Open Archives Initiative
Object Re-Use & Exchange

Herbert Van de Sompel

Acknowledgments

Carl Lagoze, Michael L. Nelson

(1) Digital Library Research & Prototyping Team, Research Library, Los Alamos National Laboratory
herbertv@lanl.gov
http://public.lanl.gov/herbertv

(2) Information Science, Cornell University

(3) Computer Science, Old Dominion University

ORE is supported by the Andrew W. Mellon Foundation
with additional support of the National Science Foundation
General information about OAI-ORE
OAI Object Re-Use and Exchange

- OAI-ORE is a new interoperability effort conducted under the umbrella of the OAI
- Supported by the Andrew W. Mellon Foundation; additional support from the National Science Foundation; likely extra support from Microsoft

- International effort; October 2006 - September 2008:
  - Coordinators: Carl Lagoze & Herbert Van de Sompel
  - ORE Technical Committee: 13 international members
  - ORE Liaison Group: 8 international members
  - ORE Advisory Committee: 16 international members
  - Representing: scholarly publishers and aggregators, eScience, eHumanities, education, search engines, various repository systems, digital library efforts, related standardization efforts, etc.

- See http://www.openarchives.org/ore/
OAI is not just about metadata anymore

<table>
<thead>
<tr>
<th>OAI-PMH</th>
<th>OAI-ORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repository structure</td>
<td>Object structure</td>
</tr>
<tr>
<td>Repository centric</td>
<td>Web centric</td>
</tr>
<tr>
<td>Metadata centric</td>
<td>Resource centric</td>
</tr>
<tr>
<td>Metadata harvesting</td>
<td>Object re-use (obtain, harvest, register)</td>
</tr>
</tbody>
</table>

OAI-PMH and OAI-ORE are complementary:
- you can do one without the other
- you can do them together
Context of OAI-ORE Standards & Protocols
An Early Formulation of the Problem

- First noticed in how people would populate their Dublin Core records
  - people need the HTML splash page
  - crawlers need the PDF file
- Ad-hoc conventions and methods used to expose the repository’s knowledge about the structure of the object
- Next three slides taken from “Resource Harvesting Within the OAI-PMH Framework”
  - http://www.dlib.org//dlib/december04/vandesompel/12vandesompel.html
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/

pdf

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

splash page

locator of resource
Dublin Core Encoding Type 2

...<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
And more recently ...

“Are repositories successfully exposing the full-text of articles (the PDF file or whatever) to Google rather than (or as well as) the abstract page?”

“Are we consistent in the way we create hypertext links between research papers in repositories?”

(from Andy Powell’s eFoundations blog)
As the objects get more complex, things get worse

Rather than continue down that path, let’s back up and restart...
Compound Information Objects

Units of scholarly communication are compound information objects:

**Identified, bounded** aggregations of related information units that form a logical whole.

Components of compound object may vary according to:

- Semantic type: book, article, moving image, dataset, ...
- Media type: PDF, HTML, JPEG, MP3, ...
- Internal relationship: parts, views, ...
- External relationships
Scholarly Examples

http://citeseer.ist.psu.edu/lagoze01open.html


OAI Object Re-Use and Exchange
ELAG 2007, University of Barcelona, Barcelona, Spain, May 10th 2007
Herbert Van de Sompel
And more scholarly examples ...

- Scholarly publication with an article and supporting information including dataset, video, etc.
- Digitized book with multiple chapters, each chapter containing multiple scanned pages.
- Archaeological assemblies of images, maps, charts, and find lists.
- An ARTstor image object that is the aggregation of various renderings of the same source image.
- ...
But these things are not only scholarly ...
Access Repositories

Compound objects are made accessible by a variety of scholarly repositories:

- Institutional repositories
- Discipline-oriented repositories
- Publisher repositories
- Dataset repositories
- Cultural heritage repositories
- Learning object repositories
- Digitized book and manuscript collections
- Research-group and managed personal (ePortfolio) repositories
- ...
Access Repositories

Repositories expose compound objects in manners specific to the repository architecture:

- Interfaces (API & user-oriented)
- Identification schemes
- Representation of compound objects
- Publication of compound objects and components to the Web
Structure of Compound Objects is invisible when published to the Web
Systems that manage digital objects

- Institutional repositories
- Discipline-oriented repositories
- Publisher repositories
- Dataset repositories
- Cultural heritage repositories
- Learning object repositories
- Digitized book and manuscript collections
- Image repositories
- ...

Systems that leverage managed digital objects

- All repositories from left column
- Search engines
- Authoring tools
- Citation management tools
- Collaborative environments
- Social network applications
- Graph analysis tools
- Preservation services
- Workflow tools
- ...
OAI Object Re-Use and Exchange

• Develop, identify, and profile extensible standards and protocols to allow repositories, agents, and services to interoperate in the context of use and reuse of compound digital objects beyond the boundaries of the holding repositories.

• Aim for more effective and consistent ways:
  o to facilitate discovery of these objects,
  o to reference (link to) these objects (and parts thereof),
  o to obtain a variety of disseminations of these objects,
  o to aggregate and disaggregate these objects,
  o Enable processing by automated agents
Taking the Web perspective
Working with the web architecture

- **Whatever we do must be congruent with the web architecture**
  - Use existing capabilities where they are appropriate
  - Cleanly layer capabilities meeting the needs of our problem space
- Provide the infrastructure for web-based information systems that exploit/enhance and therefore overlay on the existing web.
W3C Web Architecture

URI

Resource

Represents

Identifies

Representation 1

Content Negotiation

Representation 2

Represents
W3C Web Architecture: more details

Aggregation:
- No standard way to describe finite set of resources and relationships

Resource:
- First-class object
- Linkable

Relationship:
- Usually untyped
- Link type ontologies not-standardized

Representation:
- Second-class object (identified only in context of resource)
- Not linkable
- Many representations/resource
Compound Object

astro-ph/0611775

Article in PDF
Article in PS
Splash page in HTML
Metadata in DC

Multiple Views, diverging in media-type, format, and content-type
More complexity ...

boundary, logical unit

astro-ph/0611775

Article in PDF

Article in PS

Splash page in HTML

Metadata in DC

hasPart

hasRelationshipTo

local, remote

lineage, version, citation, etc.
Let's publish it to the Web
Compound Object published to the Web

“Are repositories successfully exposing the full-text of articles (the PDF file or whatever) to Google rather than (or as well as) the abstract page?”

- **Discovery**: How does Google find all these resources that originate from the same digital object?
- **Boundary**: How does Google know these resources originate in the same digital object?
Compound Object published to the Web

“Are we consistent in the way we create hypertext links between research papers in repositories?”

- Citation: Which Resource to link to?
- Citation: How to reference the PDF version (and not the PS version)?
Thoughts about a possible approach
Observation 1
Components of a compound object must be published as resources in order to be reference-able.
Observation 2
The object “as such” (boundary, structure, relationships) is invisible to Web applications
Observation 2 bis
How about publishing a resource that makes a Resource Map available that formally expresses the boundaries of the object?
Observation 3
And now facilitate discovery of the Resource Map (and hence of the compound object) by Web applications.
Observation 3 bis
Through the Resource Map, the Web application sees the compound object
Observation 4
Resource Maps reveal compound objects in the Web graph
OAI Object Re-Use and Exchange
ELAG 2007, University of Barcelona, Barcelona, Spain, May 10th 2007
Herbert Van de Sompel
Resource Map available from ORE resource

- Expresses an **aggregation of resources and relationships** in a machine-readable manner.
- Describes a **graph**:
  - finite set of resources and relationships among the resources
  - relationships among resources that are members of the aggregation and & resources are external to the aggregation
- Can be used to express:
  - Our scholarly compound objects
  - Whichever aggregation of resources and relationships
- Having a standardized format for Resource Maps opens the door to “graph publishing” (cf. Semantic Web notion).
Use and Re-Use enabled by the ORE resource

- ORE resource has a URI: HTTP\_{ORE}
- HTTP\_{ORE} identifies a graph (cf. Semantic Web notion Named Graph)
- The Resource Map is available via HTTP GET on HTTP\_{ORE}
- HTTP\_{ORE} can become the key for object re-use: Obtain, Harvest, Register (cf. Web 2.0 mash-up)
- The \{Resource\} Map is not the Resource
  - Crawlers, agents will initially transact with the Resource Map or its identifier HTTP\_{ORE}
More About Resource Map Discovery

• A combination of two general approaches:
  o create new resources that allow batch harvest of Resource Maps
    - web crawling (cf. Sitemaps)
    - OAI-PMH harvesting with new metadataPrefix for Resource Map
    - RSS/ATOM feeds
  o instrument existing resources to “point” to these resources
    - http headers
    - html “microformats”
    - http content negotiation

• Selective discovery
  o you should never get a Resource Map unless you really asked for it; existing harvesters, crawlers will not break
  o Resource Maps are for machines, not humans
So, where does ORE stand?
OAI-ORE: Current Status

• Ongoing definition of the ORE framework
  o Reach joint problem statement
  o Issues regarding identification
  o Model for ORE resource
  o Publishing ORE resources to the Web
  o Discovering ORE resources

• Review of appropriate technologies for ORE Model and Resource Map
  o ATOM
  o DID/DIDL, IMS/CP, METS, Ramlet
  o RDF, RDF/XML
  o Dublin Core Abstract Model
  o ...
OAI-ORE : Current Status

• Explore demonstrators using these concepts in preparation of May 2007 ORE Technical Committee meeting

• Post May 2007 meeting:
  o Hopefully work towards alpha specs for ORE resource, Resource Map, discovery of ORE resource
  o Experimentation with alpha specs
OAI-ORE : Afterwards

• Look into core services Obtain, Harvest, Register, in terms of ORE resource and Resource Map.

• Note:
  o It is expected that the result of ORE will largely be an aggregation of (profiles of) existing standards/specifications, not a from-scratch specification (cf. OAI-PMH).
Questions

Further information
http://www.openarchives.org/ore/