Archidora

Introduction

Archidora is the Archivematica-Islandora Integration Module. Archivematica provides a preservation system that the Archidora module integrates into Islandora.

It was developed in a partnership between Artefactual Systems and Discovery Garden, sponsored by the University of Saskatchewan Library.

About Archivematica

Archivematica is a free and open-source digital preservation system that is designed to maintain standards-based, long-term access to collections of digital objects. It uses a micro-services design pattern to provide an integrated suite of software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model. Users monitor and control the micro-services via a web-based dashboard. Archivematica uses METS, PREMIS (events, agents, rights and restrictions), Dublin Core, the Library of Congress BagIt specification and other best practice standards and practices to provide **trustworthy**, **authentic**, **reliable**, and **interoperable** archival packages (AIPs) for storage in your preferred repository.

Archivematica provides several decision points that give the user control over choices about format identification tools, printing the original order of the directories ingested, examining contents for private and personal information, extracting contents of packages and forensic images, transcribing content, and more. Users may also preconfigure most of these options for seamless ingest to archival storage and access. Archivematica offers many ingest workflows: metadata and submission documentation import, zipped and unzipped Bag ingest, digital forensic image processing, SIP arrangement, manual normalization, and dataset management.

You may read more about Archivematica here.

About Archidora

- Presentation by Justin Simpson, Artefactual at the iCamp BC - February 16 - 18, 2015
- Archidora: Integrating Archivematica and Islandora, Code4Lib Journal, January 2018

Download

Islandora module: https://github.com/Islandora-Labs/archidora

Archivematica: Archivematica 1.6.1 and Storage Service 0.10.0 or later is recommended; download from http://www.archivematica.org.

This integration is currently (as of 1.6/0.10 release) considered a beta feature. Support for Archivematica and/or the Storage Service running on secure servers (https) will likely require Storage Service 0.11 or later.

Installation

Installation and testing is similar to any Drupal module. Please see Installing the Islandora Enhancement Modules for details.

Configuration

In the Archivematica Storage Space:

- Create a Space with access protocol FEDORA via SWORD2; and create a Location within that Space (purpose = FEDORA deposits). The Fedora URL, username and password will need to be entered here. See Archivematica documentation for more details.
- Make sure the Pipeline is configured with the correct credentials (with API username and API key matching the appropriate dashboard user)

Archivematica may also be configured to call back to Islandora to delete the high-res "OBJ" datastreams.

- The callback is configured in the Storage Service - Administration | Service callbacks

    - URI: http://{islandora-base-url}/islandora/object/<source_id>/archidora/{Islandora API}/delete
    - where the Islandora API key is the "Islandora Archivematica integration API key" generated on the Archidora admin screen
    - <source_id> should be entered as shown, with the angle brackets
    - Event: post-store
    - Method: post
    - Expected status: 200
- As a workaround for a known issue, as of Archivematica 1.6.x you must also set the checksum algorithm to SHA-512 (on the Archivematica dashboard)

Note: the OBJ datastreams are not deleted automatically, but rather are listed at the collection level (or compound object level) on the Manage | Archivematica tab. They can be deleted individually or in bulk. Note also that the callback does not currently work on objects whose access is restricted by a XACML policy.

On the Archivematica dashboard:

- Add the IP address of the storage service to the IP whitelist for the REST API. This is needed to allow transfers to be approved automatically.

Storage Service - gunicorn settings

- The default SS gunicorn worker class (gevent) is incompatible with Python's multiprocessing package, which is required for the Sword API. To resolve this:

1. Add the line `env SS_GUNICORN_WORKER_CLASS=sync` to the AM SS service config file at /etc/init/archivematica-storage-service.conf.

2. Reload the config and restart the SS service:

```
$ sudo initctl reload-configuration
$ sudo service archivematica-storage-service restart
```

3. Check the SS logs and expect the last `Using worker` line to be `Using worker: sync` and NOT `Using worker: gevent`:

```
$ sudo vi /var/log/upstart/archivematica-storage-service.log
```

Archivematica automation tools:

- It is recommended that automation tools be used to regularly clean up the Archivematica dashboard (transfers and ingests). With too many transfers/ingests remaining on the dashboard, response time becomes very slow. A script is available to hide completed transfers and ingests. If a transfer/SIP is rejected or fails, it will remain on the dashboard, but otherwise (even if there are other errors) it will be hidden.
- This script is part of Artefactual Systems' automation-tools repository (transfers/amclient.py). Crontab can be used to invoke it regularly.

```
sudo python amclient.py close-completed-transfers --am-user-name <username> <archivematica-api-key>
```

```
sudo python amclient.py close-completed-ingests --am-user-name <username> <archivematica-api-key>
```

where archivematica-api-key is the API key for the dashboard user.

In Islandora:

- configure Archidora, at `admin/islandora/archidora`
    - Archivematica Storage Service Base URL - normally http://archivematica-url:8000
    - Deposit Location - will be configured automatically once storage service URL is entered
    - Archivematica User - Archivematica dashboard user to be used for Islandora integration (not storage service)
    - Archivematica API Key - API key for the Archivematica dashboard user listed above
    - EM-IRI Solr field - used for constructing Sword API call (default is "RELS_EXT_edit_media_uri_ms")
    - AIP max age - new objects will not be added to a deposit after the specified time has elapsed
    - AIP max size - new objects will not be added to a deposit after the specified size has been reached. Note that this is really the transfer size; the AIP could be larger due to normalized objects
    - Cron time - the amount of time for which the queue of items will be allowed to process, at each cron invocation. Setting a higher time is recommended if compound objects are being ingested (especially manually), otherwise the relationships may not be included in the METS file sent to Archivematica
- Cron must be enabled. You may also need to add a rule to the firewall on the Fedora server to allow access from the Archivematica storage service (e.g. to port 8080)

As a side-effect of using Cron Queues, the submission of objects to Archivematica may not complete during any one invocation of Cron. It is also recommended that cron run at reasonably frequent intervals (e.g. every five minutes), otherwise the expected callbacks may not be triggered often enough.

- Collection-level configuration:
    - Optionally, check off "Don't Archive Children" to stop objects from being sent to Archivematica for a particular collection.

Batch processing

A sample drush script is available to ingest Islandora collections in batch (e.g. for objects created before archidora was deployed on an Islandora instance).

Usage:

```
sudo drush -u 1 archidora-send-collection-to-archivematica --target=islandora:collection1
```

or

```
sudo drush -u 1 asca --target=islandora:collection1
```

Currently, it is not recursive (but an unmerged pull request adds this functionality). It also ignores the "Don't Archive Children" setting.