THE FEEDING HABITS OF OAISTER CATCHERS, OR METADATA TO GO

Stephanie C. Haas
Assistant Director, Digital Library Center
University of Florida Libraries
Gainesville, FL 32611
haas@uflib.ufl.edu

Basically, the Open Archive Metadata Harvesting Initiative is all about indexing resources and sharing the indexing.

The traditionally sophisticated MARC records of libraries have been shared using protocols such as Z39.50. This has proven very effective, except Z39.50:

- dislikes metadata other than MARC
- requires a special view for non-MARC records, and
- is not scalable. *Meaning the more distributed record collections you search, the more likely it is that one or more will be down, and search time becomes an issue.*

Dublin Core metadata is the scarecrow version of MARC. It was originally created to describe web resources in 15 fields or less.

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

Dublin Core was intended to be used as a self-indexing mechanism for web page developers. DC records were to be added to the header of html pages, but...unscrupulous web page creators packed it with enticing non-relevant words, search engines ignored it, and you can imagine what happened.

Dublin Core was repurposed by physicists at Los Alamos to index their burgeoning e-print collections. By the way, these e-print collections were called "archives" and because they were accessible to all, they were "open."

Then it became obvious that they needed a way to share information across e-print collections. In short, they needed a way to harvest Dublin Core metadata from multiple repositories. This led us right into the Open Archives Metadata Harvesting Protocol.

The Coalition for Networked Information and Digital Library Federation provided funding to establish an Open Archives Initiative secretariat at Cornell University.

Managed by Carl Lagoze and Herbert Van de Sompel an international steering committee developed the metadata harvesting protocol.

The Key players in the OAI-PMH world

Data Providers: these are the repositories that dangle out their Dublin Core metadata in a format that is OAI compliant. Minimal OAI compliant metadata consists of the unqualified 15-field Dublin Core metadata. However, the protocol does support any metadata format that is encoded in XML.

Service Providers: these are the harvesters, or the OAIster catchers, who skim along from repository to repository gathering up the OAI compliant metadata, integrating it, and caching it.

A few more definitions:

Item: object in a repository that has a unique identifier and metadata

Identifier: unique item key

Record: metadata in a specified format: Dublin Core, MARC, FGDC

Set: a construct for broadly grouping items; potentially useful in selective harvesting

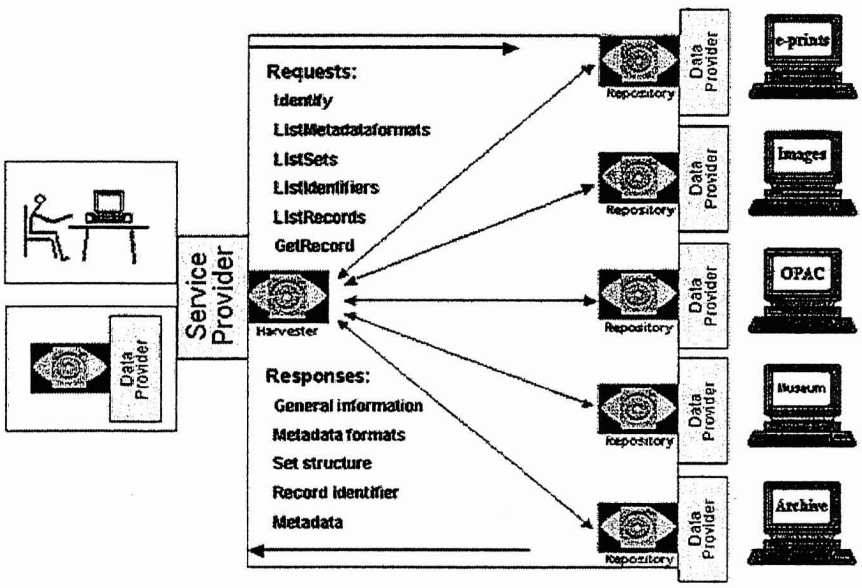
Namespace: the definitions for elements and attributes used in XML documents

Schema: defines the structure, content, and semantics of XML documents and is used to validate record formats.

How does it work?

Data providers serve the harvestable records consisting of unqualified Dublin Core in an XML wrapper.

Service provider programs use the metadata harvesting protocol to issue very simple HTTP-based requests. Basically, there are 6 request types called verbs that can be issued by the harvester applications. The diagram below shows the request-response functions and is borrowed from *OAI for Beginners*
[http://www.oaforum.org/tutorial/english/page3.htm]



Record Structure

Each record consists of three parts with an XML wrapper:

Header (mandatory)

Identifier: each item has a unique identifier that combined with a metadataPrefix acts as a key to extract a metadata record

Datestamp: last date of modification YYYY-MM-DD

setSpec elements: optional construct for grouping items for selective harvesting

Metadata Formats (mandatory)

Within a repository, metadata formats are indicated by a metdataPrefix. Prefixes are associated with namespace elements defined at a given URI. Schemas validate records for a given namespace.

About (Optional) rights & provenance statements

A sample record for The Florida aluminum phosphate zone of the Bone Valley Formation, Florida and its uranium deposits **with XML tags**

<record>
<header>
<identifier>oai:palmm.fcla.edu:AAA0019QCB</identifier>
<datestamp>2003-09-09</datestamp>
<setSpec>nsdl:feolbib</setSpec>
<setSpec>feol:feolbib</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:title>The Florida aluminum phosphate zone of the Bone Valley Formation, Florida, and its
uranium deposits.</dc:title>
<dc:date>1956</dc:date>
<dc:contributor>Jaffe, E.B.</dc:contributor>
<dc:creator>Altschuler, Z.S.,</dc:creator>
</oai_dc:dc>
</metadata>
</record>

Interpretation of the record:

Header portion
<record>
<header>
<identifier>oai:palmm.fcla.edu:AAA0019QCB </identifier>
oai=scheme
palmm.fcla.edu=repository
AAA0019QCB is the identifier for the item in the repository
<datastamp>2003-09-09</datestamp>

<setSpec>nsdl:feolbib</setSpec> is part of the National Science Digital Library, Florida Environments Online collection set
<setSpec>feol:feolbib</setSpec> is part of the Florida Environments Online collection set
</header>

Metadata portion

<metadata>
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
[says anything with the prefix "oai_dc" is part of the namespace defined by the document at the URI given. It also tags the element dc with the prefix oai_dc—this is the only ELEMENT in the record that is part of the oai_dc namespace and it is a record definition that includes all the other elements.]

xmlns:dc="http://purl.org/dc/elements/1.1/"
[says anything with the prefix dc is part of the namespace defined at the URI given]

xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
[says anything with the prefix xsi is part of the namespace defined at the URI given]

xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
[says the element schemaLocation is defined at the given URI and that the schema itself is located at the second URI. This schema is used to validate this record.]

<dc:title>The Florida aluminum phosphate zone of the Bone Valley Formation, Florida, and its uranium deposits.</dc:title>
[says this is the field defined as the Dublin core element title]

<dc:date>1956</dc:date>
[says this is the field defined as the Dublin core element date]

<dc:contributor>Jaffe, E.B.</dc:contributor>
[says this is the field defined as the Dublin core element contributor]

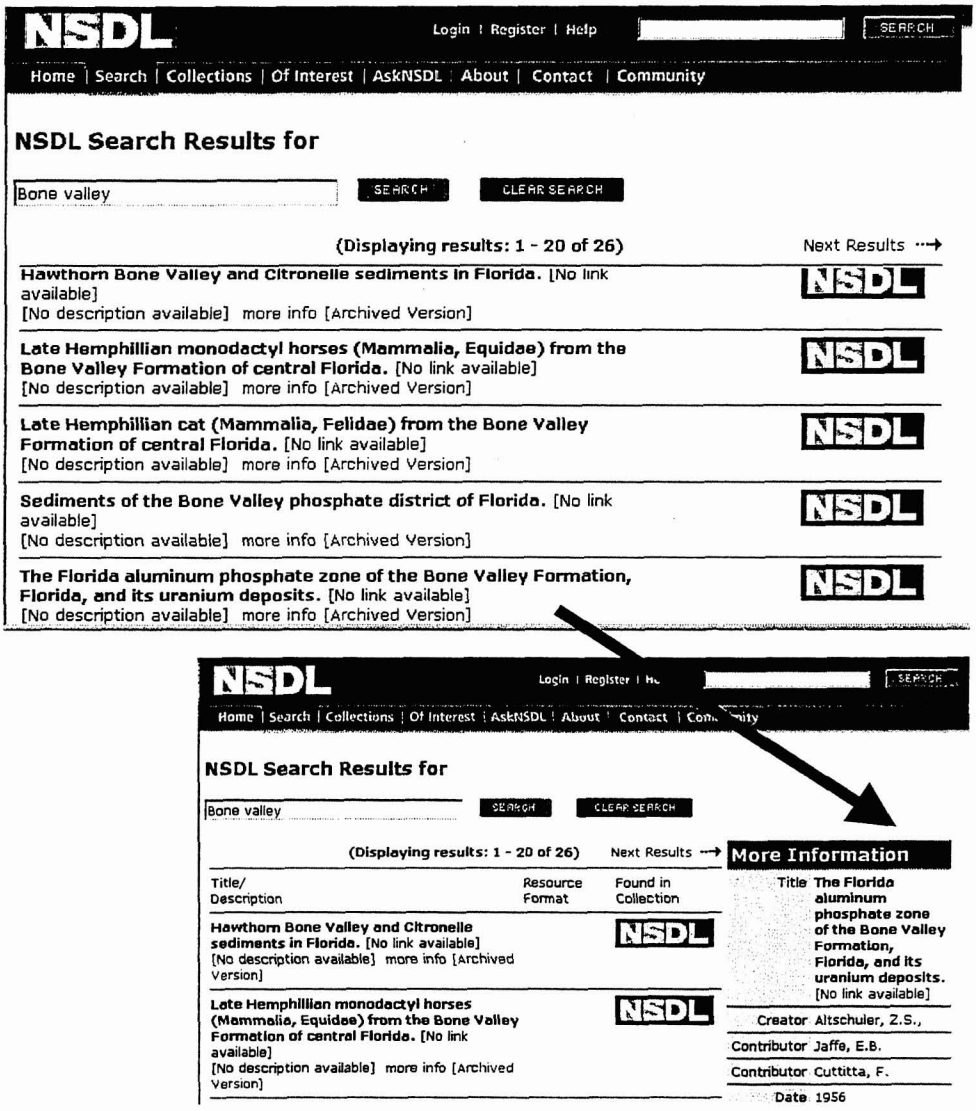
<dc:contributor>Cuttitta, F.</dc:contributor>
[says this is the field defined as the Dublin core element contributor]

<dc:creator>Altschuler, Z.S.,</dc:creator>
[says this is the field defined as the Dublin core element creator which is the MARC field author]

</oai_dc:dc>
</metadata>

</record>

Because we know that this record has been captured and cached by the National Science Digital Library, we can look for it in that repository [http://www.nsdl.org/]. The image below shows the brief record and the expanded view. Note: only five of the Dublin Core fields are actually displayed.



The Open Archives site provides lists of both existing repositories, data providers http://www.openarchives.org/Register/BrowseSites.pl and data harvesters http://www.openarchives.org/service/listproviders.html.

Good introductions to the topic of the Open Archives Initiative Metadata Harvesting can be found at the following web sites:

The Open Archives Initiative Protocol for Metadata Harvesting
http://www.openarchives.org/OAI/openarchivesprotocol.html

OAI for Beginners - the Open Archives Forum online tutorial
http://www.oaforum.org/tutorial/

NSDL (National Science Digital Library) Metadata Primer
http://metamanagement.comm.nsdlib.org/outline.html

Exposing and Harvesting Metadata Using the OAI Metadata Harvesting Protocol: A Tutorial
http://library.cern.ch/HEPLW/4/papers/3/

Acknowledgement:

With special thanks to Priscilla Caplan,Florida Center for Library Automation