OAI-PMH Best Practices

- Verifying the connection to OAI-PMH service
- Starting harvest
  - Max number of records per response
  - URL
  - Resumption token
  - Initial full harvest
  - Time required to complete
  - Monitoring harvest
- Suppressed records
- Deleted records
- Slow Performance
- Forcing record updates

Verifying the connection to OAI-PMH service

After the setup of OAI-PMH module has been done, sending Identify request to FOLIO’s OAI-PMH service is the best way to confirm the connection between the harvesting client and the service.

https://<edge-url-oai>?<api-key>?verb=Identify

If the connection is successful, the following response should be returned:

Identify response Example

```xml
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<OAI-PMH xmlns:oai-identifier="http://www.openarchives.org/OAI/2.0/oai-identifier"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <responseDate>2021-10-21T20:31:38Z</responseDate>
  <request verb="Identify">http://folio.org/oai</request>
  <Identify>
    <repositoryName>FOLIO_OAI_Repository</repositoryName>
    <baseURL>http://folio.org/oai</baseURL>
    <protocolVersion>2.0</protocolVersion>
    <adminEmail>oai-pmh@folio.org</adminEmail>
    <earliestDatestamp>1970-01-01T00:00:00Z</earliestDatestamp>
    <deletedRecord>persistent</deletedRecord>
    <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
    <compression>gzip</compression>
    <compression>deflate</compression>
    <description>
      <oai-identifier:oai-identifier>
        <oai-identifier:scheme>oai</oai-identifier:scheme>
        <oai-identifier:repositoryIdentifier>folio.org</oai-identifier:repositoryIdentifier>
        <oai-identifier:delimiter>:</oai-identifier:delimiter>
        <oai-identifier:sampleIdentifier>oai:folio.org:diku/3c4ae3f3-b460-4a89-a2f9-78ce3145e4fc</oai-identifier:sampleIdentifier>
      </oai-identifier:oai-identifier>
    </description>
  </Identify>
</OAI-PMH>
```

Starting harvest

When a harvest starts the system checks first the inventory for the updates (new or modified records) then retrieves underlying MARC Bib records from SRS. If the harvest was triggered with the metadataPrefix set to marc21_withholdings, the holdings and items data is append as described in MODOAIp MH-102.

Max number of records per response
The number of records returned in the ListRecords response is determined by the Max number of records per response setting. Its value can be configured under Settings OAI-PMH Technical and can be set between 1 and 500. The default value is 100 and this is also a recommended value for libraries that have instances with thousands of items associated with them (even if the items are spread across multiple holdings).

**URL**

The URL for harvesting consists from the following elements:

- **First request:**
  ```
  https://edge-url/oai/api-token?verb=ListRecords&metadataPrefix=<metadataPrefix>&from=<yyyy-mm-dd>&until=<yyyy-mm-dd>
  ```

  The "from" and "until" parameters can take include timestamp and the format is then: yyyy-mm-ddThh:mm:ssZ. This is especially helpful during the troubleshooting when the harvest needs to be limited to some specific hours when the records were updated. The specified time is in UTC (Coordinated Universal Time).

- **Consecutive requests (with resumption token):**
  ```
  ```

**Resumption token**

The resumption token is included in the response if the number of harvested records is larger than the configured Max records per response value.

Example of the resumption token:

```
<resumptionToken cursor="0"
>bWV0YWRhdGFQcmVmaXg9bWFyYzIxX3dpdGhob2xkaW5ncyZmcm9tPTIwMTIwODA4LTEwLjEuMTMwODM1MTI2X3dpdGhob3N0aW09IjIwMjItMjItMDUtOTYtNzAwX3dpdGhob3N0aW09IjIwMjItMjItMDUtOTYtNzAwX3dpdGhob3N0aW09IjIwMjItMjItMDUtOTYtNzAwX3dpdGhob3N0aW09IjIwMjItMjItMDUtOTYtNzAwX3dpdGhob3N0aW09IjIwMjItMjItMDUtOTYtNzAwX3dpdGhob3N0aW09IjIwMjItMjItMDUtOTYtNzAwX3dpdGhob3N0aW09IjIwMjItMjItMDUtOTYtNzAwX3dpdGhob3N0aW09IjIwMjItMjItMDUtOTYtNzAwX3dpdGhob3N0aW09i
</resumptionToken>
```

The resumption token can be decoded using online Base64decode tool. Among others, the token contains following elements:

- starting position of the harvest
  - **Example:** cursor = "0" means that it is the first request
- metadataPrefix used in the request
  - **Example:** metadataPrefix=marc21_withholdings
- number of records in the harvested batch
  - **Example:** offset=100
- unique identifier of the instance that will be harvested next:
  - **Example:** nextRecordId=0fc89818-6a91-45da-974d-9b16773a7e4e
- how long the token is valid
  - **Example:** until=2021-10-21

**Initial full harvest**

The full harvest takes a significant amount of FOLIO’s resources (inventory and SRS records) and should be carefully planned. Usually, libraries run their full harvest over weekends. Following scenarios are not supported and will lead to extremely poor performance and eventually might crash the system:

- Running concurrently multiple full harvests. Note that sharing the harvesting link with multiple users will most likely lead to starting multiple harvests at the same time
- Running full harvest during large updates to inventory and SRS records (importing data, reloading data)

**Time required to complete**

How long it will take for the harvest to complete depends on:

- Configured number of records per response. The number can be set up to 500 but the recommended value is 100 or 200 records
- Number of holdings and items associated with instance records when harvesting with metadataPrefix set to marc21_withholdings

In tests conducted on collections with 5 million records the harvest took ~10 hours with 100 records per response and ~9 hours with 200 records per response. The harvest of 8 million records took ~15 hours with 200 records per response.
Monitoring harvest

Starting with Lotus HF1 OAI-PMH provides APIs for monitoring the harvest. In prior releases this information was only available in module logs. The request GET /oai/request-metadata will provide information about currently running harvests and information about the status of retrieving records identifiers from UUIDs, number of successfully saved record, number of skipped records due to the lack of underlying SRS record and also other information:

```
{
  "requestMetadataCollection": [
    {
      "requestId": "1248ca38-2ac9-431d-bab8-2fc15ce46db7",
      "lastUpdatedDate": "2022-06-08T16:45:44.904+00:00",
      "streamEnded": true,
      "downloadedAndSavedInstancesCounter": 8305800,
      "failedToSaveInstancesCounter": 0,
      "returnedInstancesCounter": 7974441,
      "skippedInstancesCounter": 331357,
      "failedInstancesCounter": 2,
      "suppressedInstancesCounter": 0
    },
    {
      "requestId": "f6987f4c-a929-4a1b-ab06-e8398b38f482",
      "lastUpdatedDate": "2022-06-07T17:13:09.98+00:00",
      "streamEnded": true,
      "downloadedAndSavedInstancesCounter": 866,
      "failedToSaveInstancesCounter": 0,
      "returnedInstancesCounter": 860,
      "skippedInstancesCounter": 8,
      "failedInstancesCounter": 0,
      "suppressedInstancesCounter": 0
    }
  ],
  "totalRecords": 2
}
```

Sending the request with request identifier, for example GET /oai/request-metadata/(requestId)/failed-to-save-instances, will return the list of UUIDs for the records that failed to save. The list might be used for troubleshooting existing data issues in FOLIO inventory or SRS. For more information see mod-oai-pmh README.md file or FOLIO API documentation.

Suppressed records

When the **Suppressed records processing** setting (Settings OAI-PMH Behavior) is set to "Transfer suppressed records with discovery flag value", the records marked as suppressed are included in the response with added subfield t.

- Suppressed instance records will have the subfield t in the field 999 set to 1. If the instance is not suppressed the value will be 0

```
<marc:datafield tag="999" ind1="f" ind2="f">
  <marc:subfield code="s">d2c1534d-41da-49c5-858b-850dabe23d1fa</marc:subfield>
  <marc:subfield code="t">1</marc:subfield>
</marc:datafield>
```

- Suppressed holdings and item records will have the subfield t in the fields 856 and 952 set to 1.
Deleted records

When the Deletion record support setting (Settings OAI-PMH Behavior) is set to “Persistent” and the record is marked as deleted (MARC LDR 05 is set to “d”), the record will be a part of the response with the header status set to deleted.

Deleted record example

If you use an API calls to delete Instances the following steps are required to assure that the discovery is updated as well:

- Set ldr05 to “d” in the underlying SRS record. This can be done by editing the record in QuickMarc.
- Let the incremental harvest get the information about the deleted record.
- Delete instance record and corresponding SRS record via through API call

Slow Performance

The harvesting of 5 million records should not take more than 11 hours in Juniper and less than 10 in Kiwi. It is highly recommended to run REINDEX, VACUUM and ANALYZE after major updates to the inventory tables in PostgreSQL database. It is highly recommended to run ANALYZE on a regular basis.
REINDEX and ANALYSE command

```
REINDEX index <tenant>_mod_inventory_storage.audit_item_pmh_createddate_idx ;
REINDEX
xxxx=> REINDEX index <tenant>_mod_inventory_storage.audit_holdings_record_pmh_createddate_idx;
REINDEX
xxxx=> REINDEX index <tenant>_mod_inventory_storage.holdings_record_pmh_metadata_updateddate_idx;
REINDEX
xxxx=> REINDEX index <tenant>_mod_inventory_storage.item_pmh_metadata_updateddate_idx;
REINDEX
xxxx=> REINDEX index <tenant>_mod_inventory_storage.instance_pmh_metadata_updateddate_idx;
REINDEX
xxxx=> analyze verbose <tenant>_mod_inventory_storage.instance;
ANALYZE
xxxx=> analyze verbose <tenant>_mod_inventory_storage.item;
ANALYZE
xxxx=> analyze verbose <tenant>_mod_inventory_storage.holdings_record;
ANALYZE
```

Forcing record updates

Staff actions to Instances, Holdings, and Items through the UI automatically trigger updates. However API based processes modifying the discovery flag, locations, or other fields to the storage endpoints does not trigger an update. To trigger records for update, you must issue a GET to any of the following /inventory endpoints and then a PUT the record retrieved to trigger an update. Although this process is slow, a maximum of 4 threads is recommended:

- /inventory/instances
- /inventory/holdings
- /inventory/items