Automating the Preservation of Electronic Theses and Dissertations with Archivematica

Mark Jordan
Simon Fraser University
8888 University Drive
Burnaby, British Columbia, Canada
1-778-782-5753
mjordan@sfu.ca

ABSTRACT
This poster describes the tools, services, and workflows that Simon Fraser University is using to automate the movement of its ETDs (Electronic Theses and Dissertations) from its user-facing Thesis Registration System to the Archivematica digital preservation platform. The poster also describes Simon Fraser University's plans to expand its digital preservation services using Archivematica, including integration of LOCKSS as a distributed storage network for content managed by Archivematica.

Categories and Subject Descriptors
H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *Distributed systems* and H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Standards*

General Terms
Management, Standardization

Keywords
Case studies, Digital Preservation, ETDs, workflows, automation, microservices, OAIS, Drupal, Archivematica, LOCKSS

1. INTRODUCTION
Simon Fraser University (SFU) has been accepting theses, dissertations, and graduate project reports from students in digital form since 2004. In late 2012, the Library initiated a set of microservices to transfer electronic theses and dissertations (ETDs) theses from its Theses Registration System (TRS)[1] to its institutional repository, Summit,[2] without human intervention apart from sign off by Library staff that the thesis has become ready for publication. Shortly after the initiation of that automated workflow, the Library started moving theses from the TRS into the Archivematica[3] digital preservation platform, a process which is also fully automated.

This poster describes the rationale for automating the ingestion of ETDs into Archivematica, the various tools and services that are used in this automation, and how they work together. It also describes areas of active development the SFU Library is pursuing to expand this set of digital preservation services.

2. GOALS AND GUIDING PRINCIPLES
Theses and dissertations are one of the most important types of scholarly works created by universities. Even though copies of ETDs are frequently distributed in commercial services such as Proquest Dissertation Publishing[4] or in national aggregations such as Theses Canada,[5] many educational institutions that produce ETDs take on the responsibility for long-term preservation of these works. However, this commitment will require considerable resources over time.

Simon Fraser University has decided to act on this responsibility but to do so with the goal of reducing costs as much as possible. Many of the costs associated with digital preservation are difficult to predict,[6] but one aspect of this activity in which it is relatively easy to minimize costs is human labor. To that end, SFU is striving to automate as many aspects of the ETD lifecycle as completely as possible. Three guiding principles led to the development of a set of services and processes to achieve this goal.

First, the preservation of ETDs should adhere to proven, robust, standards-based digital preservation practices such as compliance with the OAIS Reference Model,[7] use of PREMIS[8] preservation metadata, use of the BagIt[9] content packaging format, and support for standard descriptive metadata such as Dublin Core Terms.

Second, any processes involved in the preservation of ETDs that can be automated should be. Human intervention will be required at certain points in preservation workflows, but the amount of human intervention and localized decision making should be reduced to a practical minimum.

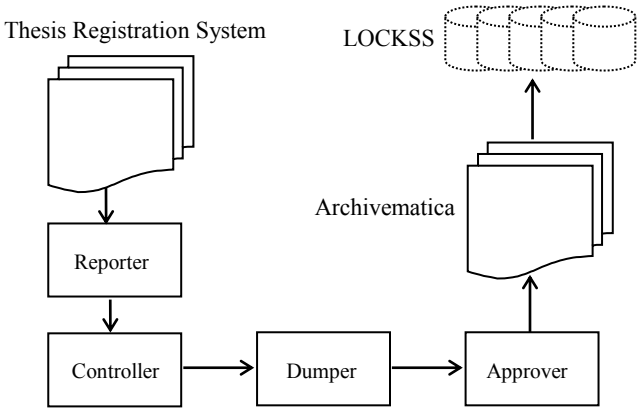
Third, any specific service or tool used in these processes should be easily replaceable. Over the long term, tools considered best in class will invariably change. It is important that any new tool that improves a process, or performs the same process at less cost, should replace the existing tool, as long as doing so is not overly disruptive to other processes that depend on the exiting tool. In addition, the ability to replace services and tools facilitates easy adaptation of the remaining components to other digital preservation processes.

These three principles informed the development of the preservation architecture described below.

It is important to note that an ETD is not only a simple textual document. Many ETDs have raw or application-specific data, multimedia content, or additional textual documents associated with them. This additional content is commonly referred to as "supplemental" content or files. In addition, an ETD will typically have at least one metadata description identifying its title, date of completion, subject content, and so on, usually expressed in the ETD-MS[10] element set. The preservation of an ETD is therefore not as simple as making sure that the thesis document is stored in a single PDF file. Long-term preservation of ETDs must take all of these types of content into account. [11]

3. ARCHITECTURE

Simon Fraser University's ETD preservation architecture is comprised of three main components, 1) its Thesis Registration System, 2) a set of microservices, and 3) the Archivematica digital preservation platform. A fourth component, a Private LOCKSS Network, is currently under development. The following is a visual overview of this architecture:



3.1 Thesis Registration System

The Thesis Registration System enables students to register their thesis and upload any associated files in what is referred to as a "submission." Once the student has completed a submission, Library staff audit the thesis before they approve it for publication in the University's institutional repository. This process involves verifying that the thesis adheres to publication standards set by the University, that any documentation such as ethics review approval has been obtained, and that all licenses for publication have been accepted by the student.

When all audit requirements have been met, Library staff record this decision within the submission record for a thesis by simply checking a box titled "Ready for publication." This attribute of the thesis submission is then used in a query, run nightly, to identify all submissions that have been approved for publication during the previous day.

The Thesis Registration System is built using Drupal,[12] an open-source Content Management Framework. Drupal manages user accounts and permissions, provides mechanisms for structuring the thesis submission, and handles the various types of files the student must upload. A custom Drupal module, developed by SFU Library staff, manages the workflows involved in auditing the submission, and sends out email messages to the student when the audit staff perform specific actions or make specific decisions. Each submission is instantiated within the Thesis Submission System as a "node," the basic content structure within Drupal.

3.2 Microservices

Moving the ETD content out of the Thesis Registration System and into Archivematica is accomplished using a small series of microservices. Each microservice is a shell or PHP script that performs one task or one group of related tasks.

The first microservice (called the "reporter") queries the Thesis Registration System for all submissions that have been approved during the previous day. This is the script that performs the query described in section 3.1, above. The script writes the Drupal node IDs (which serve as the unique identifier of each thesis submission in the Thesis Registration System) to a data file with the current date encoded in its filename.

The second microservice (the "controller") wraps two task-specific scripts; in other words, it runs each of the two scripts from within itself. This approach allows for robust handling of errors in each script, and also allows for easy cleanup of temporary files created by the wrapped scripts. The controller is scheduled to run each day after the reporter microservice runs, and uses the data file created by the reporter as its input. In effect, the controller loops through all of the submission node IDs in the current day's data file and runs the two wrapped microservices, the "dumper" and the "approver," on the submission corresponding to each node ID.

The "dumper" microservice takes a submission node ID as a parameter, queries the Thesis Registration System for the corresponding submission node, and creates Dublin Core and ETD-MS descriptive metadata files for the thesis using information in the submission record. In addition, the dumper microservice determines what files are associated with the thesis (the thesis PDF, any supplemental files, and specific licenses and other administrative documents) and writes those out to disk as well. Finally, the dumper ensures that all of the files are arranged in a subdirectory structure compliant with Archivematica's "transfer" package format (described in the next section) and creates a Bag containing the all the submission's files.

The final microservice is the "approver," which copies the Bag created by the dumper to the Archivematica server and, after confirming that the Bag has been copied successfully, issues an HTTP request to Archivematica's transfer approval API (also described in the next section).

3.3 Archivematica

Archivematica is an open-source digital preservation platform. It normalizes files into preservation-friendly formats using what it calls "format policies",[13] and stores content in OAIS-compliant Archival Information Packages (AIPs). Archivematica integrates a number of open-source tools such as FITS,[14] OpenOffice,[15] FFmpeg,[16] and Clam Antivirus[17] using its own internal microservices framework, and it employs open, standardized formats such as METS,[18] PREMIS, and BagIt to ensure long-term, standards-based management and access to the content and metadata stored in the AIPs it produces.

Content is ingested into Archivematica as a "transfer," which contains the files to be preserved, metadata describing those files, "submission documentation" (licenses and other administrative documents), and, optionally, a "processing configuration" file. The transfer structures the content in preparation for repackaging into an OAIS Submission Information Package (SIP) and then, into an Archival Information Package (AIP) for long-term management. If the content is to be made available to a given user community, Archivematica allows the creation of Dissemination Information Packages (DIPs) for that purpose.

Archivematica's user interface breaks down the workflow for processing a given set of files from transfer to SIP to AIP to DIP into a series of structured tasks, most of which are instantiated internally as microservices. Within each group of tasks, a human operator must make a number of decisions, such as whether to normalize the incoming files for preservation, access (or both),

whether to approve the results of the normalization or not, whether to apply additional descriptive metadata to the transfer, and where to store the AIP. How specific file types are normalized is determined by the format policies; for example, the format policy for audio files might dictate that they should be normalized into WAV format for preservation and MP3 format for end user access.

Workflow tasks can be automated using a processing configuration file, which encodes in a machine-readable format each of the decisions that a human operator would make if he or she were manually processing a transfer. The ability to automate workflow decisions is useful if Archivematica is to process large quantities of similar transfers in batches, or if local policy dictates that a given workflow decision should always be made.

For the processing of Simon Fraser University's ETDs, the processing configuration file specifies that the files should be normalized for preservation only (since we are not asking Archivematica to generate Dissemination Information Packages), which format identification tool Archivematica should use, and where to store the AIP.

The processing configuration file only removes the need for a human operator after a transfer package has been ingested into Archivematica. To automate the ingestion itself, Archivematica provides a REST API[19] for approval of transfers. Since the API uses REST, it is possible to interact with this API from within a script running on a different server (in this case, the "approver" microservice running on server hosting the Thesis Submission System).

It is the combination of this REST API and the processing configuration file that allows for the complete automation of moving content from a source such as SFU's Thesis Registration System into Archivematica, then through Archivematica's digital preservation microservices to produce an OAIS-compliant Archival Information Package. In the case of SFU's architecture for preserving ETDs, this process is instantiated in the dumper and approver microservices described earlier, which combined, hand over the ETD content to Archivematica's internal microservices as defined by the processing configuration file.

3.4 Long-term management of the ETDs

Over time, Archival Information Packages can be retrieved and re-ingested into Archivematica as SIPs (Submission Information Packages) when the content needs to be updated or migrated to new formats. The need to update an ETD after its publication is rare but not unheard of, and SFU's Faculty of Graduate Studies has a policy in place for that situation.

Archivematica supports the authenticity of content it preserves by storing all original documents that are included in a transfer in addition to any normalized versions created by its microservices (or by normalization external to Archivematica). It also generates and stores checksums for all files to allow auditing and verification of bit-level integrity over time. Finally, in SFU's implementation, all license agreements signed by the author of the ETD are preserved in the same Archival Information Package as the ETD document and supplemental files, complete with checksums.

3.5 Public access to the ETDs

The version of the ETD content that is transformed by Archivematica into an OAIS Archival Information Package is not intended to be accessed by end users. In fact, the AIP contains licenses and other sensitive information that should not be exposed to end users.

In SFU's implementation, the ETD and its associated metadata are transferred directly from the Thesis Registration System to the University's institutional repository, Summit, for public access. This transfer is automated and happens at the same time as the transfer of the ETD from the Thesis Registration System to Archivematica. In effect, the two processes are run in parallel. Once in the institutional repository, end users access the theses through a variety of discovery tools such as the Library's unified discovery layer and the search and browse capabilities of Summit itself.

Archivematica is capable of creating an OAIS Dissemination Information Package (DIP) and transferring the DIP to a variety of public-access content management systems and repository platforms, including AtoM, CONTENTdm, and DSpace. SFU's implementation does not use this feature because a process to move ETDs from the Thesis Registration System to Summit was already in operation when the Library began using Archivematica. It would be possible to create new Archivematica microservices to produce a DIP for SFU's Summit, but the Library has chosen an alternative approach to integrating Archivematica and its institutional repository, described in section 4.3, below.

4. DEVELOPMENT ROADMAP

The SFU Library is actively working to expand the integration of its current ETD preservation services with several other tools.

4.1 LOCKSS integration

Work is under way to allow Archivematica to store its AIPs in a Private LOCKSS Network (PLN).[20] This development will enable the automated movement of AIPs into a holding queue, from which LOCKSS will harvest them and ingest them into the PLN. Storing the AIPs in a Private LOCKSS Network will ensure that identical copies are managed in a geographically distributed, secure fashion. SFU Library and a group of partner institutions are working closely with the developers of Archivematica to ensure that this work is compliant with a new Storage API that is being developed for Archivematica. This API will allow it to use a variety of storage platforms for AIPs it generates.

4.2 Academic review of theses using Open Journal Systems

Although not directly related to preservation of ETDs, the Faculty of Graduate Studies at SFU is planning to use Open Journal Systems (OJS)[20] for the academic review of theses. Open Journal Systems provides a toolset for manuscript submission, peer review, and editorial workflow for journal articles that is easily adaptable to the review of theses by academic adjudication committees. SFU Library will be working closely with the Faculty to ensure that ETDs will move from OJS to the Library's Thesis Submission System seamlessly, and from there, through the digital preservation architecture described in this poster.

4.3 Automating preservation of content in SFU's institutional repository

The tools and workflows described in this poster can also be applied to automating the preservation of content submitted to SFU's institutional repository, Summit. Work is under way to

implement such a process. Essentially, all that is required is to modify the dumper microservice to convert non-ETD items in the institutional repository into Archivematica transfer packages. "Non-ETD items" in SFU's repository include journal and book chapter preprints, conference papers, reports, and other works submitted directly by end users and by Library staff as a service to the University community. Automating the movement of content from Summit to Archivematica will provide robust digital preservation services lacking from many institutional repositories.

The ability to replace one component of SFU's digital preservation architecture (the Thesis Registration System) with another (the institutional repository) and make only minor modifications to a single microservice (the dumper) illustrates an important guiding principle of the architecture: "any specific service or tool used in these processes should be easily replaceable." This pattern can also be applied to other sources of content the SFU Library needs to preserve, such as locally digitized manuscript collections and newspapers, research data sets, and archived websites.

5. REFERENCES

[1] Simon Fraser University's Thesis Registration System. https://theses.lib.sfu.ca

[2] Summit, Simon Fraser University's Institutional Repository. http://summit.sfu.ca

[3] Archivematica. https://archivematica.org

[4] Proquest Dissertation Publishing. http://www.proquest.com/en-US/products/dissertations/

[5] Theses Canada. http://www.collectionscanada.gc.ca/thesescanada/index-e.html

[6] Wheatley, Paul. 2012. *Digital Preservation Cost Modelling: Where did it all go wrong?* Blog post. http://openplanetsfoundation.org/blogs/2012-06-29-digital-preservation-cost-modelling-where-did-it-all-go-wrong

[7] Reference Model For An Open Archival Information System (OAIS). http://public.ccsds.org/publications/archive/650x0m2.pdf

[8] PREMIS Data Dictionary for Preservation Metadata. http://www.loc.gov/standards/premis/

[9] The BagIt File Packaging Format. http://tools.ietf.org/html/draft-kunze-bagit-09

[10] ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations. http://www.ndltd.org/standards/metadata/

[11] Shreeves, Sarah L. 2013. *Supplemental Files in Electronic Theses and Dissertations: Implications for Policy and Practice.* Poster presented at the 8th International Digital Curation Conference, Amsterdam, Netherlands, January 14-17, 2013. http://hdl.handle.net/2142/35314

[12] Drupal. http://drupal.org/

[13] Archivematica Format Policies. https://www.archivematica.org/wiki/Media_type_preservation_plans

[14] File Information Tool Set (FITS). http://code.google.com/p/fits/

[15] OpenOffice. http://www.openoffice.org/

[16] FFmpeg. http://www.ffmpeg.org/

[17] ClamAV. http://www.clamav.net/lang/en/

[18] Metadata Encoding and Transmission Standard (METS). http://www.loc.gov/standards/mets/

[19] Approving a transfer. https://www.archivematica.org/wiki/Administrator_manual_0.10#Approving_a_transfer

[20] Lots of Copies Keep Stuff Safe (LOCKSS). http://www.lockss.org/

[21] Open Journal Systems (OJS). http://pkp.sfu.ca/ojs