Describing OAI Sets: A Discussion Paper

► Introduction to issues around OAI set descriptions

The Digital Library Federation and the National Science Digital Library gathered in July 2004 to lead a joint initiative to review current issues with OAI implementations and define best practices for OAI and sharable metadata.[1] The DLF / NSDL best practices for OAI and sharable metadata group agrees that:

- Set and collection descriptions are not the same although there may be overlap between the two;
- Service providers find set descriptions useful;
- Service providers also find collection descriptions useful as distinct from set descriptions;
- The OAI implementation guidelines acknowledge that there is a difference between set and collection descriptions but do not explicitly explain the differences. The implementation guidelines refer to both collection description formats, including EAD and the Dublin Core Collection Description Application Profile, and service description formats such as the ePrints schema and UDDI when talking about set description[2];
- Data providers currently provide many different types of information within the set description. Some of this information describes collections of resources and some of which describes the OAI set as distinct from a collection of resources;
- In general, data providers do not want to repeat the same information for sets and collections;

Therefore, we believe there needs to be:

1) a clarification of the difference between a description of an OAI set and a description of a collection of resources;
2) a metadata format or an application profile to specifically describe an OAI set;
3) a recommendation on the mechanism to convey both the set description and collection description.

This discussion paper proposes a framework for both and seeks comments about the content presented here. We are actively seeking comments and feedback and welcome anyone who might want to participate in the further development of this proposal. Please contact Muriel Foulonneau (mfoulonn@uiuc.edu) and Sarah Shreeves (sshreeve@uiuc.edu) with comments, questions, and feedback.

[1] http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl?OAI_Best_Practices
[2] http://www.openarchives.org/OAI/2.0/guidelines-repository.htm#setDescription

► What is a collection?

A collection is "any aggregation of physical or digital items"[3]. Two types of collections are discussed here: the collection of metadata (an OAI set) and a collection of resources.

► What is an OAI set?

Within the OAI environment, "a set is optional construct for grouping items for the purpose of selective harvesting." [4] An OAI set is both:

- a collection of metadata records in an OAI repository; and
- the technical mechanism by which the OAI repository (the service) makes these metadata records available.

An OAI set may include a set description. **We propose that a set description should describe both the collection of metadata records as well as the appropriate elements of the technical mechanism.**

► Elements of a set description

We propose to use a Dublin Core application profile to describe OAI sets. The list below was created from the list of elements that are in actual use (study from the UIUC OAI registry[5]) and are also drawn from the last version of the Dublin Core Collection Description Application Profile (CD AP).[6] In addition, the OAI Best Practices Working Group has considered the elements that would be useful for service providers for the purposes of selective harvesting and processing of metadata records.

The sections in italics and red are areas where we are particularly interested in comments from the community.

(Note that not all elements indicated below have been officially accepted into the Dublin Core Collection Description Application Profile; the cld and gen namespaces used below are temporary until the CD AP is formalized.)

Description	Potential Metadata Element	Example	Comment
A unique identifier in the OAI world	dc:identifier	- lcoa1.loc.gov:brhc	This is the repository identifier + oai setspec
Set name	dc:title	- Brady-Handy Collection (Photographs)	This is a repetition of the oai:setname

[3] http://www.dublincore.org/groups/collections/
[4] http://www.openarchives.org/OAI/openarchivesprotocol.html#Set
[5] http://gita.grainger.uiuc.edu/registry/
[6] http://www.ukoln.ac.uk/metadata/dcmi/collection-ap-summary/

Metadata formats in which the set is available	- dc:format OR - cld:itemFormat	- MODS - oai_dc	Please note that the use of dc:format within the DC Collection Description AP is under discussion. See the DC CD listserv in May-July 2005 for further discussion.[7]
How records are added to the set	dct:accrualMethod	- Updated automatically from a database - Updated through manual export	
Whether the set is updated	dct:accrualPolicy	- No longer adding items - Closed - Records are modified	This should be coherent with the update frequency.
Update frequency	dct:accrualPeriodicity	- Monthly - Weekly	Allows service providers to know how often to reharvest.
Earliest and latest datestamps within the set	cld:contentDateRange	- 2004-11-03/2005-03-05	Allows a service provider to know whether to reharvest.
Size of the set	dct:extent	- 1000 records	Allows a service provider to double check number of harvested records
Person in charge of the metadata records / responsibility	marcrel:own	- John Smith <jsmith@collectionsinstitute.org>	This might be useful if the institution that created the records is in a different institution than the institution that hosts the OAI data provider.
Person in charge of the set maintenance / responsibility	dc:creator	- Elizabeth Taylor <etaylor@oaiservices.net>	Presumably this should be the same as the administrator in the OAI Identify response.
Access rights	dct:accessrights	- HTTP authentication required, by prior agreement – contact Elizabeth	Could be used if an authentication is required to harvest that specific set.

[7] http://www.jiscmail.ac.uk/lists/DC-COLLECTIONS.html

		Taylor <etaylor@oaiservices.net>	
Rights	- *dc:rights* *OR* - *<rightsManifest>*	- *See <rights> container for rights held over the metadata in this set.* - *<rightsManifest>… </rightsManifest>*	*Within the OAI context, rights for the metadata records (NOT resources) are defined in the <rights> container. The rights element should refer users to this container or could alternatively duplicate the rights container. Please note that there is no way to express rights over the set (aggregation) of metadata records; only the individual records.*
A reference to a collection description of the resources described by the metadata, if applicable	*dct:references*	- http://imlsdcc.grainger.uiuc.edu/collections/FullDisplay.asp?cid=2391 - <cld:collection> … </cld:collection>	*This is where a reference to a collection description should be contained through any number of ways (see below for a further discussion). URLs to collection descriptions? oai:identifier of collection description records in the repository? Embedding collection descriptions?*
Summary statement about the content of the set	*dct:abstract*	- *Records for cookbooks digitized as part of the Feeding America: Historic American Cookbooks Project*	
Documentation about the metadata records including cataloging and mapping info	*dc:description*	- *mapped to MODS from MARC records* - *LCSH terminology used in MODS format*	*May also include a URI pointing to documentation.*

Relationship to other sets	- dct:isPartOf (supercollection) - dct:hasPart (subcollection) - dc:relation (associated collection)	- Overlaps with set XXX and YYY - Belongs to superset DDD - Is composed notably although not exclusively of subsets RRR and UUU	
Language of records in set	dc:language	- en - fr - ru - ge	
Audience	dct:audience	- Created for the National Science Digital Library <http://nsdl.org/>	Note specific service providers sets have been created for.
Logo	*- cld:logo* *OR* *- <branding>*	*- see <branding> container for logo for this set* *- <branding>* *…* *</branding>*	*Similar problem as with rights (though rather less serious!)*
Institution which hosts the set	gen:isLocatedAt	- Oaiservices Inc.	Could be a static repository gateway for example (also the repository though)
Service that makes the metadata records available	gen:isAccessedVia	- http://myOAIrepository/?verb=ListRecords&metadataPrefix=oai_dc&set=brhc	Either the baseURL OAI query into the set and alternate ways to get into a set

► Set – collection relationships

There is no *a priori* relationship between a collection of resources and an OAI set. However, it can be useful if there is, in fact, a relationship between an OAI set and a collection of resources for a variety of reasons, including:

- Service providers can use collection information to provide richer context for the harvested metadata;
- Information that service providers might find useful for selective harvesting purposes such as access rights may be attached to a specific collection of resources but not to an entire OAI repository; and

- The metadata for resources exposed via OAI may be managed at a collection level and thus organizing sets based on collections may ease the related management of the metadata in the OAI repository.

We propose that when possible data providers should create sets corresponding to individual collections and optionally should provide other hierarchical or overlapping sets. A collection of resources can be defined in any way imaginable (for example, by topic, type of material, author, provenance, etc). Regardless, data providers might think about the following when defining sets in the OAI context:

1) The set creates a granularity unit that differentiates it from the other groupings within the context of a specific application.

   For example, the Library of Congress provides multiple collections via OAI that broadly fit under the category of American History; however, these have been usefully broken apart into individual sets that correspond to specific defined collections of resources. While one service provider might be interested in the "Ansel Adams's Photographs of Japanese-American Internment at Manzanar", it may not be interested in "Glass negatives from the Papers of Wilbur and Orville Wright".

   At the opposite extreme, the OAI repository for the Illinois aerial photograph collection is divided into sets by both year and county. These could be grouped into sub-collections based on the year the photographs were taken (i.e. to create a collection of aerial photos of Illinois in 1940, in 1941, etc). In our experience, these are not different enough from one another to warrant representing these in distinct sets.[8]

2) The collection of resources was created through the same digitization project or process. This often means that the metadata were at least reviewed and processed at the same time, and the formats of resources are similar. These collections would be good units for aggregators to reprocess and manage because of their potential consistency.

Obviously, it is not possible for data providers to organize resources into collections that will be useful for all harvesters, but data providers should spend some time thinking about how their sets are organized in relation to collections.

When a set corresponds to a collection of resources, the collection description corresponding to that collection of resources should be either embedded in or referred to in the set description in the <dct:references> element.

If sets represent multiple collections, the set description could embed or refer to multiple collection descriptions. If sets are not designed according to a concept of

[8] http://cicharvest.grainger.uiuc.edu/documents/usingcollectiondescriptions.pdf

collection, the set description might leave the <dct:abstract> blank or include some other sort of description such as: Contains metadata records with subject:physics.

► Sharing collection descriptions in the OAI environment

The mechanism by which data providers should share collection description is not certain. While we have not fully explored how collection descriptions should be shared within the OAI environment, several alternatives present themselves each with pros and cons:

1) An oai record with a collection description is included in the repository, but in a set containing only collection descriptions rather than in the same set as the records describing its component items. This is the example followed by the National Science Digital Library. Potentially the collection description could be located by a GetRecord link in the set description.
2) A link to a collection description on a web page. This could be embedded in the set description and/or the item level metadata records; or
3) A metadata record that could be embedded in the set description.

An additional solution consists of creating an oai record with a collection description contained within the same set as its [the collection's] component parts. The collection description could be located by a GetRecord link in the set description. The relationship between the item level metadata records and the collection description record could be noted in the <dc:relation> field. This solution is the least palatable from both a data and service provider perspective as it requires additional maintenance of not only a collection description record, but also item level records.

Therefore, the present proposal considers the descriptions of collections of resources as part of the <dct:references> field of the set description and suggests using either an embedded collection description or a URI to reference the collection description. The URI can be either a Webpage http://myWebSite/mycollectiondescription.html or an oai record http://myOAIrepository.org/?verb=GetRecord&metadataPrefix=dccoll&identifier=oai:MyCollectionDescriptionIdentifier.

We would welcome comments and thoughts on these methods.

► Examples

From the Library of Congress

Information about the set is included in a <dc:description> tag within a full collection 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>
```

From Michigan State University:

Models the proposal in this paper:

```
<set>
      <setSpec>fap</setSpec>
      <setName>Feeding America: The Historic American Cookbook Project
      </setName>
      <setDescription>
            <dc:identifier>lib.msu.edu:fap</dc:identifier>
            <dc:title>Feeding America: The Historic American Cookbook
            Project</dc:title>
            <dc:format>oai_dc</dc:format>
            <dct:accrualmethod>dynamically generated from local
            database</dct:accrualmethod>
            <dct:accrualpolicy>Records may be added as new cookbooks
            are digitized; records may be modified if
            necessary</dct:accrualpolicy>
            <dct:accrualperiodicity>Infrequently</dct:accrualperiodicit
            y>
            <cld:contentdaterange>2005-05-20/2005-05-
            20</cld:contentdaterange>
            <dct:extent>77 records</dct:extent>
            <marcrel:own>Anne Karle-Zenith</marcrel:own>
            <dct:collector>Michael Seadle</dct:collector>
            <dct:accessrights>No restrictions</dct:accessrights>
            <dc:rights>No restrictions</dc:rights>
            <dct:references>
                  <cld:collection>
                        <dc:identifier>http://digital.lib.msu.edu/project
                        s/cookbooks</dc:identifier>
                        <dc:title>Feeding America: The Historic American
                        Cookbook Project</dc:title>
                        <dct:abstract>Online collection of some of the
                        most important and influential American cookbooks
                        from the late 18th to early 20th
                        century.</dct:abstract>
                        <dct:extent>77 items</dct:extent>
                        <dc:language xsi:type="ISO639-
                        2">eng</dc:language>
                        <dc:type xsi:type="cldtype">Collection of
                        Texts</dc:type>
                        <dct:accessRights>No
                        restrictions</dct:accessrights>
                        <cld:accrualMethod
                        xsi:type="DCCDAccrualMethod">ItemCreation</cld:ac
                        crualMethod>
```

```
                    <cld:accrualPeriodicity>Unknown</cld:accrualPerio
                    dicity>
                    <cld:accrualPolicy
                    xsi:type="DCCDAccrualPeriodicity">Passive</cld:ac
                    crualPolicy>
                    <dc:subject xsi:type="LCSH">Cookery,
                    American</dc:subject>
                    <dc:subject xsi:type="LCSH">Cookery -- United
                    States -- 19th century</dc:subject>
                    <dc:subject xsi:type="LCSH">Cookery -- United
                    States -- 20th century</dc:subject>
                    <dc:creator>Michigan State University Libraries.
                    Digital &amp; Multimedia Center.</dc:creator>
                    <marcrel:own>Michigan State University Libraries.
                    Digital &amp; Multimedia Center.</marcrel:own>
                    <gen:isLocatedAt>Michigan State University, 100
                    Library, East Lansing, MI, 48224,
                    USA</gen:isLocatedAt>
                    <gen:isAvailableVia
                    xsi:type="URI">http://digital.lib.msu.edu/project
                    s/cookbooks</gen:isAvailableVia>
               </cld:collection>
          </dct:references>
          <dct:abstract >Records for cookbooks digitized as part of
          the Feeding America: Historic American Cookbooks
          Project</dct:abstract>
          <dc:description>Simple Dublin Core records created from
          documentation for each cookbook; oai_dc subject uses LCSH;
          oai_dc:type uses DCMIType vocabulary; oai_dc:coverage (1)
          uses TGN</dc:description>
          <dc:language xsi:type="dct:ISO639-2">eng</dc:language>
          <gen:isLocatedAt>Michigan State University Libraries
          Digital &amp; Multimedia Center</gen:isLocatedAt>
          <gen:isAccessedVia xsi:type="dct:URI">
          http://oai.lib.msu.edu/OAIHandler </gen:isAccessedVia>
     </setDescription>
</set>
```