CNI Task Force Meeting
December 7th 2004, Portland, OR

An Update from the OAI
<http://www.openarchives.org>
Herbert Van de Sompel <herbertv@lanl.gov>
Carl Lagoze <lagoze@cs.cornell.edu>
Michael Nelson <mln@cs.odu.edu>
Simeon Warner <simeon@cs.cornell.edu>

Outline

(1) OAI-PMH refresh
(2) OAI-rights effort
(3) OAI-PMH for Resource Harvesting
(4) mod_oai

OAI-PMH

resource
repository
OAI-PMH request
OAI-PMH response
provides services using harvested metadata
exposes metadata pertaining to resources

OAI-PMH data model

resource
item
OAI-PMH identifier
OAI-PMH sets
entry point to all records pertaining to the resource
OAI-PMH identifier metadataPrefix datestamp
Dublin Core metadata
MARCXML metadata
records metadata pertaining to the resource

OAI-PMH

repository
OAI-PMH request
OAI-PMH records
provides services using harvested metadata
exposes metadata pertaining to resources

Outline

(1) OAI-PMH refresh
(2) OAI-rights effort
(3) OAI-PMH for Resource Harvesting
(4) mod_oai
December 7, 2004 – CNI Task Force Meeting, Portland, OR

An Update from the OAI

Creative Commons as example language

- Felt we should pick one language as an example
  - RoMEO aligned with Create Commons (CC)
  - CC fits well with interests of many of the original OAI participants (e.g. arXiv considering use of CC)
  - CC is a “good thing” to promote
- Picking CC turned out to be a little complicated because of RDF formulation. Schema version may be forthcoming
- CC really is just an example, can use any XML rights expression language (REL)
  - Will likely add appendices with other example languages later
  - Ongoing collaboration with the ODRL community to define ODRL-OAI guidelines document (again, metadata first)

Who?

The OAI-rights group:
- Caroline Arms (Library of Congress), Chris Barlas (Rightscom), Tim Cole (University of Illinois at Urbana-Champaign), Mark Doyle (American Physical Society), Henk Ellerman (Erasmus Electronic Publishing Initiative), John Erickson (Hewlett Packard & DSpace), Elizabeth Gaid (Loughborough University & RoMEO), Brian Green (EDIEUR), Chris Gutteridge (Southampton University & eprints.org), Carl Lagoze (Cornell University & OAI), Mike Linksvayer (Creative Commons), Uwe Müller (Humboldt University), Michael Nelson (Old Dominion University & OAI), John Ober (California Digital Library), Charles Oppenheim (Loughborough University & RoMEO), Sandy Payette (Cornell University), Andy Powell (UKOLN, University of Bath), Steve Proberts (Loughborough University & RoMEO), Herbert Van de Sompel (Los Alamos National Laboratory & OAI), and Simeon Warner (Cornell University, arXiv & OAI)

Why OAI-rights?

OAI has matured beyond e-prints and is used to convey metadata about resources for which the ability to express rights is a factor limiting dissemination

- Encourage participation by allowing assertion of rights and restrictions

Even in the open access world it may be important to express permissions

⇒ Work inspired by the RoMEO project (Oppenheim, Probets, Gadd, 2002-2003)

How?

“The usual OAI way”:
- Assemble group of knowledgeable and interested parties (the OAI-rights group)
- Distribute first-stab white paper
- Discuss via conference call, scope work
- Email and conference call discussions, develop alpha specification (Jun 2004), revise
- Release beta specification (Nov 2004)
- Release specification (end 2004)

http://www.openarchives.org/OAI/2.0/guidelines-rights.htm

Scope

- No new rights expression language
- Don’t restrict to specific language(s)
- Don’t get bogged down in rights vs permissions vs enforcement, OAI-PMH is about transferring XML data
- Rights about metadata a separate problem from rights about resources
  - Tackle rights about metadata first
  - Postpone work on rights about resources (note overlap with resource harvesting work)
- Issues with rights expressions for aggregations of items (OAI sets; whole repositories)

OAI-PMH data model

Data model elements:
- repository
- item - all metadata about a resource, has identifier
- record - metadata in a particular format, plus header and information about the metadata
- set - optional, overlapping, hierarchical groupings of items
- resource outside scope of OAI-PMH
Different aggregation levels

- **record** - Rights about an individual record
- **repository** - Manifests of rights about all records (all metadata formats from each item) in a repository
- **set** - Manifests of rights about all records (all metadata formats from each item) in a set

Record level expression is authoritative. Other levels are optional.

---

Record level rights expressions

- Actual rights expression may be in-line (must be valid XML) or by-reference (at given URL, XML recommended)
- In-line method recommended for truly static rights expressions. Avoids possible ambiguity with delayed de-referencing

```xml
<record>
  <header> id, datestamp, sets </header>
  <metadata> metadata: DC, MARCXML, ... </metadata>
  <about> <rights> </rights> </about>
  <about> provenance, branding etc. </about>
</record>
```

---

Set and repository level expressions

- These are optional and non-authoritative
- W3C XML schema defines `<rightsManifest>` package which contains a sequence of `<rights>` elements (as used at the record level)
- `<rightsManifest>` included in
  - For repository level: `<description>` in Identify
  - For set level: `<setDescription>` in ListSets response
- Useful when there is a small set of expressions within the particular aggregation
- Should be accurate and complete but this is not enforced by specification

---

Rights about resources

- Can already be done: use an appropriate metadata format as one of the parallel metadata formats from an item. But:
  - Too much choice: need profile
  - Issues with identification of resources
- Overlap with resource harvesting work

---

Outline

1. OAI-PMH refresh
2. OAI-rights effort
3. OAI-PMH for Resource Harvesting
4. mod_oai

---

http://www.openarchives.org/OAI2.0/guidelines-rights.htm
Resource Harvesting: Use cases

• Discovery: use content itself in the creation of services
  o search engines that make full-text searchable
  o citation indexing systems that extract references from the full-text content
  o browsing interfaces that include thumbnail versions of high-quality images from cultural heritage collections
• Preservation:
  o periodically transfer digital content from a data repository to one or more trusted digital repositories
  o trusted digital repositories need a mechanism to automatically synchronize with the originating data repository

Existing OAI-PMH based approaches

Typical scenario:

1. An OAI-PMH harvester harvests Dublin Core records from the OAI-PMH repository.
2. The harvester analyzes each Dublin Core record, extracting dc.identifier information in order to determine the network location of the described resource.
3. A separate process, out-of-band from the OAI-PMH, collects the described resource from its network location.

Existing OAI-PMH based approaches : Issue 1

• Locating the resource based on information provided in dc.identifier
  - dc.identifier used to convey a variety of identifier: (simultaneously) URL, DOI, bibliographic citation, … Not expressive enough to distinguish between identifier, locator.
  - Several dereferencing attempts required
    o URI provided in dc.identifier is commonly that of a bibliographic “splash page”
    o How to know it is a bibliographic “splash page”, not the resource?
    o If it is a bibliographic “splash page”, where is the resource?

Existing OAI-PMH based approaches : Issue 2

• Using the OAI-PMH datestamp of the Dublin Core record to trigger incremental harvesting:
  - Datestamp of DC record does not necessarily change when resource changes
    - DC record datestamp: no change
    - DC record datestamp: change

| no resource update | OK
| resource update | missed resource update |
| no metadata update | OK
| metadata update | unnecessary resource download |

Existing OAI-PMH based approaches : Conventions

• Conventions address issue 1; Issue 2 can not really be addressed.
• First dc.identifier is locator of the resource
  - what if the resource is not digital?
• Use of dc.format and/or dc.relation to convey locator
Existing OAI-PMH based approaches : Conventions

- Conventions
  - metadata formats that were specifically created for representation of digital objects:
    - Complex Object Formats as OAI-PMH metadata formats
      - MPEG-21 DIDL, METS, ...

Existing OAI-PMH based approaches : Other attempts

- dc.identifier leads to splash page & splash page contains special purpose XHTTP link to resource(s)
  - What if there is no splash page?
  - How does a harvester know he is in this situation?
- OA-X: protocol extension
  - OK in local context
  - Strategic problem to generalize
  - How to consolidate with OAI-PMH data model
- Qualified Dublin Core
  - Could bring expressiveness to distinguish between locator & identifier
  - But what with datestamp issue?

Proposed OAI-PMH based approach

- Use metadata formats that were specifically created for representation of digital objects:
  - Complex Object Formats as OAI-PMH metadata formats
    - MPEG-21 DIDL, METS, ...

OAI-PMH data model

- OAI-PMH identifier
  - entry point to all records pertaining to the resource
- Dublin Core metadata
- MARCXML metadata
- MPEG-21 DIDL
- METS

metadata pertaining to the resource

item

resource

records

simple

more expressive

highly expressive

highly expressive
Complex Object Formats: characteristics

- Representation of a digital object by means of a wrapper XML document
- Represented resource can be:
  - simple digital object (consisting of a single datastream)
  - compound digital object (consisting of multiple datastreams)
- Unambiguous approach to convey identifiers of the digital object and its constituent datastreams
- Include datastream:
  - By-Value: embedding of base64-encoded datastream
  - By-Reference: embedding network location of the datastream
  - not mutually exclusive; equivalent
- Include a variety of secondary information
  - By-Value
  - By-Reference
  - Descriptive metadata, rights information, technical metadata, …

Complex Object Formats & OAI-PMH

- Resource represented via XML wrapper => OAI-PMH
- Uniform solution for simple & compound objects
- Unambiguous expression of locator of datastream
- Disambiguation between locators & identifiers
- OAI-PMH datestamp changes whenever the resource (datastreams, secondary information) changes
- OAI-PMH semantics apply: "about" containers, set membership

OAI-PMH based approach using Complex Object Format

Typical scenario:

1. An OAI-PMH harvester checks for support of a complex object format using the ListMetadataFormats verb
2. The harvester harvests the complex object metadata. Semantics of the OAI-PMH datestamp guarantee that new and modified resources are detected.
3. A parser at the end of the harvesting application analyzes each harvested complex object record:
   - The parser extracts the bitstreams that were delivered By-Value.
   - The parser extracts the unambiguous references to the network location of bitstreams delivered By-Reference.
4. A separate process, out-of-band from the OAI-PMH, collects the bitstreams delivered By-Reference from the extracted network locations.

Complex Object Formats & OAI-PMH: existing implementations

- LANL Repository
  - Local storage of Terrabytes of scholarly assets
  - Assets stored as MPEG-21 DIDL documents
  - DIDL documents made accessible to downstream applications via the OAI-PMH
- Mirroring of American Physical Society collection at LANL
  - Maps APS document model to MPEG-21 DIDL Transfer Profile
  - Exposes MPEG-21 DIDL documents through OAI-PMH infrastructure
  - Includes digests/signatures
- DSpace & Fedora plug-ins
  - Maps DSpace/Fedora document model to MPEG-21 DIDL Transfer Profile
  - Exposes MPEG-21 DIDL documents through OAI-PMH infrastructure
  - mod_oai
Complex Object Formats & OAI-PMH: issues

- Which Complex Object Format(s)
- How to Profile Complex Object Format(s) for OAI-PMH Harvesting
- Large records
- Making resources re-harvestable
- Because the resource is represented as `<metadata>`, can rights pertaining to the resource be expressed according to the "rights for metadata" OAI-rights guideline?
- Tools:
  - Software library to write compliant complex objects
  - Integration of this library with repository systems (Fedora, DSpace, eprints.org, …)

Outline

1. OAI-PMH refresh
2. OAI-rights effort
3. OAI-PMH for Resource Harvesting
4. mod_oai

Web crawlers

A more efficient way

mod_oai approach

- Goal: integrate OAI-PMH functionality into the web server itself...
- mod_oai: an Apache 2.0 module to automatically answer OAI-PMH requests for an http server
  - written in C
  - respects values in .htaccess, httpd.conf
- Result: web harvesting with OAI-PMH semantics (e.g., from, until, sets)
  - http://www.foo.edu/modoai?
    verb=listIdentifiers &
    metaDataPrefix=oai_dc &
    from=2004-09-15 &
    set=mime:video:mpeg

mod_oai approach

- Install on an Apache 2.0 server
  - compile & edit httpd.conf

http://www.foo.edu/

now has an OAI-PMH baseURL of:

http://www.foo.edu/modoai

Launch OAI effort
OAI proposal to Library of Congress NDIIP submitted
OAI-PMH data model

mod_oai : OAI-PMH concepts

<table>
<thead>
<tr>
<th>concept</th>
<th>mod_oai implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>OAI-PMH Identifier</td>
<td>URL of resource</td>
</tr>
<tr>
<td>set</td>
<td>MIME type of resource</td>
</tr>
<tr>
<td>datestamp</td>
<td>change time of resource</td>
</tr>
<tr>
<td>deleted records</td>
<td>“no” deleted records</td>
</tr>
</tbody>
</table>

mod_oai use cases

- Regular Web Crawling
  - use ListIdentifiers to discover URLs
  - add new URLs to the list of URLs to be crawled
- Harvesting Resources with OAI-PMH
  - use ListRecords to extract the entire resource as an MPEG-21 DIDL AIP
Regular Web Crawling: ListIdentifiers

- issues a ListIdentifiers,
- finds URLs of updated resources
- does HTTP GETs updates only
- can get URLs of resources with specified MIME types

OAI-PMH Resource Harvesting

- issues a ListRecords,
- gets updates as MPEG-21 DIDL documents (HTTP headers, resource By Value or By Reference)
- can get resources with specified MIME types

Datestamps and Etags

is:
- a simple way to more efficiently harvest web pages
- a possible impact on robots.txt
- fully OAI-PMH compliant
  - works with existing harvesters
- Funded by the Andrew W Mellon Foundation

-is not:
- yet suitable for dynamic files
- a replacement for
  - DSpace
  - Fedora
  - eprint.org
  - other digital libraries / repositories / cms

Discussion: at 10:30, here

(*) OAI-rights effort
(*) OAI-PMH for Resource Harvesting
(*) mod_oai
(*) NSDL validation effort
(*) DLF OAI Best Practice
(*) …

Datestamp and Etag Example