OAI

OAI Interfaces

OAI-PMH Server

In the following sections and subpages, you will learn how to configure OAI-PMH server and activate additional OAI-PMH crosswalks. The user is also referred to OAI-PMH Data Provider for greater depth details of the program.

The OAI-PMH Interface may be used by other systems to harvest metadata records from your DSpace.

OAI-PMH Server Activation

DSpace's OAI-PMH server is enabled by default. However, you can choose to enable/disable it in your local.cfg using these configurations:

```
# Enable (true) or disable (false) OAI-PMH server
oai.enabled = true

# When enabled, OAI-PMH server is available at this path
oai.path = oai
```

If you modify either of these configuration, you must restart your Servlet Container (usually Tomcat).

- You can test that it is working by sending a request to: `[dspace.server.url]/[oai.path]/request?verb=Identify` (e.g. http://localhost:8080/server/oai/request?verb=Identify)
- The response should look similar to the response from the DSpace 7 Demo Server: https://api7.dspace.org/server/oai/request?verb=Identify

If you're using a recent browser, you should see a HTML page describing your repository. What you're getting from the server is in fact an XML file with a link to an XSLT stylesheet that renders this HTML in your browser (client-side). Any browser that cannot interpret XSLT will display pure XML. The default stylesheet is located in `[dspace-source]/dspace-oai/src/main/resources/static/style.xsl` and can be changed by configuring the `stylesheet` attribute of the `Configuration` element in `[dspace]/config/crosswalks/oai/xoai.xml`.

Relevant Links

- OAI 2.0 Server - basic information needed to configure and use the OAI Server in DSpace
- OAI-PMH Data Provider 2.0 (Internals) - information on how it's implemented
- http://www.openarchives.org/pmh/ - information on the OAI-PMH protocol and its usage (not DSpace-specific)

OAI-PMH Server Maintenance

After activating the OAI-PMH server, you need to also ensure its index is updated on a regular basis. Currently, this doesn't happen automatically within DSpace. Instead, you must schedule the `[dspace.dir]/bin/dspace oai import` commandline tool to run on a regular basis (usually at least nightly, but you could schedule it more frequently).

Here's an example cron that can be used to schedule an OAI-PMH reindex on a nightly basis (for a full list of recommended DSpace cron tasks see Scheduled Tasks via Cron):

```
# Update the OAI-PMH index with the newest content at midnight every day
# NOTE: ONLY NECESSARY IF YOU ARE RUNNING OAI-PMH
# (This ensures new content is available via OAI-PMH)
0 0 * * * [dspace.dir]/bin/dspace oai import > /dev/null
```

More information about the `dspace oai` commandline tool can be found in the OAI Manager documentation.

OAI-PMH / OAI-ORE Harvester (Client)

This section describes the parameters used in configuring the OAI-ORE / OAI-ORE harvester. This harvester can be used to harvest content (bitstreams and metadata) into DSpace from an external OAI-PMH or OAI-ORE server.

Supported in 7.1 or above

OAI Harvesting was not available in DSpace 7.0. It was restored in DSpace 7.1. See DSpace Release 7.0 Status

Harvesting from another DSpace

If you are harvesting content (bitstreams and metadata) **from** an external DSpace installation via OAI-PMH & OAI-ORE, you first should verify that the external DSpace installation allows for OAI-ORE harvesting.

If the external DSpace is running v6.x or below, it must be running both the OAI-PMH interface and the XMLUI interface to support harvesting content from it via OAI-ORE.

If the external DSpace is running v7.x or above, it just needs to be running the OAI-PMH interface.

You can verify that OAI-ORE harvesting option is enabled by following these steps:

1. First, check to see if the external DSpace reports that it will support harvesting ORE via the OAI-PMH interface. Send the following request to the DSpace's OAI-PMH interface: `http://[full-URL-to-OAI-PMH]/request?verb=ListRecords&metadataPrefix=ore`
    - The response should be an XML document containing ORE, similar to the response from the DSpace Demo Server: http://demo.dspace.org/oai/request?verb=ListRecords&metadataPrefix=ore
2. For 6.x or below, you can verify that the XMLUI interface supports OAI-ORE (it should, as long as it's a current version of DSpace). First, find a valid Item Handle. Then, send the following request to the DSpace's XMLUI interface: `http://[full-URL-to-XMLUI]/metadata/handle/[item-handle]/ore.xml`
    - The response should be an OAI-ORE (XML) document which describes that specific Item. It should look similar to the response from the DSpace Demo Server: http://demo.dspace.org/xmlui/metadata/handle/10673/3/ore.xml

OAI-PMH / OAI-ORE Harvester Configuration

There are many possible configuration options for the OAI harvester. Most of these are contained in the `[dspace]/config/modules/oai.cfg` file (unless otherwise noted below). They may be updated there or overridden in your `local.cfg` config file (see Configuration Reference).

Configuration File:	**`[dspace]/config/modules/oai.cfg`**
Property:	`oai.harvester.eperson`
Example Value:	`oai.harvester.eperson = admin@myu.edu`
Informational Note:	The EPerson under whose authorization automatic harvesting will be performed. This field does not have a default value and must be specified in order to use the harvest scheduling system. This will most likely be the DSpace admin account created during installation.
Property:	`oai.url`
Example Value:	`oai.url = ${dspace.server.url}/${oai.path}`
Informational Note:	The base url of the OAI-PMH disseminator webapp (i.e. do not include the /request on the end). This is necessary in order to mint URIs for ORE Resource Maps. The default value of `${dspace.baseUrl}/oai` will work for a typical installation, but should be changed if appropriate. Please note that `dspace.baseUrl` is defined in your `dspace.cfg` configuration file.
Property:	`oai.ore.authoritative.source`
Example Value:	`oai.ore.authoritative.source = oai`
Informational Note:	The webapp responsible for minting the URIs for ORE Resource Maps. If using oai, the `oai.url` config value must be set. • When set to 'oai', all URIs in ORE Resource Maps will be relative to the OAI-PMH URL (configured by `oai.url` above) • The URIs generated for ORE ReMs follow the following convention for either setting: http://\[base-URL\]/metadata/handle/\[item-handle\]/ore.xml
Property:	`oai.harvester.autoStart`
Example Value:	`oai.harvester.autoStart = false`
Informational Note:	Determines whether the harvest scheduler process starts up automatically when DSpace webapp is redeployed.
Property:	`oai.harvester.metadataformats.PluginName`

Example Value:	`oai.harvester.metadataformats.PluginName = \` `http://www.openarchives.org/OAI/2.0/oai_dc/, Simple Dublin Core`
Informational Note:	This field can be repeated and serves as a link between the metadata formats supported by the local repository and those supported by the remote OAI-PMH provider. It follows the form `oai.harvester.metadataformats.PluginName = NamespaceURI,Optional Display Name` . The pluginName designates the metadata schemas that the harvester "knows" the local DSpace repository can support. Consequently, the PluginName must correspond to a previously declared ingestion crosswalk. The namespace value is used during negotiation with the remote OAI-PMH provider, matching it against a list returned by the ListMetadataFormats request, and resolving it to whatever metadataPrefix the remote provider has assigned to that namespace. Finally, the optional display name is the string that will be displayed to the user when setting up a collection for harvesting. If omitted, the PluginName:NamespaceURI combo will be displayed instead.
Property:	`oai.harvester.oreSerializationFormat.OREPrefix`
Example Value:	`oai.harvester.oreSerializationFormat.OREPrefix = \` `http://www.w3.org/2005/Atom`
Informational Note:	This field works in much the same way as `oai.harvester.metadataformats.PluginName` . The `OREPrefix` must correspond to a declared ingestion crosswalk, while the Namespace must be supported by the target OAI-PMH provider when harvesting content.
Property:	`oai.harvester.timePadding`
Example Value:	`oai.harvester.timePadding = 120`
Informational Note:	Amount of time subtracted from the from argument of the PMH request to account for the time taken to negotiate a connection. Measured in seconds. Default value is 120.
Property:	`oai.harvester.harvestFrequency`
Example Value:	`oai.harvester.harvestFrequency = 720`
Informational Note:	How frequently the harvest scheduler checks the remote provider for updates. Should always be longer than *timePadding* . Measured in minutes. Default value is 720.
Property:	`oai.harvester.minHeartbeat`
Example Value:	`oai.harvester.minHeartbeat = 30`
Informational Note:	The heartbeat is the frequency at which the harvest scheduler queries the local database to determine if any collections are due for a harvest cycle (based on the *harvestFrequency*) value. The scheduler is optimized to then sleep until the next collection is actually ready to be harvested. The *minHeartbeat* and *maxHeartbeat* are the lower and upper bounds on this timeframe. Measured in seconds. Default value is 30.
Property:	`oai.harvester.maxHeartbeat`
Example Value:	`oai.harvester.maxHeartbeat = 3600`
Informational Note:	The heartbeat is the frequency at which the harvest scheduler queries the local database to determine if any collections are due for a harvest cycle (based on the *harvestFrequency*) value. The scheduler is optimized to then sleep until the next collection is actually ready to be harvested. The *minHeartbeat* and *maxHeartbeat* are the lower and upper bounds on this timeframe. Measured in seconds. Default value is 3600 (1 hour).
Property:	`oai.harvester.maxThreads`
Example Value:	`oai.harvester.maxThreads = 3`
Informational Note:	How many harvest process threads the scheduler can spool up at once. Default value is 3.
Property:	`oai.harvester.threadTimeout`
Example Value:	`oai.harvester.threadTimeout = 24`
Informational Note:	How much time passes before a harvest thread is terminated. The termination process waits for the current item to complete ingest and saves progress made up to that point. Measured in hours. Default value is 24.
Property:	`oai.harvester.unknownField`
Example Value:	`oai.harvester.unkownField = fail | add | ignore`

Informational Note:	You have three (3) choices. When a harvest process completes for a single item and it has been passed through ingestion crosswalks for ORE and its chosen descriptive metadata format, it might end up with DIM values that have not been defined in the local repository. This setting determines what should be done in the case where those DIM values belong to an already declared schema. *Fail* will terminate the harvesting task and generate an error. **Ignore** will quietly omit the unknown fields. **Add** will add the missing field to the local repository's metadata registry. Default value: **fail**.
Property:	`oai.harvester.unknownSchema`
Example Value:	`oai.harvester.unknownSchema = fail | add | ignore`
Informational Note:	When a harvest process completes for a single item and it has been passed through ingestion crosswalks for ORE and its chosen descriptive metadata format, it might end up with DIM values that have not been defined in the local repository. This setting determines what should be done in the case where those DIM values belong to an unknown schema. **Fail** will terminate the harvesting task and generate an error. **Ignore** will quietly omit the unknown fields. **Add** will add the missing schema to the local repository's metadata registry, using the schema name as the prefix and "unknown" as the namespace. Default value: **fail**.
Property:	`oai.harvester.acceptedHandleServer`
Example Value:	`oai.harvester.acceptedHandleServer = \` `hdl.handle.net, handle.test.edu`
Informational Note:	A harvest process will attempt to scan the metadata of the incoming items (identifier.uri field, to be exact) to see if it looks like a handle. If so, it matches the pattern against the values of this parameter. If there is a match the new item is assigned the handle from the metadata value instead of minting a new one. Default value: *hdl.handle.net* .
Property:	`oai.harvester.rejectedHandlePrefix`
Example Value:	`oai.harvester.rejectedHandlePrefix = 123456789, myeduHandle`
Informational Note:	Pattern to reject as an invalid handle prefix (known test string, for example) when attempting to find the handle of harvested items. If there is a match with this config parameter, a new handle will be minted instead. Default value: *123456789* .

Setting up a harvest to import content into a collection

There are two options to set up a collection for harvesting. One is by using the DSpace scripts "harvest", the other is by setting up the content source of a collection through the UI.

Using the "harvest" script

The harvest script can be called from both the CLI and REST API by calling "harvest". It uses the paramaters as defined in the following table.

Short option	Long option	Argument	Explanation
-p	--purge	*[none]*	Delete all the items in the collection provided with the `-c` parameter.
-r	--run	*[none]*	Run the standard harvesting procedure for the collection provided with the `-c` parameter.
-g	--ping	*[none]*	Verify that the server provided through the `-a` parameter and the set provided through the `-i` parameter can be resolved and work.
-s	--setup	*[none]*	Set the collection provided with the `-c` parameter up for harvesting. The server will need to be provided through the `-a` parameter, and the oai set id needs to be provided by the `-i` parameter.
-S	--start	*[none]*	Start the harvest loop for all collections.
-R	--reset	*[none]*	Reset the harvest status on all collections.
-P	--purgeCollections	*[none]*	Purge all harvestable collections.
-o	--reimport	*[none]*	Reimport all items the items in the collection provided by the `-c` parameter. This is the equivalent of running both the `-p` and the `-r` command for the provided collection.
-c	--collection	[id-or-handle]	The harvesting collection (handle or id)
-t	--type	[type-code]	The type of harvesting: 0 for no harvesting, 1 for metadata only, 2 for metadata and bitstream references (requires ORE support), 3 for metadata and bitstreams (requires ORE support)
-a	--address	[url]	The address of the OAI-PMH server to be harvested
-i	--oai_set_id	[set-id]	The id of the PMH set representing the harvested collection. In case all sets need to harvested the value "all" should be provided.

-m	--metadata_format	[format]	The name of the desired metadata format for harvesting, resolved to namespace and crosswalk in the dspace.cfg
-h	--help	*[none]*	Print the help message
-e	--eperson	[email]	**(CLI ONLY)** The eperson that performs the harvest. When the command is used from the REST API, the currently logged in user will be used.

Examples of harvesting a collection through CLI commands

1. Verify whether the harvester source can be reached

```
dspace/bin/dspace -g -a https://harvest.source.org -i harvest-set
```

Replace `https://harvest.source.org` with the source you want to use, the `harvest-set` with the set/sets you want to harvest or `all` in case you want to harvest all sets.

2. Set up a collection for harvesting

```
dspace/bin/dspace harvest -s -c 123456789/123 -a https://harvest.source.org -i harvest-set -m dc -t 1
```

Replace the `123456789/123` with your collection, `https://harvest.source.org` with the source you want to use, the `harvest-set` with the set /sets you want to harves or `all` in case you want to harvest all sets. The `-m` parameter indicated the metadata format to be used and the `-t` parameter indicates the harvest type to be used. When the value `0` is used for `-t` , harvesting will be disabled.

3. Run the harvest for the set up collection

```
dspace/bin/dspace harvest -r -c 123456789/123 -e harvest-user@dspace.org
```

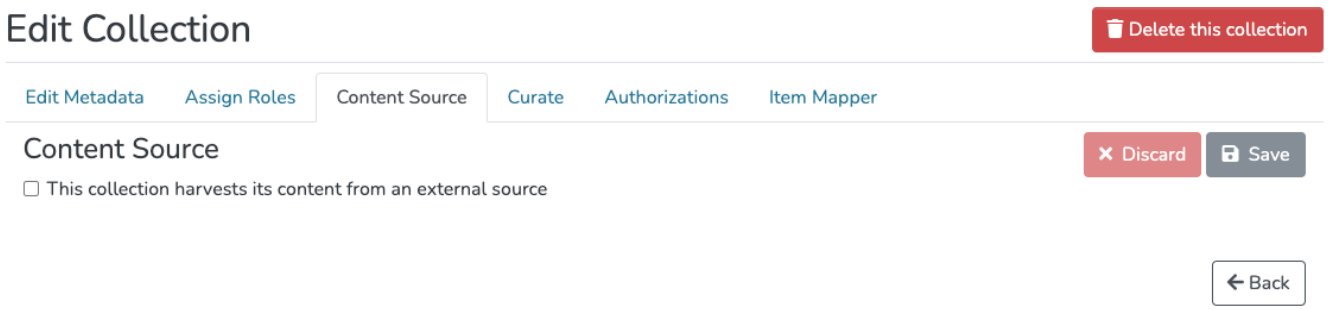
Replace the `123456789/123` with your collection, the `harvest-user@dspace.org` with an existing user in DSpace that has sufficient rights to perform the ingestion.

Setting up a harvest content source from the UI

A collection can be configured to retrieve its content from an external source. This can be done from the "Edit Collection" UI by using the following steps.

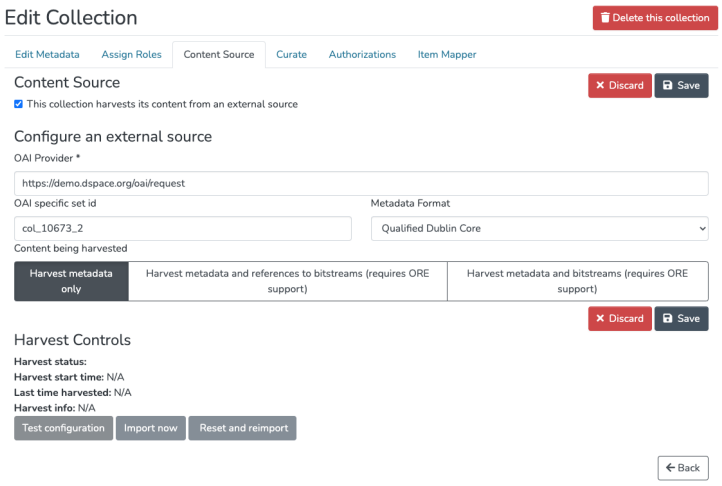
1. Configure the collection to harvest its content from an external source

Navigate to the "Edit collection" > "Content Source" tab. Tick the checkbox "This collection harvests its content from an external source".



2. Configure the harvest source

Once the checkbox has been ticket, the OAI provider, set id and metadata format can be configured. An example of the configuration can be found in the image below.

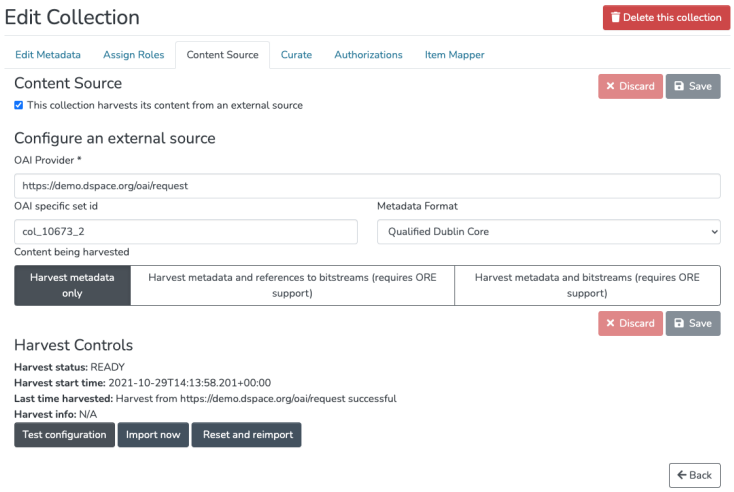


When all sets need to be harvested, the field can be left empty.

The server configuration will be tested upon clicking the "Save" button.

3. Start the harvest

Click the "Import Now" button to start the import. When the import has started, the button will indicate that the import is in progress, however, there is no need to remain on this page as the harvest will continue to run after leaving this page.



If the current server configuration needs to be retested at a later point, the "Test configuration" button can be used. To fully reset the collection by purging all items and starting a reimport, click the "Reset and reimport" button.

DSpace 7 Demo - OAI-PMH

- https://demo.dspace.org/server/oai/request?verb=Identify