Simon Fraser University Archives
Digital Preservation Strategy

Project Report

Date: 18 April 2012
Prepared by: artefactual systems inc.

Table of Contents

1. Executive summary

The SFU Archives Program supports teaching, research and university administration by acquiring and protecting university records of historic value. SFU Archives also collect records of private individuals and organizations, including faculty and university-related groups. Increasingly, the records which SFU Archives is acquiring or planning to acquire are in digital formats. Maintaining the long-term accessibility, usability and authenticity of records in digital form is challenging due to a number of factors, which including storage media degradation, lack of adequate metadata, proprietary file formats, technology incompatibility and obsolescence, and the absence of a coordinated digital preservation strategy and implementation plan.

In 2011 Artefactual Systems Inc. was contracted by Simon Fraser University Archives to assist in developing the technical capacity to address these challenges and extend its preservation mandate and services to include digital records. The work undertaken consisted primarily of running preservation pilot projects for selected digital record types (unstructured documents, digitized audio files and email) using Archivematica, an open-source digital preservation system developed by Artefactual Systems. This report presents the findings of these pilot projects, including detailed recommendations for digital preservation workflows and suggested improvements to Archivematica. The report concludes with recommended steps to move SFU Archives into a production digital preservation environment by 2013-2014.

2. Digital preservation program requirements

The ISO 14721 Reference Model for An Open Archival Information System (OAIS) is the recognized standard for defining the preservation processes and entities that make up a comprehensive digital preservation system. Using OAIS as a baseline reference, core requirements for the establishment of a digital preservation program at SFU Archives can be summarized as follows:

1. Ability to transfer a Submission Information Package (SIP), consisting of the digital objects to be preserved and descriptive metadata, from Producers to archival storage.
2. Verification of successful transfer (i.e. no data corruption/loss during transfer from one system and/or storage media to another).
3. File format identification and validation.
4. Extraction of technical metadata.
5. Implementation of preservation plans (such as normalization to preservation-friendly file formats).
6. Generation and capture of preservation metadata using recognized standards, namely PREMIS and METS.[1]
7. Packaging of the Archival Information Package (AIP) consisting of the original objects, normalized objects and descriptive, preservation and technical metadata.

[1] PREMIS Data Dictionary for Preservation Metadata, http://www.loc.gov/standards/premis/; and Metadata Encoding and Transmission Standard (METS), http://www.loc.gov/standards/mets/METSOverview.v2.html.

8. Secure storage and backup of the AIP.
9. Periodic integrity checking of the AIP and restoration from backup if needed.
10. Generation of a Dissemination Information Package (DIP) consisting of access copies of the digital objects and descriptive metadata, linked back to the stored AIP.
11. Upload of the DIP into a web-based access system.
12. Management of changes to the descriptive and preservation metadata over time.

3. Archivematica and SFU Archives

One of the key components of this project was testing Archivematica, a digital preservation system developed by Artefactual Systems which is currently in alpha development.

Archivematica provides an integrated suite of free and open-source tools that allows users to process digital objects from ingest to access in compliance with the OAIS functional model and other digital preservation standards and best practices. Archivematica implements a micro-services approach to digital preservation, in which the OAIS information entities (Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Information Package (DIP)) move through a series of services using a Unix pipeline design pattern. The services, which are modelled on OAIS functional requirements, are provided by a combination of Archivematica software scripts and one or more of the free and open-source tools bundled into the system. The micro-services can be distributed to processing clusters for highly scalable configurations. Archivematica implements normalization preservation plans on file formats for which there is a well-established conversion path and for which a reliable, tested open-source tool is available.[2]

Although Archivematica can be customized to work with external access systems, it also comes bundled with ICA-AtoM, a web-based access system which can be used to manage accessions, authority records, hierarchical archival descriptions, taxonomies and linked digital objects. ICA-AtoM is designed to be compliant with the Canadian Council for Archives' Rules for Archival Description. It makes the descriptions available to the public via search and browse interfaces and can export them as Dublin Core, EAD and EAC XML. In a separate but related project, SFU Archives is migrating descriptions from its FileMakerPro Archives Information System (AIS) to ICA-AtoM with assistance from Artefactual Systems. SFU Archives will use the search and browse components of ICA-AtoM to provide web-based, public access to its holdings. Through integration with Archivematica this can include public access to records in digital form accessioned by SFU Archives.

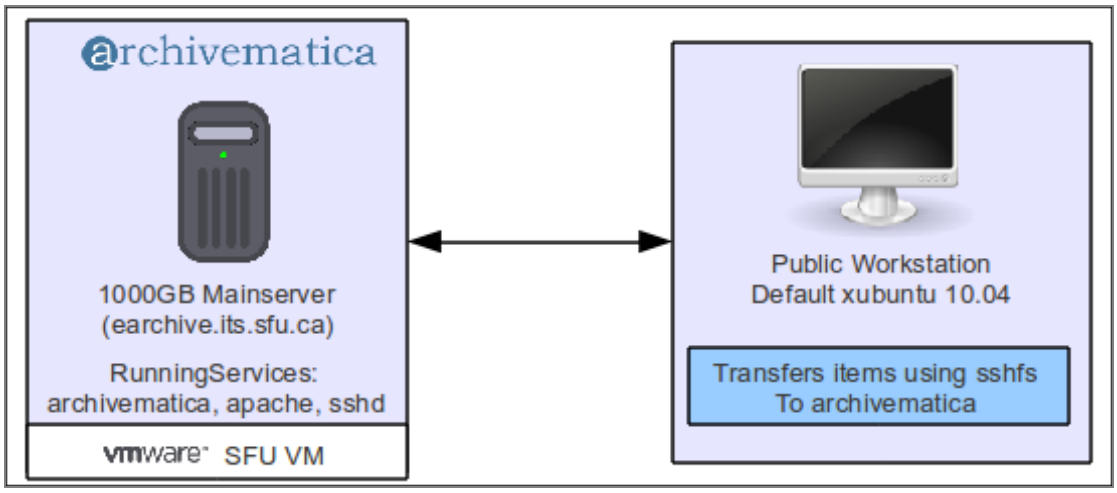
The main focus of the Archivematica pilot project testing was to identify enhancements and new features needed in Archivematica to allow SFU to move into a production digital preservation environment using future releases of the software. **Section 5** of this project report describes the pilot projects in detail, and **Appendix A** provides a summary of recommendations for development. It is important to note that the software improvements and digital preservation knowledge being generated by the SFU Archives project can be freely re-used by other institutions under the terms of the Archivematica project's AGPL and Creative Commons open-source licenses. Therefore, the financial and technical resources invested in this project are being leveraged well beyond SFU Archives' internal

[2] http://archivematica.org/preservation.

strategy to the benefit of the archival community at large.

4. Archivematica installation

In September 2011, Artefactual Systems installed Archivematica 0.7.1 as a VMware VM hosted by the University's IT department. The Archivematica web dashboard is available from SFU and public networks (secured with Apache authentication). As part of SFU IT's VM hosting service, a fresh full backup of the VM is performed once a week, with nightly incremental backups. A PC client workstation was set up to connect to the VM server's 1 TB of storage via SFTP. In February 2012, the SFU Archives' campus network connection was upgraded from 10 mbps to 1 gbps and the Archivematica installation was upgraded to Archivematica 0.8. Nearly all of the of the pilot project testing has taken place using the upgraded installation.



5. Pilot projects

The pilot projects originally included email preservation and a private fonds, although digitized audio was added as a third pilot in early 2012. The activities undertaken for each pilot project are described below, along with analysis and recommendations for new features and feature enhancements in Archivematica and ICA-AtoM.

5.1 Telearning Network fonds and Columbia River Treaty collection

The Telelearning Network fonds consists of approximately 300 MB of digital records received by the Archives via transfer of the creating organization's server. The records, which consist mainly of office and still image formats, have been copied to network servers and have been fully appraised and arranged. The Columbia River Treaty collection consists of approximately 17 GB of digitized audio

files in .wav and .aiff format. The pilot projects consisted of SFU archivists submitting the records to Archivematica for processing and uploading the DIPs into ICA-AtoM. The archivists tested the functionality and usability of both software tools, but for the purposes of this project limited comments and recommendations to Archivematica. Most of the comments and recommendations relate to user interface, workflow, batch processing, error handling and overall management and system administration.

In general, the archivists found that processing transfers and SIPs was relatively straightforward and that the system accomplished the task of preserving the ingested digital objects: "Overall, SFU Archives staff are impressed with Archivematica's functionality. The system provides us with the technical basis for the preservation of electronic records. There are no 'show-stoppers' that would prevent SFU Archives from using Archivematica as-is in a production setting." However, they also noted that the system "relies too much on the user needing to remember the correct steps in the correct order (e.g. when to restructure the object for compliance, when to appraise/arrange/delete items from the object, when to add metadata or rights information)." They also felt that although the system works well when a job can be completed immediately, a more typical scenario would be a backlog in which transfers would be minimally processed, then placed in storage and retrieved for full processing at a later date. Their recommended enhancements for managing transfers, backlogs and ingest on a large scale include:

- Initiation of transfer via the dashboard;
- An improved interface for arranging the transfer into SIPs;
- Integration of Archivematica with ICA-AtoM's accession module (at a minimum, to allow the user to link a transfer to a fonds);
- Configurable workflow options that can be accessed via the dashboard;
- Improved error reporting and procedures for rejecting a transfer and starting again;
- A repository management layer supporting deletion, export and reprocessing of objects, filtered views of repository objects (by e.g. fonds, transfer date, filesize, file formats), statistical reporting; and storage space planning.

A complete list of recommendations is provided in **Appendix A**.

5.2 Email

5.2.1 Pilot project description

This pilot project involved acquiring a snapshot of the email account of former President Stevenson, who retired from SFU in 2010. The account had been active for 10 years and no other email from it had been sent to the Archives in electronic form in the past.

SFU currently uses Zimbra Network Edition to send and receive email.[3] The Zimbra administrator's manual does not include information on how to export email from Zimbra for use in other email programs.[4] However, the University's IT department backs up the email accounts using a default directory structure specific to Zimbra, and has expressed a willingness to deliver email to the Archives in the form of these backups. However, these backups are in a format which is intended to be used to restore email to Zimbra accounts, not to migrate the accounts' contents into other systems. Furthermore, documentation of its structure is somewhat limited (see **Appendix B** for more information about the Zimbra backup format and directory structure). After analyzing the Zimbra backup and conducting research on email preservation standards and practices, the project team reached the conclusion that Zimbra email accounts need to be converted to a standard, well-documented, widely-used format that can be opened in a variety of open-source email programs or other tools such as web browsers.

Two formats which were explored as part of this project are Maildir and mbox.[5] Maildir is a text-based format which stores each folder in an email account as a separate directory (inbox, sent items, subfolders etc) and each email as an individual text or .eml[6] file; attachments are included in the text files as base64 encoded ascii text. Mbox is a single large text file with attachments included as base64 content; each folder in an account is saved as a separate mbox file. Both formats can be imported into and rendered by numerous email programs, proprietary and open-source, and both can be converted into other formats using open-source tools and scripts. Although Maildir and mbox can be rendered in a variety of email programs, mbox has more potential as an access format because it is easier to develop tools to render it that are not necessarily email programs. For example, a program called Muse, developed by Stanford University, is designed to render mbox files using only a web browser.[7] In addition, mbox is the source format for import into tools like the CERP email parser, which was developed by the Rockefeller Archive Center and the Smithsonian Institution Archives to convert email messages to hierarchically arranged XML files.[8] In essence, mbox is emerging as a de facto standard for which the digital curation community is beginning to build tools for rendering and manipulation.

---

3 http://www.zimbra.com/.

4 See http://www.zimbra.com/docs/ne/6.0.10/administration_guide/.

5 http://en.wikipedia.org/wiki/Maildir and http://en.wikipedia.org/wiki/Mbox. Note that this report refers specifically to the .mbox extension, the standard Berkeley mbox implementation of this format. For a discussion of the role of mbox in email preservation, see *Preserving Email*, Christopher J. Prom, DPC Technology Watch Report 11-01 December 2011, http://dx.doi.org/10.7207/twr11-01.

6 EML is a common email format encoded to the RFC 822 Internet Message Format standard (http://tools.ietf.org/html/rfc822) for individual emails. Messages in Maildir backups are encoded to this standard, although they lack the .eml file extension. For a discussion of the role in the eml format in email preservation, see Prom, *Preserving email*.

7 http://mobisocial.stanford.edu/muse/.

8 http://siarchives.si.edu/cerp/parserdownload.htm. The CERP XML format is designed to be a neutral, software-independent format for email preservation, but as yet there are no tools available to display the XML files as email messages that can easily be searched and navigated.

However, Maildir is preferable as a preservation format because it stores each message as a separate text file; thus any corruption to one or more text file would not cause an entire directory of messages to be lost, which is a risk with a format such as mbox.

Artefactual Systems tested the use of a tool called OfflineImap[9] to back up a test Zimbra email account to Maildir and converted the Maildir backup to mbox using a freely available python script.[10] Following these preliminary tests, the Zimbra backup of President Stevenson's email account was restored to Zimbra and captured using OfflineImap. The resulting Maildir backup was converted to mbox files (Inbox, Sent and Eudora/out) which were imported into an open-source email program called Evolution. The total message count for each folder was found to be the same in Evolution as it had been in Zimbra (71, 2544 and 7628 messages, respectively), and randomly sampled emails were opened to ascertain that the conversion and import were successful. Sample emails from the Zimbra and Maildir backups were also compared to ensure that the significant characteristics of the Zimbra version were captured in the Maildir version.[11]

A critical component of SFU's email preservation strategy is management of access based on British Columbia's Freedom of Information and Protection of Privacy Act. In any given user's account, some email messages must necessarily be excluded from public access based on the presence of personal information or other information which falls under exceptions to disclosure under the Act. SFU's archivists and FOIPPA management personnel will need to be able to view email messages, flag those with restrictions, and provide public access to only those emails which are not restricted. Preliminary tests of Muse have shown it to be capable of importing mbox files, rendering the individual messages in a web browser, allowing tagging of restricted messages, and exporting the remainder in mbox format. We have noted that tagging one message as restricted automatically tags the same email message in other threads containing the same message.

5.2.2 Proposed plan for preserving email at SFU Archives

Based on our analysis of SFU's email system and management practices and of preservation formats and conversion tools, we recommend the following for acquiring, preserving and providing access to email at SFU Archives:

Acquisition and Preservation

We recommend that SFU Archives acquire accounts in Maildir format, for the following reasons:

- The Maildir directory structure is well-documented and transparent;
- Maildir is widely used and can be created and rendered by a large number of software tools, both proprietary and open-source;

[9] http://offlineimap.org/. According to the documentation for this tool, it is possible to specify the folders to be captured, which would permit capturing folders designated specifically for archival retention. OfflineImap can also be run as a cron job, capturing email automatically at specified intervals. These features open up a number of possibilities for email archiving workflows at SFU.

[10] md2mb.py, available from https://gist.github.com/1709069.

[11] See http://www.archivematica.org/wiki/index.php?title=Zimbra_to_Maildir_using_OfflineImap for an example of the analysis of significant characteristics.

- The contents of a Maildir directory are plain ascii text messages which can be read easily in any text editor (except for attachments);
- The text-based messages are based on an open and widely-used specification;[12]
- Because each message is saved individually, accidental corruption or deletion of one or more messages would not result in the entire Maildir backup becoming unreadable (by comparison, corruption of a small amount of data in an mbox file could render the entire mbox file, with its multiple messages, unreadable);
- Maildir is easily converted to mbox for access purposes (see **Provision of access**, below).

The archivists would submit the Maildir backup into Archivematica, where it would be retained as the preservation master in the AIP.[13] The attachments would be extracted and normalized to standard open formats for preservation purposes, with links between messages and their normalized attachments being managed through UUIDs and/or filename.[14]

SFU Archives will need to work closely with its IT Department to acquire the Maildir backups. There are at least three options for acquiring the backups:

Option 1. IT makes a Zimbra backup account and restores it to a designated Archives account. The Archives then runs OfflineImap to capture the account contents as Maildir.

- Pro: Working with the restored Zimbra account in Zimbra would allow Archives staff to compare the original Zimbra account with the Maildir/mbox versions to ensure accuracy and completeness of the conversion.[15]
- Pro: Working with the restored Zimbra account would allow Archives staff to review the account for appraisal purposes prior to submission to Archivematica. If desired, Archives staff could cull email (for example, to remove junk mail or personal email) prior to conversion to Maildir. Since there would be no way to automatically capture information on the deleted files, Archives staff would then prepare an appraisal report for inclusion in the Submission Information Package.
- Con: backing up the Zimbra account and then restoring it to another account introduces the possibility of data loss.

Option 2. IT provides a backup of a Zimbra account and transfers it to the Archives, which then runs a custom script to convert it to Maildir (or IT runs the script and delivers the Maildir backup to the Archives).

[12] RFC # 822, Standard for the Format of ARPA Internet Text Messages, http://tools.ietf.org/html/rfc822.

[13] Note that Maildir backups do not capture calendars or contact lists. However, SFU Archives staff have indicated that such records would probably not be considered archival.

[14] Attachments must be extracted and normalized because they can never be read as base 64 ascii encoded text. They will always need to be rendered in a software program. In other words, even though the user may be able to open an email message in an email program he or she typically has to open the attachment separately using a software program that can render it.

[15] The assumption is that the Zimbra backup and restore have been conducted in a manner that ensures accuracy and completeness. SFU Archives should discuss with IT how the backups are made and restored and how success is measured.

- Pro: Zimbra is relatively easy to convert to Maildir and SFU IT has indicated its willingness to create and maintain such a script;
- Pro: the Zimbra account does not have to be restored before conversion to Maildir, which reduces the chances of data loss;
- Con: Archives staff would not have the opportunity to review the email in Zimbra for comparison purposes (see option 1, above). However, if desired, appraisal could still be conducted on the Maildir backup by opening it in an open-source email program such as Evolution, deleting unwanted messages, then re-exporting the email as Maildir.

Option 3. IT runs OfflineImap or a similar program against an active Zimbra account and transfers the Maildir capture to the Archives.

- Pro: OfflineImap is easy to install and run and can be used against either complete Zimbra accounts or selected folders within accounts;
- Pro: Capturing the Maildir backup directly from Zimbra reduces the chances of data loss associated with creating a Zimbra backup and converting it to Maildir.
- Con: Archives staff would not have the opportunity to review the email in Zimbra for comparison purposes (see option 1, above). However, if desired, appraisal could still be conducted on the Maildir backup by opening it in an open-source email program such as Evolution, deleting unwanted messages, then re-exporting the email as Maildir.

Provision of access

We recommend conversion of the Maildir backup directories to mbox files for providing access. Archivematica can be programmed to do this conversion automatically during processing and to generate a Dissemination Information Package (DIP) containing the mbox files. For an email account that consisted of an inbox with subfolders plus draft and sent items, the DIP would look something like this:

Inbox.mbox
Inbox.TravelCttee.mbox
Inbox.ExecCttee.mbox
Inbox.Workshops.mbox
Drafts.mbox
Sent.mbox

Provision of access must necessarily incorporate access and restriction management to comply with FOIPPA requirements. The only known open-source tool that facilitates large-scale review and tagging of email account contents is Muse. More testing will be required to determine how usable and scalable the process of email tagging and exporting is with this tool. However, it should be noted that Muse is still in active development, and the Muse project team is interested in continuing to develop and refine the tool for use by libraries and archives. This bodes well for future feature development and may

present an opportunity for SFU Archives to actively beta test the tool and provide user feedback to the Muse project team.

The options presented here all assume conversion of the ingested email to mbox files (one mbox file per email folder) and using Muse to tag and remove restricted emails from the versions to be made available to the public:

Option 1: Provide on-site access to mbox files imported into an open-source email program such as Evolution, and/or send users copies of mbox files using an ftp site, DropBox or some other means of exchanging large files over the Internet.[16]

- Pro: This would be simple to implement, especially over the short term while the email preservation program at SFU is still in the early stages.
- Pro: Mbox files are easily imported and rendered in Evolution and other email programs, where they can be navigated in ways users are likely to be familiar with.
- Pro: The email program would provide keyword search functionality.
- Con: The emails would not be provided in the context of an archival description.

Option 2: Upload the mbox files to the relevant descriptions in ICA-AtoM. Users would navigate to the description and download the mbox file, importing it into an email program of their choosing.

- Pros: Same as option 1, above.
- Con: Although the emails would be provided in the context of an archival description, they would not be viewable without being downloaded and opened in a separate program.
- Con: Although the emails would be provided via ICA-AtoM, they would not be full-text searchable within that system.

Option 3: Provide web-based access through a publicly available Muse site.

- Pro: Muse allows rapid paging through emails and attachments and provides a number of useful search and browse options. For example, when viewing an email message in Muse the user can click on the sender's email message and automatically retrieve all messages from that sender.
- Con: Muse has a poorly-designed user interface which would require improvements before this could be considered a viable option.
- Con: The current alpha version of Muse is unstable and poorly documented. Further development of the tool and documentation would be needed before using Muse could be considered a viable option.

[16] Ftp and DropBox are not particularly secure means of transferring files over the Internet; however, the files being transferred would be considered unrestricted and the person to whom they were sent would be free to use and disseminate them publicly.

- Con: The emails would not be provided in the context of an archival description.

Option 4: Convert email messages to individual text or pdf files and upload them with their attachments as items to ICA-AtoM.

- Pro: The emails would be presented individually, along with their attachments, within the context of a hierarchical archival description.
- Pro: The emails would be full-text searchable from within ICA-AtoM.[17]
- Con: This is not currently a viable option because it is difficult to use ICA-AtoM to page through large numbers of textual digital objects in rapid succession. However, future enhancements to the tool's digital object browse functionality (i.e. in 2013 or later) might eventually make this a more realistic option.
- Con: It is not clear how ICA-AtoM would manage and display threads.

Note that the adoption of one or more of these options does not preclude taking advantage of new email rendering and manipulation tools which will likely emerge within the digital curation community in the next few years. The most critical component of SFU's email preservation strategy is the regular acquisition of emails and their conversion to standard formats such as Maildir and mbox; taking these actions will prepare the Archives to provide access to them over the long term.

5.2.3 Required Archivematica development

In order to implement an email preservation plan based on these recommendations, Archivematica will need to be able to ingest Maildir backups and normalize the contents of each Maildir folder to mbox for access purposes. It will also need to strip out email attachments, normalize the attachments to preservation formats where necessary, and maintain relationships between email messages, attachments and normalized versions of the attachments.

[17] Note that the current version of ICA-AtoM does not include full-text search of uploaded digital objects. This feature will be included in version 2.0, scheduled for release in late 2012.

6. Recommendations for establishing a production digital preservation environment

6.1 Summary and analysis

Project objectives and activities

The scope and objective of this Archivematica pilot project was to identify SFU Archives requirement gaps for workflow processing, metadata and file format preservation plans and to incorporate these into the Archivematica release roadmap. The requirements and corresponding roadmap and issues numbers are listed in detail in **Appendix A**.

Artefactual Systems has extended its commitment to include an upgrade of the SFU Archivematica installation to release 0.9 when it becomes available in June 2012, scheduled to coincide with completion of the SFU Archives AIS data migration to ICA-AtoM 1.2. This will give SFU Archives the opportunity to test a fully-integrated ingest-to-access workflow using its own production version of ICA-AtoM and the Archivematica 0.9-beta release. The Archives is planning a second series of pilot tests, to take place during the remainder of the 2012/2013 fiscal year. During this time the Archives will prepare a business case to secure permanent funding for a digital preservation program using Archivematica and ICA-AtoM as its preservation and access components. Establishment of the production digital preservation program would take place in 2013 and 2014.

Preparing for a production digital preservation program: storage requirements

A business case to secure funding for a production digital preservation program must necessarily include cost estimates for storage requirements. However, storage needs will vary greatly depending on the types and quantities of objects to be acquired. If acquisitions were limited mainly to office documents, still images and email accounts, storage requirements could be relatively small. For example, the email account used for testing during this project was just under 1 GB and the mainly textual records from the Telelearning network were only 300 MB. However, if the Archives acquired large bodies of audio and video files, storage requirements would be much larger. The 14 Columbia River Treaty digitized audio files tested during this project totalled approximately 20.3 GB in size; no video files were tested, but it should be noted that they can be extremely large. A relatively small number of full-length industry-standard HD video files, for example, can require several terabytes of storage.

Another factor to consider in estimating storage requirements is the size of the AIP and DIP once the files have been ingested and processed. If the ingested files require no preservation normalization, the AIP will be roughly the same size as the original SIP (or smaller, since the AIP is losslessly compressed). However, if the SIP consists of files requiring normalization, the AIP will store both the original object and its normalized version. For textual records and images, this can result in an AIP that is two or three times the size of the original. For audio and video files, if the ingested file is heavily

compressed the normalized version can be much larger than the original. In many cases, particularly with large video files, a careful analysis of the incoming formats may result in a decision to delay normalization and monitor files for a number of years before any conversion actions are undertaken. In nearly all cases, however, a DIP must be generated for upload into an access system. Given that DIPs typically consist of web-ready compressed versions of original files, it is reasonable to estimate an average DIP size of approximately one-third of the original SIP size.

Given the considerations outline above, it is clear that SFU Archives will need to develop a digital records acquisition plan prior to presenting its business case for establishing the Archives, in order to provide a realistic estimate of storage requirements. This plan should cover, at a minimum, projected acquisitions for the first year of operation.

Preparing for a production digital preservation program: software development and maintenance

The business case should also include cost estimates for software development and maintenance. Software development may include the development of new features to meet any requirement gaps identified during the next phase of software testing. For maintenance, we recommend that SFU Archives enter a maintenance agreement with Artefactual Systems for at least the first year of production. The standard maintenance agreement would include:

- performance tuning the Archivematica installation, including the option to configure and test multi-VM processing to scale processing of large collections;[18]
- configuring and test the ACE auditing tool[19] for checksum verification and error notification for stored AIPs;
- upgrading SFU Archives' Archivematica installation to any new software versions that are released while the agreement is still active;
- applying SFU Archives' prioritized software patches (i.e. to address bugs or required enhancements that are identified between software releases);
- taking advantage of bundled ICA-AtoM technical support;
- participating in the Archivematica Customer Advisory Board to provide direct input on its development roadmap and project priorities;
- having Artefactual Systems brief SFU IT on assuming Archivematica maintenance in the event that SFU Archives decides not to renew the maintenance agreement for another year.

A sample maintenance brochure has been added as **Appendix D** to this report.

[18] Multi-VM processing refers to the use of several virtual machines to add additional processing clients to the Archivematica installation, to allow for efficient processing of large bodies of digital objects.

[19] See https://wiki.umiacs.umd.edu/adapt/index.php/Ace:Audit_Manager_User_Guide.

6.2 Next steps

Based on the results of the pilot project tests, the recommended next steps for SFU Archives are as follows:

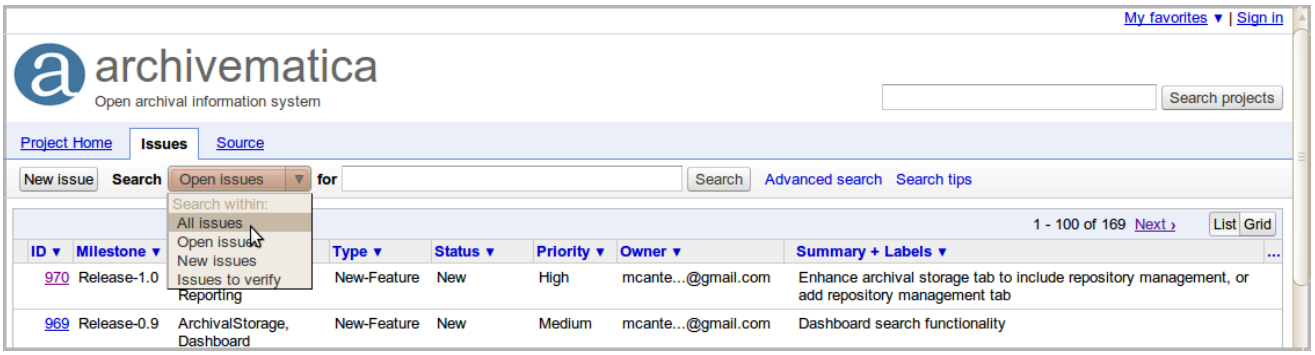
1. Install and test Archivematica 0.9 (installation will be done by Artefactual Systems);
2. Develop an acquisition plan for the first year (minimum) of a production digital preservation program in order to estimate storage requirements;
3. Liaise with SFU IT, with Artefactual's assistance, to determine its charge-back costs for storage (e.g. per GB per month) and computing resources (e.g. per virtual machine);
4. Prepare a business case to secure funding for the establishment of a production digital preservation program;
5. Install Archivematica 1.0 in early 2013 (to include the production installation and a separate test installation for pilot testing new records and filetypes);
6. Conclude an annual maintenance agreement with Artefactual Systems to cover the first year of production.

Appendix A: Recommendations for Archivematica development

1. From Telelearning Network fonds / Columbia River Treaty collection pilot projects.

These recommendations are taken from the report "SFU Archives Archivematica pilot project: findings and recommendations" provided by SFU archivists Richard Dancy and Paul Hebbard. The numbering is the same here as it is in that report.

The Archivematica issues list is located at http://code.google.com/p/archivematica/issues/list. Note that this page defaults to showing open issues; if an issue doesn't appear on the list it may be closed (i.e. fixed), in which case the Search should be changed to "All issues":



Recommendation	Development type	Release	Issue number
Interface/workflow: general			
R1 Manage all user interaction in one place, i.e. the Dashboard.	Enhancement	0.9	952
R2 Build all user actions into the workflow as decision points; i.e. user does not have to remember, but is always prompted for all required actions.	Enhancement	1.0	953
R3 Build the interface around a use case that assumes a time lag between transfer and processing	Enhancement	0.9	951
Icons			
R4 Create unambiguous icons for different error use cases.	Enhancement	0.9	947
R5 Use progress bars to indicate system is processing a microservice.	New Feature	0.9	949
R6 Create an unambiguous icon for job completed.	Enhancement	0.9	934, 947
Number flags			
R7 Make number flags sync with jobs appearing in the Dashboard and in Archivematica directories.	Bug	0.9	773

Recommendation	Development type	Release	Issue number
Decision menus			
R8 At decision points, make action menus larger; and/or require user to click a "Go" button to confirm choice before microservice continues.	Enhancement	0.9	950
Batch processing			
R9 Provide an interface to make it easy to set (and un-set) decision points to default values.	New feature	1.0	912
R10 Provide watched folders in which objects can be queued and run as a batch through microservices set to default values.	New feature	1.0	957
Error handling			
R11 Provide more clear instructions for the use case where an archivist wants to "start again from scratch" and delete the unsuccessful job.	Documentation	0.9	958
R12 Parse error messages in archivist-friendly terms.	New feature	post-1.0	none[20]
R13 Provide decision points when errors occur that set out options and implications.	Documentation	0.9	958[21]
R14 Send error reports to an Archivematica error repository, assigning report IDs and capturing user's system information.	New feature	post-1.0	960
R15 Send error reports (or parsing or Archivematica error report ID) to transfer's metadata folder; provide an interface to make these easily accessible.	New feature	post-1.0	960
Transfer			
R16 Revise the interface to initiate transfer from the Dashboard and automate copying and restructuring.	New feature	0.9	952, 954
R17 Revise the workflow to include a decision point to upload submission documentation and automate copying to appropriate folder.	New feature	0.9	952
R18 Integrate Archivematica transfer with ICA-AtoM accessioning as an option for institutions using the two systems as a bundle; at a minimum, user should be able to link a transfer to a fonds.	New feature	0.9	952[22]
R19 Build into the workflow the creation of standard diagnostic reports following transfer (eg filesize, formats); store these e.g. in transfer's metadata folder and design an interface to easily access them.	New feature	0.9	923
SIP creation			
R20 (1) Create tools that allow representation of the hierarchy	New feature	0.9	952

[20] This is problematic because most of the error messages are stack traces and other output required to diagnose the issue. We may address this in the future on a case-by-case basis, with separate issues being filed for each case.

[21] This should be handled through improvements to the user manual, since error explanations and descriptions of possible choices and implications would be too complex to describe in the dashboard itself.

[22] The link will be made through the accession number which will be added during transfer.

Recommendation	Development type	Release	Issue number
of the directory structure (both before and after arrangement)			
R20 (2) Provide access to these representations through ICA-AtoM.	New feature	1.0	380, 964
R21 Provide an interface to support archival arrangement activities.	New feature	0.9	952
R22 Automate as much as possible the movement of objects from one Archivematica directory to another so that the onus is not on the user to remember how to do it correctly.	Enhancement	0.9	952
Metadata and rights			
R23 Add microservices decision points into the workflow for Add metadata and Add rights information steps.	Enhancement	1.0	953
R24 Allow institutions to select a Metadata template based on DC, RAD or ISAD; or, alternatively, provide tool-tip help that maps the existing DC fields to RAD and ISAD.	New feature	1.0	963
R25 Provide a field on the Metadata screen that maps to Level of Description.	Enhancement	1.0	964
R26 Auto-populate the Dates field from the modification dates of items included in the object.	New feature	post-1.0	965
R27 Auto-populate the Format field with an extent statement based on items in the objects (e.g. Records in electronic form. Size: 23 MB. Total digital objects: 4,582. Original file formats: .doc .pdf .ppt").	Enhancement	1.0	157[23]
Normalization			
R28 (1) Build into the workflow a microservice that automatically generates a normalization report and files it to the transfer's metadata folder.	Existing feature (normalization reports are saved to logs directory).		
R28 (2) Alternatively, use progress bars to indicate to user when the system is still processing a request for a report.	New feature	0.9	949
DIP upload			
R29 Create an interface that would allow the user within Archivematica to search or navigate to ICA-AtoM descriptions and select the appropriate parent.	New feature	1.0	966
R30 Re-use data entered on the Metadata screen in the Part of field (if entered) for DIP upload so user need only enter it once.			none[24]
R31 Build into the preservation plan for file formats the	New feature	post-1.0	967[25]

[23] Issue 157 deals with the object level only. It may not be feasible to populate the Format field with aggregate values.

[24] The "IsPartOf" field is being removed from the DC template in Archivematica 0.9, to be replaced by the broader term "Relation". This will allow users to add different types of relation values in the field, making the metadata easier to map to different access systems. However, it will prevent the value from being used to auto-populate a DIP upload destination template, which is deisgned for a specific parent-child relationship.

[25] Note that metadata extraction is not related to preservation planning, since metadata are extracted from the original file.

Recommendation	Development type	Release	Issue number
identification of descriptive metadata that can be automatically extracted and mapped to ISAD / RAD fields; use this to populate descriptive fields on DIP upload.			
Repository management			
R32 Create an interface and tools for repository management to support: Deletion of transfers, SIPs, AIPs and DIPs, Export of AIPs, Reprocessing of AIPs, Filtered views of transfers, Filtered views of AIPs (by e.g. Fonds, transfer date, filesize, file formats), (6) Storage management (e.g. tools allow users to answer queries like “what are storage requirements for ingesting X number of wav files?”).	New feature	1.0	968, 970

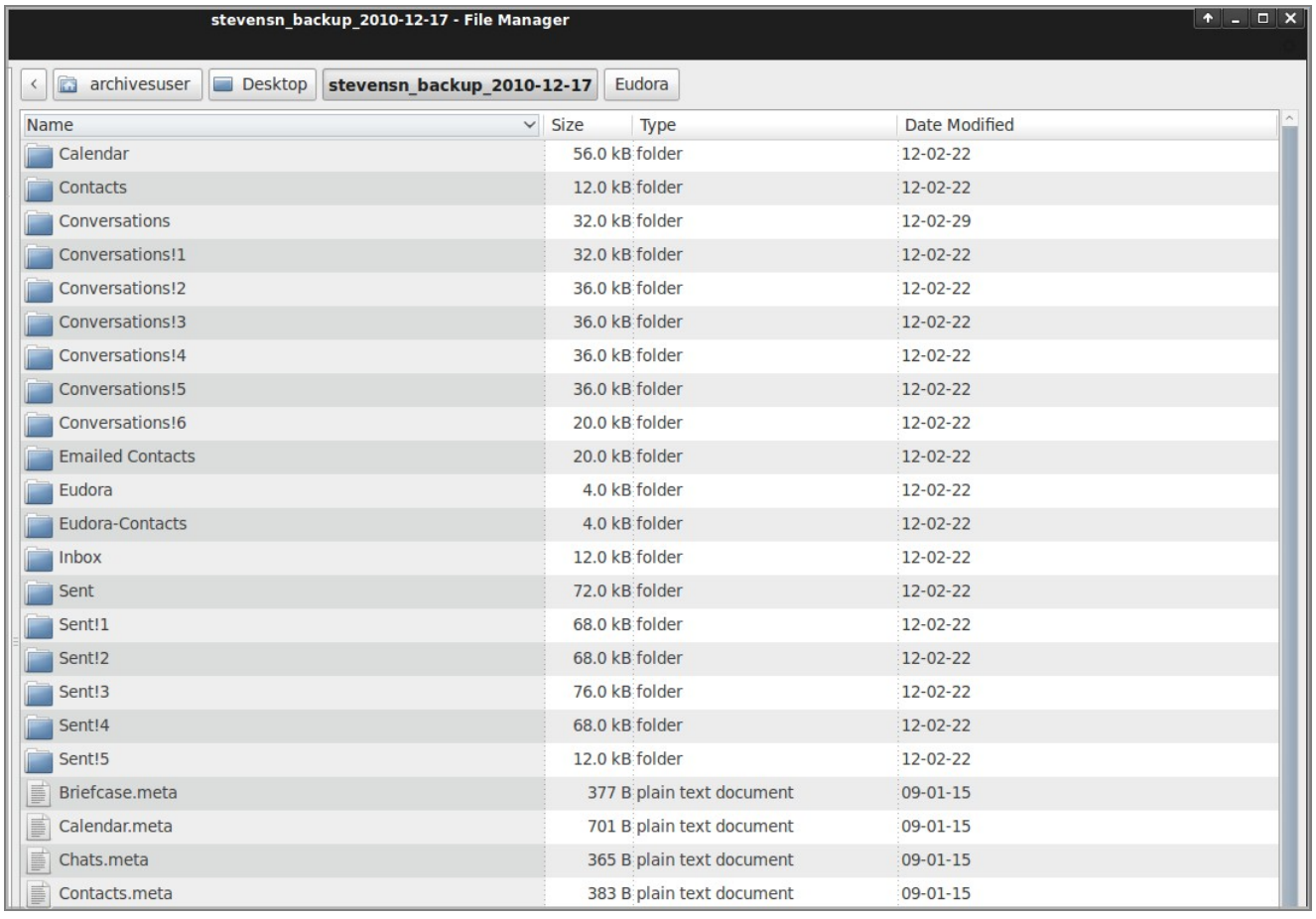
2. From email pilot project

Recommendation	Development type	Release	Issue number
Convert Maildir backups to mbox, 1 mbox file per folder	New feature	0.9	962

Appendix B: Zimbra backup format

The Zimbra backup format is a non-standard format designed specifically to restore a user's email back to Zimbra. In contrast, Maildir can be used to restore email either to the originating email account or to a wide variety of other email programs, both proprietary and open-source.[26]

When a Zimbra account is backed up, the account's directory structure is broken up to limit the size of each directory. For example, in the screenshot below, the Sent folder has been separated into Sent, Sent1!, Sent!2, etc. Moreover, although the email messages are saved as .eml files, some metadata are separated from them and stored separately as .meta files. The purpose of the .meta files is not documented in Zimbra's administrator's manual; they contain information about flags and tagging (e.g. read vs unread), directory location (Sent, Inbox etc) and relationship to other emails (threads). In addition to Inbox, Sent and other folders which would be visible to the Zimbra end-user, a separate set of directories is created called Conversations, Conversations!1, Conversations2! etc. which contain .meta files designed to facilitate the organization of Zimbra messages in the "conversation view" pane in Zimbra.



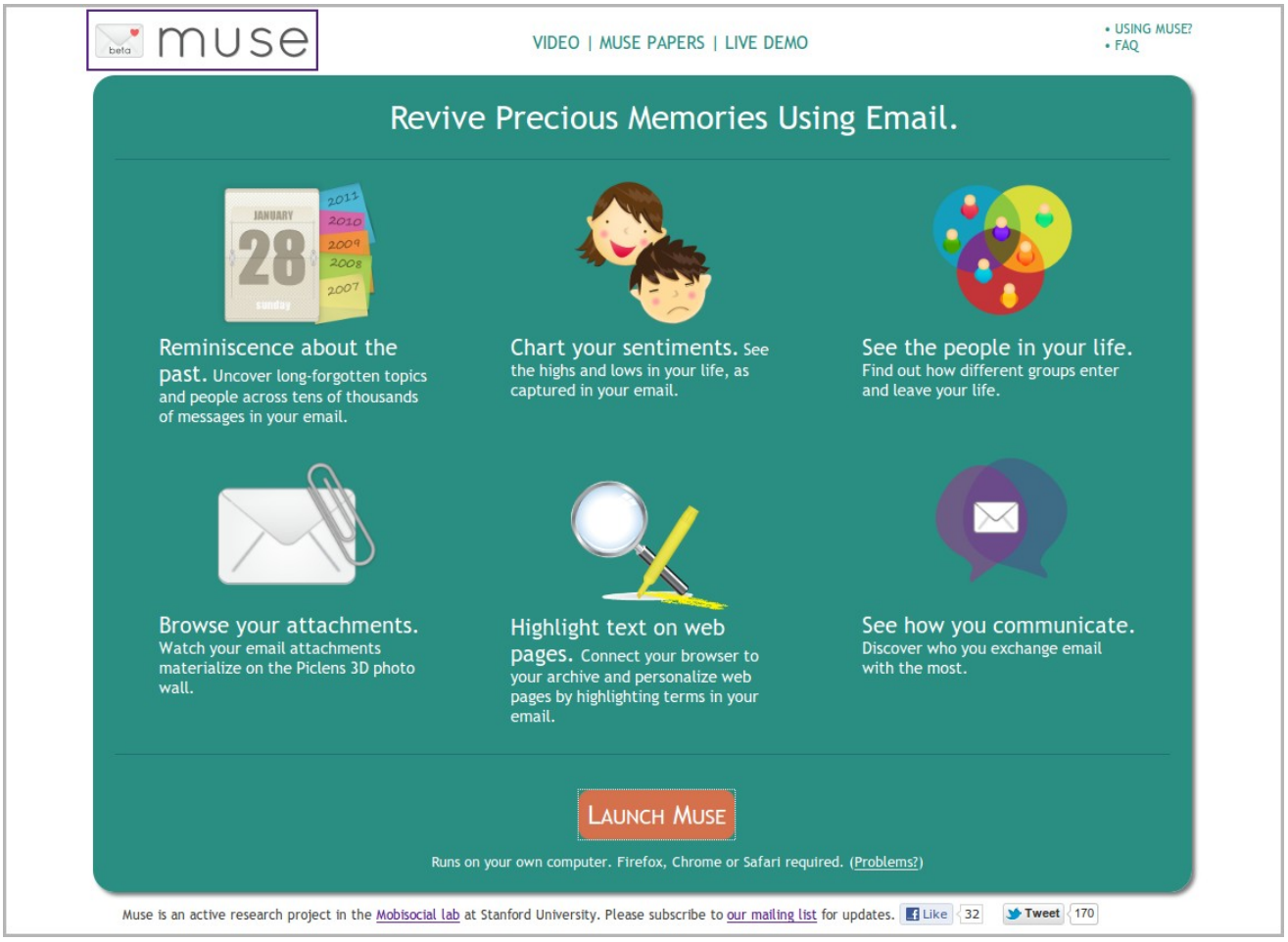
[26] See http://en.wikipedia.org/wiki/Maildir#Software_that_supports_Maildir_directly.

SFU's IT department provided the project team with a script to convert the Zimbra backup to mbox. We noted that the script is designed to convert .eml files only, not the .meta files, which would present a problem because the .meta files indicate whether the email was read or unread. Also, because mbox does not not store information about folder hierarchies, an mbox file must typically be created for each folder; this would be problematic using Zimbra because of the Zimbra folder structure shown in the screenshot. For example, while Maildir saves all sent messages in one Sent folder, the Zimbra backup breaks them across multiple Sent folders. These folders would need to be combined into one folder before an mbox conversion could be run.

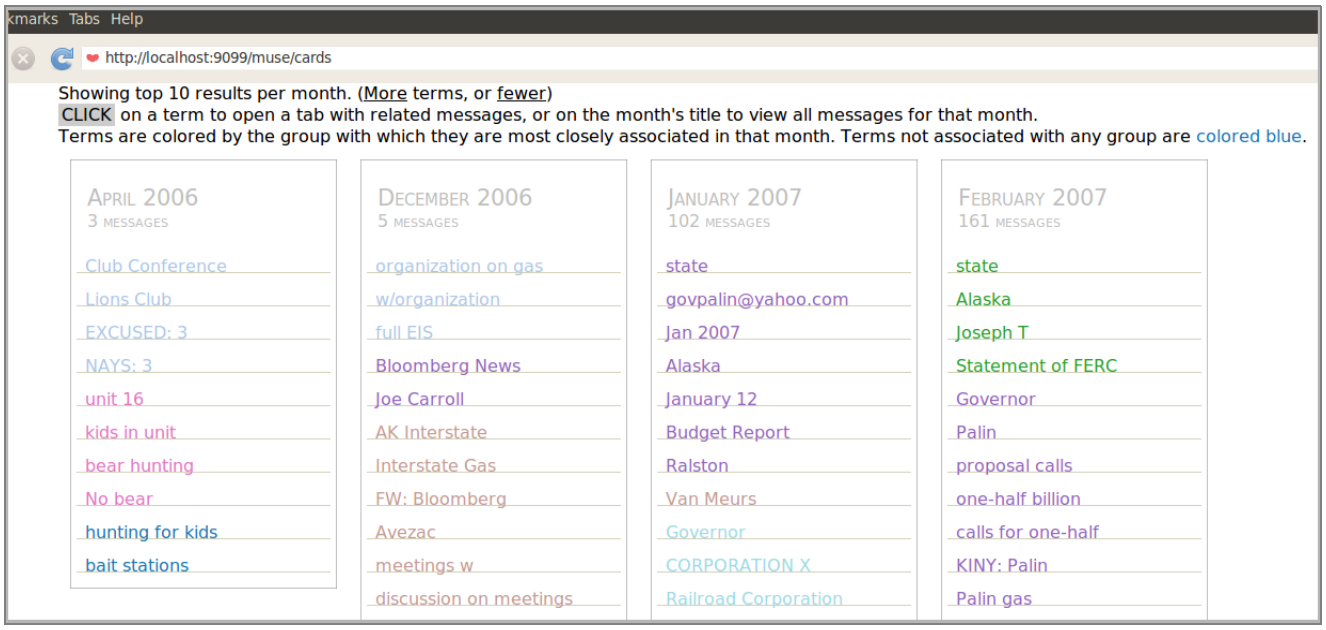
Appendix C: Muse screen captures

Muse allows the user to browse the contents of mbox files using a web interface. The following screenshots demonstrate some of the tool's functionality. Note that Muse is still in alpha development.

1. Muse homepage: the page is designed to highlight Muse as primarily a social media tool:



2. Message list page showing messages organized by automatically generated keywords:



3. Email message view showing a message with a "restricted" tag:

Appendix D: Artefactual Systems annual maintenance brochure

See next page.

Annual Maintenance Program

Nobody understands Archivematica better than Artefactual Systems.

We are the lead developers of the Archivematica software and provide best practice, best knowledge support.

Our highly skilled and experienced engineers help institutions around the world to deploy and maintain their Archivematica systems.

Through our dedicated technical support services we can help you resolve issues of all levels of complexity.

Gold Support — $14,950 / year

* Direct 1-800 phone access to our engineers during business hours
* Guaranteed one business day response to email and voicemail messages
* Our engineers will provide advice to your technical team to resolve the issue in the shortest possible time
* Installation, configuration, integration, upgrade and performance tuning advice resulting in one of:
  * answer to question * solution to problem * issue filed in bug tracking system with release priority
* Includes ICA-AtoM support services
* Five support incidents per year
* One complimentary seat per year for online training course

Platinum Support — $29,950 / year

* Direct 1-800 phone access to our engineers during business hours
* Guaranteed one business day response to email and voicemail messages
* All support incidents will result in one of:
  * answer to question * solution to problem * issue filed in bug tracking system with release priority
* Our engineers will analyze and design the Archivematica configuration most suitable to your institution's technical environment so that we can offer proactive support geared specifically to your needs
* Remote login (VPN/SSH) by our engineers for installation, integration, upgrade, troubleshooting, application of patches, performance tuning
* Entitlement to 3 pre-release code patches (enhancement/bug fix) implemented on your system by our engineers
* Includes ICA-AtoM support services
* Unlimited support incidents per year
* Three complimentary seats per year for online training course
* **Customer Advisory Board membership:**
  * biannual telecon/meeting * direct input on development roadmap and project priorities

Archivematica is open-source software. It is available free of charge, 'as-is' to all interested parties at http://archivematica.org. This Annual Maintenance Program is offered to those institutions that would like the assurance of expert technical support for their Archivematica installation. Additionally, the purchase of an Annual Maintenance Agreement helps to fund ongoing Archivematica research and development thereby contributing to project sustainability for the benefit of all Archivematica users. Pricing in effect as of January 1, 2012. Pricing is in Canadian dollars (CAD$).

artefactual systems inc.