OAI-PMH for Resource Harvesting

Herbert Van de Sompel
Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory

Michael Nelson
Computer Science Department
Old Dominion University
Tutorial Outline

- OAI-PMH for Resource Harvesting: problem statement and conceptual solution
- Accurate mirroring the collection of the American Physical Society using OAI-PMH-based Resource Harvesting
- mod_oai: An OAI-PMH-based model for Web Resource Harvesting
- OAIResource: A software tool for OAI-PMH-based Resource Harvesting
Resource Harvesting: Use cases

- **Discovery:** use content itself in the creation of services
  - search engines that make full-text searchable
  - citation indexing systems that extract references from the full-text content
  - browsing interfaces that include thumbnail versions of high-quality images from cultural heritage collections

- **Preservation:**
  - periodically transfer digital content from a data repository to one or more trusted digital repositories
  - trusted digital repositories need a mechanism to automatically synchronize with the originating data repository
Resource Harvesting: Use cases

- Discovery:
  - Institutional Repository & Digital Library Projects: UK JISC, DARE, DINI
  - Web search engines: competition for content (cf Google Scholar)
- Preservation:
  - Institutional Repository & Digital Library Projects: UK JISC, DARE, DINI
  - Library of Congress: NDIIP Archive Export/Ingest, e-deposit

**OAI-PMH is well-established. Can OAI-PMH be used for Resource Harvesting?**
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 derferencing attempts required
  - URI provided in dc.identifier is commonly that of a bibliographic “splash page”
    - How to know it is a bibliographic “splash page”, not the resource?
    - 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

<table>
<thead>
<tr>
<th>No resource update</th>
<th>DC record datestamp no change</th>
<th>metadata update</th>
<th>DC record datestamp change</th>
</tr>
</thead>
<tbody>
<tr>
<td>OK</td>
<td>missed resource update</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>unnecessary resource download</td>
<td></td>
<td>OK</td>
<td>OK</td>
</tr>
</tbody>
</table>
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
A Simple Parallel-Plate Resonator Technique for Microwave. Characterization of Thin Resistive Films

Vorobiev, A.

ING-INF/01 Elettronica

A parallel-plate resonator method is proposed for non-destructive characterisation of resistive films used in microwave integrated circuits. A slot made in one ...

Microwave engineering Europe

2002

Documento relativo ad una Conferenza o altro Evento

PeerReviewed

http://amsacta.cib.unibo.it/archive/00000014/

http://amsacta.cib.unibo.it/archive/00000014/01/GaAs_1_Vorobiev.pdf

OAI-PMH for Resource Harvesting Tutorial
OAI4, October 20th 2005, CERN, Geneva, Switzerland
Existing OAI-PMH based approaches: Conventions

...<dc:identifier>http://amsacta.cib.unibo.it/archive/00000014/</dc:identifier>
<dc:relation>
  http://amsacta.cib.unibo.it/archive/00000014/01/GaAs_1_Vorobiev.pdf
</dc:relation>
...

splash page

locator of resource
Existing OAI-PMH based approaches: Conventions

...<dc:identifier>http://amsacta.cib.unibo.it/archive/00000014/</dc:identifier>
<dc:relation>
  http://resolver.unibo.it/00000014/
</dc:relation>
<dc:relation>
  http://amsacta.cib.unibo.it/archive/00000014/01/GaAs_1_Vorobiev.pdf
</dc:relation>
...

splash page

locator of resource

splash page

OAI-PMH for Resource Harvesting Tutorial
OAI4, October 20th 2005, CERN, Geneva, Switzerland
Existing OAI-PMH based approaches: Other attempts

- dc.identifier leads to splash page & splash page contains special purpose XHTML 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

metadata pertaining to the resource

Dublin Core metadata
MARCXML metadata
MPEG-21 DIDL
METS

simple
more expressive
highly expressive
highly expressive

resource
item
records
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, …
<didl:DIDL>
  <didl:Item>
    <didl:Descriptor><didl:Statement mimeType="text/xml; charset=UTF-8">
      <oai_dc:dc>
        <dc:title>A Simple Parallel-Plate Resonator Technique for Microwave. Characterization of Thin Resistive Films</dc:title>
        <dc:creator>Vorobiev, A.</dc:creator>
        <dc:identifier>http://amsacta.cib.unibo.it/archive/00000014/</dc:identifier>
        <dc:format>application/pdf</dc:format>
      </oai_dc:dc>
    </didl:Descriptor>
    <didl:Component>
      <didl:Resource mimeType="application/pdf"
        ref="http://amsacta.cib.unibo.it/archive/00000014/01/GaAs_1_Vorobiev.pdf"/>
    </didl:Component>
  </didl:Item>
</didl:DIDL>
Complex Object Formats & OAI-PMH

- Resource represented via XML wrapper => OAI-PMH
  `<metadata>`
- Uniform solution for simple & compound objects
- Unambiguous expression of locator of datastream
- Disambiguation between locators & identifiers
- OAI-PMH datetstamp 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”
- Compound objects with multiple datastreams. What if only 1 datastream gets updated?
- 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, ….)
  - Software to harvest resources based on OAI-PMH model
Readings