
OAI and Metadata Harvesting

Mukesh Pund
Principal Scientist,
NISCAIR
New Delhi

Acknowledgements

- ◆ While preparing this presentation, I have used material from several sources on OAI-PMH by other authors
- ◆ I gratefully acknowledge these sources

Digital Repositories: Current Situation

- ◆ Mushrooming number and variety of distributed digital repositories (archives, digital libraries)
- ◆ Use of variety of hardware, software, database solutions
- ◆ Use of different search and retrieval interfaces
- ◆ Most of the content are not indexed by web search engines
- ◆ Content resides in backend databases – not picked up by web search engines

Problems faced by Users

- ◆ How to identify and retrieve relevant information from different repositories?
- ◆ Visiting and searching individual repositories is very expensive
- ◆ Key Requirement: How do we support cross searching?

Current Solutions

- ◆ Federated/ distributed searching
 - Z39.50 Information Retrieval protocol
- ◆ Metadata harvesting
 - OAI-PMH protocol

Federated/ distributed searching

- ◆ Protocol: "Information Retrieval (Z39.50): Application Service Definition and Protocol Specification", (ISO/ ANSI standard) (v1-1991, v2-1992, v3-1995)
- ◆ Client-Server model (TCP/IP Service)
- ◆ Process:
 - Client ('Origin') sends queries, formatted according to Z39.50, to repository Server ("Target").
 - Server translates this to local query format, searches the database, sends the results to the client, formatted according to Z39.50
 - Client translates the results and presents it to the user
- ◆ Client can send queries to as many related z39.50 compliant servers as possible

Z39.50 protocol ...

- ◆ Example implementation: Distributed searching of library catalogues/ bibliographic databases
- ◆ Problem - performance
 - Implementation not easy
 - Does not scale well (if nodes > 100)
 - Network bandwidth
 - Z39.50 implementation at client (“Origin”) end
- ◆ Z39.50 resources:
<http://lcweb.loc.gov/z3950/agency/> (Z39.50 International Maintenance Agency, Library of Congress)

OAI-PMH Vs. Z39.50

- ◆ OAI-PMH: Indexed Search much similar to general search Engines. Requires Service Providers and data providers
- ◆ Z39.50: Concurrent Search, No service providers only data providers

OAI-PMH

- ◆ **Open Archive Initiative-Protocol for Metadata Harvesting**
- ◆ **Protocol Version 2.0 of 2002-06-14**
<http://www.openarchives.org>


Open Archives Initiative (OAI)



The protocol is openly documented, and metadata is “exposed” to at least some peer group (note: rights management can still apply!)

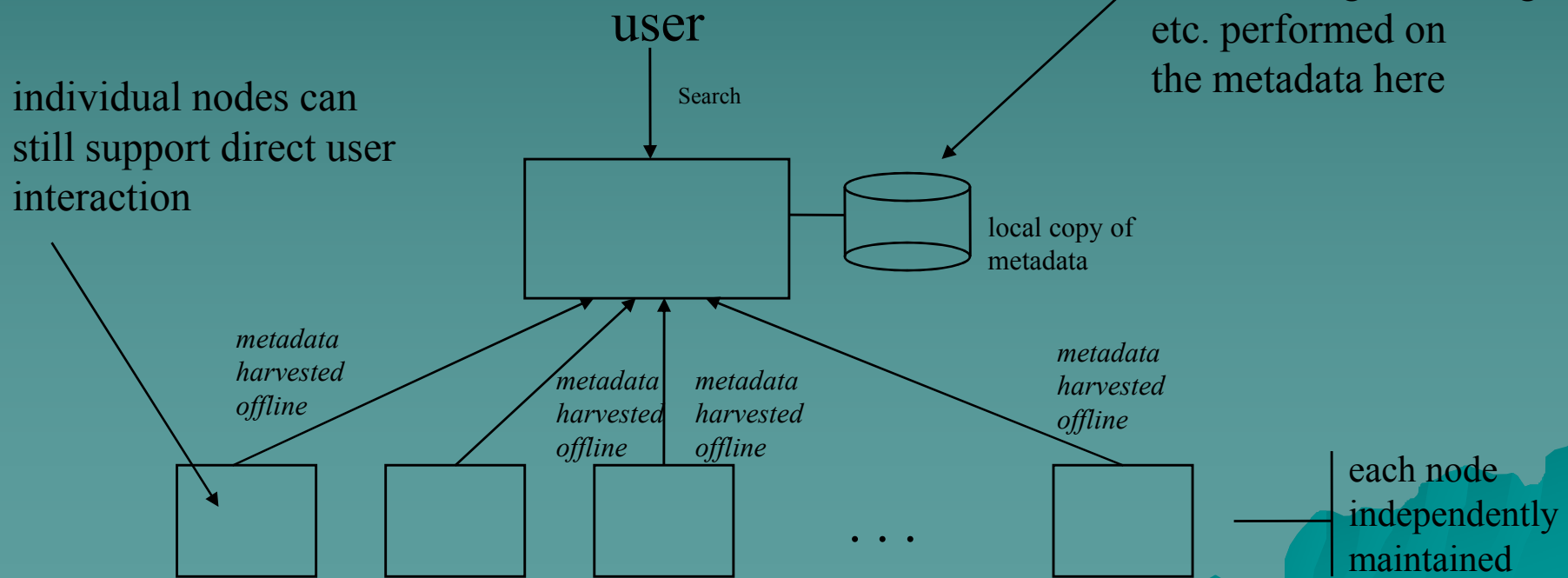
Archive defined as a dynamic “collection of stuff” -- not the archivist’s definition of “archive”. “Repository” used in most OAI documents.

OAI is happening at break-neck speed...



Metadata Harvesting

- ◆ Move away from distributed searching (e.g., Z39.50)
- ◆ Extract metadata from various sources
- ◆ Build services on local copies of metadata
 - Resources remain at remote repositories

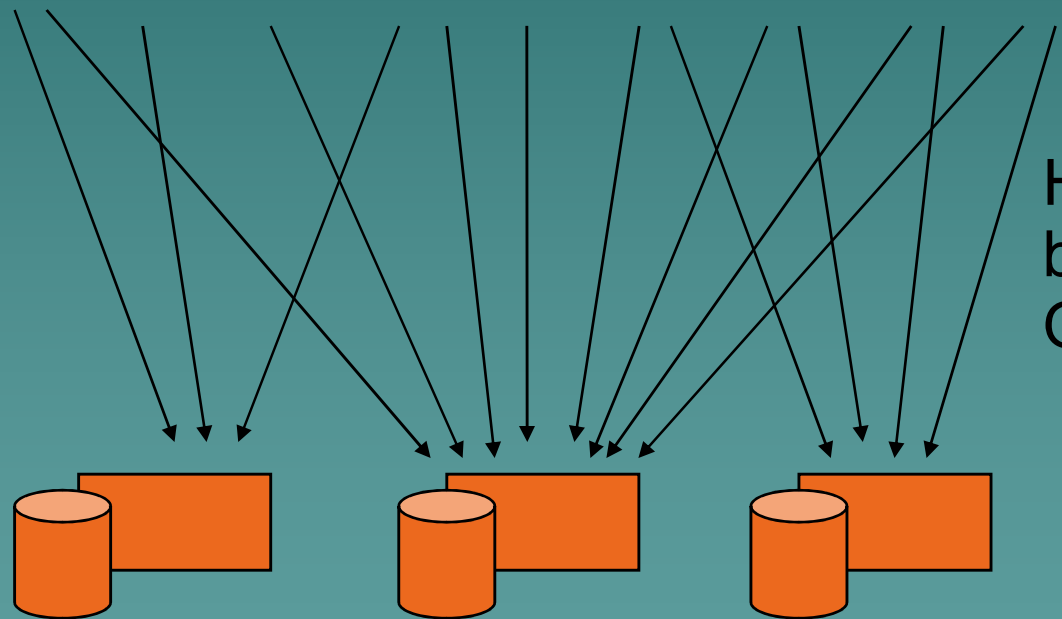


Data and Service Providers

- ◆ Data Provider
 - Creators and keepers of the metadata as well as repositories of resources
 - Give free access of metadata (not necessarily: free access to full texts / resources)
- ◆ Service Provider
 - Harvest and store metadata (no live requests!)
 - May select certain subsets from Data Providers (set hierarchy, date stamp) for selective harvesting
 - May enrich metadata
 - Offer (value-added) service on the basis of the metadata
- ◆ One 'service' can play both roles (Aggregators)

Multiple Data and Service Providers

Data providers

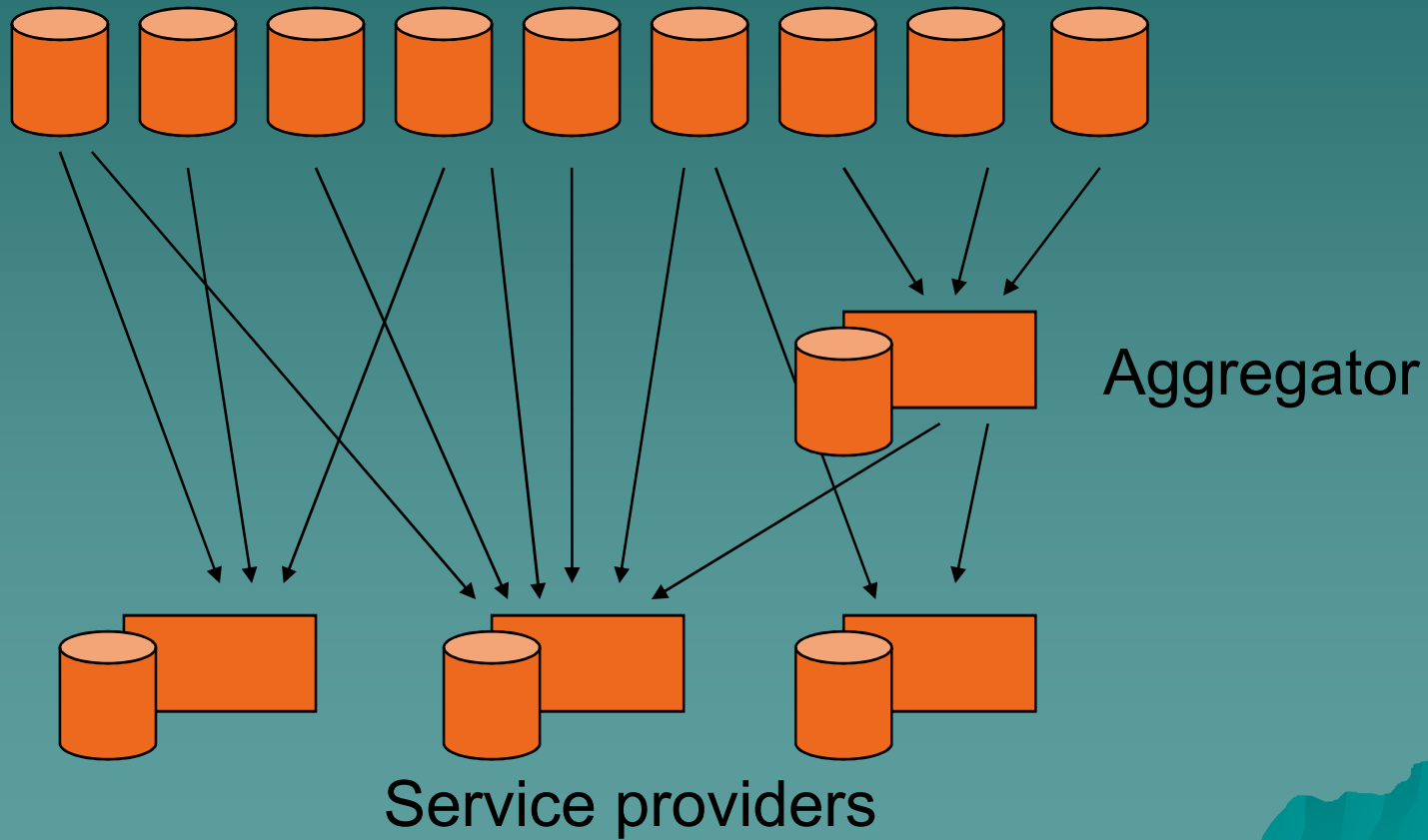


Harvesting
based on
OAI-PMH

Service providers

Aggregators

Data providers



OAI-PMH v.2.0 [06/2002]

- ◆ Low-barrier interoperability specification
- ◆ Metadata harvesting model: data provider / service provider
- ◆ Metadata about resources
- ◆ Autonomous protocol
- ◆ Not a search protocol!
- ◆ HTTP based
- ◆ XML responses
- ◆ Unqualified Dublin Core
- ◆ Stable: backward compatible

OAI Data Model:

Resources / Items / Records



← resource

item = identifier

all available metadata
about *Mona Lisa*

← item

Dublin Core
metadata

MARC
metadata

SPECTRUM
metadata

← records

record = identifier + metadata format + datestamp

Harvesting: How it works

Six OAI “Verbs”

Identify

ListMetadataFormats

ListSets

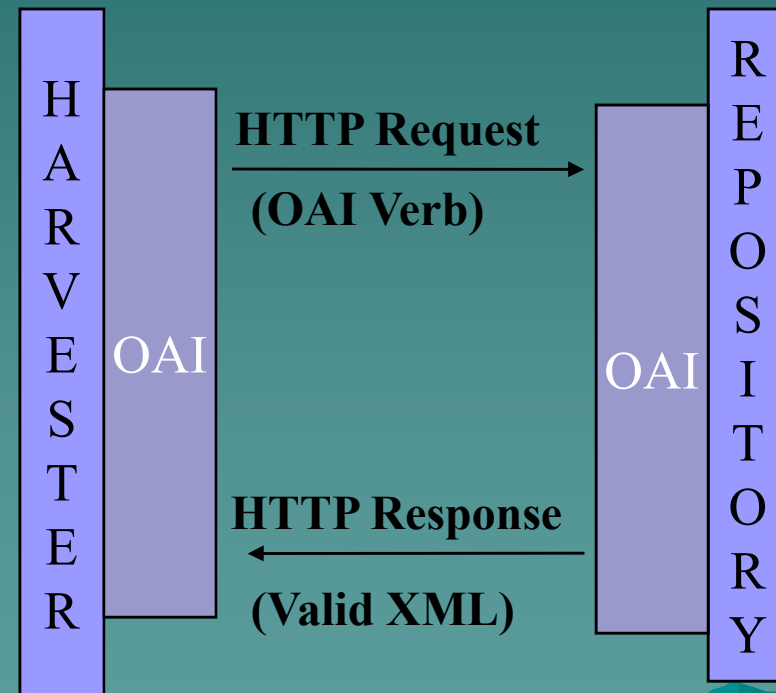
ListIdentifiers

ListRecords

GetRecord

Service Provider

Metadata Provider



Harvester

- ◆ A harvester is a client application that issues OAI-PMH requests.
- ◆ A harvester is operated by a service provider as a means of collecting metadata from repositories ■

Repository

- ◆ A repository is a network accessible server that can process the OAI-PMH requests.
- ◆ A repository is managed by a data provider to expose metadata to harvesters

Resource

- ◆ A resource is the object or "stuff" that metadata is "about". The nature of a resource, whether it is physical or digital, or whether it is stored in the repository or is a constituent of another database, is outside the scope of the OAI-PMH

Item

- ◆ An item is a constituent of a repository from which metadata about a resource can be disseminated.
- ◆ That metadata may be disseminated on-the-fly from the associated resource, cross-walked from some canonical form, actually stored in the repository, etc.

Record

- ◆ A record is metadata in a specific metadata format.
- ◆ A record is returned as an XML-encoded byte stream in response to a protocol request to disseminate a specific metadata format from a constituent item.

Unique Identifier

- ◆ A unique identifier unambiguously identifies an item within a repository
- ◆ The unique identifier is used in OAI-PMH requests for extracting metadata from the item.

cont...

Unique Identifier

- ◆ The format of the unique identifier must correspond to that of the URI (Uniform Resource Identifier) syntax
- ◆ Repositories may implement the oai-identifier

Role of Identifier

- ◆ Unique identifiers play two roles in the protocol:
- ◆ Response: Identifiers are returned by both the ListIdentifiers and ListRecords requests.
- ◆ Request: An identifier, in combination with a metadataPrefix , is used in the GetRecord request as a means of requesting a record in a specific metadata format from an item

OAI-PMH Verbs

- ◆ Identify
- ◆ ListSets
- ◆ ListMetadataFormats
- ◆ ListIdentifiers
- ◆ GetRecord
- ◆ ListRecords

Identify

- ◆ Returns general information about the:
 - ◆ Archive and its policies
 - ◆ Datestamp
 - ◆ Granularity

- ◆ Ex:

<http://192.168.0.12/dspace-oai/request?verb=Identify>

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Home Mail Print Send To Favorites Home Mail Print Send To Favorites Home Mail Print Send To

Address <http://192.168.0.12/dspace-oai/request?verb=Identify> Go

Y! Search Web My Web Mail My Yahoo! Shopping Games Music Persona

Google Go Bookmarks 0 blocked Check AutoLink AutoFill Send to

```
<?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>2008-02-13T05:19:43Z</responseDate>
  <request verb="Identify">http://192.168.0.12/dspace-oai/request</request>
- <Identify>
  <repositoryName>National Science Digital Library</repositoryName>
  <baseURL>http://192.168.0.12/dspace-oai/request</baseURL>
  <protocolVersion>2.0</protocolVersion>
  <adminEmail>mukeshpund@niscair.res.in</adminEmail>
  <earliestDatestamp>2001-01-01T00:00:00Z</earliestDatestamp>
  <deletedRecord>persistent</deletedRecord>
  <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
  <compression>gzip</compression>
  <compression>deflate</compression>
- <description>
- <toolkit xsi:schemaLocation="http://oai.dlib.vt.edu/OAI/metadata/toolkit
  http://oai.dlib.vt.edu/OAI/metadata/toolkit.xsd"
  xmlns="http://oai.dlib.vt.edu/OAI/metadata/toolkit">
  <title>OCLC's OAI-Cat Repository Framework</title>
```

ListSets

- ◆ Provide a listing of sets in which records may be organized (may be hierarchical, overlapping, or flat)
- ◆ Example:
- ◆ <http://192.168.0.12/dspace-oai/request?verb=ListSets>

http://192.168.0.12/dspace-oai/request?verb=ListSets - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Print Mail New Tab

Address http://192.168.0.12/dspace-oai/request?verb=ListSets Go

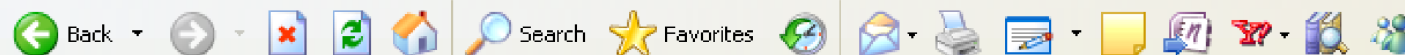
Search Web My Web Mail My Yahoo! Shopping Games Music Persona

Google Go Bookmarks 0 blocked Check AutoLink AutoFill Send to

```
<?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>2008-02-13T05:34:11Z</responseDate>
  <request verb="ListSets">http://192.168.0.12/dspace-oai/request</request>
- <ListSets>
- <set>
  <setSpec>hdl_123456789_2</setSpec>
  <setName>2007</setName>
</set>
- <set>
  <setSpec>hdl_123456789_9</setSpec>
  <setName>book</setName>
</set>
- <set>
  <setSpec>hdl_123456789_12</setSpec>
  <setName>dnms book</setName>
</set>
- <set>
  <setSpec>hdl_123456789_26</setSpec>
```

ListMetadataFormats

- ◆ Lists metadata formats supported by the archive as well as their schema locations and namespaces
- ◆ Example:
- ◆ <http://192.168.0.12/dspace-oai/request?verb=ListMetadataFormats>



```
<?xml version="1.0" encoding="UTF-8" ?>
```

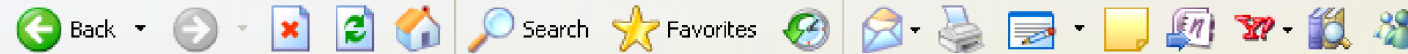
```
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"  
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"  
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/  
    http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">  
  <responseDate>2008-02-13T05:42:22Z</responseDate>  
  <request verb="ListMetadataFormats">http://192.168.0.12/dspace-  
    oai/request</request>  
  - <ListMetadataFormats>  
  - <metadataFormat>  
    <metadataPrefix>oai_dc</metadataPrefix>  
    <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd</schema>  
  
    <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadat  
  </metadataFormat>  
</ListMetadataFormats>  
</OAI-PMH>
```


ListIdentifiers

- ◆ List headers for all items corresponding to the specified parameters
- ◆ http://192.168.0.12/dspace-oai/request?verb=ListIdentifiers&metadataPrefix=oai_dc

http://192.168.0.12/dspace-oai/request?verb=ListIdentifiers&metadataPrefix=oai_dc - Microsoft Internet Explorer

File Edit View Favorites Tools Help



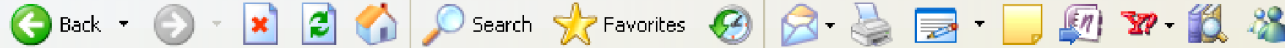
Address http://192.168.0.12/dspace-oai/request?verb=ListIdentifiers&metadataPrefix=oai_dc Go



```
<?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>2008-02-13T05:58:21Z</responseDate>
  <request metadataPrefix="oai_dc"
    verb="ListIdentifiers">http://192.168.0.12/dspace-oai/request</request>
- <ListIdentifiers>
- <header>
  <identifier>oai:192.168.0.12:123456789/3</identifier>
  <timestamp>2007-08-31T10:29:07Z</timestamp>
  <setSpec>hdl_123456789_2</setSpec>
</header>
- <header>
  <identifier>oai:192.168.0.12:123456789/10</identifier>
  <timestamp>2008-01-22T06:00:03Z</timestamp>
  <setSpec>hdl_123456789_9</setSpec>
</header>
- <header>
  <identifier>oai:192.168.0.12:123456789/13</identifier>
  <timestamp>2008-01-22T06:00:03Z</timestamp>
```

GetRecord

- ◆ Returns the metadata for a single item in the form of an OAI record
- ◆ Example:
- ◆ http://192.168.0.12/oai/request?verb=GetRecord&identifier=oai:192.168.0.12:123456789/3&metadataPrefix=oai_dc



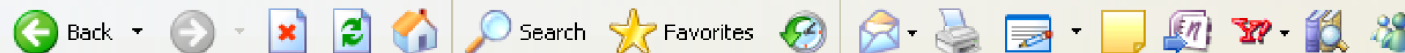
Address http://192.168.0.12/dspace-oai/request?verb=GetRecord&identifier=oai:192.168.0.12:123456789/3&metadataPrefix=oai_dc Go Links >>

Search Web Search Web My Web Mail My Yahoo! Shopping Games Music Personals >>
Go Bookmarks 0 blocked Check AutoLink AutoFill Send to Settings >>

```
<?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>2008-02-13T07:10:32Z</responseDate>
  <request identifier="oai:192.168.0.12:123456789/3" metadataPrefix="oai_dc" verb="GetRecord">http://192.168.0.12/dspace-
    oai/request</request>
  - <GetRecord>
  - <record>
    - <header>
      <identifier>oai:192.168.0.12:123456789/3</identifier>
      <timestamp>2007-08-31T10:29:07Z</timestamp>
      <setSpec>hdl_123456789_2</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:creator>Mukesh, Pund</dc:creator>
      <dc:date>2007-08-31T10:27:13Z</dc:date>
      <dc:date>2007-08-31T10:27:13Z</dc:date>
      <dc:date>2007-08-31T10:27:13Z</dc:date>
      <dc:identifier>http://hdl.handle.net/123456789/3</dc:identifier>
      <dc:format>436 bytes</dc:format>
      <dc:format>text/html</dc:format>
      <dc:language>en</dc:language>
      <dc:title>Test Document</dc:title>
      <dc:type>Book chapter</dc:type>
    </oai_dc:dc>
  </metadata>
</record>
</GetRecord>
</OAI-PMH>
```

ListRecords

- ◆ Retrieves metadata records for multiple items
- ◆ http://192.168.0.12/dspace-oai/request?verb=ListRecords&metadataPrefix=oai_dc



Address http://192.168.0.12/dspace-oai/request?verb=ListRecords&metadataPrefix=oai_dc&from=2002-12-01 Go

Search Web My Web Mail My Yahoo! Shopping Games Music Persona

Google Go Bookmarks 0 blocked Check AutoLink AutoFill Send to

Click this button to always allow popups on 192.168.0.12
To let an individual popup through, press the 'Ctrl' key while clicking on a link.

```
<?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>2008-02-13T05:50:22Z</responseDate>
  <request metadataPrefix="oai_dc" verb="ListRecords" from="2002-12-01">http://192.168.0.12/dspace-oai/request</request>
- <ListRecords>
- <record>
+ <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:creator>Mukesh, Pund</dc:creator>
  <dc:date>2007-08-31T10:27:13Z</dc:date>
  <dc:date>2007-08-31T10:27:13Z</dc:date>
  <dc:date>2007-08-31T10:27:13Z</dc:date>
  <dc:identifier>http://hdl.handle.net/123456789/3</dc:identifier>
  <dc:format>436 bytes</dc:format>
  <dc:format>text/html</dc:format>
  <dc:language>en</dc:language>
  <dc:title>Test Document</dc:title>
  <dc:type>Book chapter</dc:type>
  </oai_dc:dc>
</metadata>
</record>
- <record>
- <header>
  <identifier>oai:192.168.0.12:123456789/10</identifier>
  <timestamp>2008-01-22T06:00:03Z</timestamp>
  <setSpec>hdl_123456789_9</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">
```

ListIdentifiers

- ◆ To get a list of identifiers
- ◆ `http://192.168.0.12/oai/request?verb=ListIdentifiers&metadataPrefix=oai_dc&from=2002-12-01`

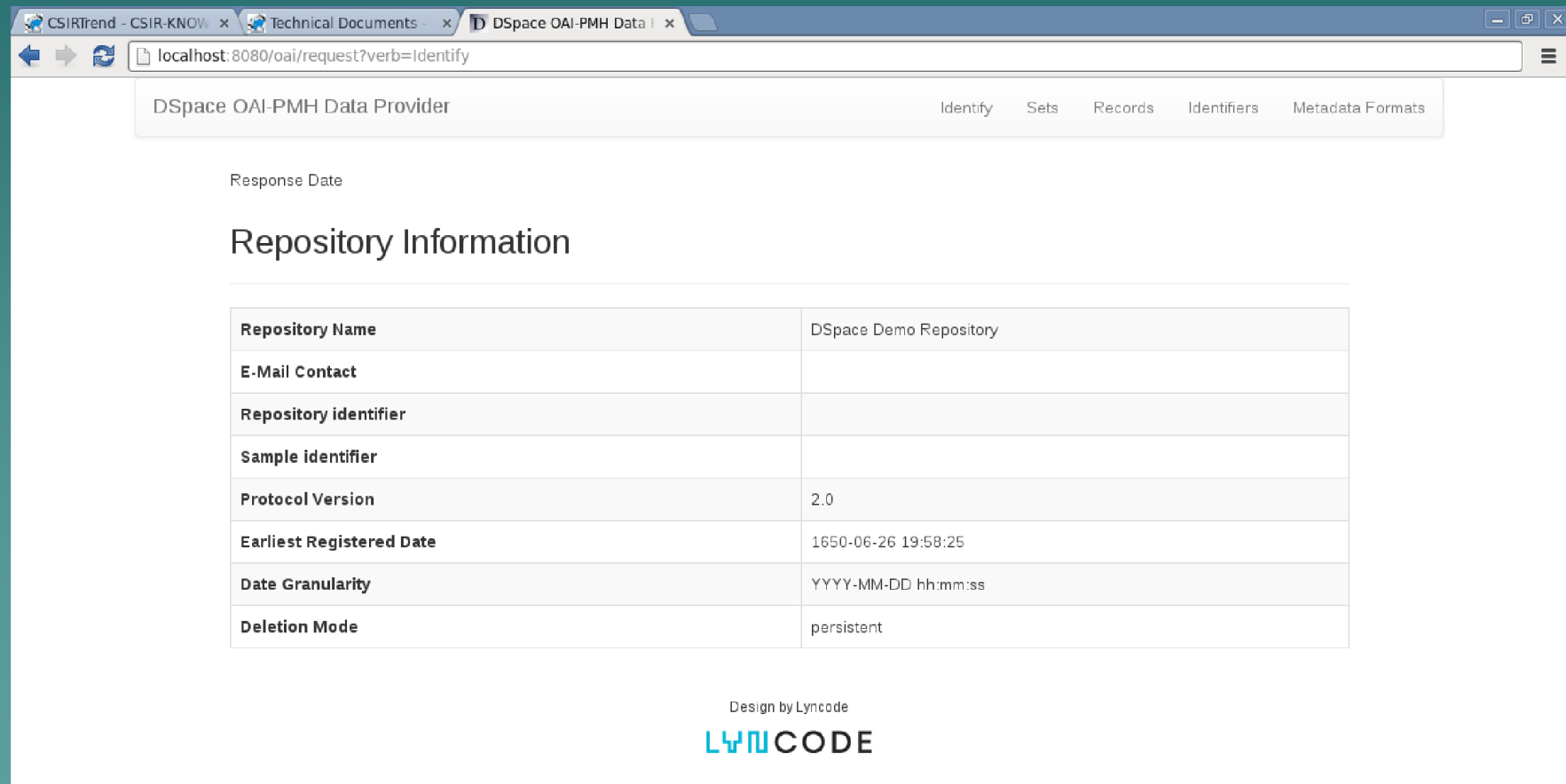


```
<?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2008-02-13T07:15:01Z</responseDate>
  <request metadataPrefix="oai_dc" verb="ListIdentifiers" from="2002-12-01">http://192.168.0.12/dspace-oai/request</request>
- <ListIdentifiers>
- <header>
  <identifier>oai:192.168.0.12:123456789/3</identifier>
  <timestamp>2007-08-31T10:29:07Z</timestamp>
  <setSpec>hdl_123456789_2</setSpec>
</header>
- <header>
  <identifier>oai:192.168.0.12:123456789/10</identifier>
  <timestamp>2008-01-22T06:00:03Z</timestamp>
  <setSpec>hdl_123456789_9</setSpec>
</header>
- <header>
  <identifier>oai:192.168.0.12:123456789/13</identifier>
  <timestamp>2008-01-22T06:00:03Z</timestamp>
  <setSpec>hdl_123456789_12</setSpec>
</header>
- <header>
  <identifier>oai:192.168.0.12:123456789/27</identifier>
  <timestamp>2008-01-24T08:59:47Z</timestamp>
  <setSpec>hdl_123456789_26</setSpec>
</header>
- <header>
  <identifier>oai:192.168.0.12:123456789/30</identifier>
  <timestamp>2008-01-24T09:08:32Z</timestamp>
  <setSpec>hdl_123456789_29</setSpec>
</header>
- <header>
  <identifier>oai:192.168.0.12:123456789/32</identifier>
  <timestamp>2008-01-24T10:00:20Z</timestamp>
  <setSpec>hdl_123456789_31</setSpec>
</header>
- <header>
  <identifier>oai:192.168.0.12:123456789/34</identifier>
```


Selective Harvesting

- ◆ By date
- ◆ &from=2002-12-01 OR
- ◆ &from=2002-12-01&until=2003-12-01
- ◆ By set (collection in Dspace)
- ◆ &set=hdl_1849_2

OAI-PMH user interface for Dspace – 5.0



The screenshot shows a web browser window with the URL `localhost:8080/oai/request?verb=Identify`. The page title is "DSpace OAI-PMH Data Provider". A navigation bar contains links for "Identify", "Sets", "Records", "Identifiers", and "Metadata Formats". Below the navigation bar, the text "Response Date" is visible. The main section is titled "Repository Information" and contains a table with the following data:

Repository Name	DSpace Demo Repository
E-Mail Contact	
Repository identifier	
Sample identifier	
Protocol Version	2.0
Earliest Registered Date	1650-06-26 19:58:25
Date Granularity	YYYY-MM-DD hh:mm:ss
Deletion Mode	persistent

Design by Lyncode
LYNCODE

Useful Sites

- ◆ OAI-PMH Official Site:
 - <http://www.openarchives.org/>
- ◆ Testing your OAI-PMH compatibility
 - <http://oai.dlib.vt.edu/cgi-bin/Explorer/2.0-1.45/testoai>
- ◆ Registering your Digital Repository
 - <http://www.openarchives.org/data/registerasprovider.html>

OAI Service Provider Software (Harvesters)

- ◆ PKP Harvester:
 - University of British Columbia, Canada
 - <http://www.pkp.ubc.ca/pkp-harvester/>
- ◆ DLESE
 - Digital Library for Earth System Education
 - <http://sourceforge.net/projects/dlese-oai/>
- ◆ ARC
 - Old Dominion University, Virginia
 - <http://arc.cs.odu.edu/>

OAI Data Provider Software

- ◆ OAICat
 - OCLC
 - <http://www.oclc.org/research/software/oai/catt.htm>
- ◆ DLESE
 - Digital Library for Earth System Education
 - <http://sourceforge.net/projects/dlese-oai-dfs>

How do baseURLs look like

- ◆ DSpace repositories

- [NSDL : 202.54.99.9/dspace](http://202.54.99.9/dspace)
- <http://202.54.99.9/dspace-oai/request>

OAI Tools

- ◆ <http://www.openarchives.org/tools/tools.html>

*Thank
You*