Selective Harvesting:
Creating and ingesting selected data without OAI-PMH sets

Never Stand Still

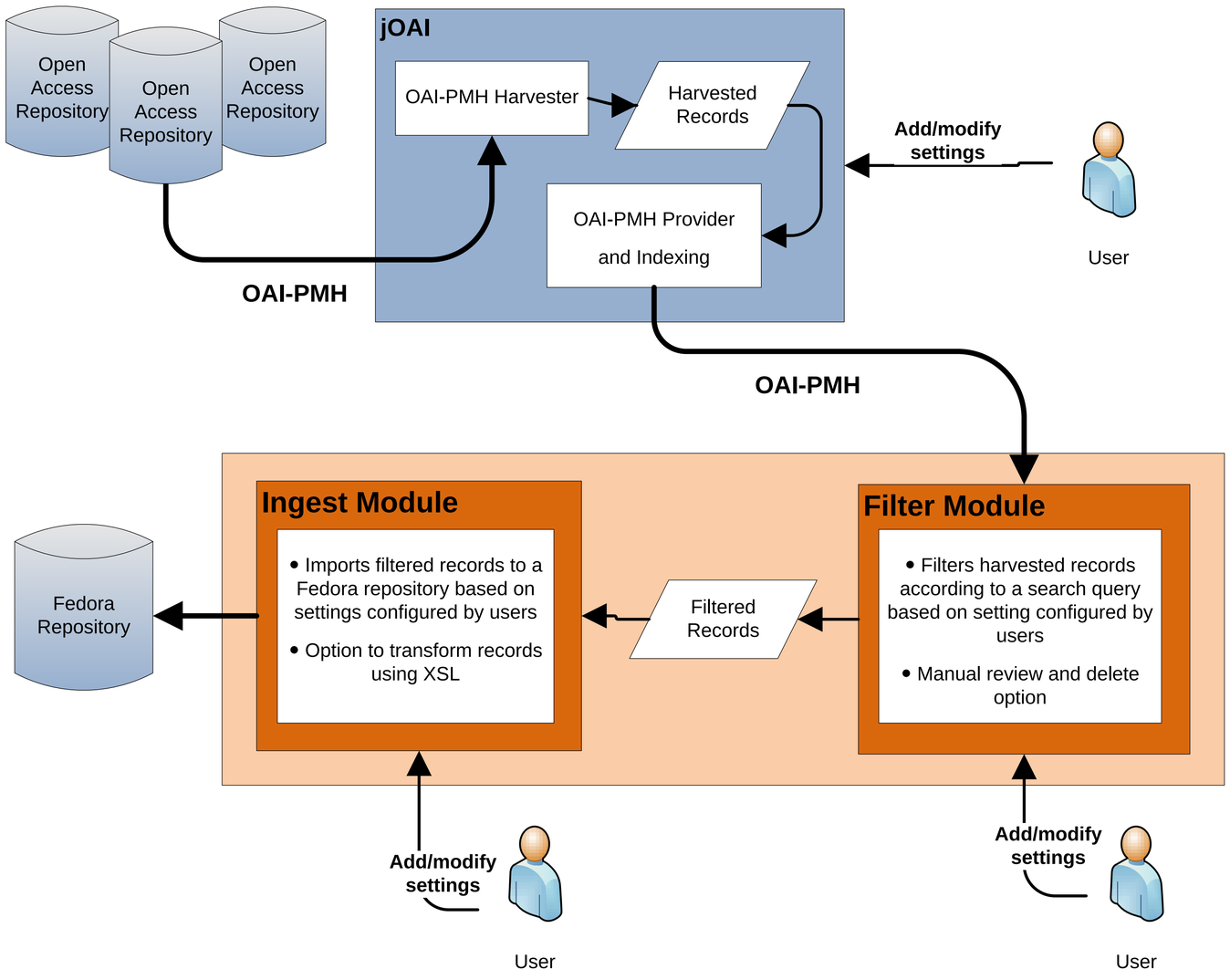
UNSW Library

Harry R. Sidhunata, Joanne L. Croucher, Maude Frances; University Library, University of New South Wales, Sydney, Australia

DEFINITION

The ***Selective Harvester*** supports scholarly communication and eResearch by providing a flexible and customisable mechanism to select and re-use metadata records from open access repositories. This workflow assists subject-based repositories in the selection and addition of relevant content.

Figure 1 *Selective Harvester* Model



PURPOSE

Conventionally, harvesting selected records from open access repositories using OAI-PMH[1] requires OAI-PMH Sets, a setting that is dependent on the configuration by the data provider. Set definition is also reliant on a degree of standardisation within the source metadata. Supply-side modifications can be a burden for the data provider, especially since a single repository may be harvested by many different harvesters.

The ***Selective Harvester*** was developed by University of New South Wales (UNSW) Library to solve these issues by providing harvesters with more flexible harvesting options and effectively reducing the burden of management and dependency on configuration by data providers. ***Selective Harvester*** allows OAI-PMH harvesters to be less dependent on supply-side changes by OAI-PMH data providers to harvest specific records.

FEATURES

- Open source Java-based web-applications
- Flexible configuration
- Harvests one or more open access OAI-compliant repositories
- Supports different XML metadata formats
- Option to define advanced search criteria to filter harvested resources
- Preview of filtered resources with option to manually review resources
- Scheduling facility (components can be set up to operate automatically at set intervals)
- Option to upload and use XSL files to transform the harvested metadata
- Ingests harvested and filtered resources to Fedora-based repository

IMPLEMENTATION

The ***Selective Harvester*** has been implemented on the NCHSR Clearinghouse[2], a subject-based repository developed jointly by UNSW Library and researchers at the National Centre in HIV Social Research (NCHSR) to harvest and ingest records from UNSWorks[3], the UNSW institutional repository, based on relevant keywords and affiliations.

It also has been implemented on the Membrane Research Environment (MemRE)[4], a component infrastructure project of the Advanced Membrane Technologies for Water Treatment Research Cluster funded by the CSIRO Water for a Healthy Country Flagship, to harvest and ingest records from DRIVER[5]. Keywords related to membrane research are used to filter harvested records from DRIVER.

Figure 2 *Filter Module* Model

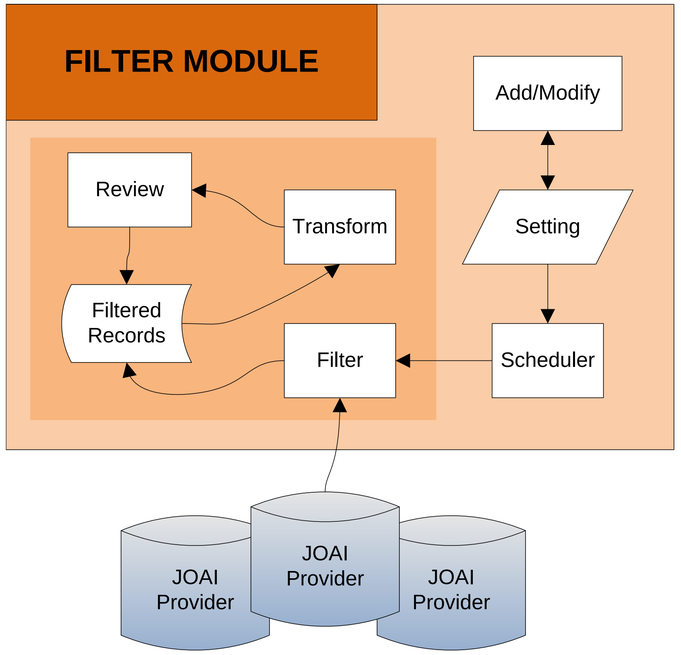
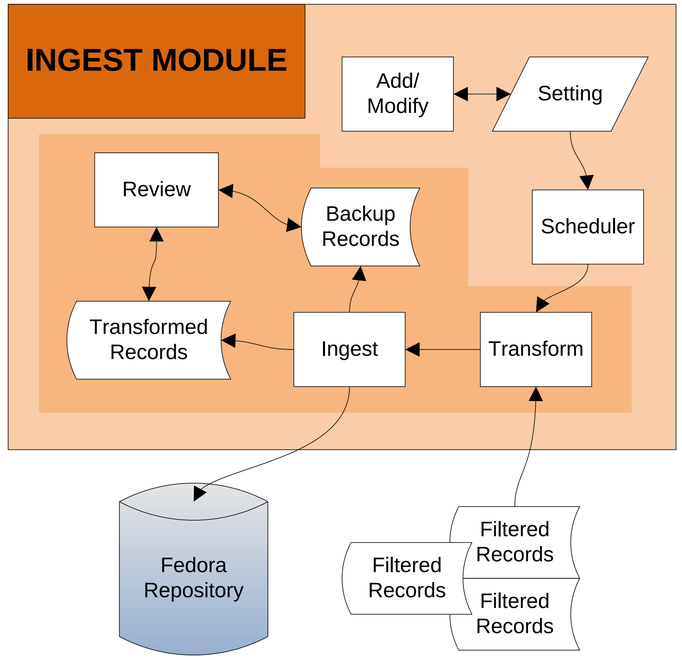


Figure 3 *Ingest Module* Model



Filter Module supports:

- Filtering harvested records according to jOAI search query
- Storing filter settings
- Filtering from multiple jOAI providers
- Option to manually review and delete filtered records
- Transformation using XSL support to simplify reviewing
- Deleting filtered records manually
- Filter scheduling

Ingest Module supports:

- Ingesting records to Fedora repository
- Ingesting from many source directories
- Transforming records to various metadata for ingestion
- Backup of all ingested records
- Simple duplicate checking at ingest
- Ingest scheduling

DESIGN

The ***Selective Harvester*** enables filtering and manages the import of selected records, which have been harvested from other open access repositories, into a Fedora repository. The model consists of two separate Java-based open source applications, each of which can be deployed and configured independently.

1) jOAI

This OAI harvester and provider was developed by Digital Learning Sciences (DLS) at the University Corporation for Atmospheric Research[6]. Highly configurable, the harvester enables the harvesting of multiple open access repositories. The provider is used to index and expose the harvested resources to the filter module.

2) Filter & Ingest Modules

Two independent modules have been developed at the UNSW Library. The *Filter Module* is used to filter and review harvested records. The *Ingest Module* is able to transform the harvested records prior to import into a Fedora repository[7]. JOAI, *Filter Module* and *Ingest Module* can also operate independently, as required. Figures 2 and 3 illustrate the workflows of the Filter and Ingest modules.

References

1. *Open Archives Initiative: Protocol for Metadata Harvesting (OAI-PMH)*, Available from: http://www.openarchives.org/pmh
2. *NCHSR Clearinghouse*, Available from: http://ssrm.nchsr.arts.unsw.edu.au
3. *UNSWorks*, Available from: http://www.unsworks.unsw.edu.au
4. *MemRE*: Colloborative development of an integrated research environment, Available from: http://membranes.edu.au
5. *DRIVER*: a cohesive pan-European infrastructure of Digital Repositories, offering sophisticated functionality services to both researchers and the general public, Available from: http://www.driver-repository.eu
6. *jOAI Overview: The Java-based Open Archives Initiative Data Provider & Harvester*, Available from: http://www.dlese.org/oai
7. *Fedora Commons Repository Software*, Available from: http://fedora-commons.org

More Information

Selective Harvesting: Creating and ingesting selected data without OAI-PMH sets, Available from: http://handle.unsw.edu.au/1959.4/50821

DOWNLOAD

The ***Selective Harvester*** Filter and Ingest Modules are written in Java, and available as Open Source software via SourceForge: http://sourceforge.net/projects/selectharvest

ACKNOWLEDGEMENTS

The ***Selective Harvester*** was developed at the University Library, University of New South Wales. This product includes software developed by Digital Learning Sciences (DLS) http://www.dlsciences.org at the University Corporation for Atmospheric Research http://www.ucar.edu

Contact

Harry Sidhunata
Technical Support Officer
Library Repository Services, University Library, UNSW
Email: h.sidhunata@unsw.edu.au

Jo Croucher
Support Officer
Library Repository Services, University Library, UNSW
Email: j.croucher@unsw.edu.au