harvesting protocol, see www.openarchives.org.

An American Ballroom Companion: Dance Instruction Manuals, ca. 1490-1920
- **Description**: A collection of over two hundred social dance manuals at the Library of Congress published from ca. 1490 to 1929. Along with dance instruction manuals, this collection also includes a significant number of antidance manuals, histories, treatises on etiquette, and items from other conceptual categories. Many of the manuals also provide historical information on theatrical dance.

  *setSpec*: mсудиbib
  *setName*: LC Dance Manuals
Technical issues for a repository

- System and architecture

- Integrating in local network architecture: system may not be open

- Data delivery
  - provenance
  - rights
  - valid XML
  - accessibility
Communication and management

An « OAI repository project »

- Budget
- Competences
- Register the repository to test it and find possible partners
- Publish information on the repository : for professional use
- Project evaluation

=> project impact on collaborations, openness, standardisation, fundraising…

Muriel Foulonneau

OAForum – September 2003
Community-based services

A charter

- Schema, terminologies under which data are available (+ guidelines)
- IPR and usage of metadata
- Data representation (in / out of frame, de-dupping, branding, ..) and service quality criteria
- Update frequency, harvesting frequency, repository availability
- New members acceptance
Access to aggregated resources

- Cross-collection content is always heterogeneous
- Access points
  - Search interfaces
  - The Dublin Core Culture: a retrieval-oriented format
  - Exploiting the wealth of terminologies
Display issues: Do not disappoint user
Approach to resource aggregation

- A common metadata schema
  - Service-specific
- Granularity issue
  - Display level information
- Metadata crosswalks
  - Examine how metadata are exploited
- Terminology issues
  - Use of a pivot terminology?
- A multilingual environment
  - Language attribute in elements?
(Meta)data harvesting

- Harvesting frequency to improve synchronisation
  - Record repository changes (ARC)
  - Ask repository
  - Push system?
- Harvesting profiles

Major problem is XML validity and servers’ availability: improvement with aggregators
Normalizing and managing content

- Reprocessing content
  - De-dupping
  - Translation
  - Relation to pivot terminology
- New collection
  - Service does not control evolution of its collection (growth)
  - New repositories?

⇒ Identity of the service collection?
Displaying information

- Information reliability
- Integrated access
  - Ambiguity on responsibilities
  - Logical browsing conditions between data provider’s Website and service

⇒ Quality criteria
Send user back to data provider’s Website

<table>
<thead>
<tr>
<th>Title</th>
<th>1942-01. WESTERN DESERT. AN UNSHAVEN FLYING OFFICER R. H. (BOBBY) GIBBES, LATE OF MANLY, NSW, NOW FLYING WITH NO. 3 SQUADRON RAAF. (NEGATIVE BY DAMIEN PARER)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td>DAMIEN PARER</td>
</tr>
<tr>
<td>Subject</td>
<td>Second World War</td>
</tr>
<tr>
<td>Description</td>
<td>1942-01. WESTERN DESERT. AN UNSHAVEN FLYING OFFICER R. H. (BOBBY) GIBBES, LATE OF MANLY, NSW, NOW FLYING WITH NO. 3 SQUADRON RAAF. (NEGATIVE BY DAMIEN PARER)</td>
</tr>
<tr>
<td>Publisher</td>
<td>Australian War Memorial</td>
</tr>
<tr>
<td>Image number</td>
<td>011962</td>
</tr>
<tr>
<td>Format</td>
<td>image/jpeg</td>
</tr>
<tr>
<td>Date or place</td>
<td>Manly; Western Desert; 1999-1945</td>
</tr>
<tr>
<td>Rights</td>
<td>Copyright Australian War Memorial 1998</td>
</tr>
<tr>
<td>Go to</td>
<td>Original image</td>
</tr>
</tbody>
</table>
Quality grid on cultural Websites

• Definition of
  ➢ institution responsibility
  ➢ services proposed
  ➢ who the service represents
  ➢ audience
  ➢ representation of the identity of the service

• Set a policy on
  ➢ IPR
  ➢ maintenance
  ➢ content selection

• Usability: Navigation
  ➢ context evidence: where user is
  ➢ link evidence: where user is goes
  ➢ backtracking soundness

• Identification of
  ➢ Accessibility: access conditions, necessary plug-ins if any…
  ➢ content sources
  ➢ currency
How to spread OAI-PMH ?

• The role of Heritage **Aggregators**
• **Industrial** partners
• Digitisation issue for taking into account the full process of digital content creation and « metadata gardening »
• **Downstream services** have more focused semantic and syntaxic interoperability difficulties
• Build the **Tools**

=> Several areas to **test** and good practices to find
Further information

Soon on OAForum Website http://www.oaforum.org

Contact mfoulonneau@relais-culture-europe.org