OAI Data Providers
http://gita.grainger.uiuc.edu/registry/Stanford-2006-08-24

By Thomas G. Habing
thabing@uiuc.edu
Grainger Engineering Library Information Center
University of Illinois at Urbana-Champaign
Outline

• Brief Overview of OAI-PMH
• Anatomy of an OAI Data Provider
• OAI Static Repositories
• UIUC’s OAI FileMakerPro Gateway
• Other Tools
• Validating
Overview: OAI-PMH

- [http://www.openarchives.org/](http://www.openarchives.org/)

- Technologies (RESTful Web Service)
  - HTTP
  - URIs
  - XML
  - Mostly stateless
Overview: Definitions and Concepts

• **Harvester** (client that issues OAI-PMH requests)

• **Repository** (server that responds to OAI-PMH requests)
  - Items (OAI Identifier)
    contain metadata about a resource
  - Records (OAI Identifier + Metadata Prefix)
    contain metadata in a specific format about a resource

• **Selective Harvesting**
  - Sets
  - Datestamps
    - From and Until Dates
Overview: Metadata

- Metadata
  - Dublin Core is required (oai_dc)
  - Many others (MODS, MARC, Qualified DC, etc.)
  - Adoption of richer metadata formats is highly encouraged, especially within communities
  - Can be used for complete digital resources, not just metadata
Overview: Verbs

• Find out about the repository
  - ?verb=Identify
  - ?verb=ListSets
  - ?verb=ListMetadataFormats&identifier=iii

• Harvest records
  - ?verb=ListIdentifiers&metadataPrefix=mmm&from=yyyy-mm-dd&until=yyyy-mm-dd&set=sss
  - ?verb=ListRecords&metadataPrefix=mmm&from=yyyy-mm-dd&until=yyyy-mm-dd&set=sss
  - ?verb=GetRecord&metadataPrefix=mmm&identifier=iii

Examples from the Library of Congress OAI Data Provider
Overview: Flow Control

• Resumption Tokens
  – ?verb=ListSets&resumptionToken=rrr
  – ?verb=ListIdentifiers&resumptionToken=rrr
  – ?verb=ListRecords&resumptionToken=rrr

• HTTP
  – 503 Service Unavailable (Retry-After)
Overview: HTTP

- 302 Found (Location)
- Compression
- Authentication
Anatomy of an OAI Data Provider

• How are OAI responses generated?
  – Static
    • OAI responses are fed from a static copy of your records; the static copy is periodically updated from your live data (daily, weekly, monthly, irregularly, etc.)
    • Staleness, minimal impact on your production system, may be amenable to certain turnkey solutions, easier to implement
  – Dynamic
    • OAI responses are generated directly from your live data
    • Up-to-date, may impact production system, must be tightly integrated to production system, may be difficult to implement depending on your current systems and workflows
Anatomy of an OAI Data Provider

• Where do the various components reside?
  – Locally
    • OAI data provider is on same server as the data, may be part of a larger monolithic system like DSpace or contentDM.
  – Distributed
    • OAI data provider is on different server than the data or data management system, may even be administered by a different organization.
Anatomy of an OAI Data Provider

• Options
  – Turnkey system that already has OAI-PMH capabilities built-in, such as DSpace or contentDM, plus many others. Can be limiting…
  – Start with an OAI-PMH toolkit and customize it to fit your needs, OCLC’s OAICat (Java), various toolkits from UIUC (ASP) or Virginia Tech (perl), and many others…
  – Build a data provider from scratch, not too difficult for a proficient web software developer
  – Use a gateway service, such as an OAI Static Repository Gateway, Emory’s Metadata Migrator, UIUC’s FileMakerPro and Z39.50 gateways.
OAI Static Repositories

The Problem

• OAI-PMH is simple, but not simple enough for:
  – Technically challenged organizations
    • Limited resources
    • No control over their web server
  – With small collections
    • 1-5000 records (10-20 MB XML File)
  – That do not change often
    • This is a pretty loose requirement (weekly?)
OAI Static Repositories

The Solution

• Static Repository
  – A single XML file containing all metadata, identifiers, and datestamps
  – Accessible from a web server via an HTTP URL, such as http://host:port/path/file.xml
  – May be created manually by an XML or simple text editor, or programmatically

• Static Repository Gateway
  – Provides intermediation for one or more Static Repositories
OAI Static Repositories

Official Specification

http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm
OAI Static Repositories

Static Repository Limitations

• Must be a single XML file (mime: text/xml)
  – No resumptionTokens
• Must be UTF-8 encoded Unicode
  – http://www.cs.cornell.edu/people/simeon/software/utf8conditioner/
• Must validate against Static Repository XML Schema
• The baseURL element must be the concatenation of the Static Gateway URL and the Static Repository URL
• ListRecords elements must conform to the OAI-PMH record format
OAI Static Repositories

Additional Limitations

- The URL of the Static Repository XML file cannot include a fragment or query string
- Sets are not supported
- Deleted records are not supported
- Response compression is not supported
- Only YYYY-MM-DD date stamp granularity is supported
- The guidelines for OAI identifiers should be followed:
  - http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm
OAI Static Repositories

Static Repository XML Sections

<Repository>
  <Identify>
    ...
  </Identify>
  <ListMetadataFormats>
    ...
  </ListMetadataFormats>
  <ListRecords metadataPrefix="oai_dc">
    ...
  </ListRecords>
  <ListRecords metadataPrefix="other">
    ...
  </ListRecords>
  ...
</Repository>
<Identify>
  <oai:repositoryName>Demo</oai:repositoryName>
  <oai:baseURL>
    http://myoai.org/oai/this.edu/col1/oai.xml
  </oai:baseURL>
  <oai:adminEmail>jondoe@oai.org</oai:adminEmail>
  <oai:earliestDatestamp>2002-09-19</oai:earliestDatestamp>
  <oai:deletedRecord>no</oai:deletedRecord>
  <oai:granularity>YYYY-MM-DD</oai:granularity>
</Identify>
<ListMetadataFormats>
  <ListMetadataFormats>
    <oai:metadataFormat>
      <oai:metadataPrefix>oai_dc</oai:metadataPrefix>
      <oai:schema>
        http://www.openarchives.org/OAI/2.0/oai_dc.xsd
      </oai:schema>
      <oai:metadataNamespace>
        http://www.openarchives.org/OAI/2.0/oai_dc/
      </oai:metadataNamespace>
    </oai:metadataFormat>
    ...
  </ListMetadataFormats>
</ListMetadataFormats>
OAI Static Repositories

<ListRecords>

<ListRecords metadataPrefix="oai_dc">
  <oai:record>
    <oai:header>
      <oai:identifier>oai:this.edu:123456</oai:identifier>
      <oai:datestamp>2001-12-14</oai:datestamp>
    </oai:header>
    <oai:metadata>
      <oai_dc:dc>
        <dc:title>Some Title</dc:title>
        ... 
        <dc:title>Some Title</dc:title>
      </oai_dc:dc>
    </oai:metadata>
  </oai:record>
  ...
</ListRecords>
UIUC’s OAI FileMakerPro Gateway

FileMakerPro Databases
- http://some.edu:591/FMPro?-db=artifacts&...
- http://this.org:591/FMPro?-db=collection&...

OAI FileMaker Gateway
- http://myoai.org/oai.aspx

OAI Harvesters
- OAIster
- reap
- http://myoai.org/oai.aspx/artifacts?verb=...
- http://myoai.org/oai.aspx/collection?verb=...
OAI FileMakerPro Gateway

The Problem

• FMP has widespread use in the museum community and is often used for special collections in libraries

• Until recently there are no easy or convenient tools for making FMP databases OAI accessible

• Could use Emory’s Metadata Migrator (or similar tools), but there could be latency problems if the database was active.
OAI FileMakerPro Gateway

Solution

• Out of the box, FMP has a built-in web server and can export XML

• This facilitates a solution similar to OAI Static Repositories

• Except it is not static; data is being fed directly from the database and not from a static copy
  – This is a slight fib: because of how datestamps are derived they only have a granularity of one day, so an incremental harvest might be up to 24 hours out of date
OAI FileMakerPro Gateway

Some Technical Details

How to Get XML From FMP

http://base_url:591/FMPro

?-db=database
&-lay=layout
&-format=format
&-max=max_records
&-skip=skip-records
&-recid=record_id
&-command

-lay=[short layout | full layout]
  for ListIdentifiers | ListRecords

-format=-fmp_xml
  -dso_xml (easier to transform)

-find
-dbnames
-layoutnames
-etc
OAI FileMakerPro Gateway

More Technical Details

• FMP XML Formats
  – The -dso_xml format:
    • Easier to transform with XSLT
    • But may be malformed in some cases (the gateway can accommodate this)
    • The XML Schema varies by database
    • Same as XML export format used by MS SQL Server
  – The –fmp_xml format:
    • Always the same XML Schema regardless of the database
    • Difficult to transform
OAI FileMakerPro Gateway

More Technical Details

• Datestamps
  – All FMP records have a RECORDID and a MODID
    <ROW MODID="2" RECORDID="12584941">
  – The MODID increments each time the record is changed, thus it can be used as a surrogate for the datestamp
  – When a new FMP database is added to the Gateway, all RECORDID and MODID are recorded locally, and each record is assigned the current date for the datestamp. Once a day, the MODID of each record are compared against the locally stored value, and the datestamp of the record is set to the current date if the MODID has changed.
<caribbeancovers>
  <add key="repositoryName" value="Caribbean Book Jacket Art Database"/>
  <add key="adminEmail" value="thabling@uiuc.edu"/>

  <!-- define the max records returned in one response -->
  <add key="MAX_ListIdentifiers" value='100'/>
  <add key="MAX_ListRecords" value='10'/>

  <!-- define the various components used to make an OAI identifier (i.e. oai:oai.library.uiuc.edu:illinet_online/AAA-1234) -->
  <add key="NamespaceIdentifier" value="lib.uic.edu.caribbeancovers"/>
  <add key="LocalIdentifierPath" value=""/>

  <!-- FileMaker Pro Parameters-->
  <add key="FMPBaseURL" value="http://libsys.lib.uic.edu:591/fmpro"/>
  <add key="FMPDatabase" value="caribbeancovers.fp5"/>
  <add key="FMPLayout_ListIdentifiers" value='Search'/>
  <add key="FMPLayout_ListRecords" value='Layout #1'/>

  <!-- build a local xml file containing datestamps deduced from the modid attribute -->
  <add key="FMPDatestampsFile" value="caribbeancovers.xml"/>
  <!-- the datestamp file will be updated with the following frequency in days -->
  <add key="FMPDatestampsFileUpdateFrequency" value='1'/>
  <!-- if there is a major change in the mappings, just delete the old datestamps file, and it will be rebuilt with all new dates -->

  <!-- transform the FMP Format into a DC Format -->
  <add key="FMPTransformation" value="caribbeancovers.xsl"/>
  <add key="DSOXMLIsNotWellFormed" value='1'/>
</caribbeancovers>
OAI FileMakerPro Gateway

Covert Implementations

• It is relatively easy to identify and intermediate FMP databases using the Gateway.
• Use Google to Find them:
  – http://www.google.com/search?q=allinurl%3A591+fmpro
• Gather configuration details like layouts, etc.
• Write an XSLT to transform –dso_xml into oai_dc
• Most FMP database owners probably don’t even realize how easy it is for someone to perform a wholesale download of their entire database
  – Good for OAI implementers,
  – But FMP database owners, be careful of sensitive data!!!
  – Make sure the web-based edit features are secured!!!
OAI FileMakerPro Gateway

An Invitation

http://cicharvest.grainger.uiuc.edu/fmpgateway/

• We are looking for FMP collections we can test with the Gateway

• We do plan to maintain the Gateway, similar to our OAI Static Gateway
Other OAI Gateways

• z39.50 <-> OAI-PMH
  – http://frasier.library.uiuc.edu/research.htm
  – ZMARCO http://zmarco.sourceforge.net/

• SRU/W <-> OAI-PMH
  – http://www.dlib.org/dlib/february05/sanderson/0
Open Source OAI Toolkits

• OCLC
  – http://www.oclc.org/research/projects/oai/default.htm

• UIUC Grainger Engineering Library
  – http://uilib-oai.sourceforge.net/

• Virginia Tech DLRL Projects
  – http://www.dlib.vt.edu/projects/OAI/

• Lots of other Open Source tools
  – http://sourceforge.net/search/?words=oai
OAI Turnkey Solutions


- Adlib
- CWIS
- ContentDM
- Digitool
- DLESE
- DLXS
- DSpace

- EPrints
- Encompass
- Fedora
- Greenstone
- Ockham
- Others…
How to Test Your OAI Provider

• Repository Explorer [http://re.cs.uct.ac.za/](http://re.cs.uct.ac.za/)
  – Good start, but does not do a complete harvest, nor does it check non-oai_dc metadata formats, so can’t find all problems
• W3C Validator for XML Schema [http://www.w3.org/2001/03/webdata/xsv](http://www.w3.org/2001/03/webdata/xsv)
  – Great for pinpointing obscure XML Schema validation errors or character encoding problems
  – Only one request at a time though
• Character Encoding Problems
• Try to harvest your OAI provider yourself
  – Use REAP, the Windows command line OAI harvester from UIUC
  – Use the U. Michigan Harvester (Kat can provide more detail)
• Ask one of us to do it 😊