Integrating Access to Digital Content

OR

OAI is easy, metadata is hard

Sarah Shreeves
University of Illinois at Urbana-Champaign

Visual Resources Association
23rd Annual Conference
Miami Beach, FL
March 6, 2005
Why Integrate Access?

- Increase access to your collections
  - 37% of visits to images of the State Library of New South Wales came through the PictureAustralia portal
- Aggregation and exposure of the ‘hidden web’
- Build a digital library out of digital collections
  - Services - (Curriculum support, exhibits, new scholarly possibilities)
- Enabling collaborations among resource developers
Search interoperability

“the ability to perform a search over diverse sets of metadata records and obtain meaningful results.”

– Priscilla Caplan

*Metadata Fundamentals for All Librarians*
Keys to Search Interoperability

- Organizational commitment
- Communication protocol (Z39.50, OAI, etc.)
- Standards, Standards, Standards
Sharing metadata: Federated search

The distributed databases are searched directly.

For Example:
Z39.50, SRU/SRW
Sharing metadata: Data aggregation

The user searches a pre-aggregated database of metadata from diverse sources.

For Example:

Search engines, union catalogs, OAI
What is the Open Archives Initiative Protocol for Metadata Harvesting?

• A tool to move metadata from one place to another

Misconceptions about OAI

• OAI ≠ Dublin Core
• OAI ≠ Search Protocol
• OAI ≠ Content (though this is changing)
The Basics of OAI

- Data providers – expose metadata
- Service providers – ‘harvest’ metadata
- All interactions based on HTTP and XML
- Requires use of simple Dublin Core BUT supports use of other metadata schemas
- Open source software available for both data and service providers
- Currently 560+ data providers and far fewer service providers

Identify
ListMetadataFormats
ListSets
ListIdentifiers
GetRecord
ListRecords

Admin. Verbs

Content Harvesting
OAI Use of Dublin Core

DC is OAI’s lowest common denominator

BUT

OAI supports & encourages use of other community-driven metadata schemas

BUT

Metadata schema **MUST** have an XML Schema (XSD) for validation purposes
Decides what metadata format to harvest

Harvests metadata in most appropriate format

Analyzes metadata for quality issues and general processing

Cleans up metadata
• Empty or useless fields
• Determine primary URL
• Processing specific types of information
• Applying encoding schemes

Maps to service metadata format

Index and make available in service
Challenges for the OAI Community

- Wide variety of domains involved
- Best practices still in development
- What is shareable metadata anyway?
The Problems for Service Providers

- Metadata written for different users and uses
- Dublin Core is not semantically complex
- The one to one rule
- Loss of contextual information (the ‘on the horse’ problem)
- Inconsistency within and across data providers
- Turnkey systems are incorporating OAI at the minimal level
"We'll have lots to eat this winter, won't we Mother?"

Grow your own
Can your own
Metadata for different communities

Identifier: MPW00250
Identifier: OCLC40637019
Identifier: http://snuffy.lib.umn.edu/image/srcb/4isu/Dispachter?mode=600&id=MPW00250
Title: "We'll have lots to eat this winter, won't we Mother?": grow your own can your own
Date: 1943
Rights: Subject to U.S. and international copyright laws. Please contact the owning repository.
Subject: World War, 1939-1945--United States--Posters
Subject: Food conservation
Subject: Agriculture
Subject: Victory gardens
Subject: War work
Language: eng
Creator: Parker, Alfred, 1906-
Contributor: United States. Office of War Information
Description: This object is held by Minneapolis Public Library
Description: "OWI Poster No. 57. Additional copies may be obtained ... from the ... Office of War Information, Washington, D.C." U. S. Government Printing Office : 1943--O-520465
Description: A mother and daughter with matching blonde ponytails and aprons process food at home, while behind them is a shelf packed with canned fruit and vegetables.
Relation: http://wwwmplbl.org/wpdlb
Relation: This object is part of a larger series in the University of Minnesota digital collections database, available at http://digital.lib.umn.edu. The specific series is War Posters (Rationing and conservation).
Type: Poster
Format: image/jpeg
Metadata for different communities

**Format:** wv20015p.jpg

**Type:** image

**Title:** We'll Have Lots to Eat this Winter, Won't We Mother?, Grow Your Own, Can Your Own

**Coverage:** 1943

**Description:** Poster, color, 16 x 22.6 in., published by the United States Government Printing Office

**Description:** Many foods, including coffee, tea, butter, meat, frozen and canned vegetables were rationed during World War II. Americans were encouraged to plant "victory gardens" to help provide food for their families and neighbors. Women often preserved the excess produce from these gardens through home canning. Canning is a food preservation system that involves precooking and then air-tight sealing of food in jars, which are then immersed in a hot water "bath" for a specified period of time. **This** hot water "bath" is intended to kill off contaminants that may have survived the processing. If the contents of the jars were very acidic, such as tomatoes, there was less danger of the canned food spoiling. Other, less-acidic, home-canned foods were protected from spoiling by the use of brines, sugar, or salt as preservatives and for flavor.

**Description:** World War II

**Description:** 14 Political systems; 16 History

**Creator:** United States. Office of War Information

**Contributor:** Parker, Albert [artist]

**Publisher:** Illinois State Library

**Language:** English

**Subject:** Canning and preserving--United States ; Food conservation--United States ; World War, 1939-1945--Food supply--United States ; War posters, American ; World War II ; Winter

**Rights:** [Link]

**Identifier:** wv20015p

**Date:** 11-20-01

**Identifier:** [Link]
Granularity of Description: Excerpt of Metadata Record
Describing “American Woven Coverlet”

<table>
<thead>
<tr>
<th><strong>Basic Information</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Artifact Identification</strong></td>
</tr>
<tr>
<td><strong>Classification</strong></td>
</tr>
<tr>
<td><strong>Visual Description</strong></td>
</tr>
<tr>
<td><strong>Artist/Maker</strong></td>
</tr>
<tr>
<td><strong>Geographic Location</strong></td>
</tr>
<tr>
<td><strong>Period/Date</strong></td>
</tr>
<tr>
<td><strong>Culture</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Physical Analysis</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dimension 1 (Length)</strong></td>
</tr>
<tr>
<td><strong>Dimension 2 (Width)</strong></td>
</tr>
<tr>
<td><strong>Dimension 3 (Depth)</strong></td>
</tr>
<tr>
<td><strong>Weight</strong></td>
</tr>
<tr>
<td><strong>Measuring Remarks</strong></td>
</tr>
<tr>
<td><strong>Materials</strong></td>
</tr>
<tr>
<td><strong>Manufacturing Processes</strong></td>
</tr>
<tr>
<td><strong>Munsell Color Information</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Research Remarks</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Published Description</strong></td>
</tr>
<tr>
<td><strong>Scholarly Notes</strong></td>
</tr>
</tbody>
</table>

| **Comparanda** | See Spurlock # 1934.01.0094A. |
| **Bibliography** | N/A |

23rd Annual Conference
Granularity of Description: Excerpt of Metadata Record Describing "Cotton coverlet with embroidered butterfly design"

Title: Cotton coverlet with embroidered butterfly design

Creator: Hayutin, Anna Ginsberg, 1893-1925.

Link: http://www.penlib.du.edu/speccoll/beck/b001detail1.htm
http://www.penlib.du.edu/speccoll/beck/b001detail2.htm
http://www.penlib.du.edu/speccoll/beck/b001full.htm

Publisher: University of Denver, Penrose Library

Description: Digital image of a single-sized cotton coverlet for a bed with embroidered butterfly design. Handmade by Anna F. Ginsberg Hayutin.

Date Original: 1912-1920?
Date Digital: 2001-04-05.

Subject(s): Bedding, Coverlets, Household linens, Interior decoration accessories, Embroidery, Cotton fabrics, Jewish crafts, Jewish embroidery, Decoration and ornament, textile crafts. [AAT]

Jews—Colorado—Social life and customs—20th century

Handicraft.

Type: image

Source: Materials: cotton and embroidery floss. Dimensions: 71 in. x 86 in. Markings: top right hand corner has 1 1/2 in. x 1/2 in. label cut outs at upper left and right hand side for head board; fabric is woven in a variation of a rib weave; color each of yellow and gray hand-embroidered cotton butterflies and flowers from two shades of each color of embroidery floss - blue, pink, green and purple single top 20 in. bordered with blue and black cotton embroidery thread; stitches used for embroidery: running stitch, chain stitch, French knot and back stitches; selvage edges left unfinished; lower edges turned under and finished with large gray running stitches made with embroidery floss.

Relation: Material Culture Collection, Ira M. Beck Memorial Archives of Rocky Mountain Jewish Historical Society, Special Collections Dept., Penrose Library, University of Denver, Denver, Colo.; Heritage Colorado

Format: Epson Expression 836 XL Scanner with Adobe Photoshop version 5.5, 300 dpi; 21-53K bytes

Rights: http://library.du.edu/About/collections/SpecialCollections/copyright.htm

Project: University of Denver
What is Shareable Metadata for the Visual Resources Community?

- Contains the necessary semantics/structure
  - VRA Core, CCO-lite, MODS, MARC, Qualified DC
- Is appropriate for its content
- Uses standards and best practices for content
  - Cataloging Cultural Objects; Controlled Vocabularies
- Provides context (the ‘on the horse’ problem)
  - Collection level description can help
- Is consistent
- Has documentation

What does your service provider need?
Community based efforts

- Open Language Archives
- Sheet Music Consortium (inactive)
- Digital Library Federation Best Practices
- National Science Digital Library

- Visual Resources?
“We are beginning to explore how to share metadata for the digitized collections, and have very good technical solutions, but this has not yet matured into a well understood set of services. This is one aspect of what I meant when I said that this activity was in the 'cottage industry' stage.”

Lorcan Dempsey

http://orweblog.oclc.org/archives/000602.html
Contact Information

Sarah Shreeves
Project Coordinator,
IMLS Digital Collections and Content Project
University of Illinois at Urbana-Champaign
sshreeve@uiuc.edu
217-244-7809