OAI-PMH Harvester best practices

- **Product:** Rosetta
- **Product Version:** 5.3+

Question

What are the OAI-PMH Harvester best practices, limitations and known issues?

Answer

Prior to setting up OAI-PMH Harvester for migration legacy repository see article on the developer network:

https://developers.exlibrisgroup.com/blog/Migrating-from-Your-Legacy-Digital-Repository-to-Rosetta

Make sure you understand the purpose of the OAI-PMH Harvester use. You can use it:

1. To synchronize the metadata of IEs already ingested in Rosetta with an external system:
    - Objects originating from Rosetta
    - Objects created in external system which still manage the metadata
    - From v5.3. the OAI-PMH harvester can match records by any DC or DCTERMS identifier, not only based on external origin identifier (OAI Header ID) or the Rosetta origin (dc:identifier)
    - Schedule an Update Metadata Job to perform the on-going updates

2. To load new objects (files + metadata) into Rosetta for the first time:
    - The OAI-PMH Harvester will create SIPs (Dublin Core or METS xml with metadata)
    - Use Do not match (duplicate) in Match parameter in the OAI-PMH Harvester configuration
    - When user name and password are left empty in the OAI-PMH Harvester configuration no authentication will be performed
    - The Submission job will upload the files referenced in these SIPs into Rosetta
    - File references in the SIP xml must point to an actual file (not a resolver, e.g. DigiTool's Delivery Manager).
      If using an URL, it must contain a legitimate filename.
      Note that the "*Migrating Your Digital Repository to Rosetta*" article includes XSL transformation examples for the dc:identifier (e.g. DSpace, DigiTool, DigitalCommons, ContentDM)
    - In the Content structure be sure to define a stream source origin (typically dc:identifier)

Testing phase:

- Use the OAI-PMH Harvester configuration Test area to verify the connectivity and your XSL Transformation
- Use 'ignore last run time' checkbox when repeating test ingests
- Test the submission job on larger sets to see if the source files can be downloaded

Other recommendations:

- Downloading files via http can take time.
  For larger migrations the file references in SIP xml should point to a NFS location mounted to Rosetta
- Rosetta will access all URLs provided the SIP xml stream origin (dc:identifier).
  Be sure that there are valid objects on these URLs.
  Think about how the source system handles versioning or delete of the files.
- Placing the set and metadata prefix configuration into the BaseURL is not recommended.
  The set and metadata prefix will be appended to the base URL from the configuration form fields.
- Rosetta cannot harvest multiple sets at a time.

Note: OAI Transformation xsl currently supports XSLT 1.0

- **Article last edited:** 26-Nov-2018