Lessons in Cross-Repository Interoperability learned from the aDORe effort

Herbert Van de Sompel & Jeroen Bekaert
Research Library
Los Alamos National Laboratory, USA
The repository model

"Pattern Recognition: The 2003 OCLC Environmental Scan"
http://www.oclc.org/membership/escan/toc.htm
Value chains starting in repositories

• New knowledge is really being created when allowing for non-anticipated use of stuff.

• These repositories are not about creating services for local users (only)
• These repositories are not about creating a service (user interface) for all users

• These repositories are about facilitating the use of materials in many contexts
• These repositories are the starting point of value chains
Value chains emerging from RSS feeds

http://www.technorati.com
Value chains starting in repositories
Rethinking Scholarly Communication
Building the System that Scholars Deserve

Herbert Van de Sompel
Los Alamos National Laboratory, Research Library
<herbertv@lanl.gov>

Sandy Payette
Cornell University, Computing and Information Science
<payette@cs.cornell.edu>

John Erickson
Hewlett-Packard Laboratories, Digital Media Systems Lab
<john.erickson@hp.com>

Carl Lagoze
Cornell University, Computing and Information Science
<lagoze@cs.cornell.edu>

http://dx.doi.org/10.1045/september2004-vandesompel
Value chains starting in repositories

To allow value chains to emerge on the basis of materials in repositories, those repositories need clear/clean machine interfaces that allow downstream applications to consume materials, aggregate them, build services, ...

⇒ Disconnection of repository content and service: allows for creation of both local and remote services
⇒ On-Web: Protocol-oriented interfaces
⇒ These value chains are about the real stuff not (only) about metadata
Credits

The reported material is based on the following work:

- The LANL aDORe repository effort
- The upcoming PhD thesis by Jeroen Bekaert (Advisor Herbert Van de Sompel) regarding protocol-based interfaces for Open Archival Information Systems (OAIS)
- The NSF-funded Pathways project in collaboration with the Information Science group at Cornell University (Carl Lagoze, Sandy Payette, Simeon Warner)
Outline

aDORe
A few words about the aDORe architecture

A Federation of Repositories
A new level of cross-repository interoperability

Pathways InterDisseminator
A context-sensitive service overlay for a federation of repositories
aDORe effort

aDORe is 2 things:

- Standards-based, modular repository architecture
  - Distributed architecture
  - Protocol-based interactions between modules
  - Applicable to create interoperable federations of heterogeneous repositories
- Actual implementation of the architecture at LANL for local storage of digital assets (currently in its 2\textsuperscript{nd} version)

aDORe is not a product

- Components of aDORe software, usable in other environments, will be released
aDORe effort

- Standards
- Distributed architecture
- Protocol-based communication

Insights in Cross-Repository Interoperability
aDORe effort

- Standards used in aDORe include:
  - XML,
  - XML Schema,
  - MPEG-21 Digital Item Declaration,
  - MPEG-21 Digital Item Identification,
  - W3C XML Signatures,
  - OAI-PMH,
  - NISO OpenURL Framework for Context-Sensitive Services,
  - Internet Archive ARC file format,
  - OAIS concepts
## OAI-PMH Federator & OpenURL Resolver

<table>
<thead>
<tr>
<th></th>
<th>Interface standard</th>
<th>Identifier</th>
<th>OAIS Access Type</th>
<th># items in response</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OAI-PMH Federator</strong></td>
<td>OAI-PMH</td>
<td>Package Identifier</td>
<td>OAIS DIP</td>
<td>1 or more</td>
</tr>
<tr>
<td><strong>OpenURL Resolver</strong></td>
<td>NISO OpenURL</td>
<td>Content Identifier, Package Identifier (with XML ID fragment)</td>
<td>OAIS DIP &amp; Result Set</td>
<td>1</td>
</tr>
</tbody>
</table>
CORDRA (Content Object Repository Discovery and Registration/Resolution Architecture):
An open, standards-based model for how to design and implement software systems for the purposes of discovery, sharing and reuse of learning content through the establishment of interoperable federations of learning content repositories.

CORDRA is designed to be an enabling model to bridge the worlds of learning content management and delivery, and content repositories and digital libraries. CORDRA aims to identify and specify (not develop) appropriate technologies and existing interoperability standards that can be combined into a reference model used to enable a learning content infrastructure.

CORDRA is:
- A formal model that can be used to design federations of repositories (the CORDRA reference model);
- A collection of operational systems built from the CORDRA model, including:
  - a prototype implementation of a repository federation; and
  - an operational federation of federations used to combined different CORDRA federations.
- The activities and projects surrounding the definition of the CORDRA model and creation of the operational systems.

The CORDRA activities are being coordinated by the Advanced Distributed Learning Initiative, (ADL) the Corporation for National Research Initiatives, (CNRI) and the Learning Systems Architecture Lab (LSAL).

http://cordra.net
Outline

aDORe
A few words about the aDORe architecture

A Federation of Repositories
A new level of cross-repository interoperability

Pathways InterDisseminator
A context-sensitive service overlay for a federation of repositories
The interoperable repository model

I will try to show that:
• a significantly higher level of cross-repository interoperability can be achieved with relatively modest means
• those means are largely available and agreed upon in our community

I will introduce:
• Repository-level requirements
• Infrastructure-level requirements
Part 1 : Requirements for a repository in a federation
Repositories & Units of Communication

• Data-oriented research
  => not only textual materials, but also datasets, software, simulations, dynamic knowledge presentations, ...

• Research results represented by variety of digital media
  ⇒ these media must receive status similar to that of text in current system

• Materials in various stages of certification:
  ⇒ units of communication not only 'papers' but also preprints, raw datasets, prototype simulations, ...

• Facilitate collaboration
  ⇒ re-use of units of communications
Repositories & Units of Communication

• Handling this requires:
  o a compound object view of a unit of communication
  o stop thinking in terms of metadata versus content

• Compound object:
  o Has a persistent identifier
  o Contain materials and metadata about those materials
  o Can contain other compound objects
Compound objects

URIs:
• minted by different repositories
• from different namespaces
• not (necessarily) locators
XML-based representation of compound objects

compound object

MPEG-21 DIDL
METS
IMS/CP
RDF

XML-based representation
Repository Interop Interface 1: OAI-PMH & CO

- machine consumption
- batches of compound objects
- OAI-PMH datestamp ~ new version of object
OAI-PMH interface to OAIS (Jeroen Bekaert)

```
baseUrl(OAIPMH_CIID)?
verb=ListMetadataFormats

list of DIP formats (ListMetadataFormats response)

baseUrl(OAIPMH_CIID)?
verb=ListRecords&
metadataPrefix=info:pathways/svc/dip.rdf

list of DIPs (derived from most recent AIPs)
```
Repository Interop Interface 1: OAI-PMH & CO

- include provenance ~ version of compound object
Repository Interop Interface 2: OpenURL & CO

- machine (& human) consumption
- single object dissemination ~ identifier of compound object
Repository Interop Interface 2: OpenURL & CO

- **ServiceType** = Request a representation of the DO expressed using a compound object format
  - Example:
    - `svc_id = info:pathways/svc/dip.didl` (request MPEG-21 DIDL representation)
    - `svc_id = info:pathways/svc/dip.mets` (request METS representation)
    - `svc_id = info:pathways/svc/dip.rdf` (request RDF representation – see later)

- Other Entities could be added to Interface #2 (think Requester)
Repository Interop Interface 2: OpenURL & CO

- independent of nature of identifiers
- ‘resolution’ independent of scheme-specific mechanisms
- conceptual interface is **persistent** over time
  - KEV & HTTP (REST)
  - XML & SOAP
  - ...

```
repository_n
```

```
OpenURL baseURL_o
```
OpenURL interface to OAIS (Jeroen Bekaert)

Herbert Van de Sompel & Jeroen Bekaert
Research Library, Los Alamos National Laboratory
CNI Task Force Meeting, December 5-6 2005, Phoenix, Arizona
Part 2: Requirements for an infrastructure supporting a federation of repositories
Repository Registry: **Who** is part of the Federation?

Per Repository:

- Repository identifier
- baseURL of OAI-PMH interface
- baseURL of OpenURL interface
- whichever kind of information that helps downstream applications understand about the nature of the repository
Object Registry: What is part of the Federation?

Per compound object:

- Object identifier
- Object datetime ~ OAI-PMH datestamp
- OAI-PMH identifier
- Repository identifier

of the object itself, and of its contained objects
OAI-PMH & OpenURL access to objects in federation

- List of existing copies
- Per copy:
  - OAI-PMH access info
  - OpenURL access info
Part 3 : Summary of requirements
## Summary of requirements

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Repository</th>
<th>Infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound Object model support</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>XML-based representations support</td>
<td>X</td>
<td>?</td>
</tr>
<tr>
<td>OAI-PMH CO support</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>OpenURL CO support</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Repository Registry</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Object Registry</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Summary of requirements

Many variations on the design possible, yet most of this can be achieved with:

- Off-the-shelf tools
  - OAI-PMH tools
  - Handle system, SRU/W tools
  - OpenURL tools
  - Tools to generate XML-based representations of objects
- Surprisingly little effort
- A feasible amount of coordination/specification
- Some shared infrastructure
The Good News™

• Microsoft & Mellon Foundation interested in taking interoperability across repositories to a new level
• Meeting is being planned to consult with the major stakeholders on identifying, specifying, and implementing concrete ways forward
• Tony Hey has (t)asked me to call that meeting
Outline

aDORe
A few words about the aDORe architecture

A Federation of Repositories
A new level of cross-repository interoperability

Pathways InterDisseminator
A context-sensitive service overlay for a federation of repositories
Pathways InterDisseminator Service Overlay

- Pathways InterDisseminator: Dynamic Service-Oriented Overlay upon the federated architecture
- Assumes the existence of:
  - OpenURL Interface to all repositories in the federation
  - Object Registry (given an identifier, at which OpenURL interface is the object available?)
  - Availability of an RDF-based representation of DO compliant with a Pathways OWL core ontology
- Is itself exposed as a different OpenURL Resolver
Pathways InterDisseminator: core ontology
Pathways InterDisseminator Service Overlay

• Part of the dissemination OpenURL Application is an engine that dynamically decides upon services for a given object from a repository (in a federation).
  o It grabs the (RDF) representation of the DO from its origin repository
  o It introspects on the properties expressed in that (RDF) representation
  o It compares these properties with its knowledge database
  o It returns a list of possible services/disseminations

• There can be many of these engines in a federation. The result is the ability to provide context-sensitive disseminations of DOs in (a federation of) repositories.
Pathways InterDisseminator Service Overlay

- There can be many of these engines in a federation. The result is the ability to provide context-sensitive disseminations of DOs in (a federation of) repositories.
# Pathways InterDisseminator Demo

<table>
<thead>
<tr>
<th>aDORe Digital Object in Demo</th>
<th>Type</th>
<th>MIME</th>
<th>identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital Object</td>
<td>scholarly paper</td>
<td>N/A</td>
<td>DOI</td>
</tr>
<tr>
<td>Constituent Datastream 1</td>
<td>metadata record</td>
<td>application/xml (MARCXML)</td>
<td>aDORe datastream id (info URI)</td>
</tr>
<tr>
<td>Constituent Datastream 2</td>
<td>metadata record</td>
<td>application/xml (original metadata)</td>
<td>aDORe datastream id (info URI)</td>
</tr>
<tr>
<td>Constituent Datastream 3</td>
<td>fulltext file</td>
<td>application/pdf</td>
<td>aDORe datastream id (info URI)</td>
</tr>
</tbody>
</table>
Demo

- Install TSCC coded (http://www.techsmith.com)
- Launch movie Pathways_InterDisseminator.avi in same path as this presentation
Comments, Flames, Questions