

DLF Forum 2006
Sharing resources by collection:
OAI sets and set descriptions

Muriel Foulonneau (mfoulonn@uiuc.edu)
Caroline Arms (caar@loc.gov)
Sarah L. Shreeves (sshreeve@uiuc.edu)

Grainger Engineering Library, University of Illinois at Urbana-Champaign
Library of Congress

April 2006 Austin



An overview of existing practices



What is an OAI set?

- Optional means to organize OAI repository
 - Sets may be hierarchical and may overlap
 - Items may belong to multiple sets or no sets
- Used by service providers to selectively harvest data providers
- Widely used (74% of repositories have sets)
- Sets **must** have setSpec and setName and **may** have a setDescription



Snapshot of Current Practice

Highest number of sets: 4806*

Lowest number of sets: 1*

Median: 14*

Average number of sets: 102*

*Skewed by empty sets (491 repos have empty sets)

Highest number of empty sets: 4610

Lowest number of empty sets: 1

Median: 12

Average number of sets: 2



What do sets represent?

**Journals:
issues**

**Institutional
repositories:
Departments,
research centers, etc.**

**EPrint Archives:
Subject,
Publication Status**

**Cultural Heritage
Repositories:
Collections
with Intent**

**Set
representations
may be
constrained by
the software
package used.**

How should <setDescription> be used?

■ OAI Protocol implementation guidelines:

(<http://www.openarchives.org/OAI/2.0/guidelines-repository.htm#setDescription>)

“Set descriptions may be included in the setDescription of the ListSets response. If sets correspond to some notion of "collections" then this allows collection description. If the whole repository represents one collection then it may be more appropriate to use the description container in the Identify response to describe the collection.”

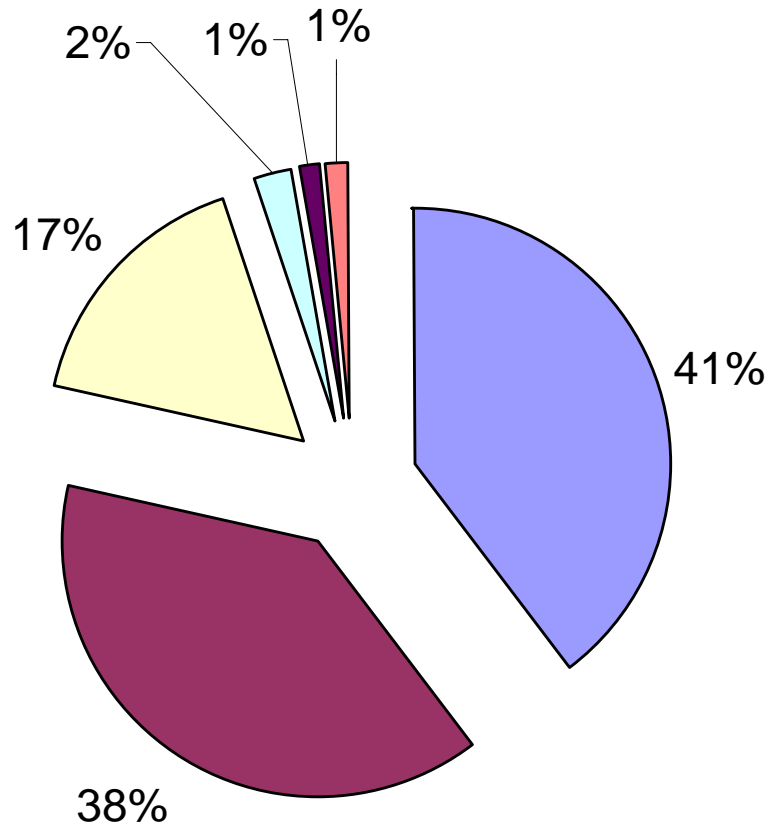
■ OAI Best Practices for Data Provider Implementations

(<http://oai-best.comm.nsd.gov/cgi-bin/wiki.pl?SetPractices>)



Set Description – Current Practice

143 (20%) of OAI repositories with sets include a <setDescription> element



- Single dc:description of collection/repository/set
- URL for a journal issue in dc:description only
- Empty <setDescription />
- Title/Publisher of collection and an identifier
- dc:description and at least one other element
- Single element (not description)

Premise:

There is a difference between the description of the OAI set as a collection of metadata records and technical construct and the description of the collection of resources represented by the metadata records in the set?

- Can you do both in the <setDescription>?
- Do you want to do both in the <setDescription>?



***Usage of OAI sets
Perspective of a Data Provider: the
Library of Congress***



Library of Congress Digitized Historical Collections

- Most of the content digitized is in "special" formats, not published, printed volumes.
- Physical collections organized along varied lines:
 - Provenance (creator, collector)
 - William Brumfield; Hans & Hanni Kraus,
 - Category of content (type of resource, genre)
 - Daguerreotypes; sheet music
 - Theme of high user interest
 - Pictures of presidents; civil war drawings



Responsibility for digital "collections"

- LC practice aligns responsibility for digitized collections with curatorial responsibility for original content.
- Responsibility for description also aligned with curatorial responsibility
 - Collection-level description is routinely done
- Aligning online collections with source collections has advantages
 - Description responsibilities coincide
 - Domain expertise coincides
- More themed, virtual assemblies online, but description and curatorial responsibilities still aligned for items
 - Collection-level records are created for these online-only "collections"



Library of Congress: OAI sets

- Aligned with record sets that support online collections
 - Tied to routine update process
 - Can take advantage of collection-level description
- Primarily by content category
 - Seemed to correspond to the requests received for bulk access
- Provide finer set structure where feasible





OAI-harvestable records

Home

- [Books](#)
- [Ephemera, Pamphlets](#)
- [Maps, Atlases](#)
- [Photos](#)
- [Posters](#)
- [Other Still Visual](#)
- [Motion Pictures](#)
- [Sheet Music](#)

➤ [Sample OAI-PMH requests](#)

➤ [OAI-PMH related resources](#)

Learn More

➤ About OAI and the protocol: www.openarchives.org.

➤ About [this implementation](#).

OAI-harvestable records for digitized historical collections

Last Updated: March 23, 2006

Item-level metadata for selected Library of Congress collections of historical materials is available for harvesting using Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) version 2.0. Available are records from selected collections presented online through [American Memory](#), [Global Gateway](#) and the [Prints & Photographs Division Online Catalog](#).

For an article about how and why the Library of Congress implemented OAI-PMH, see [Available and Useful: OAI at the Library of Congress, 2003](#).

The baseURL for harvesting these records is http://memory.loc.gov/cgi-bin/oai2_0. To learn more about the Open Archives Initiative (OAI) and the harvesting protocol, see www.openarchives.org.

|| Records are harvestable as sets organized by content type ||

- [Books](#) (11 individual sets, combined set)
- [Ephemera, Pamphlets](#) (1 set)
- [Maps, Atlases](#) (1 set)
- [Photos](#) (26 individual sets, combined set)
- [Posters](#) (2 individual sets, combined set)
- [Other Still Visual](#) (4 individual sets, combined set)
- [Motion Pictures](#) (1 set)
- [Sheet Music](#) (1 set)

For an article about how and why the Library of Congress implemented OAI-PMH, see [Available and Useful: OAI at the Library of Congress](#), 2003.

The baseURL for harvesting these records is http://memory.loc.gov/cgi-bin/oai2_0. To learn more about the Open Archives Initiative (OAI) and the harvesting protocol, see www.openarchives.org.

|| Records are harvestable as sets organized by content type ||

- [Books](#) (11 individual sets, combined set)
- [Ephemera, Pamphlets](#) (1 set)
- [Maps, Atlases](#) (1 set)
- [Photos](#) (26 individual sets, combined set)
- [Posters](#) (2 individual sets, combined set)
- [Other Still Visual](#) (4 individual sets, combined set)
- [Motion Pictures](#) (1 set)
- [Sheet Music](#) (1 set)

Sets vs. Collections

- A set is a "collection" of records
- Harvesters want information about the underlying content
 - Scope (description), subject terms, coverage, type of resource, rights for content, link to online presentation of collection
- **and** about the set of records
 - What entity is responsible for the set?
 - Is the set being added to or updated routinely?
 - E.g., because more is being described /digitized
 - What quality of cataloging to expect?
 - What is the native format for the records?
 - How are non-native formats derived?
 - How does this set relate to other sets?



Possible solutions

- For unambiguous machine–parsable semantics
 - Develop an XML Schema that allows embedding of collection description inside a set description
- For simplicity and for human readers
 - Use oai_dc schema
 - Add second Dublin Core <description> element
- Develop community conventions for using other schemas in predictable ways
 - e.g., MODS for DLF Aquifer
- Etc., ...
- Meanwhile, taking the easy approach



Sample set description using oai_dc schema

```
<setDescription>
<oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title xml:lang="en">Records for California As I Saw It (books)</dc:title>
<dc:creator>Library of Congress</dc:creator>
<dc:description>Transcribed text with illustrations of 186 works documenting the formative era of California's history
through eyewitness accounts. The collection covers the dramatic decades between the Gold Rush and the turn of the
twentieth century. It captures the pioneer experience; encounters between Anglo-Americans and the diverse peoples who
had preceded them; the transformation of the land by mining, ranching, agriculture, and urban development; the often-
turbulent growth of communities and cities; and California's emergence as both a state and a place of uniquely American
dreams.</dc:description>
<dc:description>Set characteristics for calbkbib: Source records are MARC (from LC catalog); MODS or oai_dc records
are dynamically generated using generic transformation when harvested.  dct:accrualPolicy: Closed. Contains about 200
records. Records in set calbkbib are also in set lcbooks.</dc:description>
<dc:type xml:lang="en">text</dc:type>
<dc:type xml:lang="en">books</dc:type>
<dc:type xml:lang="en">printed materials</dc:type>
<dc:type xml:lang="en">collection</dc:type>
<dc:coverage xml:lang="en">1849-1900</dc:coverage>
<dc:coverage xml:lang="en">California</dc:coverage>
<dc:subject xml:lang="en">Ethnic groups--California.</dc:subject>
<dc:subject xml:lang="en">Law and politics--California.</dc:subject>
<dc:subject xml:lang="en">California--History.</dc:subject>
<dc:subject xml:lang="en">California--Biography.</dc:subject>
<dc:subject xml:lang="en">California--Gold discoveries.</dc:subject>
<dc:contributor xml:lang="en">Library of Congress, General Collections</dc:contributor>
<dc:rights xml:lang="en">http://memory.loc.gov/ammem/cbhtml/cbres.html</dc:rights>
<dc:relation xml:lang="en">http://memory.loc.gov/ammem/cbhtml/cbhome.html</dc:relation>
</oai_dc:dc>
</setDescription>
```

Two description elements

For content (underlying collection)

<dc:description>

Transcribed text with illustrations of 186 works documenting the formative era of California's history through eyewitness accounts. The collection covers the dramatic decades between the Gold Rush and the turn of the twentieth century. ...

</dc:description>

For set (collection of records)

<dc:description>

Set characteristics for calbkbib: Source records are MARC (from LC catalog); MODS or oai_dc records are dynamically generated using generic transformation when harvested. dct:accrualPolicy: Closed. Contains about 200 records. Records in set calbkbib are also in set lcbooks.

</dc:description>



Dublin Core collection description elements

AccrualPeriodicity

Element Description: The frequency with which items are added to a collection.

Guidelines for content creation:

Terms from controlled vocabularies may be developed for the use of a particular project or in general use in a cultural materials context.

Examples: Annual, Irregular

AccrualPolicy

Element Description: The policy governing the addition of items to a collection.

Examples: Active, Closed

Note: These can have different values for set and source collection. LC has many large image collections for which cataloging and digitization are proceeding over years



***Usage of OAI sets
Perspective from a Service Provider:
CIC Metadata Portal***




The CIC metadata portal and OAI sets

- Selective harvesting in some cases
 - Because sets not part of collection development policy
 - Because harvest in multiple chunks
 - E.g., when sets harvested in multiple formats
 - E.g., when large metadata collections
- Including the concept of collections in the item level portal
 - Every item in the CIC collection belongs to one “reference” collection such as defined by the data provider (either OAI repository or OAI set)



Filtering results, adding context, browsing

CIC Search CIC Library Resources

Click [here](#) to install the Firefox plug-in 

Simple search **Advanced search** Browse collections Project Details

kw: "russian image illinois" Search Refine search

56 records.

Results by Sort by [Next 10 Records](#)


Jump to Records: [1](#) | [11](#) | [21](#) | [31](#) | [41](#)

Indiana University:
Charles W. Cushman Photograph Collection
[4 records](#)


University of Illinois at Urbana-Champaign:
Russian Publics
[56 records](#)

University of Michigan:
Art, Architecture and Engineering Library
[3 records](#)


Record 1 of 56

Title	Frantsuzskii voiazher v 1812om godu	
Author/Creator	I. Terebens__	
Type	Image	
Collection	Russian Publics	

Record 2 of 56

Title	(Katal'naia gora v Oranienbaumie.)	
Author/Creator	(unknown)	
Type	Image	
Collection	Russian Publics	

Record 3 of 56

Title	Kolonisty. Der Colons. Kolonisten.	
Author/Creator	(unknown)	
Type	Image	

Harvesting the set descriptions

```
<set xmlns='http://www.openarchives.org/OAI/2.0/'>
  <setSpec>oaiall:ummubib</setSpec>
  <setName>Art, Architecture and Engineering Library</setName>
  <setDescription>
    <oaic:dc xmlns:oaic="http://www.openarchives.org/OAI/2.0/oai_dc/"
      xmlns="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">
      <description> Images from this collection are part of the Art, Architecture and
        Engineering Library Visual Resources Collection which is housed in the lower
        level of the Media Union Library at the University of Michigan. The collection
        primarily serves the College of Architecture and Urban Planning and the
        School of Art and Design. (Collection Access: restricted;
        http://images.umdl.umich.edu/u/ummu/) </description>
    </oaic:dc>
  </setDescription>
</set>
```



The Dublin Core Collection

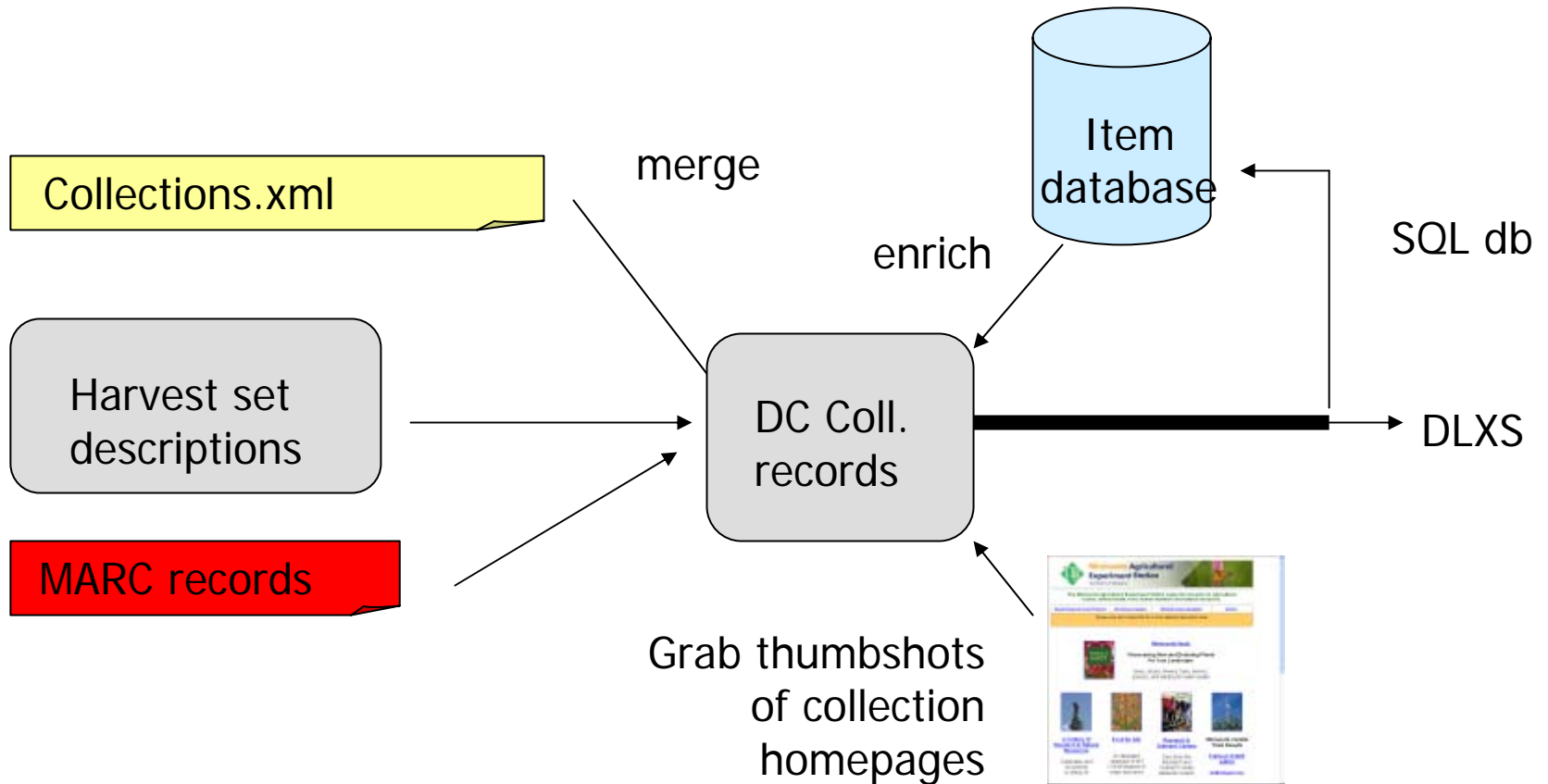
- Several issues still under discussion
 - Relation between collections and access services
 - Collection identifiers

Collection Working Group	
> InstanceTemplate	
Title of Collection	
Collection Description	
Resource URI	http://resourceURI/
Property QName	Value String
dc:identifier	value string
dc:title	value string
dcterms:alternative	value string
dcterms:abstract	value string
dc:format	value string
dcterms:extent	value string
dc:language	value string (term)
dc:type	value string (term)
dc:type	value string (term)
dc:rights	value string
dcterms:accessRights	value string

<http://dublincore.org/collectionwiki/InstanceTemplate>



1 additional workflow for collections



Some lessons the CIC experience

- Splitting repositories into sub-collections to enrich an item-level digital library is promising
 - Data providers' division of content is in many cases relevant
 - Data providers tend to define collections when they share content – it is easier to ask for a collection level description at that moment of the workflow
- Using sets to identify those collections is an easy and convenient mechanism vs use of DC relation for eg.
- Using set descriptions to convey CLD is potentially problematic
 - They include information about the set for selective harvesting
 - Do not have mechanisms for metadata sharing (eg. datestamp)



***Conclusions on a distributed
architecture to share collections***



Concluding Thoughts

- Conclusions
 - Set descriptions are recommended
 - DLF/NSDL best practices documents identify the information needs for selective harvesting
 - Set descriptions and collection descriptions are not the same
 - Some harvesters also want collection level descriptions to add context to items, particularly in the cultural heritage arena.
- Using set descriptions to convey collection level descriptions may raise difficulties in the long run
 - No identification scheme for collections
 - Since set description records have no datestamp, there is no trigger for automatic update

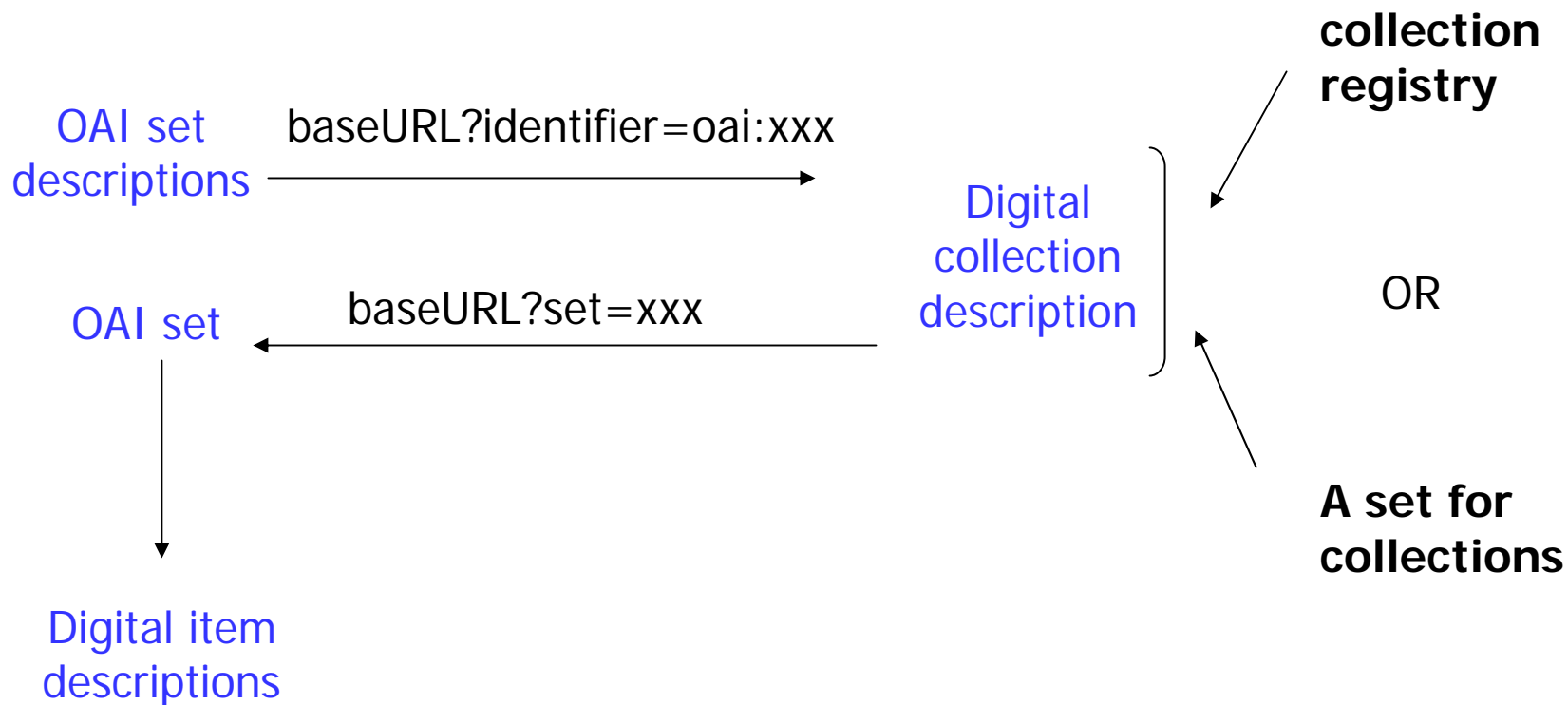


An architectural challenge for the future

- In registries networks : collections and services
- Representing relationships in a distributed environment
 - Items
 - Collections
 - Item descriptions and
 - Collection descriptions



OAI framework for sharing collections?



Questions and challenges

- Encourage the development of collection registries for communities?
- Encourage cultural heritage data providers to provide OAI sets of collection descriptions?
 - Future work to include descriptions from OAI registries
 - Relating different granularities of description
- Relating collection descriptions and set descriptions
 - Without raising the barrier to entry too high for data providers?
- Potential areas for exploration by DLF-Aquifer?



References

- OAI implementation guidelines

<http://www.openarchives.org/OAI/2.0/guidelines-repository.htm#setDescription>

- DLF/NSDL best practices for shareable metadata

<http://oai-best.comm.nsd.org/cgi-bin/wiki.pl?PublicTOC>

- Describing OAI sets: a discussion paper

http://comm.nsd.org/download.php/624/set-description_ver6.pdf

