7-2016

Digital Commons and OAI-PMH: Harvesting Repository Records

bepress

Digital Commons and OAI-PMH: Harvesting Repository Records

Version: July 2016
Available at http://digitalcommons.bepress.com/reference/80

Introduction

Digital Commons supports the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) for the sharing of repository records. Administrators have a number of options to ensure metadata is properly exposed to harvesters, making it discoverable through a variety of other platforms and services.

This document describes the metadata formats available in Digital Commons for mapping to Dublin Core elements—a key component of OAI harvesting—and explains how to form common OAI requests for Digital Commons content.

If you have questions at any time, please contact Consulting Services at dc-support@bepress.com or 510-665-1200, option 2. Frequently asked questions also appear at the end of this document.

For additional background, you can find more about OAI-PMH on the Open Archives Initiative website.

Contents

Metadata Formats in Digital Commons

Digital Commons exposes metadata fields for harvesting through four different metadata formats, or prefixes. In each case, the Digital Commons metadata fields for the given content remain the same, and are simply mapped to Dublin Core elements in a particular way. The following metadata formats correspond to the OAI metadataPrefix parameter.

Default Prefix

oai_dc
Fixed mappings to select simple Dublin Core elements.

Additional Prefixes

simple-dublin-core
Simple Dublin Core, flexible mappings. Alternate format: dcs.

qualified-dublin-core
Qualified Dublin Core, flexible mappings. Alternate formats: dcq, qdc.

oai_etdms
Generally used by Library and Archives of Canada (LAC).

The oai_dc Metadata Prefix

When a repository is created, the Dublin Core elements below are already mapped to certain fields within the default metadata prefix, “oai_dc.” These are exposed for harvesting when a submission is published on the site. We are able to reconfigure some of these fields to expose different types of information, as specified below.

- <title> - 'title' field. One instance per record.
- <creator> - 'author' field. Repeatable for co-authors, with a separate <creator> element exposed for each author.
- <subject> - Digital Commons maps data to this element from three fields: 'keywords', 'disciplines', and 'subject_area'. A separate <subject> element will be created for each term.
- <description> - 'abstract' field. Books and image galleries will show a second instance that links to the thumbnail image.
- <publisher> - Name of repository by default. This default can be replaced with a specific publisher or multiple publishers at your request. We can also create a blank text field for a value to be entered, or use the value from a particular metadata field.
- <date> - 'date' field. Uses ISO 8601 format (YYYY-MM-DD). This contains the publication date of the work. One instance per record.
- <type> - Currently this exports as DCMI class "text." This can be configured to export the value for the ‘document_type’ field upon request.
- <format> - Returns the MIME type of the primary file, if one is available. This is most commonly “application/pdf.” If an object is not uploaded, and instead a full-text link is provided directly to a file, the system will also return a MIME type if present. No <format> element will be created for a link to a webpage or other resource that does not have a MIME type.
- <identifier> - There are two instances of the <dc:identifier> tag in an “oai_dc” record.
    - URL of the article index page (e.g., http://arrow.dit.ie/sciendoc/3/).
    - A direct link to the full-text version of the primary file.
- <source> - Exports name of series (or other Digital Commons publication type) by default.

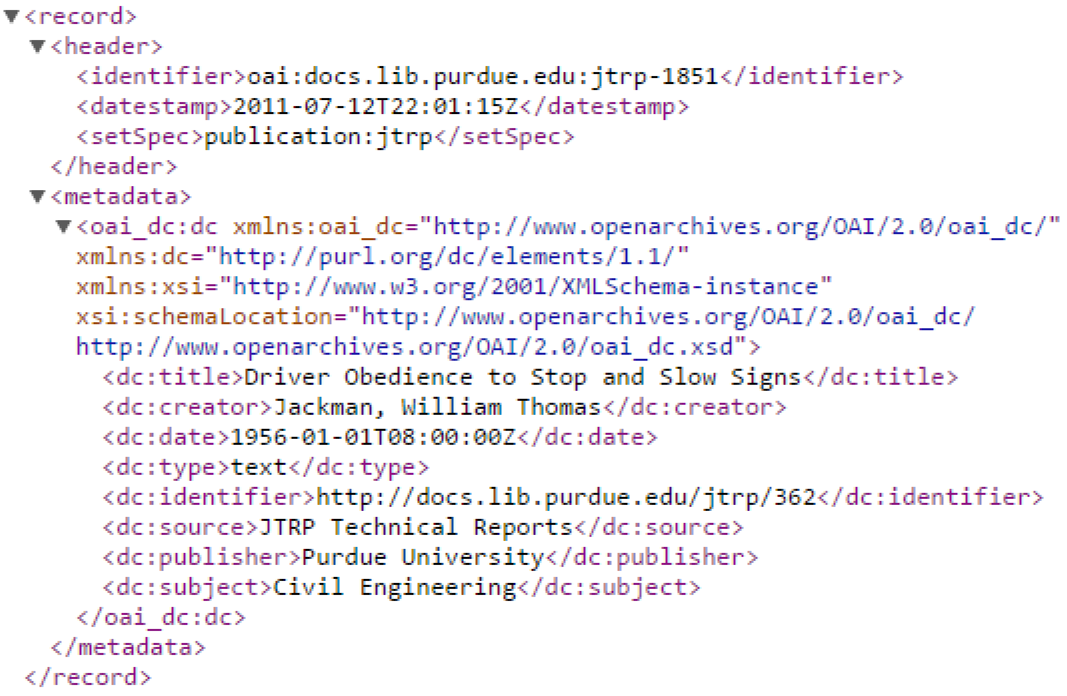
Some elements do not initially appear in “oai_dc,” but will appear if certain metadata fields are added or customized.

- <contributor> - Can be mapped to a single field or multiple fields by request. Typical fields requested are ‘contributor’ or ‘advisor’.
- <rights> - Will automatically map to the ‘rights’ field or the Creative Commons field ‘distribution_license’ if these are added to the relevant structure. Will map to both if present.
- <language> - Will map to the ‘language’ field when added to the relevant structure.
- <relation> - Will map to the ‘relation’ field or the ‘response_to_url’ field if these are added to the relevant structure. Will map to both if present.
- <coverage> - Will map to the ‘coverage’ field when added to the relevant structure.

Standard bepress fields that do not map to elements in oai_dc

Several fields are typically included on most submission forms, but do not map to Dublin Core elements in “oai_dc”:

- **Comments** is a bepress-specific field often included in default metadata, but not exposed via “oai_dc.” If you would like to use it for metadata that needs to accompany harvested content, you could use a different metadata prefix (see below). Alternatively, the metadata may be included in other fields, such as the abstract or keywords, depending on the type of information.
- **Author email address and author institution:** Harvesters generally focus on article metadata, rather than author details. Since author names are available in “oai_dc,” these extra components are not included. If you would like to export this information periodically, please ask Consulting Services about quarterly archive deliveries for your IR.
- **Other metadata:** If you need to expose different fields or types of metadata through OAI, the “simple-dublin-core” and “qualified-dublin-core” metadata prefixes described below provide a broader element set and mapping of custom fields.



```
▼<record>
  ▼<header>
     <identifier>oai:docs.lib.purdue.edu:jtrp-1851</identifier>
     <datestamp>2011-07-12T22:01:15Z</datestamp>
     <setSpec>publication:jtrp</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>Driver Obedience to Stop and Slow Signs</dc:title>
       <dc:creator>Jackman, William Thomas</dc:creator>
       <dc:date>1956-01-01T08:00:00Z</dc:date>
       <dc:type>text</dc:type>
       <dc:identifier>http://docs.lib.purdue.edu/jtrp/362</dc:identifier>
       <dc:source>JTRP Technical Reports</dc:source>
       <dc:publisher>Purdue University</dc:publisher>
       <dc:subject>Civil Engineering</dc:subject>
     </oai_dc:dc>
   </metadata>
 </record>
```

Sample OAI record using the “oai_dc” metadata prefix.

Flexible Harvesting: simple-dublin-core, qualified-dublin-core, and oai_etdms

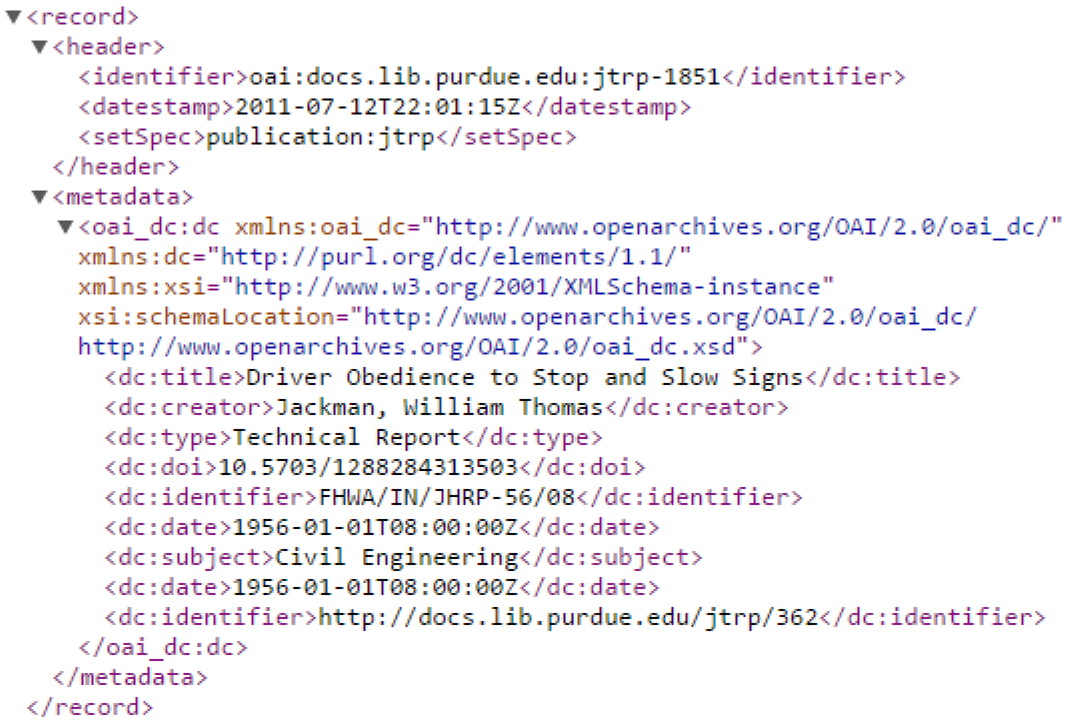
The “simple-dublin-core” and “qualified-dublin-core” metadata prefixes in Digital Commons offer additional options for exposing metadata through OAI.

With these prefixes, you are able to specify which metadata fields map to which elements, and also choose to map elements multiple times. Should you need even greater control, you may request to expose a field using a uniquely labeled element of your choice. This flexibility allows DC repositories to participate with a variety of harvesters.

The simple-dublin-core format

This metadata prefix may be formatted as “simple-dublin-core” or “dcs” in OAI requests.

The “simple-dublin-core” prefix exposes Digital Commons metadata using the 15 main elements of the Dublin Core Metadata Element Set. (See the full list in the accompanying document, “Dublin Core Elements in Digital Commons.”) Mappings are customizable. Some defaults are assigned in each publication type, but these may be modified by request, and additional fields mapped as well. This prefix will also expose any custom Dublin Core labels requested. See below for more about custom labels.



```
▼<record>
  ▼<header>
     <identifier>oai:docs.lib.purdue.edu:jtrp-1851</identifier>
     <datestamp>2011-07-12T22:01:15Z</datestamp>
     <setSpec>publication:jtrp</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>Driver Obedience to Stop and Slow Signs</dc:title>
       <dc:creator>Jackman, William Thomas</dc:creator>
       <dc:type>Technical Report</dc:type>
       <dc:doi>10.5703/1288284313503</dc:doi>
       <dc:identifier>FHWA/IN/JHRP-56/08</dc:identifier>
       <dc:date>1956-01-01T08:00:00Z</dc:date>
       <dc:subject>Civil Engineering</dc:subject>
       <dc:date>1956-01-01T08:00:00Z</dc:date>
       <dc:identifier>http://docs.lib.purdue.edu/jtrp/362</dc:identifier>
     </oai_dc:dc>
   </metadata>
 </record>
```

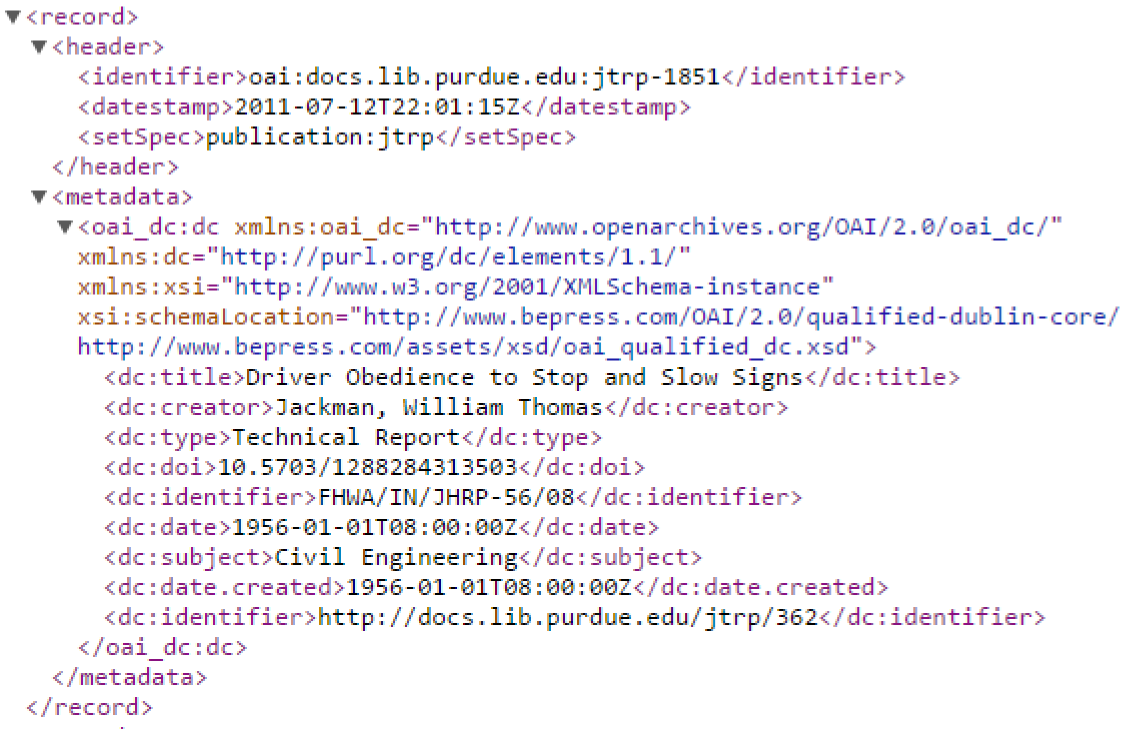
Same OAI record as above using the “simple-dublin-core” metadata prefix.

The qualified-dublin-core format

The metadata prefix “qualified-dublin-core” has two aliases: “dcq” and “qdc.” It is commonly formatted in OAI requests using one of these shortened versions.

This prefix employs additional qualifiers to further refine the elements available in “oai_dc” or “simple-dublin-core.” (A full list is available in the accompanying document, “Dublin Core Elements in Digital Commons.”) As with the “simple-dublin-core” prefix, mappings are customizable and this prefix will also expose any custom labels assigned by request.

A metadata field mapped to a qualified Dublin Core element, such as <format.medium>, will export in "qualified-dublin-core" through this element. It will also export in "simple-dublin-core" through the simple version of that element; in this case, <format>.



```
▼<record>
  ▼<header>
     <identifier>oai:docs.lib.purdue.edu:jtrp-1851</identifier>
     <datestamp>2011-07-12T22:01:15Z</datestamp>
     <setSpec>publication:jtrp</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.bepress.com/OAI/2.0/qualified-dublin-core/
     http://www.bepress.com/assets/xsd/oai_qualified_dc.xsd">
       <dc:title>Driver Obedience to Stop and Slow Signs</dc:title>
       <dc:creator>Jackman, William Thomas</dc:creator>
       <dc:type>Technical Report</dc:type>
       <dc:doi>10.5703/1288284313503</dc:doi>
       <dc:identifier>FHWA/IN/JHRP-56/08</dc:identifier>
       <dc:date>1956-01-01T08:00:00Z</dc:date>
       <dc:subject>Civil Engineering</dc:subject>
       <dc:date.created>1956-01-01T08:00:00Z</dc:date.created>
       <dc:identifier>http://docs.lib.purdue.edu/jtrp/362</dc:identifier>
     </oai_dc:dc>
   </metadata>
 </record>
```

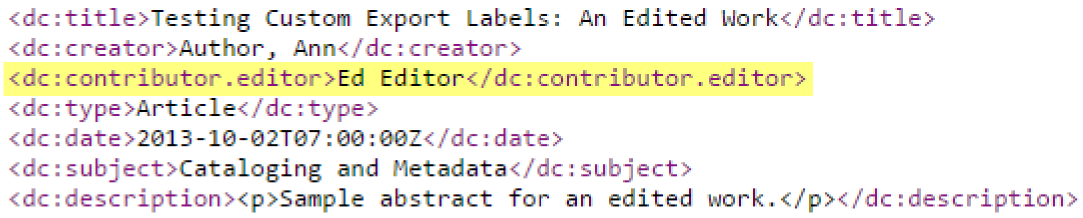
Same OAI record as examples above using the "qualified-dublin-core" metadata prefix.

The oai_etdms format

In the "oai_etdms" metadata prefix, submission-level metadata is mapped according to the NDLTD ETD-MS standard by default. However, mappings can be configured by request in the same way as "simple-dublin-core" and "qualified-dublin-core." This format is used to harvest into Library and Archives of Canada.

Custom Export Labels

Digital Commons supports custom export labels for instances where mapping to a specific term outside of standard Dublin Core is necessary. Terms may correspond to a different element set, such as one used in an archival or museum context, or simply provide additional clarification. One possibility is to add a custom refinement to an element like <contributor>, such as <contributor.editor>:



```
<dc:title>Testing Custom Export Labels: An Edited Work</dc:title>
<dc:creator>Author, Ann</dc:creator>
<dc:contributor.editor>Ed Editor</dc:contributor.editor>
<dc:type>Article</dc:type>
<dc:date>2013-10-02T07:00:00Z</dc:date>
<dc:subject>Cataloging and Metadata</dc:subject>
<dc:description><p>Sample abstract for an edited work.</p></dc:description>
```

Sample "simple-dublin-core" record exposing an Editor field with a custom <contributor.editor> label.

Custom terms display just like regular Dublin Core elements when viewing OAI results. As shown in the above example, when including a qualifier in a custom export label, the qualifier will display in both "simple-dublin-core" and "qualified-dublin-core."

Consulting Services can provide more information about custom export labels if these seem appropriate in your case.

Testing your Mappings

If you have asked to apply a large number of Dublin Core mappings or labels for exposure through any of the customizable prefixes above, you can review the output by retrieving OAI results for a single context or submission. Please see "List Records for Individual Publication Contexts" or "Get Record for a Single Item" below for details.

Formulating Common OAI Requests in Digital Commons

Each Digital Commons repository is ready for harvesting out of the box using the "oai_dc" metadata prefix, described earlier. Whether you choose the standard, or customize the fields you expose with one of the other available metadata prefixes, it's easy to get started retrieving records. Once you know the basics, requests follow a predictable pattern.

Most OAI requests are built using 1) a base URL specific to your repository, 2) a verb specifying the action requested, and 3) a metadata prefix that retrieves a particular metadata format.

Base URL Format for Your Repository

Base URLs in Digital Commons support OAI 2.0. The OAI base URL for a Digital Commons site is:

```
http://[Site URL]/do/oai/
```

For example, http://digitalcommons.wpi.edu has the base URL http://digitalcommons.wpi.edu/do/oai/.

Entering only the base URL in a browser will return an error message, since the OAI protocol requires that a verb be associated with each request. For harvesting repository content, the most common requests utilize the List Records verb.

List Records for the Repository Level

The List Records verb allows you to view all of the records and associated metadata exposed through OAI for the entire repository. This URL will display the first 100 records in a repository-level request:

```
http://[Site URL]/do/oai/?verb=ListRecords&metadataPrefix=[Enter oai_dc or other prefix]
```

Example: http://digitalcommons.wpi.edu/do/oai/?verb=ListRecords&metadataPrefix=oai_dc

You may view records beyond the first 100 using a Resumption Token, described below.

List Records for Individual Publication Contexts (Sets)

The List Records verb also allows you to view all of the records and associated metadata exposed through OAI for a particular publication, considered a "set" in OAI. This URL will return the first 100 records in a publication-level request:

```
http://[Site URL]/do/oai/?verb=ListRecords&metadataPrefix=[Insert
prefix]&set=publication:[Series/journal/other URL label]
```

Example: http://digitalcommons.wpi.edu/do/oai/?verb=ListRecords&metadataPrefix=oai_dc&set=publication:mechanicalengineering-pubs

Records beyond the first 100 are easily retrieved using a Resumption Token, described below.

Resumption Tokens

Once you have run a List Records query for the repository or any large collection in the repository, you can continue to view the records 100 at a time by using a Resumption Token. You can find the Resumption Token at the bottom of the first page of List Records XML results. Example:

**<resumptionToken completeListSize="3871" cursor="0">2523496/oai_dc/100//
</resumptionToken>**

In this case, the Resumption Token specified is: **2523496/oai_dc/100//**

To view the next 100 records, insert the token into the URL.

```
http://[Site URL]/do/oai/?verb=ListRecords&resumptionToken=[Insert resumption token]
```

Example:
http://digitalcommons.wpi.edu/do/oai/?verb=ListRecords&resumptionToken=2523496/oai_dc/100//

Resumption Tokens will continue to advance in number and appear at the bottom of each results page, allowing you to view all records in the repository or publication. To see a particular record number or section of records, you can adjust the ending "cursor" number to specify the record number where you want to start.

List Records for One Record

You can also test your mappings on a single item using a different verb, Get Record, with the following structure:

```
http://[Site URL]/do/oai/?verb=GetRecord&metadataPrefix=[Insert prefix]&identifier=oai:[Site
URL]:[Publication]-[MS #]
```

Example: http://digitalcommons.unf.edu/do/oai/?verb=GetRecord&metadataPrefix=dcq&identifier=oai:digitalcommons.unf.edu:etd-1419

List Records for a Particular Time Period

Adding a date range to an OAI request will include records that were either posted to the site or revised during that time period. This is not based on the actual publication date, except where the date posted happens to be the same as the publication date.

A date range request includes "from" and/or "until" parameters, formatted with a year, month, and date:

```
http://[Site URL]/do/oai/?verb=ListRecords&metadataPrefix=[Insert prefix]&from=YYYY-MM-DD&until=YYYY-MM-DD
```

Example:
http://digitalcommons.wpi.edu/do/oai/?verb=ListRecords&metadataPrefix=oai_dc&from=2014-01-01&until=2014-12-31

Frequently Asked Questions

Are Digital Commons repositories OAI-compliant?

While there is no official standard for OAI compliance, we fully support OAI-PMH. Digital Commons sites pass the validation tests for inclusion on the Open Archives list of OAI-conforming repositories, available at http://www.openarchives.org/Register/BrowseSites.

Why did my repository fail the test for validation?

We pass OAI's own tests for validation, and Digital Commons repositories are included as registered data providers on http://www.openarchives.org/Register/BrowseSites. Other tests may include variables that are optional to the OAI-PMH, and if you encounter difficulty, please inform Consulting Services so we can help assess the need for adjustments.

Do you expose the direct URL of the published full-text document for harvesting?

Yes. This link is available in the second <dc:identifier> element in the OAI record, and appears automatically when using the "oai_dc" metadata prefix. For the other prefixes, the full-text URL can be enabled by request.

In the case of image galleries, the direct URL of the image thumbnail is also exposed with the <dc:description> element.

How do Digital Commons metadata field labels and Dublin Core elements differ?

In Digital Commons, metadata field labels are flexible to accommodate the needs of a specific collection or institution. A field label can be unique and granular, or it can be more generic if that is appropriate. The possibilities are wide-ranging, and Consulting Services can help create fields to capture just about any metadata you want.

For the sake of interoperability, Dublin Core uses specific, pre-defined elements to describe the types of data captured by metadata fields. These elements are mapped to fields in Digital Commons. Following best practices, we then expose those fields with the Dublin Core mappings so the information can be harvested via the OAI-MPH protocol.

Example: A field labeled "advisor" would likely map to the Dublin Core element <contributor>. Information stored in the field would appear with a <dc:contributor> tag when exposed for harvesting.

How do items that are delimited by commas and semicolons appear in the OAI record?

Keywords and authors appear on multiple lines, with one line per metadata value. Other delimited items, which are entered into a text field and separated by a comma or semicolon in the metadata record, generally appear together on a single line in the OAI record.

Examples:

Each keyword/keyword phrase appears on its own line.

```
<dc:subject>Music History</dc:subject>
<dc:subject>Jazz</dc:subject>
```

Delimited fields besides authors and keywords will be displayed in a single line.

```
<dc:coverage>New York, 20th century, 1920s</dc:coverage>
```

Can I use OAI to harvest from other sources into my Digital Commons repository?

OAI does not function in Digital Commons as an import method, but as a means to expose metadata that is already published to the repository. You may consider using batch upload to import metadata records into your Digital Commons repository. More information is available in the batch import guide at http://digitalcommons.bepress.com/reference/14/.

Does OAI expose supplemental content?

No, URLs of additional files in Digital Commons are not exposed through OAI.

Do removed items appear in OAI results?

If items in the repository are withdrawn by an administrator, basic metadata (title, author, URL) will continue to display on the article page for permanent citation purposes. However, following an update, withdrawn items will no longer be exposed for harvesting.

For additional information, please contact Consulting Services at dc-support@bepress.com or call us weekdays at 510-665-1200, option 2, 8:30 a.m.–5:30 p.m. Pacific time.