The Community-driven Evolution of the Archivematica Project

Peter Van Garderen
President, Artefactual Systems, Inc.
202-26 Lorne Mews
New Westminster, BC, Canada
1.604.527.2056
peter@artefactual.com

Courtney C. Mumma
Systems Archivist, Artefactual Systems Inc.
202-26 Lorne Mews
New Westminster, BC, Canada
1.604.527.2056
courtney@artefactual.com

ABSTRACT

In this paper, we discuss innovations by the Archivematica project as a response to the experiences of early implementers and informed by the greater archival, library, digital humanities and digital forensics communities. The Archivematica system is an implementation of the ISO-OAIS functional model and is designed to maintain standards-based, long-term access to collections of digital objects. Early deployments have revealed some limitations of the ISO-OAIS model in the areas of appraisal, arrangement, description, and preservation planning. The Archivematica project has added requirements intended to fill those gaps to its development roadmap for its micro-services architecture and web-based dashboard. Research and development is focused on managing indexed backlogs of transferred digital acquisitions, creating a SIP from a transfer or set of transfers, developing strategies for preserving email, and receiving updates about new normalization paths via a format policy registry (FPR).

General Terms

Documentation, Performance, Design, Reliability, Experimentation, Security, Standardization, Theory, Legal Aspects.

Keywords

archivematica, digital preservation, archives, OAIS, migration, formats, PREMIS, METS, digital forensics, agile development, open-source, appraisal, arrangement, description, acquisition

1. INTRODUCTION

The ISO 14721-OAIS Reference Model [1] gave the archives community a common language for digital archives architectures. One such architecture is the Archivematica suite of tools which was based on an extensive requirements analysis of the OAIS functional model [2]. The Archivematica project is nearing its first beta release. Project partners and independent implementers have been testing alpha releases using real-world records. These activities have identified some OAIS requirement gaps for digital archives systems.

The project has found that, while it serves as an excellent foundation and framework for long-term preservation strategies, the OAIS model proves inadequate to address some functions unique to archives. In particular for the areas of appraisal, arrangement, description, and preservation planning there were clear gaps between the model and the way that archivists actually process records. The Archivematica project has added requirements to its development roadmap to fill those gaps in its micro-services architecture and web-based dashboard. Other research and development is focused on managing a backlog of indexed digital acquisitions, creating a Submission Information Package (SIP) from a transfer or set of transfers, developing strategies for preserving email, and receiving updates about new normalization paths via a format policy registry (FPR).

2. ABOUT THE ARCHIVEMATICA PROJECT

The Archivematica system uses a micro-services design pattern to provide an integrated suite of free and open-source software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model [3]. It allows archivists and librarians to process digital transfers (accessioned digital objects), arrange them into Submission Information Packages (SIPs), apply media-type preservation plans and create high-quality, repository-independent Archival Information Packages (AIPs). Archivematica is designed to upload Dissemination Information Packages (DIPs) containing descriptive metadata and web-ready access copies to external access systems such as DSpace, CONTENTdm and ICA-AtoM. Users monitor and control the micro-services via a web-based dashboard.

A thorough use case and process analysis identified workflow requirements to comply with the OAIS functional model. Through deployment experiences and user feedback, the project has expanded beyond OAIS requirements to address analysis and arrangement of transfers into SIPs and allow for archival appraisal at multiple decision points. The Archivematica micro-services implement these requirements as granular system tasks which are provided by a combination of Python scripts and one or more of the free, open-source software tools bundled in the Archivematica system.

Archivematica uses METS, PREMIS, Dublin Core and other recognized metadata standards. The primary preservation strategy is to normalize files to preservation and access formats upon ingest when necessary (for example, when the file is in a format that is proprietary and/or is at risk of obsolescence). The media type preservation and access plans it applies during normalization are based on format policies derived from an analysis of the significant characteristics of file formats [4]. The choice of access formats is based on the ubiquity of viewers for the file format as well as the quality of conversion and compression. Archivematica's preservation formats are all open standards [5]. Additionally, the choice of preservation and access formats is based on community best practices and availability of open-source normalization tools.

Archivematica maintains the original files to support future migration and emulation strategies. However, its primary preservation strategy is to normalize files to preservation and access formats upon ingest. The default normalization format policies can be edited and disabled.

All of the software, documentation and development infrastructure are available free of charge and released under AGPL3 and Creative Commons licenses to give users the freedom to study, adapt and re-distribute these resources as best suits them. Archivematica development is led by Artefactual Systems, a Vancouver based technical service provider that works with archives and libraries to implement its open-source solutions as part of comprehensive digital preservation strategies. All funding for Archivematica development comes from clients that contract Artefactual's team of professional archivists and software developers to assist with installation, integration, training and feature enhancements. The majority of Archivematica users take advantage of its free and open-source license without additional contracting services.

3. ACQUISITION AND BACKLOG MANAGEMENT

Early implementers of the Archivematica suite of tools have consistently struggled with the mechanics of acquiring digital materials. Analogue records are delivered to the repository or are picked up from the donor's storage location, but digital acquisition can be more varied. Digital materials can arrive via digital transfer over a network such as email, FTP or shared directories. The archives may have to send an archivist to acquire the digital materials onsite, and even then, there are several options for acquisition including pickup, copying, or imaging. Depending on the type of acquisition, should the archivist photograph the condition of the materials in their original location? What steps must be taken to ensure that the digital objects copied or imaged retain their integrity during transfer to the archives? Finally, when digital materials are donated to the archives onsite, how do processes differ from pickup and digital network transfer?

Archivists who deal primarily with analogue materials are well accustomed to the need to maintain a backlog. Acquisitions regularly occur for which there are limited or no resources to process them immediately. For this reason, it is imperative that the archives achieve a minimum level of control over the material so that it can be tracked, managed, prioritized and, if necessary, subjected to emergency preservation actions.

Archivematica runs through a set of transfer actions in the dashboard to establish initial control of the transfer. It verifies that the transfer is properly structured or structures it if necessary. Then, it assigns a unique universal identifier (UUID) for the transfer as a whole and both a UUID and a sha-256 checksum to each file in its /objects directory. Next, Archivematica generates a METS.xml document that captures the original order of the transfer and that will be included in any SIP(s) generated from this transfer. Any packaged files are unzipped or otherwise extracted, filenames are sanitized to remove any prohibited characters, and file formats are identified and validated. Finally, technical metadata is extracted from the files and the entire transfer content and metadata is indexed. At this point in the process, the transfer is ready to be sent to a backlog storage location that should be maintained in much the same way as the archival storage. The transfer is ready for future processing. These features will be added and evaluated in forthcoming releases of the Archivematica software.

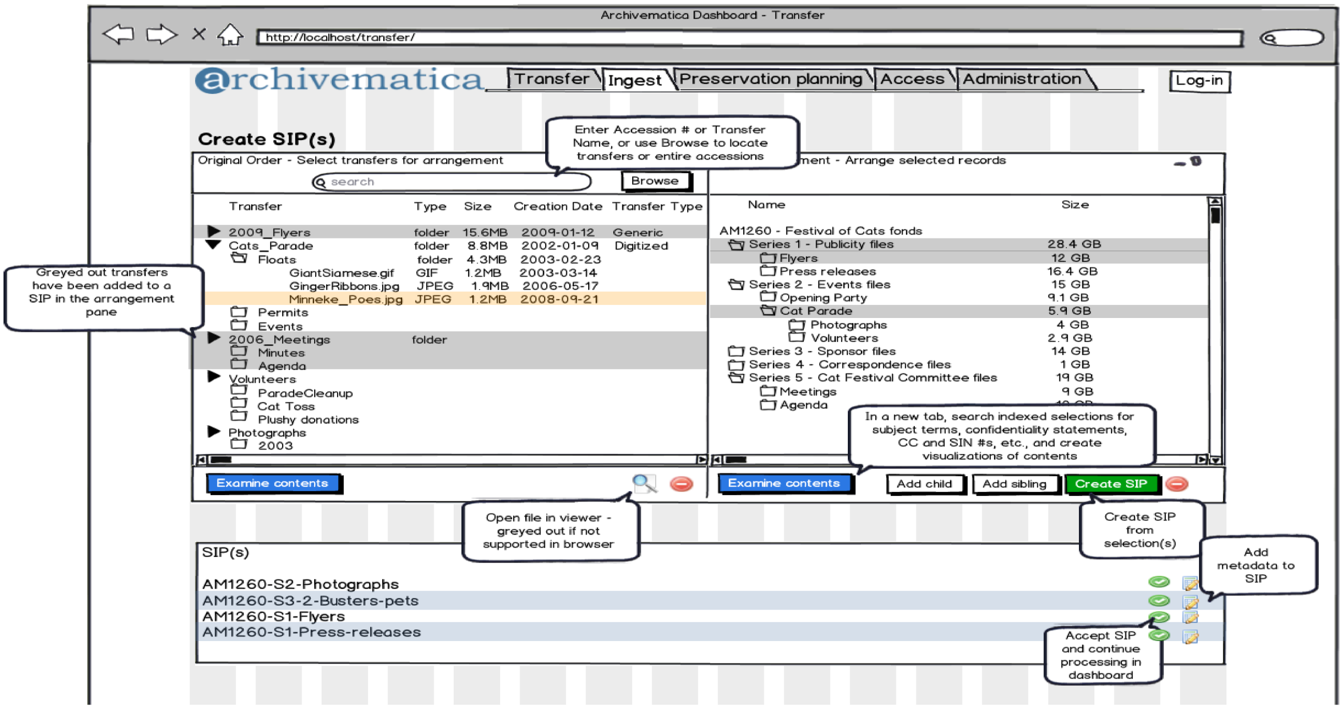
4. ARRANGEMENT AND DESCRIPTION

Once an archives is ready to process one or more digital acquisitions, the next challenge comes from making a SIP from disparate parts of an acquisition. For example, in a situation in which an acquisition arrives on multiple digital media, the archives may have accessioned transfers from each media type and/or broken a very large hard drive into two or more transfers. Presumably, archivists will want their SIPs to be formed so that the resultant AIPs and DIPs conform to some level of their archival description, so SIP content could derive from one or more transfers or parts of transfers.

Arrangement and description do not neatly occur at one specific point during processing. Archivists arrange and describe analogue records intermittently. Arrangement is based upon the structure of the creator’s recordkeeping system, inherent relationships that reveal themselves during processing and compensations made to simplify managing records and/or providing access. Archivists document their arrangement decisions and add this information, along with additional descriptive information gathered about the records during processing, to the archival description. Further, documentation of arrangement decisions and actions supports respect des fonds by preserving information about original order. Digital records must be arranged and described in order to effectively manage and provide access to them. Analogue functionality is very difficult to mimic in a digital preservation system such as Archivematica, because any interaction that allows for analysis of the records can result in changing original order and metadata associated with the records.

The OAIS model assumes that a digital archives system receives a fully formed SIP. However, this is often not the case in practice. Early Archivematica implementers were often manually compiling SIPs from transfers in the Thunar file browser bundled with the system. After transfer micro-services are completed successfully, Archivematica allows transfers to be arranged into one or more SIPs or for one SIP to be created from multiple transfers. The user can also re-organize and delete objects within the SIP(s). The original order of the transfer is maintained as its own structMap section in the transfer METS file, a copy of which is automatically added to each SIP. Additionally, the archivist can use dashboard functionality to add basic descriptive metadata to the SIP at this point, including information about rights and restrictions.

The Archivematica project is now working on the ability to call up a transfer into a file browser interface in the dashboard's Ingest tab, examining its contents and forming it into SIPs for processing (See Figure 1).


Archivematica Dashboard - Transfer
http://localhost/transfer/
archivematica
Transfer
Ingest
Preservation planning
Access
Administration
Log-in
Create SIP(s)
Original Order - Select transfers for arrangement
search
Browse
Enter Accession # or Transfer Name, or use Browse to locate transfers or entire accessions
ment - Arrange selected records
Greyed out transfers have been added to a SIP in the arrangement pane
Examine contents
Open file in viewer - greyed out if not supported in browser
Examine contents
Add child
Add sibling
Create SIP
In a new tab, search indexed selections for subject terms, confidentiality statements, CC and SIN #s, etc., and create visualizations of contents
Create SIP from selection(s)
Add metadata to SIP
Accept SIP and continue processing in dashboard

Transfer	Type	Size	Creation Date	Transfer Type
▶ 2009_Flyers	folder	15.6MB	2009-01-12	Generic
▼ Cats_Parade	folder	8.8MB	2002-01-09	Digitized
Floats	folder	4.3MB	2003-02-23	
GiantSiamese.gif	GIF	1.2MB	2003-03-14	
GingerRibbons.jpg	JPEG	1.9MB	2006-05-17	
Minneke_Poes.jpg	JPEG	1.2MB	2008-09-21	
Permits				
Events				
▶ 2006_Meetings	folder			
Minutes				
Agenda				
▶ Volunteers				
ParadeCleanup				
Cat Toss				
Plushy donations				
▶ Photographs				
2003				

Name	Size
AM1260 - Festival of Cats fonds	
Series 1 - Publicity files	28.4 GB
Flyers	12 GB
Press releases	16.4 GB
Series 2 - Events files	15 GB
Opening Party	9.1 GB
Cat Parade	5.9 GB
Photographs	4 GB
Volunteers	2.9 GB
Series 3 - Sponsor files	14 GB
Series 4 - Correspondence files	1 GB
Series 5 - Cat Festival Committee files	19 GB
Meetings	9 GB
Agenda	

SIP(s)

AM1260-S2-Photographs
AM1260-S3-2-Busters-pets
AM1260-S1-Flyers
AM1260-S1-Press-releases

Figure 1. Create SIP dashboard interface mockup.

Much of the inspiration for such an interface came from digital forensics software and the Curator's Workbench at the University of North Carolina, Chapel Hill. The UNC Libraries had developed Curator’s Workbench [6], a tool that, among other things, allows for arrangement of digital records without losing the original order. The Archivematica team considered including the tool in their suite, but because of concerns about integration and ongoing support, they opted instead to mimic its arrangement functionality. Archivematica’s 0.8 alpha release uses the Xubuntu file browser Thunar to arrange records and METS to keep a record of the original order within each SIP formed from a transfer. In future releases, the METS file is still generated while file browser functionality has been moved to the dashboard. Future developments could see expanded METS and/or PREMIS profiles that includes information about selection actions undertaken during SIP creation and at the various appraisal stages.

The limitations for analyzing and forming SIPs using only the file browser were clear in earlier Archivematica releases. Transfers may contain restricted material, passwords, personal information or other content that is unsuitable for continued preservation. For insight into this problem, the project explored the possibilities of using digital forensics techniques. Digital forensics experts must review massive sets of digital records and compile selections from them as evidence. Clearly, the set of records presented as evidence must be verifiably authentic. Since archives are held to the same standards of authenticity there is much to be learned from the digital forensics field, which for over thirty years has been developing tools for processing evidence that guarantees its acceptance in courts. Such tools allow for auditing an investigator’s actions, recording information about the set of records and its origin while adding descriptive metadata and grouping portions of the set into discrete evidence packages, indexing and examining the file system structure and contents, and ensuring integrity. Many of the software tools used by digital forensics experts are proprietary, but in recent years open source tools have been developed to perform the same functions.

Despite their availability, open source digital forensics tools can be difficult to understand by non-experts. Serendipity’s role in open source software development cannot be overstated. Just when Archivematica's systems analysts realized that they could not possibly decipher the entire canon of digital forensics software in time for the next release, digital humanities scholars and archivists in the United States were conceptualizing the BitCurator Project. From the BitCurator website [7]: “The BitCurator Project is an effort to build, test, and analyze systems and software for incorporating digital forensics methods into the workflows of a variety of collecting institutions.” Artefactual Systems is closely involved with the BitCurator Project, with its president, Peter Van Garderen, on the Development Advisory Group and Courtney Mumma, systems analyst, on the Professional Experts Committee. Ideally, BitCurator will result in a set of open source tools that allow for arrangement, description and other valuable functionalities that integrate well into the Archivematica suite.

Since open source digital forensics tools for archivists like BitCurator are not yet ready to be integrated into the Archivematica suite, the team looked for other ways to provide the necessary services to satisfy their workflows. One possible solution is using Apache Tika [8] and ElasticSearch [9] to index and search transfers in a dashboard file browser window to determine which part(s) to include in the SIP and to create visualizations of the transfer and SIP contents.

Requirements for future releases include indexing and reporting on all text content, file embedded metadata and file formats. Using Tika, ElasticSearch and other tools, Archivematica will provide keyword and pattern matching for privacy/security sensitive information (e.g. social insurance numbers/social security numbers, credit card numbers, email addresses and security keywords) and reports of such things as PDFs that have not been OCR'ed, password protected and encrypted files and duplicates with their full file paths. Reports for all of these indexing requirements will be available via the Examine Contents windows, accessible from the Create SIP browser window in the Ingest tab of the dashboard.

The Examine Contents reports will include a search box for the indexed transfer content, general information about the transfer or selected file group (e.g. number of files, size, name, UUID, and accession number), a pie graph visualization showing file type distribution overall and a bargraph visualization showing file type by folder and ordered by size. Clickable links will open to sub-reports on all contents of a specified format in context of the entire transfer, duplicates with their locations, privacy and confidentiality keywords and numbers and password protected files with their distribution across the entire contents visualized as a graph (See Figure 2).

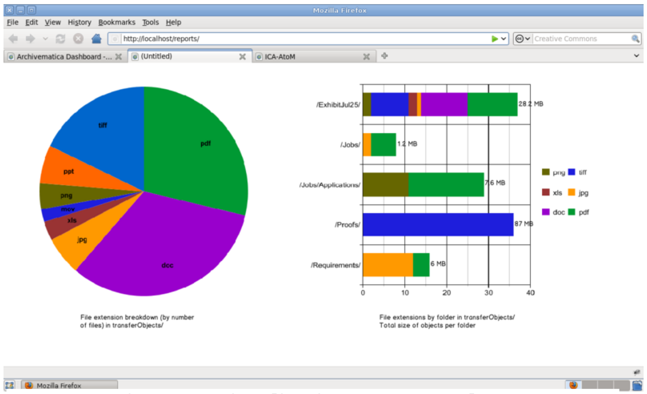

Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/reports/
Creative Commons
Archivematica Dashboard -...
(Untitled)
ICA-AtoM
tiff
pdf
ppt
png
mov
xls
jpg
doc
File extension breakdown (by number of files) in transferObjects/
/ExhibitJul25/
28.2 MB
/Jobs/
1.2 MB
/Jobs/Applications/
7.6 MB
/Proofs/
87 MB
/Requirements/
6 MB
0
10
20
30
40
png
tiff
xls
jpg
doc
pdf
File extensions by folder in transferObjects/ Total size of objects per folder
Mozilla Firefox

Figure 2. Visualization report mockup.

Such functions are being developed iteratively as part of the 2013 Archivematica 1.0 and subsequent releases. Should BitCurator or something else come along that can fulfill or expand on any of these functions, Archivematica's microservices architecture is such that the same requirements can be accomplished by these other tools with only minor changes to the code.

5. APPRAISAL

Originally intended for the long-term preservation of scientific data, OAIS does not address archival appraisal. To advise in the formation of appraisal requirements, the team consulted with the InterPARES 3 Project [10] to conduct a gap analysis between OAIS and the InterPARES 1 Project's Chain of Preservation (COP) Model [11]. Review of the model, along with consultations with archivists about processing analogue records, revealed that appraisal occurs in a few different stages during archival processing. Archivists make an acquisition decision based on a preliminary appraisal, then reassess iteratively when they discover more about the records during accessioning actions and processing. Project partners including archivists and Archivematica developers built workflows around these different appraisal functions, which resulted in constructing three opportunities for appraisal in Archivematica: Selection for Acquisition, Selection for Submission and Selection for Preservation. The three appraisal opportunities as they were manifested in Archivematica 0.7.1 are discussed in detail in a recent Archivaria article [12], so the following is a brief summary of their functions and the associated Archivematica micro-services.

Selection for Acquisition occurs before records are accepted into an archives' custody for processing and preservation. Common practice in archives is to gather and review information about the records creator, the recordkeeping system(s) and the records to make an acquisition decision. For digital records, this includes learning as much as possible about the technological context of the records [8]. Because of limited access to originating technological environments for various reasons, it may become necessary for archives to acquire many more records than they might from an analogue body of records. Therefore, steps must be taken to ensure integrity of the records acquired while appraisal decisions are made over time.

Selection for Submission is the process of forming Submission Information Packages (SIPs) from acquired digital records or "transfers". In Archivematica, a transfer is any set of digital records acquired but not yet processed. Each SIP derives from one or more transfers. However, the SIP cannot be formed until the archivist has some information about the content of the transfer. For this reason, the transfer undergoes several micro-services first so that the archivist can review the results and assess how the received contents compare to the initial Selection for Acquisition expectations.

In the 0.8 alpha iteration of Archivematica, the archivist starts by adding a transfer to a specified folder in the file browser. The transfer begins processing in the Transfer tab of the web-based dashboard, where it is verified to be compliant for ingest in the system. Then, it is renamed with a transfer UUID and is assigned file UUIDs and checksums. If checksums already exist in the transfer, they are verified. A METS.xml file is added to the transfer, the transfer can be quarantined, and any packages are extracted. After a virus scan, prohibited characters are removed from filenames, formats identification process are run and metadata is characterized and extracted. All of the information generated from these micro-services allow the archivist to decide which parts of the transfer are archival materials ready for further processing.

In the 0.9 beta iteration of Archivematica, only one transfer can become one SIP, which is deprecated from the functionality of 0.8. In 0.8, one or more transfer(s) could become one or more SIP(s), but the arrangement was done in the file browser. The reason for this deprecation is so that the 1.0 release can move all the Selection for Acquisition functions to the web browser and improve the tools to create SIPs as discussed in the Arrangement and Description section of this paper.

Selection for Preservation results in forming an Archival Information Package (AIP). A SIP is subjected to several micro-services, displayed in the Ingest tab, before the archivist has an opportunity to review the resulting AIP. Micro-services include verifying SIP compliance, renaming SIP with a SIP UUID, sanitizing file, directory and SIP name(s), checking integrity, copying metadata and logs from the transfer, and normalization. Once normalization and all other processing micro-services have run, the archivist can review the AIP contents and metadata in another browser window or download it to review using the file browser. At that point, they can either reject or accept the AIP and upload it into designated archival storage.

At every stage of appraisal, archivists may choose to destroy or deselect a record or set of records. Archivematica keeps logs of these changes by adding a text file listing excluded records to the logs directory in the transfer or SIP. This may even allow for richer and more transparent descriptive information about archival processing than is accomplished in analogue archives. It is important to note that the aforementioned steps are optional choices for the user. If the user has limited time or knows a great deal about the contents of a SIP, for instance, if the SIP is made up of described digitized videos, Archivematica can be configured to allow for automatic ingest.

In forthcoming releases, these appraisal processes will be incrementally moved to a web browser interface in the dashboard. Elastic Search indexing of the transfer and the AIP should also contribute to a richer, more informed selection process. Other development may include an automated process for “flagging” transfer content that may require further appraisal review based on a predefined set of indexing results.

6. PRESERVING AND PROVIDING ACCESS TO EMAIL

Several Archivematica project partners targeted email preservation as a priority in their digital archives planning. One pilot project involved acquiring a snapshot of the email account of a former university president. The account had been active for 10 years and no other email from the account had been sent to the university archives in electronic form in the past.

The university was using Zimbra Network Edition to send and receive email [13]. The Zimbra administrator's manual does not include information on how to export email from Zimbra for use in other email programs.[14] However, the university's IT department backs up the email accounts using a default directory structure specific to Zimbra, and was willing 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. 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 were Maildir and mbox [15]. 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 file [16]; 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 [17], is designed to render mbox files using only a web browser. 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 [18]. 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. 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.

The project team tested the use of a tool called OfflineImap [19] to back up a test Zimbra email account to Maildir and converted the Maildir backup to mbox using a freely available python script [20]. Following these preliminary tests, the Zimbra backup of the sample 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 [21].

A critical component of the University's email preservation strategy is management of access based on compliance with Freedom of Information and Protection of Privacy legislation. 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. The University'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.

Based on our analysis of pilot project email systems, email management practices, and preservation formats and conversion tools, we have summarized Archivematica requirements for acquiring, preserving and providing access to email. Ideally, email is acquired, per account, 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;
- OfflineIMAP is proving to be a useful tool for capturing email accounts in maildir format. Acting as an IMAP client, it can interact with a wide number of mail server programs, avoiding the need to add support for other mail server or email archive format conversions.
- The contents of a Maildir directory are plain 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 [22];
- 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.

The archivists would submit the Maildir backup into Archivematica, where it would be retained as the preservation master in the AIP. Note that Maildir backups do not capture calendars or contact lists. However, University Archives staff have indicated that such records would probably not be considered archival. 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. Attachments must be extracted and normalized because they pose a usability risk as base 64 ascii encoded text. They will always need to be rendered in a software program for human cognition of its content. 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.

For access, Archivematica will automatically generate a Dissemination Information Package (DIP) containing mbox files generated from the maildir preservation master. 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
```

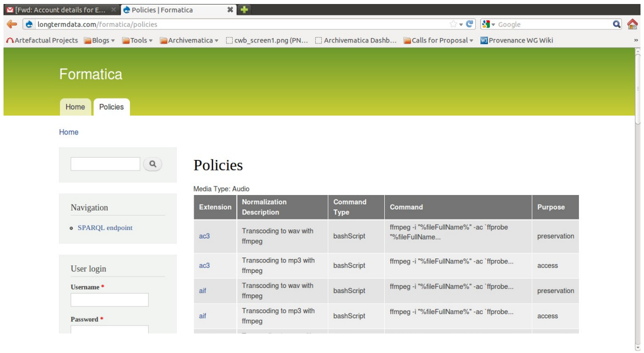
For most university and public repositories, provision of access must necessarily incorporate access and restriction management to comply with freedom of information, privacy and confidentiality requirements. The only known open-source tool that facilitates large-scale review and tagging of email account 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 informed by Archivematica community members.

7. FORMAT POLICY REGISTRY - FPR

The Archivematica project team has recognized the need for a way to manage format conversion preservation plans, referred to by the project as format policies, which will change as formats and community standards evolve. A format policy indicates the actions, tools and settings to apply to a particular file format. The Format Policy Registry (FPR) will provide valuable online statistics about default format policy adoption as well as customizations amongst Archivematica users and will interface with other online registries (such as PRONOM and UDFR) to monitor and evaluate community-wide best practices. It will be hosted at archivematica.org/fpr.

An early prototype has been developed by Heather Bowden, then Carolina Digital Curation Doctoral Fellow at the School of Information and Library Science in the University of North Carolina at Chapel Hill (See Figure 3). A basic production version implementing these concepts will be included in upcoming releases. The FPR stores structured information about normalization format policies for preservation and access. These policies identify preferred preservation and access formats by media type. The choice of access formats is based on the ubiquity of viewers for the file format. Archivematica's preservation formats are all open standards; additionally, the choice of preservation format is based on community best practices, availability of open-source normalization tools, and an analysis of the significant characteristics for each media type. These default format policies can all be changed or enhanced by individual Archivematica implementers. Subscription to the FPR will allow the Archivematica project to notify users when new or updated preservation and access plans become available, allowing them to make better decisions about normalization and migration strategies for specific format types within their collections. It will also allow them to trigger migration processes as new tools and knowledge becomes available.

One of the other primary goals of the FPR is to aggregate empirical information about institutional format policies to better identify community best practices. The FPR will provide a practical, community-based approach to OAIS preservation and access planning, allowing the Archivematica community of users to monitor and evaluate formats policies as they are adopted, adapted and supplemented by real-world practioners. The FPR APIs will be designed to share this information with the Archivematica user base as well with other interested communities and projects.



Policies

Media Type: Audio

Extension	Normalization Description	Command Type	Command	Purpose
ac3	Transcoding to wav with ffmpeg	bashScript	ffmpeg -i "%fileFullName%" -ac `ffprobe "%fileFullName...	preservation
ac3	Transcoding to mp3 with ffmpeg	bashScript	ffmpeg -i "%fileFullName%" -ac `ffprobe...	access
aif	Transcoding to wav with ffmpeg	bashScript	ffmpeg -i "%fileFullName%" -ac `ffprobe...	preservation
aif	Transcoding to mp3 with ffmpeg	bashScript	ffmpeg -i "%fileFullName%" -ac `ffprobe...	access

Figure 3 FPR format policies in early "Formatica" prototype. "Formatica" has since been renamed "FPR".

8. CONCLUSION

Working with pilot project implementers, the Archivematica team has gathered requirements for managing a backlog of indexed digital acquisitions transfers, creating a SIP from a transfer or set of transfers, basic arrangement and description, preserving email, and receiving updates about new normalization paths via a format policy registry (FPR). After creating workflows that would account for real-world archival processing needs, these requirements have been added to our development roadmap for 0.9, 1.0 and subsequent Archivematica releases [23].

The Archivematica pilot project analysis and development described in this article are driven by practical demands from our early adopter community. The alpha release prototype testing sponsored by our contract clients and shared by a growing community of interested users from the archives and library professions and beyond has provided the opportunity to spearhead the ongoing evolution of digital preservation knowledge in the form of a software application that is filling a practical need for digital curators.

At the same time, the digital curation community is also evolving and maturing. New tools, concepts and approaches continue to emerge. The Archivematica technical architecture and project management philosophy are designed to take advantage of these advancements for the benefit of Archivematica users and the digital curation community at large.

The free and open-source, community-driven model provides the best avenue for institutions to pool their technology budgets and to attract external funding to continue to develop core application features as requirements evolve. This means the community pays only once to have features developed, either by in-house technical staff or by third-party contractors such as Artefactual Systems. The resulting analysis work and new software functionality can then be offered at no cost in perpetuity to the rest of the user community at-large in subsequent releases of the software. This stands in contrast to a development model driven by a commercial vendor, where institutions share their own expertise to painstakingly co-develop digital curation technology but then cannot share that technology with their colleagues or professional communities because of expensive and restrictive software licenses imposed by the vendor.

9. REFERENCES

1. ISO 14721:2003, Space data and information transfer systems – Open archival information system – Reference model (2003).
2. Artefactual Systems, Inc. and City of Vancouver, Requirements, http://archivematica.org/wiki/index.php?title=Requirements (accessed May 21, 2012).
3. Artefactual Systems, Inc., Archivematica homepage, http://archivematica.org (accessed May 24. 2012).
4. Archivematica significant characteristics evaluation, https://www.archivematica.org/wiki/Significant_characteristics (accessed August 19, 2012).
5. Wikipedia definition of open standards, http://en.wikipedia.org/wiki/Open_standard (accessed August 17, 2012).
6. Carolina Digital Repository Blog, "Announcing the Curator's Workbench", http://www.lib.unc.edu/blogs/cdr/index.php/2010/12/01/announcing-the-curators-workbench/ (accessed May 21, 2012).
7. BitCurator Tools for Digital Forensics Methods and Workflows in Real-World Collecting Institutions, http://www.bitcurator.net/ (accessed May 21, 2012).
8. Tika website, http://tika.apache.org/ (accessed May 21, 2012).
9. ElasticSearch website, http://www.elasticsearch.org/ (accessed May 21, 2012).
10. .InterPARES 3 Project, http://www.interpares.org/ip3/ip3_index.cfm (accessed May 21, 2012).
11. InterPARES 2 Project, Chain of Preservation (COP) Model, http://www.interpares.org/ip2/ip2_model_display.cfm?model=cop (accessed May 21, 2012).
12. Courtney C. Mumma, Glenn Dingwall and Sue Bigelow, "A First Look at the Acquisition and Appraisal of the 2010 Olympic and Paralympic Winter Games Fonds: or, SELECT * FROM VANOC_Records AS Archives WHERE Value='true';" (Archivaria 72, Fall 2011) pgs. 93-122.
13. Zimbra website, http://www.zimbra.com/ (accessed May 21, 2012).
14. Administration Guide to Zimbra, http://www.zimbra.com/docs/ne/6.0.10/administration_guide/ (accessed May 24, 2012).
15. Wikipedia articles describing maildir and mbox., http://en.wikipedia.org/wiki/Maildir and http://en.wikipedia.org/wiki/Mbox. Note that this paper refers specifically to the .mbox extension, the standard Berkeley mbox implementation of this format. For another discussion of the role of mbox in email preservation, see Christopher J. Prom, "Preserving Email," DPC Technology Watch Report 11-01 (December 2011), http://dx.doi.org/10.7207/twr11-01. (accessed May 23, 2012).

16. 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”.

17. Muse website, http://mobisocial.stanford.edu/muse/ (accessed May 21, 2012).

18. CERP XML format, 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.

19. Offline Imap website, 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.

20. Python script, md2mb.py, available from https://gist.github.com/1709069.

21. Significant characteristics analysis for maildir, http://www.archivematica.org/wiki/index.php?title=Zimbra_to_Maildir_using_OfflineImap for an example of the analysis of significant characteristics. (accessed May 24, 2012).

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

23. Archivematica Development Roadmap, https://www.archivematica.org/wiki/Development_roadmap/ (accessed August 21, 2012).