Digital Library Interoperability via Metadata Harvesting

Applying the Open Archives Initiative Protocol

Carl Lagoze, Cornell University
lagoze@cs.cornell.edu
Edward A. Fox, Virginia Tech
fox@vt.edu

Acknowledgements

• People
  – Herbert Van de Sompel
  – Dan Greenstein
  – Clifford Lynch
  – Hussein Suleman
  – Members of the OAI community

• Funding Organizations
  – Digital Library Federation
  – Coalition for Networked Information
  – National Science Foundation
Agenda

- Goal: to produce communities of OAI implementers and supporters
- Process:
  - History and context of the OAI
  - Definitions and concepts of the technology
  - Protocol details
  - Working with the OAI community
    - Tools
    - Mailing lists
    - Projects
  - Future Plans

Digital Library Interoperability

A Short History of Interoperability

- Naming: URNs, Handles, DOIs
- Metadata: Dublin Core, IMS, MARC
- Search and Discovery: Z39.50, Harvest, Dienst, STARTS, SDLIP
- Object Models: Kahn/Wilensky, FEDORA, Buckets
- Encoding: SGML, HTML, XML, RDF

Interoperability Trade-offs
OAI's Location in a Broader Interoperability Fabric

Data Structuring (XML, XML Schema)

Data Semantics (Dublin Core, other metadata)

Exchange of Structured Information

Object Access

Yes, it’s about resource discovery over distributed collections

Author
Title
Abstract
Identifier
Beyond resource discovery to distributed custodianship

• Traditional portal (e.g., Yahoo!)
  – linkage with limited responsibility
• Hybrid Portal
  – Goal: assertion of (some semblance) of curatorial role over linked objects
  – Mechanism: sharing structured information (metadata) amongst distributed content providers

The Library should selectively adopt the portal model for targeted program areas. By creating links from the Library’s Web site, this approach would make available the ever-increasing body of research materials distributed across the Internet. The Library would be responsible for carefully selecting and arranging for access to licensed commercial resources for its users, but it would not house local copies of materials or assume responsibility for long-term preservation.

page 5
Facilitating/Monitoring Longevity of Distributed Content

Personalization of Content

View A:
- View slides
- View video
- View synchronized presentation using applet

View B:
- Get transcript of audio
- Search for keyword
- Get slides translated to French
Cross-Repository Reference Linking

Origins of the OAI

- Increasing interest in alternative scholarly publishing solutions – e.g., LANL arXiv
- Increasing impact through federation
- UPS Mtg., Sante Fe, October 1999
  - Representatives of various E-Print, library, and publishing communities
  - Goal: definition of an interoperability framework among E-Print providers
  - Result: Santa Fe Convention, interoperability through metadata harvesting
“Open” Archives

• Political Agenda?
  – Author self-archiving of E-Prints
  – “Mission” to reformulate scholarly publishing framework
• Technical?
  – Infrastructure to facilitate interoperability across multiple domains

Other Communities of Interest

• “Cambridge” Digital Library Federation meetings
  – research library community has many materials for which they’d like to ‘expose’ metadata
• OAI workshops
  – librarians, publishers (some), researchers, others
• Museum Community
  – Museums on the Web and CIMI
Technical Umbrella for Practical Interoperability…

...that can be exploited by different communities

OAI Organizational Structure
Key Features

• Clear focus and scope
  – Developing and refining technical specification
  – Community building and evangelism limited to serving that goal and to encouraging widespread adoption

• Encouraging specialization and community-specific activities

• Division of responsibility
  – Executive (Van de Sompel and Lagoze)
  – Steering Committee
  – Technical Committee
  – Mailing Lists (community)
OAI Technical Infrastructure

Key Technical Features

- Deploy now technology – 80/20 rule
- Two-party model – providers (*data providers*) and consumers (*service providers*)
- Simple HTTP encoding
- XML schema for some degree of protocol conformance
- Extensibility
  - Multiple item-level metadata
  - Collection level metadata

The World According to OAI
Key Features of the OAI Metadata Harvesting Protocol

- definitions & concepts
  - repository
  - record
  - identifier
  - datestamp
  - set

- protocol features
  - HTTP encoding
  - metadata prefix & schema
  - flow control

- protocol requests
  - supporting requests
  - harvesting requests
repository

support data

harvester

OAI protocol

repository

harvesting data

record

<record>
  <header>
    <identifier>oai:eg:001</identifier>
    <datestamp>1999-01-01</datestamp>
  </header>
  <metadata>
    <dc xmlns="http://purl.org/dc">
      <title>My Example</title>
    </dc>
  </metadata>
  <about>
    <ea xmlns="http://www.arXiv.org/ea"
        <usage>No restrictions</usage>
    </ea>
  </about>
</record>

protocol support

format-specific metadata

community-specific record data
identifiers

locally unique key for extracting a record from a repository

\[
oai-identifier = \text{oai:archive-identifier:record-identifier}
\]

Registered URI Scheme

Archive Identifier: Registered within OAI

Unique ID within archive: (syntax is archive-specific)

example = oai:ncstrl:ncstrl.cornellcs/TR94-1418

selective harvesting - datestamps

harvest within date range
selective harvesting - sets

harvest within set

repositories define hierarchical organization
• each item in a repository may be organized in one set, several sets, or no sets at all
• meaning of sets or of set hierarchy is not defined in protocol
• individual communities may formulate common set configurations

set specifics
HTTP encoding - requests

BASE-URL ----------> an.oa.org/OAI-script
keyword arguments --> verb=ListIdentifiers&set=S1

GET
http://an.oa.org/OAI-script?verb=ListIdentifiers&set=S1

POST
POST http://an.oa.org/OAI-script HTTP/1.0
Content-Length: 78
Content-Type: application/x-www-form-urlencoded
verb=ListIdentifiers&set=S1

HTTP encoding - responses

<xml version="1.0" encoding="UTF-9" ?>
<GetRecord
xmlns="http://oai.namespace.uri"
xmlns:xsi="http://w3.namespace.uri"
<responseDate>2000-19-01T19:30:30-04:00</responseDate>
<requestURL>http://an.oa.org/OAI-script?verb=GetRecord
&identifier=oai%3AarXiv%3A0001
&metadataPrefix=oai_dc</requestURL>
<record>
record contents
</record>
additional records
</GetRecord>
metadata prefix and schema

- support for harvesting multiple metadata formats
  - metadata schema: each format must have a validating XML schema at a publicly accessible URL (communities may define shared formats and schema).
  - metadata prefix: each repository maps a prefix to the schema it supports, which is used in protocol requests.
- support for unqualified Dublin Core mandatory
  - reserved schema URL at http://www.openarchives.org/OAI/dc.xsd
  - reserved prefix oai_dc.
flow control specifics

- applies to all protocol requests that return lists: ListRecords, ListIdentifiers, ListSets
- resumptionToken is opaque
- semantics of partitioning of responses within resumption requests is undefined
- time-to-live of resumptionToken is not defined by the protocol

OAI Protocol

service provider

- Supporting protocol requests:
  - Identify
  - ListMetadataFormats
  - ListSets

data provider

- Harvesting protocol requests:
  - ListRecords
  - ListIdentifiers
  - GetRecord
Supporting Protocol Requests

Identify

• Repository name
• Base-URL
• Admin e-mail
• OAI protocol version
• Description Container

Supporting Protocol Requests

ListMetadataFormats

REPEAT
• Format prefix
• Format XML schema
/REPEAT
Supporting Protocol Requests

- ListSets

REPEAT
  - Set Specification
  - Set Name
/REPEAT

Harvesting Protocol Requests

- ListRecords

REPEAT
  - Identifier
  - Datestamp
  - Metadata
  - About Container
/REPEAT
Harvesting Protocol Requests

service provider

ListIdentifiers

* from=a
* until=b
* set=klm

REPEAT
• Identifier
• Datestamp
/REPEAT

repository

data provider

* from=a
* until=b
* set=klm

REPEAT
• Identifier
• Datestamp
/REPEAT

Harvesting Protocol Requests

service provider

GetRecord

* identifier=oai:mlib:123a
* metadataPrefix=oai_dc

• Identifier
• Datestamp
• Metadata
• About

repository

data provider
Other OAI Functions

• Registry of data and service providers
• Tool registry
• Community communication
Registering as a Data Provider

Data providers who support the OAI protocol may choose to list their repository in the OAI registry. The goals of the registry are:

- To provide a publicly accessible list of OAI-compliant repositories, making it easy for service providers to discover repositories from which metadata can be harvested
- To provide a mechanism for service providers to compare its usefulness with the OAI (protocol specification)
- To provide a means for the OAI registry to monitor the value of the protocol and plan future activities and strategies.

This page allows you to register your repository by entering your OAI-ORI in the box at the bottom of this page. Before doing so, please read all of the instructions on this page to understand what registration means and the choices you have.

Consequences of Registration

By registering your repository, you agree to the following:

- The OAI will periodically run a set of registration tests on your repository. To ensure integrity of the registry, we will only accept reports that complete these tests. For more information on how the tests will be scored, please refer to the documentation at the OAI registry. The tests will be submitted to the OAI registry only after you agree to the terms.
- You will be required to maintain your repository, including providing the necessary metadata and ensuring that it is up-to-date. If your repository fails to complete the tests, or if it is no longer available, we will remove your repository from the registry and notify you via email.
- The registry will store all the information you provide, including your ORI, repository name, and any other information provided during registration.
- The registry will use your ORI to generate a unique identifier for your repository.
- The registry will allow you to update your repository information at any time.
- The registry will allow you to remove your repository from the registry at any time.

Protocol Testing

You can test the protocol by entering your ORI into the test box. If the test is successful, you will receive a notification that your OAI-ORI is valid. If the test fails, you will receive a notification indicating the reason for the failure.

Registered Data Providers

The application allows you to browse the current list of OAI-compliant repositories. Currently, there are over 100 such repositories, which are listed alphabetically. You can search for a repository by its name, ORI, or by its repository name.

You may find more information about each repository by selecting one of the rows in the following table. You can select the registration record from the database, or if your browser can render tables, you can download the HTML file to view the selected repository and receive the current XML response.
OAI Tools

- Repository Explorer
- Servers and utilities
- Related resources
  - XML
  - Unicode
Implementation Utilities

- Protocol handlers
  - OCLC
  - Virginia Tech
  - UIUC
- Metadata Utilities
  - MARC to DC (OCLC, Virginia Tech, …)
- eprints.org

Participating in the OAI Community

- Listservs
  - oai-general – discussion of OAI related issues
  - oai-implementers – sharing technical questions and agendas
- OAI website (www.openarchives.org)
  - Post news and links to OAI related activities
- Community-specific
  - How does OAI apply to your community?
Externally funded initiatives

- European Community
  - Open Archives Forum
  - Cyclades Project
- Andrew W. Mellon Foundation
  - Funding for 7 service providers
- Digital Library Federation
  - Gateways for access to member's digital collections
- National Science Foundation
  - NSDL (www.nsdl.nsf.gov) Core Infrastructure
  - Virginia Tech awards IIS-9986089, 0086227, 0080748 with joint funding by DFG (Germany), CONACyT (Mexico)

Where do we go from here
2001-2002

- Controlling the stampede
- Technical re-evaluation leading to "final" 2.0 specification
  - OAI Technical Committee
- Strategy for standardization
- Community building focused on verification and validation
Open Archives:
Communities, Interoperability and Services
(Workshop - Sep. 13, 2001 - New Orleans)

- http://purl.org/net/oaisept01
- Session 1: Intro to OAI
- Session 2: Technical Details
- Session 3: Concurrent Group Discussions
  - Applicability of OAI to distributed community building.;
    community support needed to leverage OAI standards
  - Evaluation of tech stds; current and future directions of
    stds and services (related to the OAI protocols)
  - See details on next slide
- Session 4: Presentations of Group Findings
- Session 5: Moving Forward

---

Open Archives:
Communities, Interoperability and Services
(Workshop - Sep. 13, 2001 - New Orleans)

<table>
<thead>
<tr>
<th>Building Communities</th>
<th>Technical Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support for different types of communities</td>
<td>Protocol evaluation: experiences, efficiency, …</td>
</tr>
<tr>
<td>Developments aiding community building</td>
<td>Support for internationalization</td>
</tr>
<tr>
<td>Selective harvesting (sets)</td>
<td>Services enabled by OAI</td>
</tr>
<tr>
<td>Community building ex’s</td>
<td>Support for full-text retrieval</td>
</tr>
<tr>
<td>Social aspects of OAI-based community projects</td>
<td>Support for protocol adoption</td>
</tr>
</tbody>
</table>
Open Archives:
Communities, Interoperability and Services
(Workshop - Sep. 13, 2001 - New Orleans)

- Attendees from various institutions

<table>
<thead>
<tr>
<th>Institution</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caltech</td>
<td>U. of Illinois, U-C</td>
</tr>
<tr>
<td>CMIS, Carlton, Australia</td>
<td>U. of Oldenburg, GE</td>
</tr>
<tr>
<td>Dartmouth College</td>
<td>U. of Southampton</td>
</tr>
<tr>
<td>Emory University</td>
<td>U. of Tennessee</td>
</tr>
<tr>
<td>Los Alamos Nat’l Lab</td>
<td>US Dept. of Energy</td>
</tr>
<tr>
<td>Louisiana State Univ.</td>
<td>Virginia Tech</td>
</tr>
<tr>
<td>Michigan State Univ.</td>
<td></td>
</tr>
<tr>
<td>NASA Center for Aerospace Information</td>
<td></td>
</tr>
</tbody>
</table>

Case Study: NSDL

- National Science, mathematics, engineering and technology education Digital Library (NSF)
- Urgent need: “doors open” Oct. 2002
- Core integration track: building on collections
- Collections track: building portals and centralized repositories (metadata, learning objects / educational resources)
- Metadata: DC, LOMS/IMS
- Problems: will publishers share metadata? Will those with small repositories adopt OAI?
Case Study: NDLTD

- Metadata: MARC21 (coded in XML), ETDMS (see www.ndltd.org/standards)
- Protocols in use: Z39.50, Harvest, Dienst, OAI, as well as http (web sites)
- OCLC’s LAF (authority control) to work with RDF implementation of ETDMS
- Union collection -> VTLS’s Virtua, Virginia Tech’s MARIAN
- Phased efforts for development and testing over more than a year

Case Study: NCSTRL

- CSTR and WATERS -> NCSTRL
  - Federated search of regular sites, harvesting of lite sites
- Changes: disinterest in central service, decline in interest in dept report series, increase in interest in personal web pages (ACM allowance)
- Kepler to support personal Open Archives
- Shift from Dienst-based service to OAI-based service underway in Fall 2001 (aided by Virginia’s Internet Technology Innovation Center, through ODU, UVA, and Virginia Tech – along with others)
Case Study: SOLINET

- Mellon Foundation
- SOuteastern LLibrary NETwork (Atlanta)
- Deadline: February 2003
- 10+ univ. collections about American South
- Scholars to learn about OAI, decide how to apply, work toward controlled vocabulary
- Harvesting to central site
- New central DL services (to be developed)

Community Options

- Is DC sufficient, or is there a list of one or more metadata standards existing or that can be developed to suite community needs?
- Is there a natural set structure, or several?
  - Year? Topical areas? Location / institution?
- What are the social, economic, political issues regarding who will run an Open Archive?
- Will all share metadata or must there be federated search as well?
Community Assistance

- Awareness
- Training
- Tools
- Test and validation
- Operation
- Logging and analysis
- Sharing experiences and solutions

Conclusion

- Interoperability
- History / evolution of OAI
- Protocol for metadata harvesting
- Implementations and support
- Current situation / progress
- Community building and support by OAI