Building On OAIS: Processing Born-Digital Records Using Archivematica

Courtney C. Mumma; Artefactual Systems, Inc.; Vancouver, British Columbia/Canada

Abstract

In early 2012, the 0.8 alpha version of the Archivematica (AGPL3) open-source digital preservation system was released. This system is based on the ISO-OAIS functional model and is designed to maintain standards-based, long-term access to collections of digital objects. Using Archivematica, the City of Vancouver Archives stored its first production archival information package (AIP). Processing 50 TB of e-records from the Vancouver Organizing Committee for the Olympic and Paralympic Games revealed some limitations of the ISO-OAIS model in the areas of appraisal, arrangement and description. The project resulted in adding requirements intended to fill those gaps to Archivematica's development roadmap for its micro-services architecture and web-based dashboard.

Introduction

Archivists have been aware of the need for digital preservation strategies for many years and the foundation for building these strategies has been the ISO 14721-OAIS Reference Model [1]. Recently, strategies have developed to the point that they can be practically tested on real records. One such strategy is to use the Archivematica suite of tools, which was conceived as an implementation of the OAIS model and designed based on an extensive requirements analysis [2].

While it serves as an excellent foundation and framework for long-term preservation strategies, the OAIS model proves inadequate to address functions unique to archives. Practical application of Archivematica 0.8 alpha to process the records of the Vancouver Organizing Committee for the Olympic and Paralympic Games (VANOC) was an opportunity to identify and fill OAIS requirement gaps for digital archives systems.

About Archivematica

The Archivematica system uses a micro-services design pattern to provide an integrated suite of software tools that allows users to process digital objects from ingest to access in compliance with the OAIS model [3]. Users monitor and control the micro-services via a web-based dashboard.

Archivematica uses METS, PREMIS, Dublin Core and other recognized metadata standards. It implements media type preservation plans based on an analysis of the significant characteristics of file formats. To support migration and emulation preservation strategies, the original format of all ingested files is maintained.

The primary preservation strategy is to normalize files to preservation and access formats upon ingest. 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. 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.

Processing the VANOC Records

The first large-scale test of the Archivematica system was transferring and processing the VANOC records. In the course of this work, the team discovered gaps in the OAIS model and compiled new requirements to allow for appraisal, arrangement and description functions.

While a digital archivist at the City of Vancouver Archives, Courtney Mumma was responsible for managing the acquisition of the hybrid analog/born-digital records of the 2010 Winter Games. Partnered with Artefactual Systems, Inc., Archives staff had begun to gather requirements for their digital preservation system in 2008. When Mumma joined them in 2009, baseline requirements had been established.

The Archives embraced Artefactual's open-source, agile software development methodology and continued working with them to build their digital preservation system around Archivematica. Meanwhile, Mumma negotiated the donor agreement between VANOC and the City for over a year. The resulting transfer of records totalled over 200 boxes and 25 terabytes of multi-format digital records on external drives and optical media. The mass of digital records was such that the archivist sustained a repetitive stress injury while copying the recordds for safekeeping while Archivematica development continued. There is a detailed account of the acquisition in a recent Archivaria article by Mumma and other City staff [4].

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 [5] to conduct a gap analysis between OAIS and the InterPARES 1 Project's Chain of Preservation (COP) Model [6]. Review of the model, along with consultations with archivists about processing analog 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. The team of archivists and Artefactual staff 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 are discussed in detail in a recent Archivaria article

[7], 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 analog 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.

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 0.8 alpha 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 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.

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 object's 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 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 analog archives.

Arrangement and Description

Like appraisal, arrangement and description do not occur in a vacuum. Archivists arrange and describe analog records intermittently while they process a fonds. 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, along with additional descriptive information gathered about the records during processing, this will ultimately end up in 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. Analog 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. In Archivematica, a SIP has to be manually compiled from the transfer or transfers by the archivist in the file browser. After transfer micro-services are completed successfully, 0.8 alpha allows transfers to be arranged into one 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 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.

Digital Forensics Tools

Obviously, there are limitations to the level of analysis possible forming SIPs using only the file browser. Transfers may contain restricted material, passwords, personal information or other content that is unsuitable for continued preservation. For insight into this problem, the team looked to the field of digital forensics. 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. Archives are held to the same standards of authenticity, so there is much to be learned from the digital forensics field. For over thirty years, they have developed 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 as the team realized that they could not possibly decipher the entire canon of digital forensics software in time to get the VANOC records processed, digital humanities scholars and archivists in the United States were conceptualizing the BitCurator Project. From the BitCurator website [9]: "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 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.

Next Steps

There was simply not time to learn and practice using all the forensics tools that may help fill in the gaps between practical archival processing and the OAIS model in time to meet the deadline for processing the first part of the VANOC acquisition. The Archivematica team gathered appraisal, arrangement and description requirements and drew up workflows that would account for real-world archival processing needs. Since open source digital forensics tools were not ready to be integrated into the Archivematica suite for various reasons, the team looked for other tools to provide the necessary services to satisfy their workflows.

The University of North Carolina Libraries, for instance, had just developed Curator's Workbench [10], 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 ongoing support, they opted instead to mimic its arrangement functionality. Archivematica's 0.8 alpha release uses METS and the Xubuntu file browser Thunar to arrange records and keep a record of the original order within each SIP formed from a transfer.

Incorporating new tools and functionality happens iteratively within the agile development model. Because of discoveries made while processing the VANOC acquisition, several new features are included on the Archivematica 0.9 and 1.0 beta release development roadmaps [11]. In 0.9 there will be a file management interface in the dashboard to replace using the browser in the filesystem. This will allow users to more easily interface with transfers while forming them into SIPs.

Additionally, 0.9 will include indexing and visualization of transfer content so that archivists will have richer information upon which to base their appraisal, arrangement and description decisions. An example would be a pie chart illustrating the distribution of file types in a transfer. Thorough indexing could allow for many different kinds of analysis in future releases. For instance, archivists could do keyword searches to identify records related to a particular author or subject. Keyword searching might also help to identify restricted records and personal information.

Release 1.0 will include, among many other things, richer archival description functionality that includes file-level metadata entry and rights management. Additionally, it will support a distributed processing infrastructure to avoid scalability issues. Processing the VANOC acquisition revealed that sometimes, transfers could total up to several terabytes of objects. Such scale makes it difficult to arrange and describe SIPs unless there is capacity to keep all of the transfer content accessible to the browser at once. The City of Vancouver's deployment of Archivematica is currently a bare metal installation on a local area network (LAN), which causes lag time and dropped processes. Future deployments will likely be run from a Virtual Server.

Conclusion

Future releases will continue to develop new requirements as Archivematica is adopted by new users who will, no doubt, uncover limitations or suggest enhancements. Meanwhile, the archives and library communities are increasingly developing new open source tools that do some portion of the digital preservation workflow. One advantage of the agile micro-services development model is that these new tools can be evaluated and adopted by the Archivematica suite very quickly.

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 March 21, 2012).

[3] Artefactual Systems, Inc., Archivematica homepage, http://archivematica.org (accessed March 21. 2012).

[4] 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.

[5] InterPARES 3 Project, http://www.interpares.org/ip3/ip3_index.cfm (access March 21, 2012).

[6] InterPARES 2 Project, Chain of Preservation (COP) Model, http://www.interpares.org/ip2/ip2_model_display.cfm?model=cop (accessed March 21, 2012).

[7] Mumma, Dingwall and Bigelow (2011).

[8] "Technological context." InterPARES 1, Glossary. The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project (2001).

[9] BitCurator Tools for Digital Forensics Methods and Workflows in Real-World Collecting Institutions, http://www.bitcurator.net/ (accessed March 21, 2012).

[10] 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 March 21, 2012).

[11] Artefactual Systems, Inc., "Archivematica Development Roadmap", http://archivematica.org/wiki/index.php?title=Development_roadmap (accessed March 21, 2012).

Author Biography

Courtney Mumma holds Masters degrees in Archival Studies and Library and Information Studies from the University of British Columbia (2009) and is the Archivematica Community Manager at Artefactual Systems, Inc. Formerly a digital archivist at the City of Vancouver Archives, she managed the acquisition of the records of the Vancouver 2010 Winter Games, a hybrid fonds containing a large portion of born-digital records,. In the course of that work, she completed a requirements analysis for a digital archives system in the City which resulted in a partnership with Artefactual Systems to develop and migrate to Archivematica and ICA-AtoM. Courtney has contributed research to the InterPARES 2 project, the UBC Digital Records Forensics project and is currently on the Professional Experts Panel of the BitCurator project.