Version 2 of the OAI-PMH & some other stuff

2nd Workshop on the OAI, CERN Geneva, October 17th 2002

Herbert Van de Sompel
Los Alamos National Laboratory

Carl Lagoze
Cornell University
about OAI-PMH v.2.0

measures of success

future?
releasing OAI-PMH v.2.0
⇒ creation of OAI-tech

⇒ revision phase

⇒ alpha testing phase

⇒ beta phase

⇒ release of OAI-PMH v.2.0
creation of OAI-tech (06/01)

• charge:
  • review functionality and nature of OAI-PMH v.1.0
  • investigate extensions
  • release stable version of OAI-PMH by 05/02

US representatives

Thomas Krichel (Long Island U) - Jeff Young (OCLC) - Tim Cole - (U of Illinois at Urbana Champaign) - Hussein Suleman (Virginia Tech) - Simeon Warner (Cornell U) - Michael Nelson (NASA) - Caroline Arms (LoC) - Mohammad Zubair (Old Dominion U) - Steven Bird (U Penn.)

European representatives

Andy Powell (Bath U. & UKOLN) - Mogens Sandfaer (DTV) - Thomas Baron (CERN) - Les Carr (U of Southampton)
revision phase [09/01 - 02/02]

• review process by OAI-tech [09/01 - 01/02]
  • identification of issues
  • discussion of issues
  • proposals for resolution by OAI Exec
• drafting of revised protocol document [02/02]
  • Lagoze, Van de Sompel, Nelson, Warner
alpha testing phase [03/02 - 05/02]

- extension of OAI-tech with alpha testers
- continuous feedback from their implementations
- ongoing revision of protocol document
OAI-PMH 2.0 alpha testers

- The British Library
- Cornell U. -- NSDL project & e-print arXiv
- Ex Libris
- FS Consulting Inc -- harvester for my.OAI
- Humboldt-Universität zu Berlin
- InQuirion Pty Ltd, RMIT University
- Library of Congress
- NASA
- OCLC
- Old Dominion U. -- ARC, DP9
- U. of Illinois at Urbana-Champaign
- U. Of Southampton -- OAIA, CiteBase, eprints.org
- UCLA, John Hopkins U., Indiana U., NYU
- UKOLN, U. of Bath - RDN
- Virginia Tech -- repository explorer
beta phase [05/02-06/02]

• beta release on May 1st 2002 to:
  • registered data providers and service providers
  • interested parties
  • general public
• fine tuning of protocol document
• preparation for the release of 2.0 conformant tools by alpha testers

• release June 14th 2002
what’s new in OAI-PMH v.2.0
⇒ quick recap

⇒ general changes to improve solidity of protocol

⇒ corrections

⇒ new functionality
### overview of OAI Verbs

<table>
<thead>
<tr>
<th>Verb</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify</td>
<td>description of repository</td>
</tr>
<tr>
<td>ListMetadataFormats</td>
<td>metadata formats supported by repository</td>
</tr>
<tr>
<td>ListSets</td>
<td>sets defined by repository</td>
</tr>
<tr>
<td>ListIdentifiers</td>
<td>OAI unique ids contained in repository</td>
</tr>
<tr>
<td>ListRecords</td>
<td>listing of N records</td>
</tr>
<tr>
<td>GetRecord</td>
<td>listing of a single record</td>
</tr>
</tbody>
</table>

Most verbs take arguments: datestamps, sets, ids, metadata formats and resumption token (for flow control)
general changes
protocol vs periphery

- clear distinction between protocol and periphery
  - fixed protocol document
  - extensible implementation guidelines:
    - e.g. sample metadata formats, description containers, about containers
    - allows for OAI guidelines and community guidelines
OAI-PMH vs HTTP

- clear separation of OAI-PMH and HTTP
  - OAI-PMH error handling
    - all OK at HTTP level? => 200 OK
    - something wrong at OAI-PMH level? => OAI-PMH error (e.g. badVerb)
  - http codes 302, 503, etc. still available to implementers, but no longer represent OAI-PMH events
resource - item - record

set-membership is item-level property

item ~ identifier

all available metadata about David

records

record ~ identifier + metadata format + datestamp

Dublin Core metadata
MARC metadata
SPECTRUM metadata
other general changes

• better definitions of harvester, repository, item, unique identifier, record, set, selective harvesting

• oai_dc schema builds on DCMI XML Schema for unqualified Dublin Core

• usage of must, must not etc. as in RFC2119

• wording on response compression
other general changes

• all protocol responses can be validated with a single XML Schema
  
  • easier for data providers
  
• no redundancy in type definitions

• SOAP-ready

• clean for error handling
response no errors

<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH>
<responseDate>2002-0208T08:55:46Z</responseDate>
<request verb="GetRecord"... ...>http://arXiv.org/oai2</request>
<GetRecord>
  <record>
    <header>
      <identifier>oai:arXiv:cs/0112017</identifier>
      <datestamp>2001-12-14</datestamp>
      <setSpec>cs</setSpec>
      <setSpec>math</setSpec>
    </header>
    <metadata>.....</metadata>
  </record>
</GetRecord>
</OAI-PMH>

no URL encoding of the OAI-PMH request
response with error

```xml
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH>
  <responseDate>2002-02-08T08:55:46Z</responseDate>
  <request>http://arXiv.org/oai2</request>
  <error code="badVerb">ShowMe is not a valid OAI-PMH verb</error>
</OAI-PMH>
```

with errors, only the correct attributes are echoed in <request>
corrections
dates/times

• all dates/times are UTC, encoded in ISO8601, Z-notation

1957-03-20T20:30:00Z
resumptionToken

• idempotency of resumptionToken: return same incomplete list when rT is reissued
  • while no changes occur in the repo: strict
  • while changes occur in the repo: all items with unchanged datestamp

• new, optional attributes for the resumptionToken:
  • expirationDate
  • completeListSize
  • cursor
noRecordsMatch

- 1.x - if no records match, an empty list was returned
noRecordsMatch

- 2.0 - if no records match, the exception condition noRecordsMatch is returned -- not an empty list
new functionality
harvesting granularity

- harvesting granularity
  - mandatory support of YYYY-MM-DD
  - optional support of YYYY-MM-DDThh:mm:ssZ
  - granularity of from and until must be the same
Identify

- Identify more expressive

<Identify>
  <repositoryName>Library of Congress 1</repositoryName>
  <baseURL>http://memory.loc.gov/cgi-bin/oai</baseURL>
  <protocolVersion>2.0</protocolVersion>
  <adminEmail>joe@here.org</adminEmail>
  <adminEmail>jane@there.edu</adminEmail>
  <deletedRecord>transient</deletedRecord>
  <earliestDatestamp>1990-02-01T00:00:00Z</earliestDatestamp>
  <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
  <compression>deflate</compression>
**header contains set membership of item**

```xml
<record>
  <header>
    <identifier>oai:arXiv:cs/0112017</identifier>
    <datestamp>2001-12-14</datestamp>
    <setSpec>cs</setSpec>
    <setSpec>math</setSpec>
  </header>
  <metadata>
    ....
  </metadata>
</record>
```

eliminates the need for the "double harvest" 1.x required to get all records and all set information
ListIdentifiers

- ListIdentifiers returns headers

```xml
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH>
  <responseDate>2002-02-08T08:55:46Z</responseDate>
  <request verb="…" …>http://arXiv.org/oai2</request>
  <ListIdentifiers>
    <header>
      <datestamp>1999-02-23</datestamp>
      <setSpec>physic:hep</setSpec>
    </header>
    <header>
      <datestamp>1999-03-20</datestamp>
      <setSpec>physic:hep</setSpec>
      <setSpec>physic:exp</setSpec>
    </header>
  </ListIdentifiers>
```

......
ListIdentifiers

- ListIdentifiers mandates metadataPrefix as argument

http://www.perseus.tufts.edu/cgi-bin/pdataprov?

verb=ListIdentifiers

&metadataPrefix=olac

&from=2001-01-01

&until=2001-01-01

&set=Perseus:collection:PersInfo
ListIdentifiers

• the changes to ListIdentifiers are subtle, and reflect a change in the OAI-PMH data model
• Could have been named “ListHeaders” or reduced to an option for ListRecords
  - “ListIdentifiers” kept for lexicographical consistency
metadataPrefix

- character set for `metadataPrefix` and `setSpec` extended to URL-safe characters

A-Z a-z 0-9 _ ! ` $ ( ) + - . *
in the periphery
• introduction of provenance container to facilitate tracing of harvesting history

<about>
  <provenance>
    <originDescription>
      <baseURL>http://an.oa.org</baseURL>
      <identifier>oai:r1:plog/9801001</identifier>
      <datestamp>2001-08-13T13:00:02Z</datestamp>
      <metadataPrefix>oai_dc</metadataPrefix>
      <harvestDate>2001-08-15T12:01:30Z</harvestDate>
    </originDescription>
  </provenance>
</about>
introduction of friends container to facilitate web-style discovery of repositories
introduction of branding container for DPs to suggest rendering & association hints

<branding xmlns="http://www.openarchives.org/OAI/2.0/branding/
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/branding/
 http://www.openarchives.org/OAI/2.0/branding.xsd">
 <collectionIcon>
 <url>http://my.site/icon.png</url>
 <link>http://my.site/homepage.html</link>
 <title>MySite(tm)</title>
 <width>88</width>
 <height>31</height>
 </collectionIcon>
 <metadataRendering
 metadataNamespace="http://www.openarchives.org/OAI/2.0/oai_dc/
 mimeType="text/xsl">http://some.where/DCrender.xsl</metadataRendering>
 <metadataRendering
 metadataNamespace="http://another.place/MARC"
 mimeType="text/css">http://another.place/MARCrender.css</metadataRendering>
</branding>
revision of oai-identifier

<description>
  <oai-identifier xmlns="http://www.openarchives.org/OAI/2.0/oai-identifier"
                   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                   xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai-identifier
                                      http://www.openarchives.org/OAI/2.0/oai-identifier.xsd">
    <scheme>oai</scheme>
    <repositoryIdentifier>oai-stuff.foo.org</repositoryIdentifier>
    <delimiter>:</delimiter>
    <sampleIdentifier>oai:oai-stuff.foo.org:5324</sampleIdentifier>
  </oai-identifier>
</description>

domain based repository names
oai_dc

- **OAI 1.0**: oai_dc Schema defined by OAI
- **OAI 2.0**: oai_dc Schema imports from DCMI Schema for unqualified DC elements
MARC21

- OAI 1.x: oai_marc
- OAI 2.0: **LoC marxml, oai_marc**
  - http://www.loc.gov/standards/marcxml/
measures of success
⇒ registered data providers

⇒ acceptance as fundamental infrastructure for research and implementation
registered data providers
data providers highlights

- OCLC XtCat ~ thesis and dissertation
- Institute of Physics Publishing
acceptance as fundamental infrastructure

- NSDL
- Open Language Archives Community
- The European Library
- Belgian Union Catalogue
- Illinois State Union Catalogue
- CIMI
- JISC FAIR awards
- Mellon OAI-PMH service provider projects
- LOCKSS
- SPARC "institutional repository" paper
- Budapest Open Access Initiative
- JCDL & ECDL sessions on OAI-PMH
- DCMI 2002
future?
⇒ unanswered questions

⇒ OAI plans
unanswered questions

• Is OAI-PMH really low-barrier infrastructure?
  - NSDL experience indicates that significant barriers remain
  - OAI work on low-entry specs and tools
• Utility of core metadata (unqualified DC)
  - NSDL and other experience raises doubts
• Utility beyond resource discovery
  - certification, usage logs, citation data, etc.
OAI plans

return to eprints mission: work on OAI-PMH eprints profile

awareness  

interoperable grid

registration

certification

rewarding

archiving
OAI plans

return to eprints mission : work on OAI-PMH eprints profile

• e.g.
  • Specification for the exchange of references
  • Exploration of problem domain of exchange of usage log data
  • Exchange of certification metadata
  • Rights metadata
  • Others? => come to our discussion group
OAI plans

return to eprints mission: work on OAI-PMH eprints profile

• Interest from DLF and Mellon to fund the OAI to pursue this path

• Interest from NSF in the exploration of research problems related to general interoperability between eprint repositories

• Creation of OAI eprints core group: Lagoze, Van de Sompel, Nelson, Warner

  • Compile list of priorities

  • Invite relevant partners to collaborate on specific selected topics

  • Keep close contact with parties working on eprint interoperability issues related to OAI-PMH (e.g. RomEO)
questions

http://www.openarchives.org

openarchives@openarchives.org