Implementing the OAI Protocol for Metadata Harvesting at the Library of Congress

Caroline Arms

Office of Strategic Initiatives

Library of Congress

caar@loc.gov

Becoming an OAI “data provider”

- Easy and fun
  - Simple to implement, low impact
  - Quick results
  - Demonstrates benefits of interoperability
  - LC staff love sharing their treasures
- Side-effects
  - Experience with new tools
  - Persistent identification
  - Data quality

Background

- American Memory
  - – Rich resource for mining and re-purposing
  - – People want to use content in other resources
    - For particular audiences (scholarship, education)
    - More comprehensive
    - More specific in theme or genre
    - Describe more fully
    - New interfaces
  - – LC has culture of sharing
    - But not if it drains resources from local priorities

Musings on interoperability

- Hard to define
  - cooperation among independently managed systems
- Noticeable when absent
  - like reliability, customer service
- A state of mind
- “spectrum of interoperability”
  - identified by NSDL core integration team
  - three levels of agreement
  - technical, content, organizational

LC/Ameritech Competition

- 3 years, 23 collections
- Metadata harvesting
- Loose agreements at all 3 levels
- Learning experience
- Consistency is important
  - but technology can compensate in some areas
- Rich metadata is wonderful
  - but expensive
  - don’t discourage it

Immediate appeal of OAI-PMH

- Low barrier to entry for data providers
  - Easy to implement
  - Low impact on primary users
- Accommodates simple AND rich metadata
  - Unqualified Dublin Core for interoperability
  - Richer schemas for those who will use them
- Practical approach to standards development
- Dissemination of American Memory content

Appeal also based on potential

- Integrated access to resources
  - from libraries, museums, and other cultural heritage institutions
- Marketplace in metadata practices
  - What makes metadata usable in multiple contexts?
  - How to allocate scarce resources?
  - Balance costs between data providers and service providers (harvesters)

Easy to implement

- Very basic functionality
- Tools available
  - XML, HTTP
- Synergies with other American Memory directions
  - Persistent identifiers (handles)
  - Links from LC’s catalog records
  - Consistency of “non-MARC” records
- Synergies with other MARC directions
  - Character mapping -- UNICODE
  - MARC in XML

Available now

- 123,268 records
- Items from American Memory
  - Books (208)
  - Sheet music (62,976)
  - Photographs (47,767)
  - Early movies (614)
  - Broadsides and pamphlets (7,740)
  - Maps (3,963)
- More in pipeline
  - Books, photos

In several formats

- Simple Dublin Core
  - for interoperability
- MARC21
  - Marc21slim schema
  - represents structure
  - allows round trip conversion to ISO 2709
- MODS
  - new schema from MARC Standards Office

What is MODS?

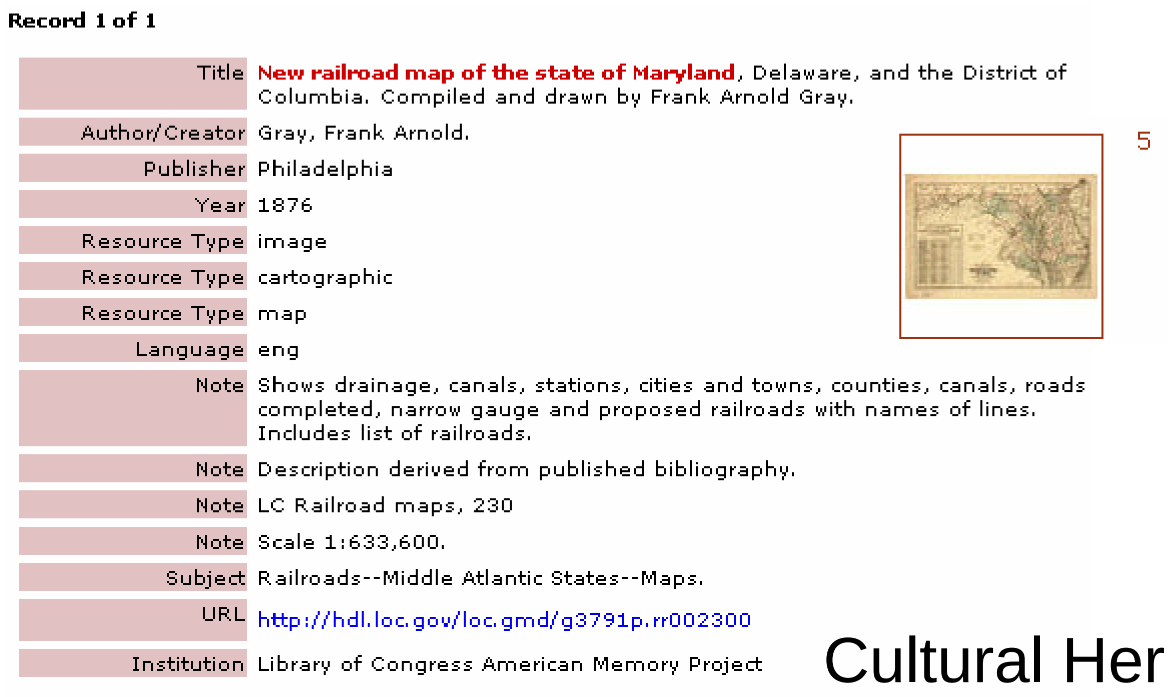
- Metadata Object Description Schema
- A subset of MARC elements, using words as element names
- Specifically for library applications, although could be used more widely
- Element set is richer than Dublin Core, simpler than MARC
- Rich structure for related items allows for description of complex digital objects and a variety of relationships

MODS high-level elements

- TitleInfo (mandatory)
- Name
- Type of resource
- Genre
- PublicationInfo
- Language
- Physical description
- Abstract
- Table of contents
- Target audience
- Note
- Cartographics
- Subject
- Classification
- Related item
- Identifier
- Location
- Access conditions
- Extension
- RecordInfo

Implementation

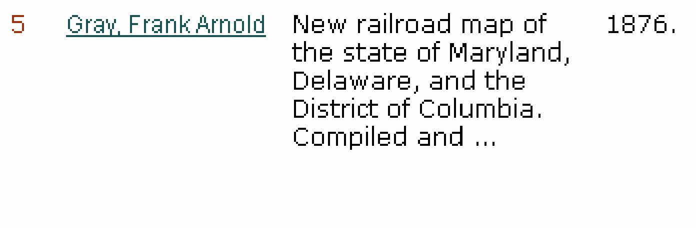
- Source records in MARC (ISO 2709)
- Indexed to retrieve individual record
- 3 MySQL tables
  - Sets, items, set-item links
  - Items: record id, date last updated
- Dynamic generation of XML
  - ISO 2709 => MARC21slim (perl)
  - ISO 2709 => DC (perl)
  - MARC21slim => MODS (XSLT- Saxon)



Record 1 of 1

Title	**New railroad map of the state of Maryland**, Delaware, and the District of Columbia. Compiled and drawn by Frank Arnold Gray.
Author/Creator	Gray, Frank Arnold.
Publisher	Philadelphia
Year	1876
Resource Type	image
Resource Type	cartographic
Resource Type	map
Language	eng
Note	Shows drainage, canals, stations, cities and towns, counties, canals, roads completed, narrow gauge and proposed railroads with names of lines. Includes list of railroads.
Note	Description derived from published bibliography.
Note	LC Railroad maps, 230
Note	Scale 1:633,600.
Subject	Railroads--Middle Atlantic States--Maps.
URL	http://hdl.loc.gov/loc.gmd/g3791p.rr002300
Institution	Library of Congress American Memory Project

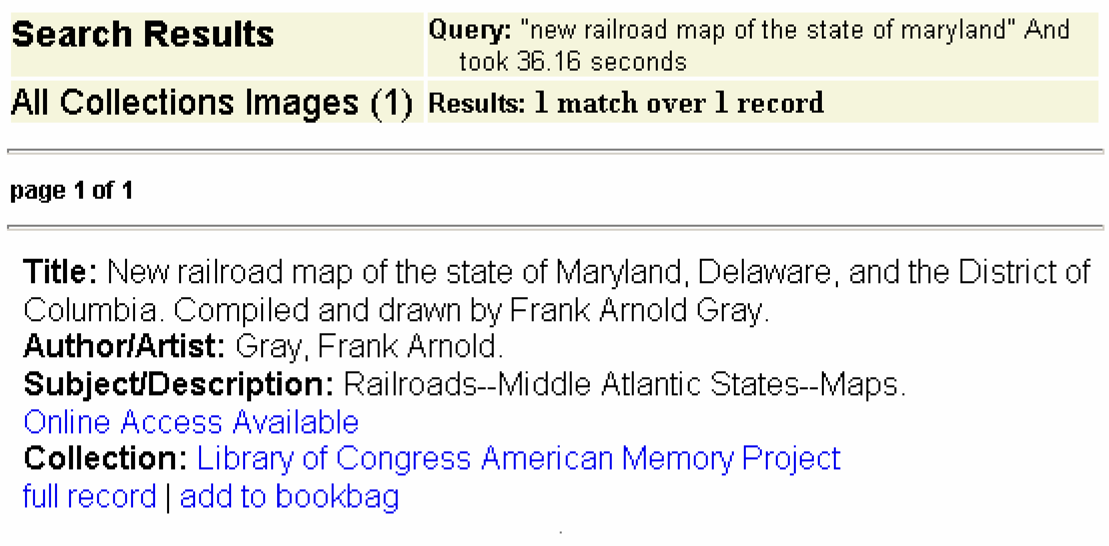
RLG’s CMI



5	Gray, Frank Arnold	New railroad map of the state of Maryland, Delaware, and the District of Columbia. Compiled and ...	1876.

Cultural Heritage Repository - UIUC

OAIster



Search Results	**Query:** "new railroad map of the state of maryland" And took 36.16 seconds
All Collections Images (1)	**Results: 1 match over 1 record**

page 1 of 1

Title: New railroad map of the state of Maryland, Delaware, and the District of Columbia. Compiled and drawn by Frank Arnold Gray.
Author/Artist: Gray, Frank Arnold.
Subject/Description: Railroads--Middle Atlantic States--Maps.
Online Access Available
Collection: Library of Congress American Memory Project
full record | add to bookbag

Thoughts on moving forward

- Experience
  - With harvesting metadata from LC/Ameritech awardees
  - With preparing metadata for harvesting
  - With considering mappings from MARC to DC and to MODS
- Leads to questions for harvesters

Will harvesters use richer schemas?

- RLG uses LC’s MARC records
- Will MODS address issues harvesters have been facing?
- MODS & some American Memory issues
  - Roles for creator/contributors
  - Broad types of resource and finer genre terms
  - Location (e.g. of original)
  - Related items
- Questions
  - Dates
  - Places

Beyond agreements on syntax to agreements on metadata content

- Expensive – bottleneck
- Some is essential
- How best to allocate scarce resources?
- What makes a good Dublin Core record for cross-domain discovery?
- Or should we focus on richer records?
- How to describe what is essential for certain categories of content?

What (descriptive) metadata does for me

- Let’s me
  - find stuff
  - understand what I am looking at
  - assess usefulness of resource for my task
- Supports interface functionality
  - Limiting/refining searches
  - Organize/manipulate search results
  - Browsing lists of terms
  - Timeline and map views
  - Links to related items or more information

How people search for stuff

- Need to understand how to provide metadata cost-effectively to support
  - A couple of terms in a search box
  - Refine (search within, limit, boolean)
  - Visual browsing (pattern-matching)
  - Identify authoritative source or relevant “collection”
  - Traverse hierarchies
  - Follow links to related items

Questions for harvesters

- What do you want most from us?
  - (for historical and non-textual materials)
  - Single date for sorting/limiting
  - Resource type for limiting/grouping
  - Uniform use of structured place names
  - Normalization of personal names
- Free text or terms for topical access?
  - How best to use interest of domain experts
- Value of controlled vocabularies
  - Precise, but not necessarily what users use
  - How close to perfection does application have to be?

Digression on names

- Virtual International Authority File
  - Early stages
  - LC, OCLC, Deutsche Bibliothek
  - Starting with personal names
  - Initially for library use
  - Will use OAI-PMH for updating
- Persistent identifiers for authority records
- Inconsistency between
  - Personal name authority record
  - Name as subject
  - Had you noticed, does it matter, should LC fix?

More questions for harvesters

- How little is too little?
  - American Memory uses text for topical access
  - Letter from A to B on date X
- How much is too much?
  - Should we include lyrics or short transcriptions (where available)
- Or do you just want more stuff?

Who does what and when?

- Tools for enhancing metadata
  - Recognizing and normalizing names
  - Gazetteers for linking placename to coordinates
  - Biographies and other supporting reference tools
  - Ontologies to map/scope/relate common words
- Who and when?
  - Data provider -- fix metadata
  - Service provider -- fix metadata or proprietary interface
  - Community – metadata manipulation tools or dynamic services

Business model

- Costs, benefits, and motivations
- Big issues
  - For common good
  - For competitive advantage
- Practical, incremental steps
  - Based on what we can now observe
  - How to structure conversations?
  - What might DLF do?