Archivematica As a Service: COPPUL's Shared Digital Preservation Platform

Bronwen Sprout
Digital Programs and Services, University of British Columbia Library
bronwen.sprout@ubc.ca

Mark Jordan
Library Systems, WAC Bennett Library,
Simon Fraser University
mjordan@sfu.ca

Abstract:
The Council of Prairie and Pacific University Libraries (COPPUL) is piloting a cloud-based preservation service using the Archivematica digital preservation system. The service is offered to COPPUL member institutions that wish to preserve digital holdings but prefer a hosted service to installing and managing local Archivematica instances. This service is a joint effort of COPPUL, Artefactual Systems (Archivematica lead developers), and University of British Columbia (UBC) Library (the cloud storage provider). COPPUL is responsible for promoting the service, signing up new institutions and seeding the one-time set- up costs. Artefactual Systems provides account administration, installation, server administration, and user technical support. UBC Library provides fee-based server hosting and digital object storage service. This article discusses COPPUL's Archivematica-as-a-service model generally and covers certain aspects of implementation in greater detail, concluding with a discussion of future directions.

Keywords: digital preservation, Archivematica, COPPUL, library consortia, cloud computing

Introduction

This article discusses a program developed by the Council of Prairie and Pacific University Libraries (COPPUL) to offer cloud-based digital preservation services to its members. The service is offered to COPPUL member institutions that wish to preserve digital holdings but prefer a hosted service to installing and managing local Archivematica instances. Archivematica is a free and open-source digital preservation system that is designed to maintain standards-based, long- term access to collections of digital objects. This COPPUL service (which is known as "Archivematica as a service") demonstrates many of the benefits of a community-based digital preservation model as well as some of the preconditions necessary for its success, including institutional cloud-based computing, experience with an open source vendor, a history of working cooperatively, and trust within the membership.

After introducing COPPUL and highlighting previous shared services of its Digital Preservation Working Group (DPWG), the article discusses Archivematica as a service, including COPPUL's rationale for the offering, details of the service, support and communication mechanisms, and a description of Archivematica and how it was chosen to fill digital preservation needs. The article con- siders the benefits for both the provider and the service users and concludes with a discussion of future directions and challenges for the service, including issues of sustainability, scale, and shared governance.

About COPPUL and the DPWG

COPPUL 'provides leadership in the development of collaborative solutions addressing the academic information resource needs, the staffing development needs, and the preservation needs of its member institutions' (COPPUL n.d.a). The consortium comprises twenty-three university libraries located in Manitoba, Saskatchewan, Alberta, and British Columbia, and fifteen affiliate members that participate in licenses for electronic resources, discounted pricing, and favourable terms on licensed resources. Beyond consortial licenses to resources, benefits for COPPUL members include networking and information sharing for directors and staff, shared expertise to advance collaborative projects, workshops and continuing education for staff at member libraries, and the opportunity to participate in working groups, including the DPWG. In its 2010–15 strategic directions framework, in addition to identifying a role for COPPUL as a research and development incubator, digital and electronic collections are identified as one of the three main areas of focus, and digitization and digital preservation are further identified as areas on which COPPUL members will work on collectively (COPPUL n.d.b).

The DPWG is one of several working groups that are active within COPPUL. Other working groups focus on scholarly communications, research data, collections, and return on investment. COPPUL also offers several programs, of which Archivematica as a service has perhaps most in common with the Shared Print Archive Network (SPAN), a distributed retrospective print repository program (COPPUL n.d.c). Twenty-one COPPUL libraries participate in this program to preserve an optimal number of printed journals and provide access to shared print archives. In many ways, including a level of comfort with cost sharing, a commitment to working together, and a shared leadership model, this history of collaborating to preserve print archives has paved the way for COPPUL to also enable its members to work together to preserve their digital materials.

As noted in its statement of purpose, the work of the DPWG is ''informed by significant developments in digital preservation currently underway in the memory institution community.'' Various COPPUL and DPWG members are engaged in related digital preservation efforts including participation in several Private LOCKSS Networks (PLNs), participation in the Global LOCKSS Network and/or Portico, local implementations and use of Archivematica, use of Archive-It to archive websites, and digital preservation policy development. Within this context, the DPWG was tasked with, among other things, developing options for a common approach to digital preservation for COPPUL libraries, with a particular focus on solutions that require consortial-level, or inter-institutional, cooperation for their effectiveness.

Prior to the DPWG (which was formed mid-2012), many of the members had experience working together to subscribe to, or develop, other shared digital preservation services. The most relevant example of this is the COPPUL PLN, which was established in December 2007, when the directors of COPPUL libraries agreed to support a two-year pilot project. The mission of the COPPUL PLN is to ''preserve digital collections of local interest to COPPUL members that are not being preserved elsewhere, other than local backup'' (COPPUL Digital Preservation Working Group n.d.). All locally created collections that are at risk of being lost are candidates for the network, and material that has been contributed includes open access journals, most created using the Open Journal Systems platform, as well as born digital government publications, theses and dissertations, and locally digitized materials. Any COPPUL full member institutions that are able to meet the COPPUL PLN membership requirements are eligible to participate as members in the COPPUL PLN. Although there must be a minimum of seven members for the COPPUL PLN to function, there is no maximum number of members, and since its inception there have been at least nine participating nodes in the COPPUL PLN. The operations

of the PLN were originally overseen by a Steering Committee and supported by a technical committee, but since the creation of the DPWG in 2012, this oversight has become a function of the DPWG.

Archivematica as a service

The examples of the SPAN and the PLN illustrate some of the experience COPPUL members share in working together to develop preservation-related programs of mutual benefit. However, although the need for digital preservation and the benefits of shared services are well understood by DPWG members, the services to support these services are not consistently in place at the local level, particularly in some of the smaller institutions. The COPPUL PLN provides redundant storage for certain types of collections, but other features and functions of preservation were missing, including format migration, preservation metadata, and more. To address the gap between preservation storage and full preservation services, some members were using the Archivematica digital preservation software, and the DPWG had formed a working group to discuss testing results and the experiences and workflows of a few local production implementations. Several other COPPUL institutions were interested in implementing Archivematica but did not have the local infrastructure in place to support it.

Although members shared a broad level of knowledge and interest in the Archivematica software generally,[1] interest in Archivematica as a service was first generated at a digital preservation workshop organized by the DPWG and held in Vancouver in March 2013. Library directors were asked to participate in the workshop along with one staff member from each institution. Many of the presentations and much of the discussion concerned members' use of Archivematica digital preservation software. Directors saw the potential for a shared service, and the DPWG was asked to follow up with a proposal. This proposal was presented to the directors at their September 2013 meeting, and after a funding model was agreed upon, several members of the DPWG began working with Artefactual Systems, the lead developers of Archivematica, to operationalize the service.

In developing the service to provide hosted instances of Archivematica, it was envisioned that hosting and digital object storage would be provided by one or more COPPUL institutions. Artefactual had identified potential commercial cloud hosts, but few Canadian options existed (a necessity for potential users in terms of privacy regulations, especially around the storage of potentially sensitive archival data). Coincidentally, the University of British Columbia's (UBC) information technology (IT) group had just launched a new cloud hosting service, branded as EduCloud, which seemed promising. The EduCloud service is a cloud-computing service based on the UBC Vancouver campus that allows self-management from a web portal and self-deployment from templates. Importantly, for BC clients, this service meets BC provincial privacy requirements under the Freedom of Information and Protection of Privacy Act.[2] In addition, it offers the benefits of a virtual server service such as server consolidation, resource pooling, high service availability, and regular backups. Multiple consumption models are available, ranging from capacity-as-you-go to reserved pools (UBCIT n.d.). EduCloud

appeared to meet the needs of the Archivematica service offering, and UBC Library offered to act as a liaison between UBCIT and COPPUL/Artefactual.

Artefactual chose EduCloud partly because high-volume discounts on EduCloud's virtual machine (VM) platform license meant that its prices were highly competitive with commercial providers. A more important factor was that all of the parties felt that partnering with UBC added a level of accountability to the service that would be missing if Artefactual, a private company, had developed its own private branded cloud service using a third-party commercial cloud provider such as Amazon Web Services. This is particularly true in the case of EduCloud, whose goal is to provide low-cost computing and storage infrastructure to scholarly institutions. This goal meshes well with COPPUL's support for university libraries and Artefactual's mission to provide open-source software to the heritage community.

With UBC identified as the cloud service provider, the roles of the three parties involved were defined: the service would be a joint effort of COPPUL, Artefactual Systems (the Archivematica lead developers and support providers), and UBC, the cloud storage provider. Responsibilities for the service have been divided along functional lines: COPPUL is responsible for promoting the service, signing up new institutions, and subsidizing Archivematica technical support; Artefactual Systems provides account administration, installation, server administration and user technical support, and end-user training with each significant upgrade; and UBC provides fee-based server hosting and digital object storage service. Artefactual deploys and manages the VMs on EduCloud with UBC Library acting as a liaison. Individual VMs for member institutions are installed on EduCloud and could be moved in-house if they decide to withdraw from the COPPUL service.

The proposal was structured to include options for members who wanted full preservation and access services and those who wanted to gain experience with Archivematica while making less of a financial commitment. The service and the fee structure were designed based on a tiered model, with a range of storage, functionality, and support available. Three different service levels were offered to member institutions (see Table 1).

Participating institutions derive substantial benefit from the service, including the ability to use an existing digital preservation platform; training and technical support services from experienced Archivematica developers and digital preservation specialists; centralized system administration at a much lower cost than paying for a local system administrator; and annual maintenance and software upgrades subsidized by COPPUL. Participating institutions also benefit by being part of an existing community of Archivematica users. The DPWG as a whole benefits from the added dimension that the service offers its preservation discussions, and, indeed, the model of the Archivematica service suggests a possible future for a shared preservation network. Since Archivematica is open source, all users also benefit from a larger community of clients and non-

Table 1: Archivematica as a service levels.

Bronze Service Level	Silver Service Level	Gold Service Level
Basic ingest and storage management:	Full ingest micro-services:	Full ingest micro-services, plus DIP upload to AtoM and full AtoM support:
• assign universal unique identifier to each object	• all services provided in tier 1	• all services provided in tiers 1 and 2
• transfer digital holdings into Archivematica	• prepare Submission Information Packages	• generate DIP (access copies)
• assign UUID (universal unique identifier to each object	• assign descriptive metadata	• upload DIP to AtoM
• calculate checksums	• normalize (generate preservation copies)	• display digital objects in AtoM
• extract packaged files	• generate PREMIS metadata	• enhance metadata and manage accessions in AtoM
• generate METS file	• generate AIP METS file	• eight CPUs, 48 GB RAM, 2 TB disk space
• scan for viruses	• index METS file	• fifteen support tickets
• clean up filenames (remove prohibited characters)	• package contents in Library of Congress BagIt format	• online Archivematica and AtoM training
• extract technical metadata	• compress AIP	
• identify format	• place AIP in secure storage	
• validate format	• periodically verify checksums of stored AIPs	
• indextransfer	• four CPUs, 32 GB RAM, 1 TB disk space	
• assign rights metadata	• ten support tickets	
• place transfer in secure storage	• online Archivematica training	
• periodically verify checksums of stored transfers		
• two CPUs, 16 GB RAM, 400 GB disk space		
• five support tickets		
• online Archivematica training		

clients alike, who share knowledge including technical support in a public forum—the Archivematica discussion list.

For Artefactual, the benefits of a centrally hosted model mean that the lower-level issues are managed without the individual clients needing to know about them. Artefactual has direct contact with UBC IT and has been able to resolve some issues before the end users are aware of them. Further, a level of standardization has been achieved that is more difficult to attain when each institution hosts its own server. From a technical perspective, EduCloud provisions a pool of compute resources (central processing unit, memory, and storage) and Artefactual provisions VMs for

each institution, with resources allocated as defined by the subscriber's service level. Artefactual then deploys Archivematica (and the AtoM archival description software) on these VMs and configures the applications for use by the subscribers.

In terms of application support, when hosted service users have questions about how to use the software, or experience technical problems when using the software, the support process is the same as for other Artefactual clients. Whereas the cost of a support contract from Artefactual may be too high for smaller institutions to take on by themselves, the COPPUL hosting service levels the playing field for these users.

Technological infrastructure

As stated earlier, the service is hosted on the UBC's EduCloud server platform. OVH, another cloud infrastructure provider, was selected as a backup host. An important goal in developing the service was to allow for the hosting of Archivematica on a variety of cloud platform providers.

After the initial research and selection of EduCloud, work started on building the infrastructure required to deploy and manage what amounts to a private cloud. After successful initial test deployments in EduCloud (with OVH as a backup and test environment), Artefactual developed a suite of deployment tools based on the Ansible automated configuration management system. Prior to the development of these tools, installing Archivematica required a high level of technical expertise and three to four hours of time. The Ansible tools brought the deployment time down to twenty to thirty minutes. Even more importantly, the tools allow all of the configuration information to be documented and easily reproducible. This has numerous benefits, including improved backup and disaster recovery processes and the ability to reproduce perfectly a production site in a test environment to replicate bugs reported by users. It is worth noting that once the COPPUL infrastructure was completed, Artefactual received funding from other institutions to improve and extend the original Ansible tools. Since Archivematica is an open-source project, these tools are being released under the AGPL3 open-source software license for others to use and enhance. In this way, COPPUL supported not only its own member institutions but also the digital preservation community at large.

The resulting infrastructure highlights the advantages of a hosted service over siloed local installations in diverse hardware and software environments. Artefactual systems has easy, standardized access to all of the client installations, which makes error diagnosis and software upgrades much simpler than they are when working with the same number of installations in diverse locations with varying hardware, network, storage, and security infrastructures. Moreover, when multiple institutions pool resources in a virtualized cloud-hosting environment, it is a simple matter to allocate resources based on processing needs. This means that institutions with large numbers of video fi for example, can purchase additional processing power at a relatively low price. A fi and very important advantage of the pooled hosted service is that all of the clients are able to rely

on UBC's IT department to deal with storage, security, backup, and other issues. This can make a tremendous difference to small institutions with limited IT resources.

Experience so far

To date (about half-way through the first year of the service at the time of writing), progress in implementing Archivematica at the subscribing institutions has varied for several reasons. Artefactual has been working with all of the subscribers on training, preservation planning, and operationalization of the service, and, overall, progress has been substantial. During this initial implementation period, several interesting issues have surfaced. First, one institution has been required to satisfy their campus legal staff that preserving organizational records using Archivematica is consistent with the university's privacy policies. Security considerations are an important part of the Open Archival Information System (OAIS) Reference Model and are outlined in Annex F of the ''Magenta Book'' (Consultative Committee for Space Data Systems 2012). These considerations are ''informative'' and not ''normative,'' which means that they simply point out security issues and define what a compliant system must do to address them and do not stipulate a specific technical security model. Archivematica, as a digital preservation system that aims to be compliant with the OAIS functional model, implements specific access controls on content under its purview, but these controls must also be consistent with local policies dealing with security, privacy, and records retention.

Another issue is that implementing Archivematica requires considerable resources that have little to do with a library's ability to provide technological infrastructure. Many of the subscribing institutions do not have comprehensive digital preservation policies or frameworks, and the lack of a digital preservation framework that defines preservation priorities and policies has forced subscribing institutions to spend staff resources addressing these questions early in their implementations of the service. The absence of a comprehensive digital preservation framework before implementing a system such as Archivematica is not necessarily negative. For many sites, implementing a system offers them a concrete opportunity to focus on their priorities and to develop policies around the operational strategies that Archivematica offers.

A third issue, related to the previous one, is that integrating Archivematica with content repositories such as DSpace requires working with their campus' central IT department. In cases where these repository platforms are hosted on behalf of the library by central IT departments, requirements arising from integrating the platforms with external applications such as Archivematica may not have been anticipated when the repositories were implemented or may not be possible given the security policies applied to the local infrastructure. For example, Archivematica can accept exports from DSpace (Artefactual Systems), but taking advantage of this feature requires access to the exported content in ways that many central IT departments may find problematic to configure, especially if they were not anticipated when DSpace was originally provisioned. Archivematica's requirements are not unreasonable or

insecure, but they may pose barriers to implementation in some situations. Nonetheless, at least one subscribing institution has started ingesting DSpace exports in their hosted instance—their access to Archivematica as a hosted service provided them the opportunity to work with their central IT staff to investigate and implement the integration.

From Artefactual's perspective, the experience so far has been positive in several ways. First, COPPUL's service has enabled them to work with a group of clients who may not otherwise have implemented Archivematica on their own because they lacked the local technical infrastructure to do so. Second, in preparation for implementing the COPPUL service, Artefactual developed a suite of deployment tools based on an open-source automated configuration management system called Ansible. Prior to the development of these tools, installing Archivematica required a high level of technical experience and three to four hours of time. The Ansible tools brought the deployment time down to twenty to thirty minutes. Even more importantly, the tools allow all of the configuration information to be documented and easily reproducible. This has numerous benefits, including improved backup and disaster recovery processes and the ability to reproduce perfectly a production site in a test environment to replicate bugs reported by users. This work has allowed Artefactual to develop hosting services with new partners. For example, in August 2014, Artefactual Systems and DuraSpace announced a collaborative service to host Archivematica on the DuraCloud platform (DuraSpace n.d.), which has been branded "Archives Direct" (Archives Direct).

Future directions for the service

In the immediate term, encouraging more COPPUL members to subscribe to Archivematica as a service is a priority. The funding and sustainability models used by COPPUL's Archivematica service, combined with a flexible model for provisioning the necessary server and storage infrastructure to meet demand, will allow the number of subscribers to the service to expand incrementally within the next few years. The DPWG is also exploring the development of additional shared digital preservation services for COPPUL members, modelled after the Archivematica service. These new services may also incorporate aspects of SPAN, where applicable. The most obvious such service would be to transform the current COPPUL PLN so that COPPUL members that do not host nodes in the network can have access to shared storage capacity. Another may be development of a shared service to use the Internet archives' Archive-It service to ensure that institutions that do not subscribe have input into collaborative web archiving initiatives. The cost-sharing and service models developed for Archivematica as a service can serve as a template for these and other shared digital preservation services within COPPUL.

Acknowledgements

The authors would like to thank Evelyn McLellan, Justin Simpson, and Courtney Mumma of Artefactual Systems for their invaluable assistance in the preparation of this article.

Notes

1. Archivematica is a free and open-source digital preservation system that is designed to maintain standards-based, long-term access to collections of digital objects. Archivematica 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 International Organization for Standardization OAIS functional model. Users monitor and control the micro-services via a web-based dashboard (Artefactual Systems).
2. Freedom of Information and Protection of Privacy Act, RSBC 1996, c 165.

References

Consultative Committee for Space Data Systems (CCSDS). 2012. *Reference Model for an Open Archival Information System (OAIS): Recommended Practice CCSDS 650.0-M-2: Magenta Book*. Washington, DC: CCSDS Secretariat. http://public.ccsds.org/publications/archive/650x0m2.pdf.

COPPUL. N.d.a. *About Us*. http://www.coppul.ca/about-us.

COPPUL. N.d.b. *2012–2015 Strategic Directions Framework.* http://www.coppul.ca/ sites/default/files/uploads/StratFramework.pdf.

COPPUL. N.d.c. *Shared Print Archive Network (SPAN)*. http://coppul.ca/programs/ shared-print.

COPPUL Digital Preservation Working Group. N.d. *PLN Subgroup*. http://coppuldpwg.wordpress.com/committees/pln-subgroup/.

DuraSpace. N.d. *DuraSpace and Artefactual Partner to Offer New Hosted Service*. http://duraspace.org/articles/2211.

UBCIT. N.d. *EduCloud Server Service*. http://it.ubc.ca/services/web-servers-storage/ educloud-server-service.