Archivematica: Lowering the Barrier to Best Practice Digital Preservation

Peter Van Garderen. Artefactual Systems Inc. Vancouver, Canada.

Abstract

In the last few years much progress has been made in digital preservation theory and technology. Yet for many small to medium sized archives and memory institutions there are still significant obstacles to implementing best practice solutions to meet the digital continuity requirements of their user communities. Three critical barriers are intellectual complexity, technical complexity and cost. Inspired by a call to action in a recent UNESCO Memory of the World report [1], the goal of the Archivematica project is to lower each of these barriers and give archivists the tools, methodology and confidence to begin preserving digital information today. Archivematica integrates a number of open-source tools to create a comprehensive digital archives system that is based on a practical analysis of the OAIS functional model. The system is free and full access is given to the source code repository and technical documentation under the open-source GPL license.

Reducing Intellectual Complexity

The Archivematica project wiki documents a detailed use case and workflow analysis of the OAIS functional model. [2] This analysis has resulted in a set of granular UML activity diagrams that help to demystify OAIS and define explicitly what basic actions must be carried out to be compliant with its functional model. The system also automates the implementation of best practice metadata standards including PREMIS, METS, Dublin Core, EAD and ISAD(G). Archivematica applies media type preservation plans based on empirical tests of significant properties and a practical assessment of currently available free and open-source normalization tools. The OAIS analysis, end-user documentation, metadata standard implementation and fully-documented preservation plans help to reduce the intellectual complexity of interpreting and applying these standards and best practices. Full access is given to this documentation under Creative Commons license to allow institutions to customize these resources to suit their own organization-specific requirements.

Reducing Technical Complexity

The Archivematica system helps to tackle technical complexity by providing a set of fully integrated tools and workflow instructions. Archivematica is available from a single install script while deployments can range from running the system on a simple USB key to threading processes over a multi-node architecture. Future plans include making the system available from standard Linux code repositories to allow installation of the system with a single 'apt-get' command.

Archivematica is an implementation of the micro-services approach which is a light-weight alternative to repository and framework based solutions. [3] Instead of relying on a repository interface to a digital object store, this approach uses loosely-coupled tools to provide granular and orthogonal digital preservation services built around file-system storage. This reduces technical complexity for development and maintenance but is also relevant as a long-term preservation strategy because it provides archivists with direct, unmediated access to archival storage. Furthermore, file system technology is long-proven and extremely robust, typically outlasting the lifespan of enterprise information systems. On top of the file system interface Archivematica integrates well established Unix utilities and Python modules with best practice digital preservation tools and web-based archival description and access software. The integration code is written as Python scripts. As an interpretive language, which is proven in large-scale integration scenarios [4], it supports easy customization and an agile development methodology that allows for testing changes in real-time while still maintaining code integrity through the use of standard code versioning and issue tracking tools (i.e. Subversion and Googlecode project hosting). [5]

From Ingest to Access

Archivematica provides a template to create Submission Information Package (SIP) profiles based on qualified Dublin Core and METS elements. However, the system will accept files for ingest with as much or as little metadata as is available. It runs the SIP through a series of ingest processes including unpacking, checksum verification and creation, unique identification, quarantine, format identification, format validation, metadata extraction and normalization. A variety of tools are used in each of these processes, including Easy Extract, Detox, UUID, CLAM AV, Thunar, Incron, Flock, JHOVE, DROID, NLNZ Metadata Extractor, File, FFident, File Information Tool Set (FITS), Xena, OpenOffice, Unoconv, FFmpeg, ImageMagick, and Inkscape. The web-based Archivematica Dashboard monitors the progress of each SIP, logs the results of each process, reports on any errors and prompts the archivist to trigger subsequent processes.

Archivematica maintains the original format of all ingested files to support migration and emulation preservation strategies. However, the primary preservation strategy is to normalize files to preservation and access formats upon ingest. Archivematica assigns each file format to a media type preservation plan (e.g. text, audio, video, raster image, vector image, etc.). Archivematica's preservation formats must all be open standards; additionally, the choice of formats is based on community best practices, availability of free and open-source normalization tools,

and an analysis of the significant properties for each media type. The choice of access formats is based largely on the ubiquity of web-based viewers for the file format.

Archivematica packages Archival Information Packages (AIPs) using qualified Dublin Core, PREMIS and METS elements and Library of Congress' Bagit format. It then prepares default Dissemination Information Packages (DIP) which are based on the designated access formats for each media type. Consumers can subsequently request AIP copies but caching access copies is a much more scalable approach that will address the majority of access requests in the most performant manner (i.e. reducing the bandwidth and time required to retrieve AIPs from archival storage and uploading them to the Consumer). The DIP access derivatives are sent via a REST interface to a web-based application such as ICA-AtoM or Archon for further enhancement of descriptive metadata (using ISAD(G), EAD, DACS, etc). These can then be arranged as accruals into existing archival descriptions to provide search and browse access to the institution's analogue and digital holdings from one common web-based interface. The Archivematica Dashboard manages the read and write operations of the AIP to file storage and also coordinates the syncing of metadata updates between the AIPs and the access system.

Reducing Cost

The primary costs for digital preservation are program and human resource management, systems analysis, hardware and storage, software licensing and system maintenance. While it will always cost money to implement and maintain technology solutions, these costs can be greatly reduced and better managed through the open-source model. In particular, open-source software eliminates software licensing costs and provides a cost-effective way to manage system maintenance expenses by freely sharing technical knowledge and documentation, providing direct access to core developers for technical support and feedback, and eliminating the need for maintenance contracts to implement release upgrades. At the same time, mature open-source communities are supported by third-party solution providers that can provide optional customization, help-desk, hosting and even service level agreements for those institutions that lack the capacity to implement or support their own digital preservation systems. Archivematica's software development has been led thus far by Artefactual Systems, a contractor based in Vancouver, Canada that provides open-source software solutions and consulting services for archives and memory institutions. Artefactual is also the lead developer of the International Council on Archives' ICA-AtoM software project.

The Archivematica project implements an agile software development model that is focused on rapid release cycles and iterative, granular updates to the requirements documentation, software code and user documentation. One of the project assumptions is that all software systems are dynamic and ever-evolving. This is particularly true for a digital preservation system that must respond to changes in the technology that creates digital information as well the technology that is available to manage it. Therefore, one of the key deliverables of the Archivematica project is a software development methodology and infrastructure which is able to evolve and manage rapid change. This methodology is developed dynamically in collaboration with the user community and is freely available for implementation and customization by institutions as a system maintenance tool.

The Archivematica project is structured in a truly open way to encourage a grass-roots, collaborative development model which makes it easy for other institutions and third-party contractors to benefit and contribute. All new software code is released under a GPL license and all the tools integrated into the system are checked for license compatibility. As Richard Stallman, the original author of the GPL, points out, this grants software users four key freedoms [6]:

1. The freedom to run the program, for any purpose
2. The freedom to study how the program works, and adapt it for your own needs (must be supported by easy access to the source code)
3. The freedom to redistribute copies to help friends, family, colleagues or society in general
4. The freedom to improve the program, and release your own improvements to the public, so that the whole community benefits (again, easy access to the source code is a precondition for this).

No license fees, membership dues or account registration is required for downloading Archivematica or checking out the source code from the public Subversion repository. Full documentation is provided on how to build the application from source. The community is encouraged to update the issues list and wiki pages and to join the discussion list and weekly development meetings in the online chat room.

The open-source development model encourages users to pool their technology budgets as well as attract external funding to develop core application features. This means the community pays only once to have features developed, either by in-house technical staff or by third-party contractors. This new functionality can then be offered at no cost in perpetuity to the rest of the user community. This stands in contrast to a development model driven by a commercial vendor, where institutions share their own expertise to painstakingly co-develop digital preservation technology but then cannot share that technology with their colleagues or professional communities because of expensive and restrictive software licenses imposed by the vendor.

The Archivematica project is only a year old but already the UNESCO Memory of the World Subcommittee on Technology has provided external funding to contribute to its core development, while both the City of Vancouver Archives and the International Monetary Fund Archives have sponsored the development of new features by deploying the system as part of their own internal proof-of-concept projects and contributing new code back to the project under GPL licenses.

Archivematica is still in the initial stages of development, having been made available as an alpha release earlier this year. However, by the end of 2010 a beta version will be implemented in

production pilots by collaborating institutions. Throughout this time period, the system's development will continue to be heavily influenced by the day to day feedback of its community. Like any newly launched open-source project, Archivematica is looking to grow its network of implementation institutions, end-users, developers, solution providers, and funding sponsors. If you think that the Archivematica technology and methodology is a good fit for your institution then we encourage you to get involved in the project. You can download the application and source code or simply get started by posting questions in the discussion list, dropping in on the developers chat room or contacting the project leads directly.

References

[1] Kevin Bradley, Junran Lei, Chris Blackall. "Towards Open Source Archival Repository and Preservation System" (2007). www.unesco.org/webworld/en/mow-open-source/

[2] http://archivematica.org

[3] See for example California Digital Library Curation Micro-Services, http://www.cdlib.org/services/uc3/curation/

[4] http://www.python.org/about/quotes/

[5] http://archivematica.googlecode.com

[6] http://www.gnu.org/philosophy/free-sw.html

Author Biography

Peter Van Garderen is the Archivematica project manager and principal of Artefactual Systems, a Canadian company providing open-source technology services for the international archival community, including development of the International Council on Archives' ICA-AtoM archival description software. Mr. Van Garderen is a graduate of the University of British Columbia's Master of Archival Studies and Software Engineering programs. He is also a Doctoral Candidate in Archival Science at the University of Amsterdam.