Increasing access to OA material through metadata aggregation

Mark Jordan
Simon Fraser University
SLAIS Issues in Scholarly Communications and Publishing
2008-04-02
We will discuss…

- Overview of metadata aggregation
- AlouetteCanada Portal
- The CARL Harvester
- Challenges in aggregating metadata
Overview of aggregation

- Bringing together of metadata from disparate sources to provide services
  - Searching
  - Clustering
  - Supplementation
  - Etc.
- Why aggregate when Google crawls it all?
OA Material

- Institutional Repositories (IRs)
- OA journals, proceedings, and books
- Local digital collections
Models

- Pull
  - Aggregation retrieves metadata from each source
- Push
  - Each source supplies its metadata to aggregation
Push: Submitting Metadata
Alouette Portal

http://alouette.ourontario.ca/
West Beyond the West

The West Beyond the West
British Columbia’s history, heritage, and culture

West Beyond the West is...

Digital images, text, audio and video materials about British Columbia history, heritage and culture. Search the collections of libraries, archives, museums, historical societies, heritage and community groups, government agencies, and private collections in British Columbia and across Canada... via a single search portal. Search results link you to the digital content on a contributing partner’s web site.

Note about the terms used in historical resources
Historical content is presented in the context in which it was originally created. Some materials may contain outdated language, terms and stereotypes that may be offensive and/or no longer in use.

http://westbeyondbowell.ca/
Workflow for Metadata Processing

- Source institution provides metadata
  - Relational database
  - MARC
  - Delimited
  - XML
- Alouette staff apply transformations, filters, etc.
- Alouette staff load processed metadata into Portal
Benefits of Pull Aggregation

- More consistent aggregated metadata
- Easier to supplement metadata
- Lower technical barrier to participation for contributors
Pull: Automated Harvesting
CARL Harvester

- “Canadian Association of Research Libraries / Association des bibliothèques de recherche du Canada's Institutional Repository Metadata Harvester”
- http://carl-abrc-oai.lib.sfu.ca/
- Launched June 2004
- Primarily a search engine for the harvested metadata
OAI-PMH Model

Data providers expose metadata

Service providers harvest metadata and do something useful with it
Benefits of Pull Aggregation

- Easy to automate
- Low barrier to participate (if technology present)
- More “standardized” than push
Challenges of Aggregating Metadata

- Inconsistent metadata
- Local vs. group practice
- Sustainability
- Cost vs. benefits
Inconsistency 1: Date

- 1998
- 1998-03
- 1998-03-14
- 1998-03-14 00:00:00.0
- 1998-03-14T14:49:04Z
- Very few invalid dates
Inconsistency 2: Type

- Electronic Thesis or Dissertation
- Thesis
- text

- Article
- Journal (On-line/Unpaginated)
- Journal (Paginated)
- Learned or Scientific Journal's article (on-line or printed)
- Preprint
Inconsistency 3: Description

- Types of values
  - Abstracts
  - Conference names/places/dates
  - Place names
  - Research network, project names/funders
  - “no abstract”
  - “none”
Metadata Application Profiles

- A set of metadata elements, policies, and guidelines defined for a particular application or implementation
- Defines best practices appropriate to the application
- Examples
  - ePrints UK “Using Simple Dublin Core to Describe Eprints”
  - “ARROW Discovery Service Harvesting Guide”
## Element: Type

**Definition:** The genre of the work.

**Obligation:** Mandatory

**Recommended Encoding:** None

### Element Guidelines:
- Repeatable.
- Prefer document types (article, thesis, etc.).
- Document formats (image, video, etc.) should be coded in the "Format" element.
- Must be one of the list of recognized types or variants for retrieval from the CARL Harvester.

**Types:**
- animation
- article (journal)
- book / book chapter
- dataset

**Examples:**
- learning object
- peer reviewed
- preprint
- presentation
- technical report
- thesis / dissertation
- working paper

[See values under Element Guidelines]
Realistic Goals

- Such a profile would
  - Be voluntary, not imposed
  - Emphasize easily achievable goals
  - Be flexible enough for the distributed creation of metadata
  - Use existing practices and standards as much as possible
Low Hanging Fruit

- Include rights
- Include publisher
- Include language
- Standardize use of date
  - Not format, but meaning
More Low Hanging Fruit

- Standardize use of identifier
  - Minimally, supply a URL to the resource/record
  - Additional local identifiers welcome
- Use DCMI Type Vocabulary
  - “provides a general, cross-domain list of approved terms that may be used as values for the Resource Type element to identify the genre of a resource”
  - Supplement with agreed-upon list of more specific genres
Fruit a Bit Higher Up

- Require OAI validation of providers
  - Software
  - XML encoding
- Identify minimal required elements, recommended elements
- Develop a metadata format specific to Canadian scholarly information
  - Bilingual elements, with language attribute
  - Coverage element
  - Controlled vocabularies
Discussion