Service Providers: Future Perspectives

Michael L. Nelson
Old Dominion University
Norfolk Virginia, USA
mln@cs.odu.edu
http://www.cs.odu.edu/~mln/

2nd Workshop on the Open Archives Initiative: Gaining Independence With E-print Archives and OAI

CERN, Switzerland
October 18, 2002
Outline

• History of the history of OAI-PMH
• (Traditional) public service providers not present for this meeting
• Why the OAI-PMH is not important
• Defining the OAI-PMH data model
• Abusing the OAI-PMH data model
• Current and nearly-current interesting services
# OAI-PMH Meeting History

<table>
<thead>
<tr>
<th>OAI Open Day, Washington DC 1/2001</th>
<th>This meeting CERN 10/2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Protocol definition, development tools</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>DPs, retrofitting existing DLs</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>SPs, new services</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Socio-Economic-Political Issues</td>
<td></td>
</tr>
</tbody>
</table>
Shift of Topics

- From the protocol itself, supporting & debugging tools and how to retrofit (existing) DLs...
- ...to building (new) services that use the OAI-PMH as a core technology and reporting on their impact to the institution/community
NTRS

- http://ntrs.nasa.gov/
- metadata harvesting replacement for http://techreports.larc.nasa.gov/cgi-bin/NTRS
  - previous NTRS was based on distributed searching
  - hierarchical harvesting
- (nigh) publicly available
Arc

- http://arc.cs.odu.edu/
- harvests all known archives
- first end-user service provider
- source available through SourceForge
- hierarchical harvesting
NCSTRL

- http://www.ncstrl.org/
- metadata harvesting replacement for Dienst-based NCSTRL
- based on Arc
- computer science metadata
Archon

- http://archon.cs.odu.edu/
- physics metadata
- based on Arc
- features:
  - citation indexing
  - equation-based searching
Torii

- http://torii.sissa.it/
- physics metadata
- features
  - personalization
  - recommendations
  - WAP access

This is the website of TORII (version 1.0), a prototype designed and implemented by the consortium TIPS (project IST-1998-10419) as part of the IST program of the European Union Fifth framework. The project is described in a booklet.

The system has been designed to serve as a single environment from where to perform all web-related work. It can also be accessed via WAP at http://torii.sissa.it/wap.wml. In this version, the website gives access to the open archives: arXiv (Physics, Mathematics, Nonlinear Sciences, Computer Science), JHEP, BioMed Central and MIDGE. To browse, search or get a document from these archives, just use the command bar at the top.

You can use the more advanced features by registering. Once registered and signed-in, you will have your own personal folder where to store documents for future reference. You will also have the possibility of defining a profile of interests, the system will then screen your browsing according to your profile by ordering documents by their relevance for you. Finally, you will be able to evaluate any document accessed and your evaluation will be stored and used as a first implementation of a network of quality control tools.
iCite

- http://icite.sissa.it/
- physics metadata
- features
  - citation based access to arXiv metadata
my.OAI

- http://www.myoai.com/
- covers all registered metadata
- features
  - result sets
  - personalization
  - many other advanced features
Cyclades

- http://www.ercim.org/cyclades
- scientific metadata
- features
  - personalization
  - recommendations
  - collaboration
- status?
citebase

- http://citebase.eprints.org/
- arXiv metadata
- citation based indexing, reporting
OAister

- http://oaister.umdl.umich.edu/
- harvests all known archives
Public Knowledge Project

- http://www.pkp.ubc.ca/harvester/
- domain-specific filtering of harvested metadata (?)
Perseus

- http://www.perseus.tufts.edu/
- they claim to harvest all DPs, but only humanities related DPs appear in the pull down menu
Service Providers

- It is clear that SPs are proliferating, despite (because of?) the inherent bias toward DPs in the protocol
  - easy to be a DP -> many DPs -> SPs eventually emerge
  - hard to be a DP -> SPs starve
  - currently 5x DPs more than SPs

- SPs are beginning to offer increasingly sophisticated services
  - competitive market originally envisioned for SPs is emerging
Why The OAI-PMH is NOT Important

• Users don’t care
• OAI-PMH is middleware
  – if done right, the uninterested user should never have to know

• Using the OAI-PMH does not insure a good SP
• OAI-PMH is *(or is becoming)* HTTP for DLs
  – few people get excited about http now
    • http & OAI-PMH are core technologies whose presence is now assumed
Other Uses For the OAI-PMH

• Assumptions:
  - Traditional DLs / SPs will continue on their present path of increasing sophistication
    • citation indexing, search results viz, personalization, recommendations, subject-based filtering, etc.
  - growth rates remain the same (5x DPs as SPs)
• Premise: OAI-PMH is applicable to any scenario that needs to update / synchronize distributed state
  - Future opportunities are possible by creatively interpreting the OAI-PMH data model
**OAI-PMH Data Model**

- **set-membership is item-level property**
- **item = identifier**
- **all available metadata about David**
- **record = identifier + metadata format + datestamp**
Typical Values

• repository
  - collection of publications
• resource
  - scholarly publication
• item
  - all metadata (DC + MARC)
• record
  - a single metadata format
• datestamp
  - last update / addition of a record
• metadata format
  - bibliographic metadata format
• set
  - originating institution or subject categories
Repositories...

- Stretching the idea of a repository a bit:
  - contextually sensitive repositories
    - “personalization for harvesters”
    - communication between strangers, or communication between friends?
  - OAI-PMH for individual complex objects?
    - OAI-PMH without MySQL?!
      - Fedora, Multi-valent documents, buckets
      - tar, jar, zip, etc. files
Resource

• What if resource were:
  - computer system status
    • uptime, who, w, df, ps, etc.
  - or generalized “system” status
    • e.g., sports league standings
  - people
    • personnel databases
    • authority files for authors
Item

- What if item were:
  - software
    - union of versions + formats
  - all forms of metadata
    - administrative + structural
    - citations, annotations, reviews, etc.
  - data
    - e.g., newsfeeds and other XML expressible content
      - metadataPrefixes or sets could be defined to be different versions
Record

• What if record were:
  - specific software instantiations / updates
  - access / retrieval logs for DLs (or computer systems)
  - push / pull model inversion
    • put a harvester on the client behind a firewall, the client contacts a DP and receives “instructions” on how to submit the desired document (e.g., send email to a specified address)
Datestamp

- semantics of datestamp are strongly influenced by the choice of resource / item / record / metadataPrefix, but it could be used to:
  - signify change of set membership (e.g., workflow: item moves from “submitted” to “approved”)
  - change datestamp to reflect access to the DP
    - e.g., in conjunction with metadataPrefixes of “accessed” or “mirrored”
metadataPrefix

- what if metadataPrefix were:
  - instructions for extracting / archiving / scraping the resource
    - verb=ListRecords&metadataPrefix=extract_TIFFs
  - code fragments to run locally
    - (harvested from a trusted source!)
  - XSLT for other metadataPrefixes
    - branding container is at the repository-level, this could be record- or item-level
Set

- sets are already used for tunneling OAI-PMH extensions (see Suleman & Fox, D-Lib 7(12))
- other uses:
  - in aggregators, automatically create 1 set per baseURL
  - have “hidden” sets (or metadataPrefix) that have administrative or community-specific values (or triggers)
    - set=accessed>1000&from=2001-01-01
    - set=harvestMeWithTheseARGS&until=2002-05-05&metadataPrefix=oai_marc
Interesting Services

- **DP9**
  - gateway to expose repository contents in HTML suitable for web crawlers
- **Celestial**
  - OAI “cache”, also 1.1 -> 2.0 converter
- **Static (mini-) repositories**
  - XML files, based on OLAC work
- **OpenURL metadata format registries**
  - record = metadata format
DP9 Architecture

see Liu et al., JCDL 2002; http://dlib.cs.odu.edu/dp9
DP9 Formatting

• Format of URLs

• HTML Meta tags
  - Some crawlers (such as Inktomi) use the HTML meta tags to index a Web pages; DP9 also maps Dublin Core metadata to corresponding HTML meta tags.
  - For pages that are designed exclusively for robots navigation, a noindex robots meta tag is used
  - X-FORWARDED-FOR header to distinguish between different users coming in via a proxy
Celestial

- Developed by Brody @ Southampton
  - http://celestial.eprints.org/
  - designed to complement DP9
  - see Liu, Brody, et al., D-Lib Magazine 8(11)

- Where DP9 is a non-caching proxy, Celestial caches the metadata records
  - can off-load work from individual archives, higher availability
  - can harvest 1.1, 2.0; exports in 2.0
“Static” Repositories

• Premise: a repository does not wish to have an executing program on its site, so it has a “static” XML file with some of the OAI-PMH responses in place
  - Design still being discussed
    • accessed through a proxy
    • could be a low functionality node, or the XML file could be produced by a process and moved outside a firewall

• Based on OLAC work by Bird & Simons
  - http://www.language-archives.org/
OpenURL Metadata Registry

- Registry of metadata formats for OpenURL
  - http://www.sfxit.com/openurl/
Goal:
- inform linking servers re Schema
- ease of admin for all parties involved
- limit human overhead

registrars

XML Schema
URL1

XML Schema
URL2

XML Schema
URLn

Slide from Van de Sompel
Registry:
- schemaLocation
- registration date
- mirror of Schema

central repository

registrars

XML Schema
URL1

XML Schema
URL2

... ... ...

XML Schema
URLn

registration
Poll:
• fetch schema at schemaLocation
• log failure/success
• compare fetched Schema with mirror
  • changed => replace mirror
  • removed => deregistered
OAI repo:
- record-ids = schemaLocation
- oai_dc record:
  - registration info
  - (de)registration datestamp
- xsi record:
  - mirror schema
  - schema update datestamp
- poll record:
  - process info
  - recent poll datestamp

registration
polling

Slide from Van de Sompel
Conclusions

• DPs continue to proliferate
  - and spawn SPs!
• SPs are / are becoming a competitive market
  - e.g., at least 10 different interfaces to arXiv metadata
  - growing sophistication of services
  - differentiation of SPs will be on features that have little to nothing to do with OAI-PMH
Conclusions

- **Protocol / transport gateways**
  - Dienst <-> OAI
  - Z39.50
    - ZMARCO (UIUC)
  - SOAP
    - prototypes @ VT (Suleman) & ODU (Zubair)
  - WebDAV/DASL
    - resurrect DASL?
OAI-PMH Will Have Arrived When:

- general web robots issue OAI-PMH verbs
  - ...DP9 will no longer be needed
  - requires shift in “control”: harvester or repository?
- mod_oai is developed and is included in the default Apache configuration
- OAI-PMH fades into the background
  - similar to TCP/IP, http, XML, etc.
  - next year’s workshop is on OpenURL