OAI-PMH Harvester

Overview

The OAI-PMH Harvester allows users to load objects into Rosetta directly from an OAI-PMH digital repository. It has two modes of operation:

- Loading new objects into Rosetta
- Updating objects already existing in Rosetta – This mode can be divided in to two scenarios:
  - Content originally created in Rosetta – This typically means that the metadata was harvested and enriched by an external system and is now being passed back to Rosetta for harvesting. In this case, the match is on the Rosetta PID.
  - Content originally created outside of Rosetta –This refers to content that was loaded previously into Rosetta but continues to be maintained in an external system. The latter is now sending updates to those previously loaded records. In this case, the match is on the original system ID, as stored in the DNX.

Loading New Objects

When loading new objects, the OAI-PMH Harvester job creates SIPs for loading with a Submission Job (see **Submission Jobs**). Users can choose to create SIPs in either METS or Dublin Core format and schedule harvests to run at given time intervals. You can use this job for migration purposes, or for ongoing loading from an external repository into Rosetta (for example, to preserve data managed by another repository).

Updating Existing Objects

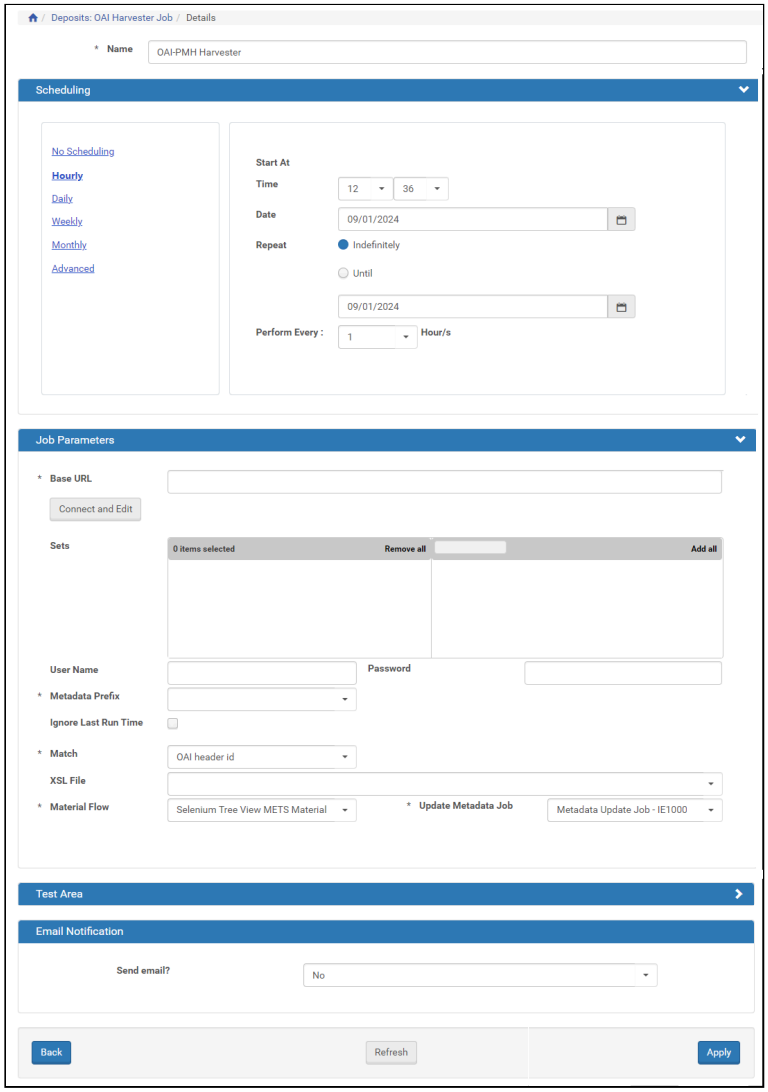
When updating existing objects, indicate if the content being harvested is of Rosetta origin or an external repository origin. If the content is of external origin, indicate the material flow to be used when loading new objects. If the content is of Rosetta origin, you can indicate a qualifier and string with which to match. Upon matching, the harvester can either create an update MD package with the new (transformed) metadata or ignore the changes. You can select one of the existing Update MD jobs for which the harvester creates a package or select an option to ignore the record.

Creating the OAI_PMH Harvest Job

This section describes the procedure for creating an OAI-PMH harvest job.

To create an OAI-PMH new harvest job:

1. Under **Deposits > Jobs** > **OAI Harvester Job** click **Add Job.**



OAI-PMH Harvester

2. Enter a name for your job (for example, Fedora Harvest)
3. Schedule your job.

Note

Rosetta appends the previously run and current timestamps to OAI-PMH harvest requests. This is not affected by your scheduling preferences.

---

4. Enter the OAI-PMH harvesting parameters as described in the following table:

OAI-PMH Harvesting Parameters

Parameter	Explanation	Mandatory?
Base URL	The OAI-PMH server base URL (for example: host:port/oaiprovider/request)	Yes
Connect and Edit	Select to populate the Set and Metadata Prefix fields with data from the base URL.	
Sets	The OAI-PMH selective harvesting setSpec element. Select all that apply.	No

Parameter	Explanation	Mandatory?
Metadata Prefix	The OAI-PMH metadataPrefix element	Yes
User Name	Used for OAI-PMH servers that require login. If defined, a basic authentication header is sent with the request	No
Password	Used for OAI-PMH servers that require login. If defined, a basic authentication header is sent with the request	No
Ignore Last Run Time	Select to disregard the last run time.	No

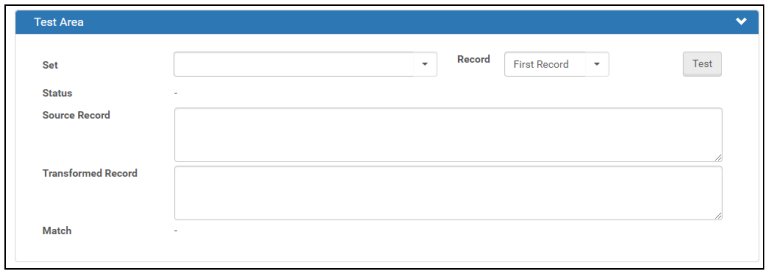
Match	Select from the drop-down list:
	• OAI Header ID – Imported records originate in an external repository
	• Other identifiers of the following type:
	  • Identifier (DC)
	  • Identifier (DCTERMS)
		Yes

Qualifier	(Available with Rosetta Origin) Identifier of the DC record.	No
Contains String	(Available with Rosetta Origin) The string at the beginning of the DC record.	No
XSL File	Transforms OAI-PMH records to Rosetta METS/Dublin Core records. If no XSL file is selected, the default XSL that transforms to DC is used.	No
Material Flow	(Available with External Repository Origin or Do Not Match) The Rosetta Material Flow used when importing new records, which is used to generate/process the SIPs	Yes
Update Metadata Job	Select the metadata job to update records or (with external origin) select to not update.	Yes

5. You can test the job to test connectability and record transformation and match per selected job configuration.
   - From the **Set** drop-down list, select the set in which the source record is located.
   - From the **Record** drop-down list, select one of the following options:
     - First – the first record
     - Random – Rosetta picks a record at random
     - By Identifier – Enter the ID of the record that you want Rosetta to test
   - Click **Test**

   If the test is successful, the following occurs:

   - **Success** appears in the Status field.
   - The source record appears in the **Source** field.
   - The transformed record appears in the **Transformed Record** field.
   - If a match is found, the IE PID of the matched record appears (unless **Do not match** was selected). If no match is found, **No match found** appears.



Test Example

Note

- A key to Rosetta OAI-PMH harvesting is the location of the reference to the filestreams in the metadata. The OAI-PMH record must provide a URI that Rosetta can access (either via HTTP or NFS) to obtain the files. Depending on the `xslt` transformation, these references are either placed in the METS fileSec or the DC stream source field (defined in the DC content structure configuration). Rosetta does not harvest the filestreams during OAI-PMH metadata harvest – this is done during the submission itself.
- The selected material flow and the `xslt` transformation file must be aligned. For example, if the material flow uses a METS content structure, you must select an `xslt` transformation that produces METS. Rosetta provides built-in examples for `xslt` transformation. To add/edit `xslt` transformers, go to **Deposits > Advanced Tools > OAI Harvester Transformation**.
- To create a dedicated submission job for processing OAI-PMH harvested records, you must also configure a dedicated material flow for the harvester and the submission job.
- The OAI-PMH harvester job places a lock file (`.locked`) in the submission folders it creates and removes the lock only after the job is completed. The submission job does not process a folder with a lock file. See (**About Submission Jobs**) for more details.
- Rosetta generates one SIP for every OAI-PMH response. If the OAI-PMH server returns a resumption token, another request is sent and another SIP is generated from each subsequent response. The number of IEs depends on the number of records returned per response. For large IEs, it is therefore recommended to configure the OAI-PMH server to return fewer records per response.
- It is generally recommended to use a Dublin Core content structure and material flow for simple objects (IEs with one representation). If you need to apply more complex logic (for example, map streams to separate representations) use a METS flow.
- Rosetta stores the OAI identifier header in the IE Original Object Identifier DNX field. This is done either by direct mapping (in the case of METS transformation) or indirect mapping. (Rosetta stored this information in a temporary DC field, which is later mapped to IE Original Object Identifier.) If you select **External Repository Origin**, Rosetta searches the repository for other records (within the same institution) based on this field and value.
- The harvesting job ignores OAI-PMH records that have a **Deleted** status.

---

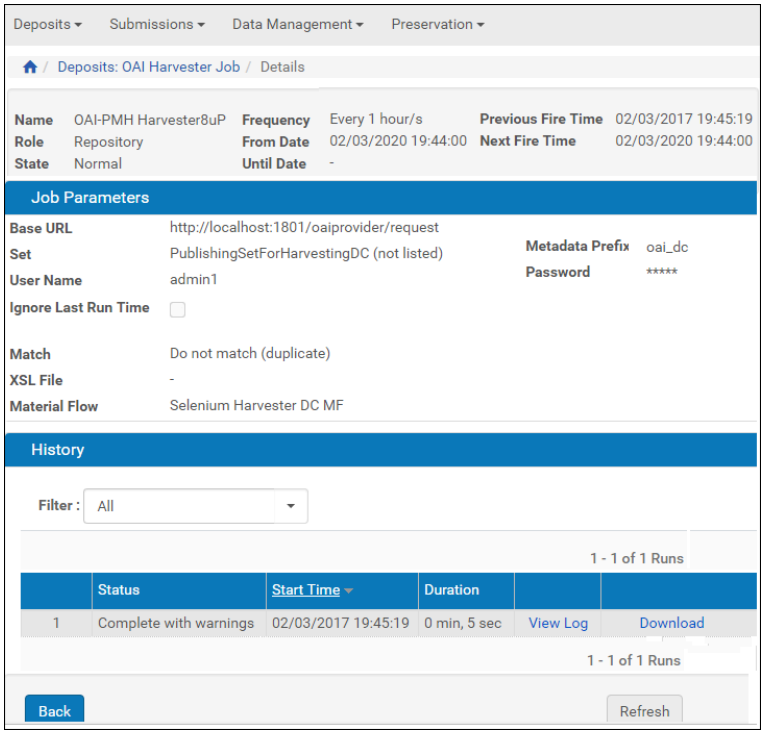
6. From the **Send Email?** drop-down list, you have the following options:
   - No (default)
   - Yes (if there was work)
   - On failure only
     The email contains a report/log file for the job.
7. Click the **Apply** button to add the job to the list of OAI-PMH harvest jobs.

Viewing the OAI-PMH Harvest Job History

You can view the history of the OAI-PMH harvest job.

To view the history of the OAI-PMH harvest job:

From the OAI Harvester Job page, click the **History** link for a job. The OAI Harvester Job History page opens:



OAI Harvester Job History

A list of times the job ran is displayed. The following actions are available:

- Click **View Log** to see the log of the job.
- Click **Download** to download the job log.