Open Archive Initiative Protocol for Metadata Harvesting

Mariano Gardellini
Max Planck Institute for Psycholinguistic - Nijmegen
garmar@mpi.nl

CLARIN centers workshop
Prague
2009-11-05/06
Index

- OAI–PMH overview
- Verbs
- Adapting OAI–PMH to specific purposes: OLAC
- Hands–on PMH: OLAC harvester
OAI–PMH overview

- **OAI:**
  
  Open Archive Initiative
  
  The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content.

- **PMH:**
  
  Protocol for Metadata Harvesting (OAI–PMH) is a low-barrier mechanism for repository interoperability.
OAI–PMH overview

Terminology

- **Data Provider**
  It is a repository that exposes structured metadata via OAI–PMH

- **Service Provider**
  It makes OAI–PMH service requests to harvest exposed metadata and provides services among those metadata

- **Repository**
  It is a network accessible server able to process OAI–PMH requests

- **Item**
  Any atomic entry (metadata) in a repository that can be harvested in a certain format

- **Record**
  An item expressed in an XML format
OAI-PMH overview

**Data Provider**

- Repository
  - HTTP Server
  - XML transformation
  - items
    - (metadata)
  - resources

**Service Provider**

- Service Provider
  - OAI-PMH responses (metadata)
  - HTTP requests
  - harvest

**IMDI BROWSER**

- Olac corpus
  - Provider A
    - Id=X1
  - Provider B
    - Id=X2

**Internal Infrastructure**

- Service
OAI–PMH overview

- A *record* is a metadata item expressed in a specific format

- It is a response (XML byte stream) to a PMH request specifying:
  - *unique identifier* of the item
  - metadata format

- It contains the *request parameters*, a *date stamp*, the *representation* of the metadata itself and a optional XML part for rights and provenance statement (*about*)

**Minimum repository requirement:**
support the representation of items in the Dublin Core xml format
Dublin Core elements

Fifteen elements:

- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights

http://dublincore.org/documents/dces/
Verbs

Data Provider (base URL)

http://an.oa.org/OAI-script?
verb=GetRecord
&identifier=oai:arXiv.org:hep-th/9901001
&metadataPrefix=oai_dc

HTTP request

OAI-PMH response (metadata)

- Content-type = text/xml
- Status Code
- XML document well formed and valid
- UTF-8 encoding
- Date and time in ISO8601 format
General Response format

```xml
<?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>2002-05-01T19:20:30Z</responseDate>

  <request verb="GetRecord"
    identifier="oai:arXiv.org:hep-th/9901001"
    metadataPrefix="oai_dc">
    http://an oa.org/OAI-script
  </request>

  <GetRecord>
    <record> ... </record>
  </GetRecord>

</OAI-PMH>
```
XML declaration

<?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>2002-05-01T19:20:30Z</responseDate>

   <request>
      http://an.oa.org/OAI-script
   </request>

   <error code="badArgument">
      The request includes illegal arguments, is missing required arguments, includes a repeated argument, or values for arguments have an illegal syntax (metadataPrefix - value missing)
   </error>

</OAI-PMH>
Verbs

GetRecord

http://baseUrl?verb=GetRecord&metadataPrefix=oai_dc&identifier=idxyz

It returns a valid XML file containing the requested record or an error:

```
<GetRecord>
  <record>
    <header>
      <identifier>oai:arXiv.org:cs/0112017</identifier>
      <datestamp>2001-12-14</datestamp>
      <setSpec>cs</setSpec>
    </header>
    <metadata>
      ...
      <dc:title>Using Structural Metadata to Localize Experience of Digital Content</dc:title>
      <dc:creator>Dushay, Naomi</dc:creator>
      <dc:subject>Digital Libraries</dc:subject>
      ...
    </metadata>
  </record>
</GetRecord>
```

Error and Exception Conditions

- badArgument
- cannotDisseminateFormat
- idDoesNotExist
ListMetadataFormats

http://baseURL?verb=ListMetadataFormats&identifier=id_xyz

This verb is used to retrieve the metadata formats available from a repository. An optional identifier argument restricts the request to the formats available for a specific item.

<metadataFormat>
  <metadataPrefix>olac</metadataPrefix>
  <schema>http://www.language-archives.org/OLAC/olac-0.2.xsd</schema>
  <metadataNamespace>
    http://www.language-archives.org/OLAC/0.2/
  </metadataNamespace>
</metadataFormat>

Error and Exception Conditions

- badArgument
- idDoesNotExist
- noMetadataFormats - There are no metadata formats available for the specific item
Verbs

Identify

http://baseURL?verb=Identify

This verb is used to retrieve information about a repository. Some of the information returned is required as part of the OAI-PMH. Repositories may also employ the Identify verb to return additional descriptive information.

The response must include one instance of the following elements:

- repositoryName
- baseURL
- protocolVersion
- earliestDatestamp
- deletedRecord
- granularity (YYYY-MM-DD or YYYY-MM-DDThh:mm:ssZ)

The response must include one or more instances of the following element:

- adminEmail

The response may include multiple instances of the following optional elements:

- compression
- description
Verbs

ListIdentifiers

http://baseUrl?verb=ListIdentifiers&metadataPrefix=<meta_prefix>

This verb retrieves only headers rather than records.

Arguments

- **from**: specifies a lower bound for datestamp-based selective harvesting (optional)
- **until**: specifies a upper bound for datestamp-based selective harvesting (optional)
- **metadataPrefix**: specifies that headers should be returned only if the metadata format matching the supplied metadataPrefix is available or, depending on the repository's support for deletions, has been deleted (mandatory)
- **set**: and a setSpec value specifying set criteria for selective harvesting (optional)
- **resumptionToken**: an exclusive argument with a value that is the flow control token returned by a previous ListIdentifiers request that issued an incomplete list

Error and Exception Conditions

- **badArgument**
- **badResumptionToken** – The value of the resumptionToken argument is invalid or expired
- **cannotDisseminateFormat**
- **noRecordsMatch** – The combination of the values of the from, until, and set arguments results in an empty list
- **noSetHierarchy** – The repository does not support sets
Verbs

ListSets

This verb is used to retrieve the set structure of a repository

Arguments

- resumptionToken

Error and Exception Conditions

- badArgument
- badResumptionToken
- noSetHierarchy
Metadata sets

Set A
Set B
Set C

IMDI tree

Corpus nodes
Session nodes
Resources

Set A
Set B
Set C
Selective harvesting

In order to limit the results returned for ListIdentifiers and ListRecords requests, harvesters can use the following optional arguments:

- **from=x:**
  return records with a date stamp equal or newer \( x \)

- **until=x:**
  return records with a date stamp equal or older \( x \)

- **set=x:**
  return records belonging to set \( x \)
Verbs

ListRecords

This verb is used to harvest records from a repository. Optional arguments permit selective harvesting of records based on set membership and/or datestamp.

Arguments

- **from**: specifies a lower bound for datestamp-based selective harvesting (optional)
- **until**: specifies a upper bound for datestamp-based selective harvesting (optional)
- **metadataPrefix**: specifies that headers should be returned only if the metadata format matching the supplied metadataPrefix is available or, depending on the repository's support for deletions, has been deleted (mandatory)
- **set**: and a setSpec value specifying set criteria for selective harvesting (optional)
- **resumptionToken**: an *exclusive* argument with a value that is the flow control token returned by a previous ListIdentifiers request that issued an incomplete list

Error and Exception Conditions

- **badArgument**
- **badResumptionToken**
- **cannotDisseminateFormat**
- **noRecordsMatch**
- **noSetHierarchy**
# Who uses OAI-PMH

OAI website lists about 1147 data providers

### Registered Data Providers

This page lists registered OAI conforming repositories, registered through our registration and validation page. Currently there are 1038 such repositories. The table may be sorted either by the Repository Name, the base URL, or the oai-identifier namespace (if defined; used in the oai-identifier scheme). For each repository you may view the registration record from the database, or alternatively, if your browser can render XML, you may issue an Identify request to the selected repository and receive the current XML response.

Service providers can get an XML formatted list of base URLs of registered data providers from http://www.openarchives.org/registration/listFriends.

<table>
<thead>
<tr>
<th>View</th>
<th>Repository Name</th>
<th>base URL (sort)</th>
<th>oai-identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record</td>
<td>Universität Stuttgart, Fakultät 5, Germany, Computer Science Archive</td>
<td><a href="http://www.informatik.uni-stuttgart.de/cgi-bin/OAI/OAI.pl">http://www.informatik.uni-stuttgart.de/cgi-bin/OAI/OAI.pl</a></td>
<td>informatik.uni-</td>
</tr>
<tr>
<td>Record</td>
<td>E-Prints Complutense</td>
<td><a href="http://eprints.ucm.es/cgi/oai2">http://eprints.ucm.es/cgi/oai2</a></td>
<td>pub.bsalut.ne</td>
</tr>
<tr>
<td>Record</td>
<td>The University Digital Archive of the University of Groningen, The Netherlands</td>
<td><a href="http://sr.ub.rug.nl/oai/">http://sr.ub.rug.nl/oai/</a></td>
<td><a href="http://www.ucm.es">www.ucm.es</a></td>
</tr>
<tr>
<td>Record</td>
<td>11th Joint Symposium on Neural Computation</td>
<td><a href="http://jasc.library.caltech.edu/peri/oai2">http://jasc.library.caltech.edu/peri/oai2</a></td>
<td>jasc.library.ca</td>
</tr>
<tr>
<td>Record</td>
<td>@rchiveSIC- - ©HAL</td>
<td><a href="http://archivesis.ccsd.cnrs.fr/oai/oai.php">http://archivesis.ccsd.cnrs.fr/oai/oai.php</a></td>
<td>archivesis.ccs</td>
</tr>
<tr>
<td>Record</td>
<td>A Celebration of Women Writers</td>
<td><a href="http://digital.library.upenn.edu/webbin/OAI-celebration">http://digital.library.upenn.edu/webbin/OAI-celebration</a></td>
<td>celebration</td>
</tr>
<tr>
<td>Record</td>
<td>Academic Archive On-line</td>
<td><a href="http://www.diva-portal.org/oai/OAI">http://www.diva-portal.org/oai/OAI</a></td>
<td>DIVA.org</td>
</tr>
<tr>
<td>Record</td>
<td>Access to Research Resources for Teachers Space</td>
<td><a href="http://academiccommons.columbia.edu:8080/ac-oai/request">http://academiccommons.columbia.edu:8080/ac-oai/request</a></td>
<td>acervo.iteso.co</td>
</tr>
<tr>
<td>Record</td>
<td>Acervo General de la biblioteca “Dr Jorge Villalobos Padilla, S. J.” del IESO</td>
<td><a href="http://docu.gdl.iteso.mx/oai/default.aspx">http://docu.gdl.iteso.mx/oai/default.aspx</a></td>
<td>acervo.iteso.co</td>
</tr>
<tr>
<td>Record</td>
<td>Aedhe - Empresarios del Henares</td>
<td><a href="http://www.aedhe.org/oai/oai2.asp">http://www.aedhe.org/oai/oai2.asp</a></td>
<td>invenia</td>
</tr>
</tbody>
</table>
Who uses OAI-PMH

Service Providers: e.g.: Scientific Commons
OAI-PMH for your needs

- OAI–PMH has to support unqualified Dublin Core as a bare minimum

- The protocol can be used with different metadata formats as well by introducing new namespaces and providing appropriate xml schemata

For instance:

- IMDI is provided via the OAI–Bridge at the MPI

- OLAC is provided by OLAC data providers, it provides 4 elements in addition to the Dublin Core ones

- CLARIN expects CLARIN metadata from those centers that have metadata
How to deal with collections

- CLARIN decided to have collections represented by metadata records
  - the collection metadata then point to resources or other collection metadata
  - enabling “structured” collections
- OAI–PMH supports the use of “sets”
  - every resource can belong to one or more sets
  - this also allows the modeling of structured collections
  - but the set has no separate existence as a metadata record
  - so CLARIN does not support it to model CLARIN collections
- CLARIN metadata providers should create collection metadata
Collections & “Legacy” metadata

- **IMDI**
  - Provide a record for the collection
  - The IMDI metadata itself contains linking elements as “CorpusLink” pointing to the collection's constituents

- **DC/OLAC**
  - Provide a record for the collection
  - Provide “hasPart” metadata elements to specify the constituents
MPI usage of OAI-PMH

Imdi Browser (Service Provider) → OAI-2 Bridge (Data Provider) → Harvester

Imdi Archive

OLAC Importer (Harvester) → OLAC Importer (Harvester) → Data Providers

OLAC
DC
IMDI
Harvester

Data Providers
OLAC Importer

• Role in the CLARIN VLW (Virtual Language World)
  − To harvest
  − Enabling interoperability between heterogeneous formats

• Implementation overview:
  − Stand-alone application (cron job)
  − ImdiAPI ↔ Imdi specification
OLAC Importer

start

Perform pre operations and load the OLAC Providers list

For each Provider

Collect ids
Harvest records

Perform post operations

stop
OLAC Importer

start

Create the main OLAC directory (if needed)

Create the main OLAC corpus node (if needed)

Load the list of the OLAC providers

config.xml

ImdiAPI
OLAC Importer

For each Provider

Data Provider

Collect all its IDs

For each ID

GetRecord

Store|Replace record in the file system
Delete *missing* nodes

Ingest harvested nodes into the archive

stop

archive
OAI2-Bridge

- Converts IMDI records to oai_dc or olac records using XSLT technology
- Those records are offered for harvesting
- Implemented as a Java Web Application
- Refreshes its cache several times a day
Thank you for your attention