Selective Harvesting: Creating and Ingesting Custom OAI-PMH Sets

Selective Harvesting:
Creating and Ingesting Custom OAI-PMH Sets

Harry R. Sidhunata[1], Joanne L. Croucher[2], Maude Frances[3]
[1]University Library, University of New South Wales, Sydney, Australia, h.sidhunata@unsw.edu.au
[2]University Library, University of New South Wales, Sydney, Australia, j.croucher@unsw.edu.au
[3]University Library, University of New South Wales, Sydney, Australia, m.frances@unsw.edu.au

INTRODUCTION

The *Selective Harvester* provides a flexible and customisable mechanism to select and re-use metadata records from open access repositories. The open-source Java-based application developed at the University Library, University of New South Wales (UNSW) integrates an existing OAI-PMH harvesting tool (jOAI) [1], with an application which filters and ingests selected records into a Fedora repository [2]. This model has applications in scholarly communication and eResearch services, especially in relation to populating subject-based repositories.

BACKGROUND

Subject-based repositories act as a search and discovery portal for resources on a specific topic, aggregating and filtering resources obtained from multiple sources. In some cases, filtered resources are also reviewed or transformed [e.g. 3, 4]. The *Selective Harverster* has been implemented by the NCHSR Clearinghouse [5], a subject-based repository developed jointly by UNSW Library and researchers at the National Centre in HIV Social Research (NCHSR) [6].

The *Selective Harvester* employs the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), a widely used standard for disseminating metadata records [7],[8]. An OAI-PMH *data* provider exposes a collection of metadata records. These records are then available to be harvested by one or more OAI-PMH *service* providers. The OAI-PMH protocol supports selective harvesting by two different criteria: by date or by set [9]. A set is defined as “an optional construct for grouping items for the purpose of selective harvesting” [10]. For example, an institutional repository may create a set which contains all records of theses.

There are some constraints on the standard operation of OAI-PMH sets. Specifically, sets are configured at the data provider side, by a system administrator at the source repository. Set definition is also reliant on a degree of standardization within the source metadata. Supply-side modifications can be a burden, especially as a single repository may be harvested by many different service providers.

SELECTIVE HARVESTING MODEL

The *Selective Harvester* design employs two separate open-source applications, each of which may be deployed and configured independently.

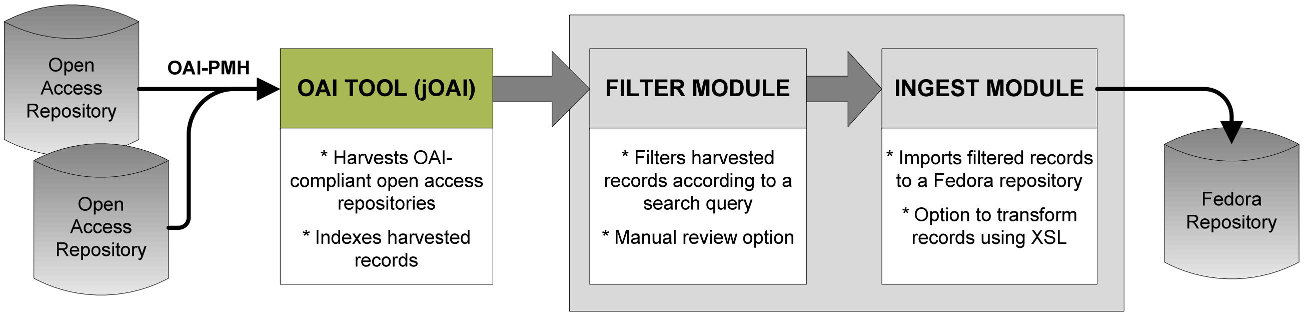
Open Access Repository
Open Access Repository
OAI-PMH
OAI TOOL (jOAI)
* Harvests OAI-compliant open access repositories
* Indexes harvested records
FILTER MODULE
* Filters harvested records according to a search query
* Manual review option
INGEST MODULE
* Imports filtered records to a Fedora repository
* Option to transform records using XSL
Fedora Repository

Figure 1: Selective Harvesting Model

Firstly, jOAI is employed as an OAI-PMH harvester and provider [1]. This open-source application was developed by Digital Learning Sciences (DLS) [11] at the University Corporation for Atmospheric Research [12]. Highly configurable, it enables the harvesting of multiple open access repositories as well as the specification of OAI-PMH sets.

The second application, developed at UNSW, is comprised of two key components: a Filter Module and an Ingest Module. The application offers a manual filter review option as well as the functionality to transform and ingest selected records into a Fedora repository. Fedora is an open-source repository software solution, and is widely-used by Australian institutional repositories.

The *Selective Harvesting* applications include a scheduling facility, making it possible to set-up schedules to automatically and routinely harvest, filter and ingest new records.

SELECTIVE HARVESTING COMPONENTS

The Filter Module is used to define and create a custom set of harvested records. This module leverages jOAI's use of Apache Lucene indexing and the ODL Search Specification [13],[14]. Complex filter criteria can therefore be constructed using Boolean operands. Via the Filter Module, users can also test the operation of a filter and review matching records. Based on the results, users may choose to update the search query or manually delete any remaining non-relevant records.

The Ingest Module is responsible for adding the filtered records to a Fedora repository. It can also be used as a stand-alone application, independent of jOAI and the Filter Module. The Ingest Module includes an option to upload custom XSL transformation files and use these to customize or transform the harvested metadata. For example, a simple transformation file could be created to insert the name of a source repository into the Dublin Core metadata. A more complex transformation could be coded to convert the harvested metadata into another metadata schema. While the *Selective Harvester* does not synchronize records with the source repositories, the Ingest Module can create a backup directory of all records ingested to Fedora.

CONCLUSION

The poster demonstrates a framework for selective harvesting which enables aggregation and filtering of records from open access OAI-compliant repositories, followed by the transformation and ingestion of filtered records into a Fedora repository. The *Selective Harvester* can be used to support subject-based repositories and facilitate the sharing of research resources across national and international eResearch systems.

REFERENCES

1. *jOAI Overview: The Java-based Open Archives Initiative Data Provider & Harvester*. Available from: http://www.dlese.org/oai/, accessed 17 September 2010.
2. *Fedora Commons Repository Software*. Available from: http://fedora-commons.org/, accessed 17 September 2010.
3. Merceur, F., *Set up an Institutional Repository and an OAI harvester for marine and aquatic sciences, at Ifremer*, in *IAMSLIC: Every continent, every ocean: proceedings of the 32nd Annual Conference of the International Association of Aquatic and Marine Science Libraries and Information Centers (IAMSLIC). 2007. IAMSLIC, Fort Pierce, Florida, USA*: p. 93-108. Available from: http://hdl.handle.net/1912/2137, accessed 16 September 2010.
4. Sanderson, R., Harrison, J., and C. Llewellyn, *A curated harvesting approach to establishing a multi-protocol online subject portal*, in *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06). 2006, ACM/IEEE, New York, NY, USA*: pp. 355. doi: 10.1145/1141753.1141850
5. *NCHSR Clearinghouse*. Available from: http://ssrm.nchsr.arts.unsw.edu.au/, accessed 17 September 2010.
6. Frances, M. & J. Croucher, *Infrastructure for problem-based collaborative research: Aligning research, policy and practice*, eResearch Australasia. 2009, Sydney, Australia. Available from: http://handle.unsw.edu.au/1959.4/44532, accessed 17 September 2010.
7. *Open Archives Initiative: Protocol for Metadata Harvesting*. Available from: http://www.openarchives.org/pmh/, accessed 17 September 2010.
8. Shreeves, S.L., Habing, T.G., Hagedorn, K. and J.A. Young, *Current Developments and Future Trends for the OAI Protocol for Metadata Harvesting*, Library Trends, 2005. **53**(4): p. 576-589.
9. Lagoze, C., and H. Van de Sompel, *The Open Archives Initiative: Building a Low-Barrier Interoperability Framework*, in *Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01). 2001, ACM, New York, NY, USA*: p. 54-62.
10. *The Open Archives Initiative Protocol for Metadata Harvesting (Protocol Version 2.0 of 2002-06-14)*. Available from: http://www.openarchives.org/OAI/openarchivesprotocol.html#Set, accessed 14 September 2010.
11. *Digital Learning Services*. Available from: http://www.dlsciences.org/, accessed 14 September 2010.
12. *University Corporation for Atmospheric Research*. Available from: http://www.ucar.edu/, accessed 14 September 2010.
13. *Lucene*. Available from: http://lucene.apache.org/, accessed 14 September 2010.
14. *ODL Search Specification*. Available from: http://www.dlese.org/oai/docs/odlsearch.do, accessed 16 September 2010.