

The Open Archives Initiative Protocol for Metadata Harvesting: Overview

Jewel Ward <jewelw@lanl.gov>

Visiting Scholar, Keio University

Lib-Sys Seminar, Keio University, Mita Campus

17 June 2003

Acknowledgements

- JCDL 2001/2002: OAI-PMH Introduction
 - Hussein Suleman (then at Virginia Tech)
- JCDL 2003: Introduction to the OAI-PMH
 - Timothy W. Cole (UIUC)
 - William H. Mischo (UIUC)
 - Thomas Habing (UIUC)

Acknowledgements

- JCDL 2003: Advanced Overview of Version 2.0 of the OAI-PMH
 - Michael L. Nelson (Old Dominion University)
 - Herbert Van de Sompel (LANL)
 - Simeon Warner (Cornell University)
- Digital Library Federation (DLF) Spring Forum 2003
 - "The OAI Static Repository: a file-based approach to exposing metadata via the OAI-PMH."
 - Herbert Van de Sompel (LANL)
 - This research was conducted by Patrick Hochstenbach (LANL), Henry Jerez (LANL) and Herbert Van de Sompel.

Outline

- Briefly: Institutional Repositories
- Background & Development
- OAI-PMH Basics
- New Developments
- Further Information
- Questions?

Institutional Repositories

- Institutional Repository: “digital collections capturing and preserving the intellectual output of a single or multi-university community” (Johnson, 2002) (DLib article)
- It’s a way to aggregate the research output of an organization into one location as opposed to the current “scatter” method

Institutional Repositories

- arXiv is *not* an institutional repository (and it is now @Cornell University)
- Current LANL institutional repository projects
 - AISTI (the Alliance for Innovation in Scientific and Technical Information)
 - Within LANL

Movement and Protocol

- The Open Archives Movement
 - Enhance public access to research output and scholarly materials
 - Reaction to commercial publisher's pricing of scholarly journals
- The Open Archives Protocol for Metadata Harvesting
 - Number of ePrint repositories and DLs growing
 - ePrint/Library community desired interoperability of scholarly archives

OAI-PMH Technical Development

- Gopher, FTP
- Union Catalogs
- Z39.50
- Kahn-Wilensky Framework
- Dienst Protocol
- Harvest
- UPS
- OAI-PMH

Overview of the OAI-PMH

- What is the OAI-PMH?
 - The protocol defines an application-independent specification for the **interoperability** [of digital libraries] through metadata harvesting.
 - The protocol is a building block that can facilitate/enable variety of services and functions.
- OAI versus OAI-PMH

Overview of the OAI-PMH

What the OAI-PMH is *not*

- The protocol is *not* a search service
- The protocol is *not* a database
- The protocol is *not* OAIS
- The protocol does *not* define a metadata specification
- The protocol does *not* equal Dublin Core

Data & Service Providers

- Data Providers (DPs) – Repositories – refer to entities who possess resources and metadata and are willing to share metadata with others via well-defined OAI protocols
- Service Providers (SPs) – Harvesters – are entities who harvest metadata from DPs in order to provide high level services to users (such as search and discovery).
- Data equals server, Service equals client

OAI-PMH Verb Set

Verb	Function
Identify	description of repository
ListMetadataFormats	metadata formats supported by repository
ListSets	sets defined by repository
ListIdentifiers	OAI unique ids contained in repository
ListRecords	listing of N records
GetRecord	listing of a single record

metadata
about the
repository

harvesting
verbs

Most verbs take arguments: dates, sets, ids, metadata formats and resumption token (for flow control).

OAI-PMH Metadata

- Repositories are required to expose their metadata as the Dublin Core Metadata Element Set (DCMES).
- Repositories are strongly encouraged to expose their metadata in more expressive formats.
- Examples of other formats in use:
 - MARC
 - RFC-1807
 - Open Languages Archives Community Metadata Set
 - Electronic Theses and Dissertation Metadata Set (Nelson, 2001)

resource – item - record



← resource

*set-membership is
item-level property*

item = identifier

all available metadata
about *David*

← item

Dublin Core
metadata

MARC
metadata

SPECTRUM
metadata

← records

record = identifier + metadata format + datestamp ¹⁴

Unique Identifiers

- Each item must have a unique identifier
- Identifiers must follow the URI syntax
 - OAI has its own format:
 - oai:<archiveID>:<recordID>
 - oai:etd.vt.edu:edt-1234567890
 - Can also use other formats
 - http
 - handle

Datestamps

- Required to support incremental harvesting
- Can be either YYYY-MM-DD or YYYY-MM-DDThh:mm:ssZ (must be GMT timezone)
- Different from dates within the metadata; this datestamp is used only for harvesting
- The datestamp is the **creation date of the metadata record itself**
 - It is *not* the publication date
 - It is *not* the creation date of the item

Sets

- Optional, depends on local DPs
- Must provide `setSpec` & `setName`, may provide `setDescription`, for each set in DP
- May be hierarchical (use “:”) to allow for harvesting of subcollections

How the OAI-PMH Works

OAI “VERBS”

Identify

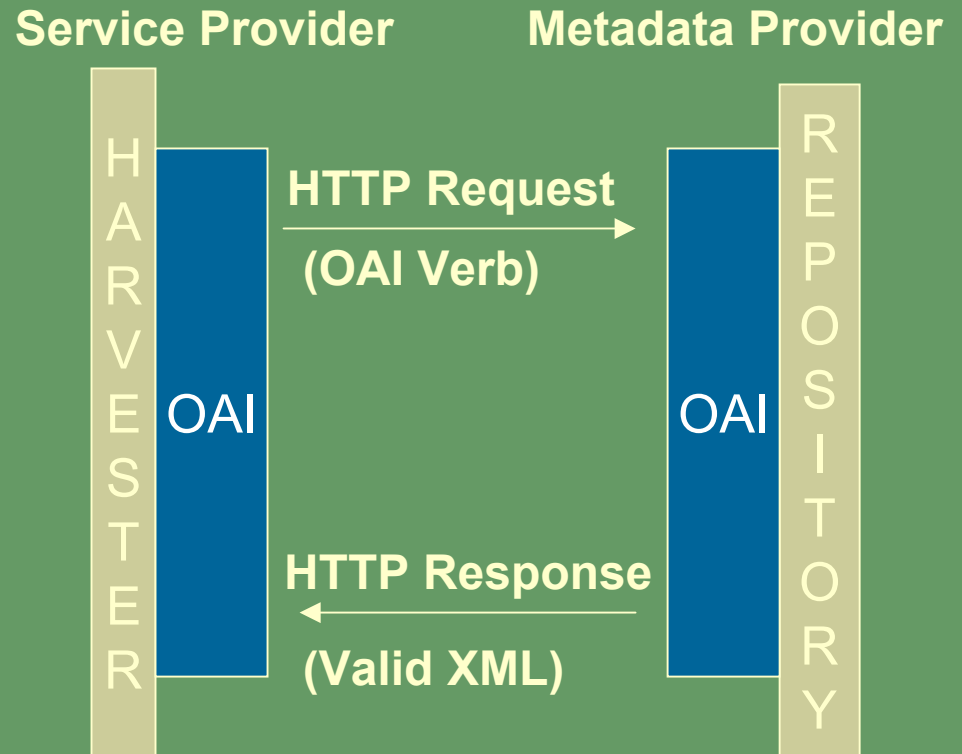
ListMetadataFormats

ListSets

ListIdentifiers

ListRecords

GetRecord



baseURL+verb

Examples

- <http://arXiv.org/oai2?verb=Identify>
- <http://arXiv.org/oai2?verb=ListSets>
- <http://arXiv.org/oai2?verb=ListMetadataFormats>
- http://arxiv.org/oai2?verb=ListIdentifiers&metadataPrefix=oai_dc
- http://arxiv.org/oai2?verb=GetRecord&identifier=<recordID>&metadataPrefix=oai_dc
- http://arXiv.org/oai2?verb=ListRecords&metadataPrefix=oai_dc

Example Response

```
<?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
  http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2003-06-03T20:13:50Z</responseDate>
  <request verb="GetRecord" metadataPrefix="oai_dc"
    identifier="oai:arXiv.org:acc-
    phys/9411001">http://arXiv.org/oai2</request>
- <GetRecord>
..
</GetRecord>
</OAI-PMH>
```

Example Record

```
- <record>
- <header>
  <identifier>oai:arXiv.org:acc-phys/9411001</identifier>
  <datestamp>2003-02-05</datestamp>
  <setSpec>physics:acc-phys</setSpec>
  <setSpec>physics:physics</setSpec>
</header>
- <metadata>
- <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
  xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
  instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
  http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:title>Symplectic Computation of Lyapunov Exponents</dc:title>
  <dc:creator>Habib, Salman</dc:creator>
  <dc:creator>Ryne, Robert D.</dc:creator>
  <dc:subject>Accelerator Physics</dc:subject>
  <dc:description>..</dc:description>
  <dc:description>Comment: 12 pages, uuencoded PostScript (figures included)</dc:description>
  <dc:date>1994-10-31</dc:date>
  <dc:type>text</dc:type>
  <dc:identifier>http://arXiv.org/abs/acc-phys/9411001</dc:identifier>
</oai_dc:dc>
</metadata>
</record>
```

Optional Container Elements

- Repository level (set)
 - <Identify><description>
 - Additional information about repository
 - oai-identifier, eprints, friends, branding, other...
 - <ListSets><setDescription>
- Metadata level
 - <about>
 - Meta-metadata, i.e. record level rights

Resumption Tokens, etc.

Resumption Tokens/Flow Control/Load Balancing

- “`resumptionToken`” is used for an incomplete response
- The client is issued a response with a token which may be presented to the server to receive more results at a later time

Resumption Tokens, etc.

Resumption Tokens/Flow Control/Load Balancing

- Options include: completeListSize, cursor, and expiration date attributes
- Combine from/until/metadataPrefix/set and a record number indicator with delimiters into a sequential token
 - from!until!metadataPrefix!set!recordnumber
 - 2000-01-01!2001-01-01!oai_dc!All!100
- Use a session manager with automatic expiry

Resumption Tokens, etc.

Resumption Tokens/Flow Control/Load Balancing

Idempotency

- Purpose is to allow harvesters to recover from lost responses or crashes without starting a large harvest from scratch
- Recover by re-issuing request using `resumptionToken` from previous request
- **IMPLICATION:** harvester must accept both the most recent `resumptionToken` issued and the previous one

Error Handling

All protocol errors are in XML format

- badVerb: illegal verb requested
- badArgument: illegal parameter values or combinations
- badResumptionToken, cannotDisseminateFormat, idDoesNotExist: parameters are in right format but are not legal under current conditions
- noRecordsMatch, noMetadataFormats, noSetHierarchy: empty response exception

Example Error Message

```
<?xml version="1.0" encoding="UTF-8" ?>  
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"  
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"  
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/  
  http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">  
  <responseDate>2003-06-03T20:32:53Z</responseDate>  
  <request>http://arXiv.org/oai2</request>  
  <error code="badArgument">Verb 'ListRecords', argument  
    'metadataPrefix' required but not supplied.</error>  
</OAI-PMH>
```

OAI-PMH Static Repository

Motivation

- OAI-PMH is a low-barrier protocol
- OAI-PMH favors to make it easy for Data Providers
 - Bias has its origins in the Santa Fe Convention

OAI-PMH Static Repository

Motivation

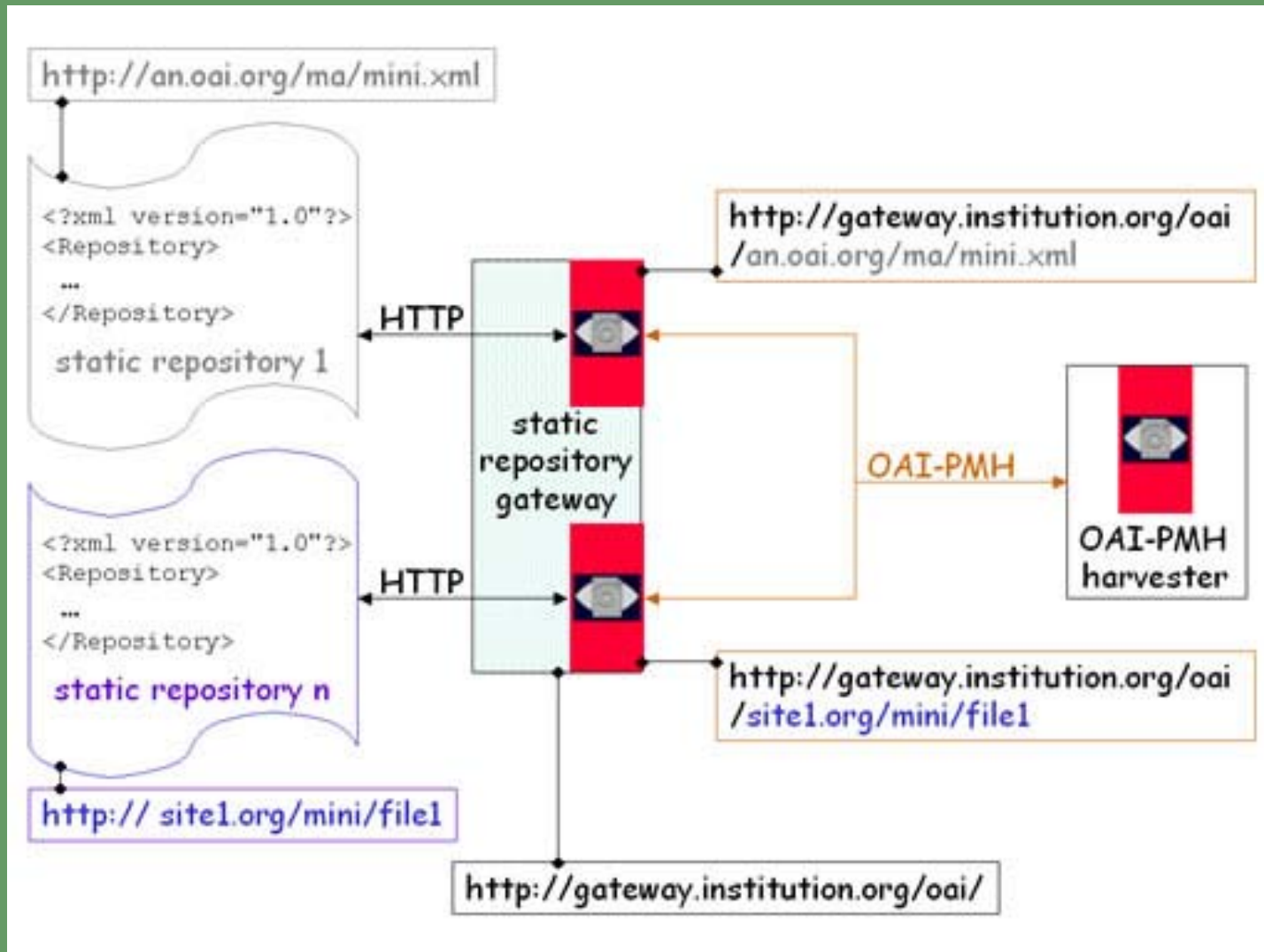
- Implementation is sometimes not trivial
 - Lack of technical expertise
 - Size of collection does not justify the investment
 - Security considerations re: database
 - ISP does not allow 3rd party software
 - Cf. OLAC, union catalogue, LoCKSS

OAI-PMH Static Repository

Motivation

Therefore: research to devise approaches to further lower the barrier to sharing metadata collections through the OAI-PMH.

OAI-PMH Static Repository



Rights Effort

- Exploring rights about:
 - Resource
 - Metadata
- Framework based on the Creative Commons (CC)
- Collaborative Effort JISC/OAI/CC
(JISC is the “Joint Information Systems Committee” involved with RoMEO.)

Further Information

- Johnson, R. (2002). Institutional Repositories Partnering with Faculty to Enhance Scholarly communication. DLib Magazine. Retrieved February, 2003 from
 - <http://www.dlib.org/dlib/november02/johnson/11johnson.html>
- SPARC Institutional Repository Checklist & Resource Guide
 - http://www.arl.org/sparc/IR/IR_Guide.html

Further Information

- Open Archives Initiative
 - <http://www.openarchives.org>
- OAI Metadata Harvesting Protocol
 - <http://www.openarchives.org/OAI/openarchivesprotocol.htm>
- OAI-PMH Tools Index
 - <http://www.openarchives.org/tools/index.html>
- Virginia Tech DLRL OAI Projects
 - <http://www.dlib.vt.edu/projects/OAI/>
- Repository Explorer
 - http://purl.org/net/oai_explorer
- Nelson, M. (2001). Better Interoperability Through the Open Archives Initiative. *The New Review of Information Networking*. 7, 133-146.
- ARC Cross-Archive Search Service
 - <http://arc.cs.odu.edu/>

Further Information

- ARC Cross-Archive Search Service
 - <http://arc.cs.odu.edu/>
- OAI-PMH Static Repository
 - Registration
 - <http://libtest.lanl.gov/registry.html>
 - Example Repository
 - http://libtest.lanl.gov/cgi-bin/gateway.cgi/lib-www.lanl.gov/%7Ehochsten/desktop.xml?verb=ListRecords&metadataPrefix=oai_dc
 - Specification
 - <http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm>

Further Information

- Creative Commons
 - <http://www.creativecommons.org/>
- JISC
 - <http://www.jisc.ac.uk/>
- Dspace
 - <http://dspace.org/news/dspace-news.html>
- E-Prints DL-in-a-box
 - <http://www.eprints.org>
- Greenstone Digital Library
 - <http://www.greenstone.org/english/home.html>

Further Information

- NDLTD
 - <http://www.ndltd.org>
- XML Schema Validator
 - <http://www.w3.org/2001/03/webdata/xsv>
- Dublin Core Metadata Initiative
 - <http://www.dublincore.org>
- XML Tools at W3C
 - <http://www.w3.org/XML/#software>

Questions?