Realizing the Archivematica vision: delivering a comprehensive and free OAIS implementation

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 Analyst, Artefactual Systems Inc.
202-26 Lorne Mews
New Westminster, BC, Canada
1.604.527.2056
courtney@artefactual.com

ABSTRACT
Archivematica began in 2008 as a working hypothesis that assumed a comprehensive yet free digital preservation system could be created by matching existing open-source software tools against the OAIS functional model. Five years later the production release of the software is ready to go into production at several major North American archives and libraries, while the beta version is already widely-deployed worldwide. In the absence of a single major funding sponsor, the project management team worked as third-party contractors to several early-implementer institutions that shared the project's architectural and open-source vision while needing to implement an effective and sustainable digital curation solution for the digital content entrusted to their care. From the outset, Archivematica's system requirements were based on an ongoing dialogue within the digital curation community about the gaps between the standards and strategies that were held up as best practice (OAIS, PREMIS, normalization, agile development and so forth) and the ability for the average archivists and librarians to implement them. The iPres conference has proven to be a critical forum for advancing this dialogue and has included papers about the Archivematica micro-services architecture [1] and community-driven development approach [2]. This paper will provide a conclusion to these earlier papers by discussing the key architectural, digital curation and sustainability challenges that the Archivematica project has addressed as it emerged from a working prototype to a full-featured digital preservation system. This includes system scalability, customization, digital repository interfaces, format policy implementation, and a business plan that stays true to the ideals of the free software community.

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

Keywords
archivematica, digital preservation, archives, OAIS, migration, formats, PREMIS, METS, agile development, open-source

1. HISTORY
In 2007, the UNESCO Memory of the World Subcommittee on Technology report entitled "Towards an Open Source Repository and Preservation System" concluded that "for simple digital objects, the solution to digital preservation is relatively well understood, and that what is needed are affordable tools, technology and training in using those systems...A practical open source system for digital preservation could, with a little work, be constructed and...this would be of enormous benefit to communities and institutions all over the world." [3] On the report's recommendation, UNESCO offered its support to fund the beginnings of what would become Archivematica, a system that would make it possible and easy to implement in one system what had until then been disparate advances in open source tools for digital preservation. With a trusted digital repository system, memory institutions could preserve the authenticity of their valuable digital records over time.

Nearly concurrently, the City of Vancouver Archives had reached a similar conclusion to that of the UNESCO report. In 2008, the City of Vancouver Archives contracted Artefactual Systems to design and develop a comprehensive digital preservation system that implements the ISO 14721 Open Archival Information System reference model [4]. In 2009, the International Monetary Fund (IMF) Archives also contracted Artefactual Systems to develop a proof-of-concept system based on the work that was being done at the City of Vancouver Archives.

In 2009, the Archivematica project and its partners first translated the OAIS functional model into use case scenarios [5], subsequently developing working prototypes demonstrating implementation of these scenarios. In 2010, Peter Van Garderen introduced Archivematica to the international digital preservation community via his iPres paper, "ARCHIVEMATICA: Using Micro-Services and Open-Source Software to Deliver a Comprehensive Digital Curation Solution" [6] and in a paper for the IS&T Archiving proceedings [7].

As the project advanced, more institutional partners offered to fund features, and what began as tools bundled together into a loose workflow reliant upon Python micro-services and active folders in an operating system became a fluid workflow operated via an elegant, web-based dashboard. Then, in 2012, Van Garderen and Courtney Mumma updated the international

community on the agile, open-source development of the project in their iPres paper "The Community-driven evolution of the Archivematica project" [8]. Now, the Archivematica project is nearing its first production release, which includes features and enhancements motivated and steered by Artefactual's core development team, client partners and users in the community at large. Several clients have successfully deployed beta versions for their individual pilot projects, providing rigorous testing and valuable feedback.

From the outset, Archivematica's system requirements were based on an ongoing dialogue within the digital curation community about the gaps between the standards and strategies that were held up as best practice (OAIS, PREMIS, normalization, agile development and so forth) and the ability for the average archivists and librarians to implement them. The iPres conference has proven to be a critical forum for advancing this dialogue. This paper will provide a conclusion to the earlier iPres papers by discussing how the system has come to fruition. As with each prior release, the first production version of Archivematica will include some new features sponsored by client institutions, enhancements and bug fixes as well as features the Artefactual team considers to be essential for a full-production system. These include system scalability, customization, digital repository interfaces, format policy implementation, and a business plan that stays true to the ideals of the free software community.

2. DIGITAL REPOSITORY INTERFACES

Since the beginning of the digital era, institutions have been building their digital capacity by investing resources and training in content management systems, storage infrastructure, and web components. Recognizing the value of this investment and wishing to bolster rather than replace the systems in existence, Archivematica was conceived as a back-end supplement to manage as-yet unaddressed preservation risks. Since its inception, the intent has been to allow for integration of Archivematica with any number of different access and storage systems. A core design principle is to work with existing collections management tools (e.g. ICA=AtoM, CONTENTdm) and storage architectures (e.g. network storage devices, LOCKSS, cloud storage).

Archivematica is intended to fill the digital preservation services gap for existing repository management applications rather than try to replace them or replicate their functionality. For example, work with our pilot project partners lead to integration of the Archivematica processing pipeline with systems like DSpace, CONTENTdm, Fedora, ICA-AtoM (which is developed in tandem and comes packaged with Archivematica), and Archivists' Toolkit. The partner institutions continue to use the same tools they were already using for collections management, cataloging and public access while Archivematica handles digital preservation services and workflows for the digital materials managed by those other systems.

In one example, Archivematica functions as a "dark archive" for DSpace, providing back-end preservation functionality while DSpace remains the user deposit and access system [9]. For this integration, Archivematica added rules for structuring the DSpace export for ingest, enhancements to the METS file, and an OAI harvesting option [10] that allows for automated ingest of updated descriptions in DSpace. In another example, CONTENTdm integration required changes to the METS structMap including user-supplied structMaps that will allow users to set upload and display order based on logical divisions like book chapters. In addition, Archivematica added a variety of CONTENTdm workflow options for DIP creation and upload [11].

During the development of these and other interfaces, it became clear that institutions' instances of identical systems were unique based on their local configurations. Artefactual's first priority is to integrate with the client-specific configurations, but the community ultimately benefits from a more generic feature that can then be tailored to local specifications. For this reason, Archivematica now includes a generic version of the feature developed alongside the sponsoring client. Moreover, with each integration, Archivematica moved closer to application programming interfaces (APIs) [12] for Ingest, Storage and Access systems. These APIs are the way forward for future integration development.

3. CUSTOMIZATION

With each iteration, Archivematica has been developing methods to make it easier for users to customize their workflows within the constraints of the system. While Archivematica will continue to build and enhance features for customization in future releases, much work has already been done to make the system more flexible.

Users can now change the workflow to skip, automate or include decision points for micro-services, adjust compression algorithm and size for the AIP and pre-select a standard AIP storage location. For instance, the user can decide to skip backing up their transfer or to quarantine the contents for 30 days prior to processing in order to allow for updates to virus definitions in the malware checking tool. If users are ingesting digitized objects, they can set Archivematica to automatically approve normalization or to detect access derivatives included in the SIP, thereby reducing processing time for digital object types that have known behaviour in the system.

Should users have a local tool, proprietary or open-source, that proves better suited within their institution to normalize a particular format to its preservation and/or access copy, they can opt to use a new manual normalization feature either at the beginning or in the middle of the Archivematica workflow. Another option in the normalization workflow is to pick from an ever-expanding set of tools which identify file formats as the basis of normalization actions. Additionally, users can choose to send their transfers to a backlog with enough metadata in the METS file and an accession number so that they can be retrieved from storage to ingest at a later date and even by a different user.

4. FORMAT POLICIES

The format problem is one that, despite the noble efforts of information professionals and hobbyists, does not appear to be solvable in the immediate future. However, advances are happening more rapidly than they were even five years ago, with groups like Open Planets Foundation [13] investing their resources into rigorously testing tools for format identification. Early on, Archivematica was developing media type preservation plans, attempting to discern best practice from the available research at the time. Unfortunately, there was not much information about what institutions were choosing as preservation formats--there still is very little, in fact. It was also the case then and is now that the information out there was in various types of unstructured formats (e.g. webpages, pdf).

Archivematica researchers garnered what they could about significant characteristics and best practices from the varied community of information professionals and from institutional policies. After this analysis, they tested open source tools to implement a two-pronged approach to preservation planning: normalization on ingest and the preservation of the original file to support future strategies such as migration and emulation. Normalization is based on *format policies*, which indicate the actions, tools and settings to apply to a file of a particular file format in order to make a preservation or access copy. The criteria for selecting default formats for normalization in Archivematica are that they must be free of licenses and patent restrictions, have freely available specifications, and be widely used and/or endorsed by major repositories. Preservation formats must also allow for no or lossless compression and there must be open source tools readily available to write and render them.

Archivematica analysts have been closely involved in the digital archives and library community monitoring advances in format identification. Conversations and workshops at events like CURATEcamp [14], national and international conferences and online in blogs and on listservs have highlighted the dearth of certainty about best practices, tools and preservation formats. Because format policies will change as formats and community standards, tools and practices evolve, a Format Policy Registry (FPR) [15] emerged as Archivematica's strategy for the treatment of format policies.

One of the primary goals of the FPR is to aggregate empirical information about institutional format policies to better identify community best practices. The FPR provides 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 practitioners. The FPR APIs are designed to share this information with the Archivematica user base as well with other interested communities and projects.The FPR lists all of Archivematica's default format policy rules and provides valuable online statistics about default format policy adoption and customizations amongst Archivematica users. In the past, the Archivematica project managed all format policy documentation on its public wiki; with the FPR, this information is captured in a structured format (SQL/JSON). Subscription to the FPR (fpr.archivematica.org) via the Archivematica dashboard provides users with notifications about new or updated preservation and access format policies, allowing them to make better decisions about normalization and migration strategies for specific format types within their collections. The FPR will evolve to interface with other online registries (such as PRONOM and UDFR) to monitor and evaluate community-wide best practices. Use of the FPR over time will enrich community understanding of format preservation practices and help to reduce the risk of technology obsolescence and incompatibility.

5. SCALABILITY

Early adopters and testers have consistently asked for scalability metrics; however, without many production betas deployed in client repositories, these metrics were slow in coming. To start, the project team tested distribution of services across multiple processors in order to maximize ingest productivity [16]. While such tests were useful, more rigorous, onsite scalability testing was clearly necessary.

For 1.0, the project staff set up a dedicated testing environment with a full matrix of test parameters [17]. The testing environment begins with several virtual machines set up in a hosted environment, where hardware resources can be scaled up and down between tests. Creating different test configurations allows the project to compile operating system (i.e. cpu and memory usage) and Mysql metrics. With these statistics documented on our public wiki, we can provide metrics to users about scalability and make more informed deployment decisions.

Archivematica has also introduced multiple installation scenarios. One option is a single node installation on a very powerful machine with large capacity. A second is multiple node installation where there is one Archivematica pipeline, running on many (potentially virtual) machines. Another option is multiple installations, which run independently (perhaps one per department or workflow) but share archival storage. Finally, Archivematica allows for multiple independent installations, each with separate archival storage.

6. SUSTAINABILITY

The first production release signals to the Archivematica community of users that they can download and use the system as-is to complete their digital curation workflow. Subsequent releases will allow for an enhanced system that can continue to get better over time. More scalability testing will help to optimize production and, if necessary, select tools that perform better to accomplish micro-service tasks. The quality of Archivematica 1.0 was especially important since the open-source development model relies heavily upon community adoption and support.

There are several open-source development funding models, but two of the most pronounced are funding by a foundation or trust and crowdsourced funding led by a third party company or organization. In its beginning, it appeared as though Archivematica would be of the first type, funding largely by UNESCO. However, as the system evolved and Artefactual partnered with the City of Vancouver, it was clear that the project was destined to follow the latter model. Clients partner with Artefactual to fund the development of features that are in turn shared with the community. Distinct benefits of this model are its agility and variety of users and clients. Archivematica deploys agile development by setting release deadlines with a prioritized list of requirements [18], which puts pressure on Artefactual to release as much as we can during each release cycle so the community can evaluate changes and comment on their value. Artefactual makes it easy for the community to contribute via its user forum [19] and public issue management system [20].

The open-source development model encourages users to stretch their investments by pooling their technology budgets. This means the digital preservation community pays only once to have features developed, either by in-house technical staff or by third-party contractors like Artefactual. Archivematica project staff provide free community support and free software release management. All the software and documentation gets released under open-source (AGPL3) license and is 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. Commercial vendors benefit from the knowledge, time and money invested in open-source tools without contributing in-kind, or worse, selling the tools back to digital preservation colleagues in one form or another. The open-source model employed by Artefactual also stands in contrast to the "freemium" style open source business model, in which code is released while documentation or some other deployment necessity is withheld from all but paying partners.

As the community grows and contributes to the system, Artefactual can focus on building up a preferred provider network of trusted service providers. Preferred partners will be those contractors that can demonstrate their ability to provide users with high-quality support, and share Artefactual's open source values--that is, they will provide code completely free (AGPL3 license [21]) as a service to the archives and library community. A widening scope of service providers beyond just Artefactual will allow the Archivematica project to focus on innovation and moving forward for the benefit of all its users.

7. CONCLUSION

Thanks to a growing community of dedicated beta testers and client pilots, the first production version of Archivematica is a full-production digital preservation system, ready to be implemented, integrated with other systems, and developed further by Artefactual and community members. Baseline requirements for system scalability will continue to be stress tested over time to allow for enhancement and improvements as new features are added. User customization options allow for flexibility in repository workflows. Each new digital repository interface will be based upon a generic version and/or API, usable beyond the sponsoring repository. Format policy implementation, while being essential in staying current with preservation planning best practices, can have a broad effect in the larger digital preservation community, allowing for quantifiable data about normalization processes, successes and failures over time. Finally, a business plan that stays true to the ideals of the free software community will allow for new feature development and enhancements over time and nurture system sustainability.

8. REFERENCES

[1] Van Garderen, P. 2010. Archivematica: Using micro-services and open source software to deliver a comprehensive digital curation solution. *iPres Proceedings.* (Sept. 2010), http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/vanGarderen28.pdf

[2] Mumma,C. and Van Garderen, P. 2012. The Community-driven evolution of the Archivematica project. *iPres Proceedings.* (Oct.. 2012), 164-171, https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf

[3] Bradley, K., Lei, J., Blackall, C. Towards An.Open Source Archival Repository and Preservation System (2007), http://www.unesco.org/webworld/en/mow-open-source/

[4] ISO 14721:2003. Space data and information transfer systems -- Open archival information system -- Reference model.

[5] Archivematica public wiki, OAIS use cases. (2009), https://www.archivematica.org/wiki/OAIS_Use_Cases.

[6] Van Garderen, P. 2010. Archivematica: Using micro-services and open source software to deliver a comprehensive digital curation solution. *iPres Proceedings.* (Sept. 2010), http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/vanGarderen28.pdf

[7] Van Garderen, P. 2010. Archivematica: Lowering the Barrier to Best Practice Digital Preservation. *IS&T Archiving proceedings* (May 2010), http://www.imaging.org/IST/store/epub.cfm?abstrid=43770

[8] Mumma,C. and Van Garderen, P. 2012. The Community-driven evolution of the Archivematica project. iPres Proceedings. (Oct.. 2012), 164-171, https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf

[9] Archivematica public wiki, DSpace export and integration. (2012), https://www.archivematica.org/wiki/DSpace_integration, https://www.archivematica.org/wiki/DSpace_exports

[10] Open Archives Initiative. http://www.openarchives.org/

[11] Archivematica public wiki, CONTENTdm integration. (2012-2013), https://www.archivematica.org/wiki/CONTENTdm_integration

[12] Application Programming Interface (API). *Wikipedia*. (accessed April 20, 2013), http://en.wikipedia.org/wiki/Application_programming_interface

[13] Open Planets Foundation. http://www.openplanetsfoundation.org/

[14] CURATEcamp public wiki. http://curatecamp.org/

[15] Archivematica public wiki. Format policy registry requirements. (2012-2013), https://www.archivematica.org/wiki/Format_policy_registry_requirements

[16] Archivematica. Video of multiple processors. *Youtube.* (2012) https://www.youtube.com/watch?feature=player_embedded&v=lOZ-Kcw4DQs.

[17] Archivematica public wiki.Scalability testing documentation. (2011-2013), https://www.archivematica.org/wiki/Scalability_testing

[18] Archivematica public wiki. Development roadmap. https://www.archivematica.org/wiki/Development_roadmap:_Archivematica_1.0

[19] Archivematica user forum. https"//groups.google.com/forum/?fromgroups#!forum/archivematica

[20] Archivematica public issues list. https://projects.artefactual.com/issues/

[21] AGPL3 license. http://en.wikipedia.org/wiki/Affero_General_Public_License