# ARCHIVEMATICA: USING MICRO-SERVICES AND OPEN-SOURCE SOFTWARE TO DELIVER A COMPREHENSIVE DIGITAL CURATION SOLUTION

**Peter Van Garderen**
Artefactual Systems
New Westminster, Canada
http://artefactual.com

## ABSTRACT

Digital curation micro-services offer a light-weight alternative to preservation systems that are developed on digital repository and framework technology stacks. These are often too complex for small and medium-sized memory institutions to deploy and maintain. The Archivematica project has implemented a micro-services approach to develop an integrated suite of free and open-source tools that allows users to process digital objects from ingest to access while applying format specific preservation policies. Inspired by a call to action in a recent UNESCO Memory of the World report, the goal of the Archivematica project is to reduce the cost and technical complexity of deploying a comprehensive, interoperable digital curation solution that is compliant with standards and best practices.[1]

## 1. DIGITAL CURATION MICRO-SERVICES

Instead of relying on a repository interface to a digital object store, the micro-services approach uses loosely-coupled tools to provide granular and orthogonal digital curation services built around file system storage. File system technology is long-proven and extremely robust, typically outlasting the lifespan of enterprise information systems. Making the file system the focal point of micro-services operations is noteworthy as a long-term preservation strategy because it provides archivists with the option of direct, unmediated access to archival storage. This might be necessary one day because the various layers and generations of digital preservation system components are just as susceptible to the risk of technology obsolescence and incompatibility as the digital objects they are attempting to preserve.

Basing services around a basic file system store or interface (e.g. NFS, CIFS) also reduces technical complexity for development and maintenance. As noted by the University of California's Curation Center: "since each service is small and self-contained, they are collectively easier to develop, deploy, maintain, and enhance. Equally as important, since the level of investment in, and concomitantly, commitment to, any given service is small, they are more easily replaced when they have outlived their usefulness." [2]

Each service and tool integrated into the Archivematica system can be swapped for another (e.g. replacing the UUID Linux utility with the NOID application to provide the unique identifier micro-service). As a matter of fact, the entire Archivematica system is disposable from one release to the next. Release upgrades are carried out by completely deleting one disk image containing the operating system and software suite with the newer release. This is possible because Archivematica is essentially a pipeline of services, built on top of a customized Xubuntu Linux distribution, that moves digital information packages through a series of file system directories. Together these steps process digital objects from ingest through to access, leaving the Archival Information Packages (along with backups of system metadata and configuration settings) in the archival storage file system. Cached copies of Dissemination Information Packages are uploaded to a web-based access system when processing is complete. The information packages exist completely independent from the software tools. This highlights the "permanent objects, disposable systems" [3] characteristic that is a distinguishing feature of micro-service based solutions as they have come to be defined over the past year couple of years in a series of articles, architecture documentation, specifications and software tools developed at the University of California Curation Center.[4] Taken together, this substantial body of work has formed the theoretical foundation for digital curation micro-services and has established it as a legitimate alternative to repository-based digital curation systems.

## 2. ARCHIVEMATICA MICRO-SERVICES

While the University of California's micro-services have their origins in the need to provide support services to their campus community, Archivematica's micro-service

[1] Bradley, K., Lei, J., Blackall, C.. Towards Open Source Archival Repository and Preservation System, 2007. http://www.unesco.org/webworld/en/mow-open-source/ (last accessed May 4, 2010)

[2] UC Curation Center / California Digital Library. UC3 Curation Foundations, Rev. 0.13 – 2010-03-25. http://www.cdlib.org/services/uc3/curation/ (last accessed May 4, 2010).

[3] Abrams, S., Cruse, P., Kunze, J., "Permanent Objects, Disposable Systems", *Proceedings of the 4th International Conference on Open Repositories*, Atlanta, U.S.A, 2009.

[4] Curation Micro-Services. http://www.cdlib.org/services/uc3/curation/ (last accessed May 5, 2010).

definitions are based on a detailed use-case and workflow analysis of the OAIS functional model and the business processes of public archival institutions.[5] These were refined through proof-of-concept projects carried out in 2009 and early 2010 at the City of Vancouver Archives and the International Monetary Fund Archives.

This process led to the specification of twenty four micro-services grouped into nine OAIS workflow categories:

| Category | Micro-Service |
|---|---|
| 1. receiveSIP | verifyChecksum |
| 2. reviewSIP | extractPackage<br>assignIdentifier<br>parseManifest<br>cleanFilename |
| 3. quarantineSIP | lockAccess<br>virusCheck |
| 4. appraiseSIP | identifyFormat<br>validateFormat<br>extractMetadata<br>decidePreservationAction |
| 5. prepareAIP | gatherMetadata<br>normalizeFiles<br>createPackage |
| 6. reviewAIP | decideStorageAction |
| 7. storeAIP | writePackage<br>replicatePackage<br>auditFixity<br>readPackage<br>updatePackage |
| 8. provideDIP | uploadPackage<br>updateMetadata |
| 9. monitorPreservation | updatePolicy<br>migrateFormat |

**Table 1.** Archivematica micro-services

Each micro-service is a set of processing steps carried out on a conceptual entity that is equivalent to an OAIS information package: the Submission Information Package (SIP), the Archival Information Package (AIP) and the Dissemination Information Package (DIP).[6]

Used together, the Archivematica micro-services make it possible to fully implement the OAIS functional model, including preservation planning. As the UC Curation Center notes, "Although the scope of any given service is narrowly focused, complex curation function can nevertheless emerge from the strategic combination of individual, atomistic services." [7]

The terminology used to define the Archivematica micro-services specifications does differ from the UC Curation micro-services specifications but there is much overlap in their scope and function. Therefore, the Archivematica project would like to do more work in the coming year to align Archivematica's micro-services more closely with the UC Curation's digital curation specifications and APIs.

## 3. THE ARCHIVEMATICA SOFTWARE

The Archivematica system is packaged as a virtual appliance that bundles a customized Xubuntu Linux operating system with a suite of open-source software tools. Using a virtual machine application (e.g. Sun VirtualBox, VMWare Player), the Archivematica virtual appliance can be run on top of any consumer-grade hardware and operating system. The same disk image used for the virtual appliance can also be used as a bootable USB key, a Live DVD version of the system or for bare-metal installs of networked Archivematica servers and workstations. Current Archivematica development is focused on coordinating these types of networked installations to thread high-volume ingest processes over multiple nodes and thereby scale the system up to support resource-intensive production environments.

The information packages ingested by Archivematica are moved from one micro-service to the next using the Unix pipeline pattern.[8] A Unix pipeline is a well-established system design pattern wherein a set of processes are chained by their standard I/O streams, so that the output of one process feeds directly as input to the next one.[9] In Archivematica this pattern is implemented using Bash and Python scripts together with the Unix *incrond* and *flock* utilities.

Micro-service functionality is provided by one or more of the open-source software utilities and applications bundled in the Archivematica system. Where necessary, these are supplemented by Archivematica integration code written as Python scripts. Python is a proven and preferred language in large-scale integration scenarios.[10] As an interpreted language, 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.[11]

[5] Micro-Services. http://archivematica.org/micro-services. (last accessed May 5, 2010).

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

[7] UC Curation Center / California Digital Library. UC3 Curation Foundations, Rev. 0.13 – 2010-03-25. http://www.cdlib.org/services/uc3/curation/ (last accessed May 4, 2010).

[8] Abrams, S., Cruse, P., Kunze, J., Minor, D. "Curation Micro-services: A Pipeline Metaphor for Repositories", *Proceedings of the 5th International Conference on Open Repositories, Madrid, Spain, 2010.*

[9] Abrams, S., Cruse, P., Kunze, J., Minor, D. "Curation Micro-services: A Pipeline Metaphor for Repositories", *Proceedings of the 5th International Conference on Open Repositories, Madrid, Spain, 2010.*

[10] See, for example, *Quotes about Python* http://www.python.org/about/quotes/ (last accessed May 5, 2010).

[11] The Archivematica Subversion repository and issue tracking list are available at Googlecode Project Hosting, http://archivematica.googlecode.com (last accessed May 5, 2010).

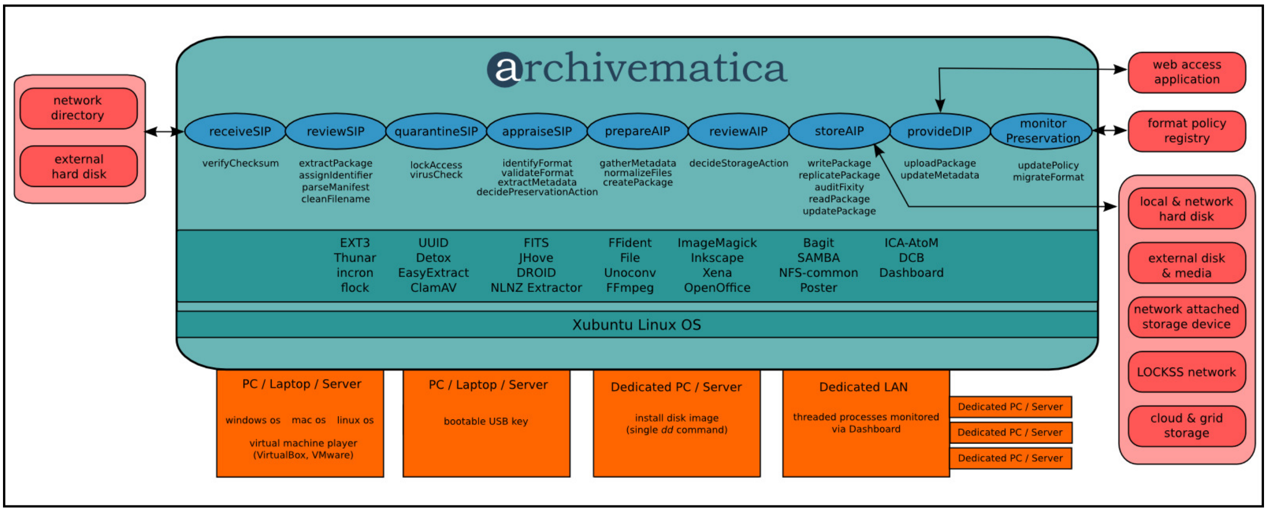

**Figure 1**. Archivematica architecture

The Archivematica 0.6-alpha release was made available in May 2010. It uses the Xubuntu desktop and Thunar file manager as its primary user interface. A web-based 'Dashboard' is currently in development to provide a more sophisticated multi-user interface that will report on the status of system events and make it simpler to control and trigger specific micro-services. The following sections describe the functionality of the current 0.6-alpha release.

### 3.1. Receiving files for Ingest

Archivematica provides a simple text template to create SIP manifests using qualified Dublin Core 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 processing steps 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), OpenOffice, Unoconv, FFmpeg, ImageMagick, and Inkscape.

### 3.2. Format specific preservation plans

Archivematica maintains the original format of all ingested files to support migration and emulation strategies. However, the primary preservation strategy is to normalize files to preservation and access formats upon ingest. Archivematica groups file formats into 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 characteristics for each media type. The choice of access formats is based largely on the ubiquity of web-based viewers for the file format.

Digital format identification registries (e.g PRONOM, UDFR) and conversion testbed services (e.g. Planets) are important components in a global, service-based preservation planning infrastructure. These services must be supported, in turn, by executable digital format policies that can be easily integrated into operational digital curation systems. In other words, after identifying formats, analyzing significant characteristics and evaluating risks we need to make some practical decisions about which preservation formats will ultimately be implemented and how to support these in currently operational systems using the available tools. While the analysis is important and interesting, most small and medium-sized institutions are simply asking for someone to summarize best practices and make it easy to implement them. Therefore, we need a method to document preservation format policies in a structured way to make them easy to implement in a system like Archivematica. Ideally, these policies are also publicly shared to better determine international best practices and enable risk assessment methodologies like those being developed for the Preserv2 file format registry.[12] To date, it has been difficult to analyze community consensus on preservation file formats policies. These are often unreported or scattered about in reports and articles with varying degrees of accessibility.

The Archivematica project publishes its format policies and media-type preservation plans on the project wiki as these are being developed and analyzed. These will be moved to a structured, online format policy registry that brings together format identification information with significant characteristic analysis, risk assessments and normalization tool information to arrive at default preservation format and access format policies for Archivematica. The goal is to make this registry

[12] Tarrant, D., Hitchcock, S., Carr, L., "Where the Semantic Web and Web 2.0 Meet Format Risk Management: P2 Registry" *Proceedings of the Sixth International Conference on Preservation of Digital Objects*, San Francisco, U.S.A., 2009.

interoperable with the upcoming UDFR registry, Planets Registry and tools like the Preserv2 registry. Alternately, if these other tools are extended to support format policy implementation requirements then Archivematica could switch to use them instead. Interoperability will be facilitated by adopting the use of standards such as the eXtensible Characterization Language (XCL) specifications and by providing an RDF interface to the registry.[13] This will also facilitate the sharing of default Archivematica format policies, which might be useful to other projects and institutions.

Archivematica installations will use the registry to update their local, default policies and notify users if there has been a change in the risk status or migration options for these formats, allowing them to trigger a migration process using the available normalization tools. Users are free to determine their own preservation policies, whether based on alternate institutional policies or developed through the use of a formal preservation policy tool like Plato. The system uses a simple digital format policy XML schema that makes it easy to add new normalization tools or customize the default media-type preservation plans.

### 3.3. Preparing files for archival storage

Archivematica creates Archival Information Packages (AIPs) using qualified Dublin Core, PREMIS and METS elements and Library of Congress' Bagit format. As well, support for the TIPR project's Repository eXchange Package specification is currently in development.[14] Archivematica is able to interact with any number of storage systems using standard protocols (NFS, CIFS, HTTP, etc.) to allow for the flexible implementation of an archival storage and backup strategy. Standard operating system utilities can be used to provide backup functionality. Archival storage options range from local hard disk, external hard disks, network attached storage devices and LOCKSS networks (e.g. MetaArchive, COPPUL). Support for storage grids (e.g. iRODS, Bycast) and cloud storage (e.g. Amazon S3, Microsoft Azure) interfaces are also being analyzed. Ideally, the storage platform provides its own fixity check functionality (e.g. Sun ZFS, LOCKSS, iRODS) but for those that do not, a fixity check daemon will be added to Archivematica.

### 3.4. Making files available for access

Archivematica 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, namely by reducing the bandwidth and time required to retrieve AIPs from archival storage and uploading them to the Consumer.[15] The DIP access derivatives are sent via a REST interface to a web-based application such as ICA-AtoM for further enhancement of descriptive metadata (using ISAD(G), Dublin Core, EAD, etc). These can then be arranged as accruals into existing archival descriptions to provide integrated search and browse access to the institution's analogue and digital holdings from one common web-based interface. The Archivematica Dashboard will coordinate the read and write operations of the AIP to file storage and the syncing of metadata updates between the AIPs and the access system.

## 4. SIMPLIFYING DIGITAL CURATION BEST PRACTICES

The project's thorough OAIS use case and process analysis has synthesized the specific, concrete steps that must be carried out to comply with the OAIS functional model from ingest to access. Archivematica assigns each of these steps to micro-services implemented by one or more of the bundled open-source tools. These, in turn, automate the use of digital curation standards (e.g. PREMIS) and best practices (e.g. Bagit). If it is not possible to automate these steps in the current Archivematica release iteration, they are incorporated and documented into a manual procedure to be carried out by the end user.

For example, in early alpha releases of the Archivematica system, some of the workflow controls (e.g. event triggering, error reporting, etc.) are handled via the Thunar file manager (e.g. drag-and-drop, desktop notifications). As the system approaches beta maturity all of the micro-services workflow will be managed and monitored via the web-based Dashboard application. Likewise, as the system matures, each service will be exposed via a command-line and/or REST API.

Focusing on the workflow steps required to complete best practice digital curation functions, instead of the technical components, helps to ensure that the entire set of preservation requirements is being carried out, even in the very early iterations of the system. In other words, the system is conceptualized as an integrated whole of technology, people and procedures, not just a set of software tools.

All software-intensive systems are dynamic, ever-evolving and, arguably, perpetually incomplete. This is particularly true for a digital curation system that must respond to changes in the technology that creates digital information, as well as the technology that is available to manage it. Therefore, the Archivematica project is a working example of the the "disposable system" concept, complemented by an agile software

[13] The Planets XCL Project. http://planetarium.hki.uni-koeln.de/planets_cms/ (last accessed May 5, 2010).

[14] Repository eXchange Package (RXP) specification http://wiki.fcla.edu:8000/TIPR/21

[15] Wright, G., Creighton, T., Stokes, R., "FamilySearch: Extending OAIS for Large-Scale Access and Archiving" *Preservation and Archiving Special Interest Group (PASIG)*, San Francisco, U.S.A., 2009.

development model that is focused on rapid release cycles and iterative, granular updates to the requirements documentation, software code and user documentation.

## 5. USING THE OPEN-SOURCE MODEL TO REDUCE COSTS AND LEVERAGE KNOWLEDGE

Archivematica is still in the initial stages of development, having been made available as an alpha release earlier this year. However, by early 2011 beta versions will be implemented in production pilots at collaborating institutions. Throughout the intervening time period, the systems development will continue to be heavily influenced by the day to day feedback of its community. 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, projects and third-party contractors to benefit and contribute. All of the software, documentation and development infrastructure is available free of charge and released under GPL and Creative Commons licenses to give users the freedom to study, adapt and re-distribute these resources as best suits them.

No software 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 Archivematica virtual appliance from the source code. 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 model 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. It also encourages users to pool their technology budgets and to 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 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.

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.

Mature open-source communities are supported by third-party solution providers that can provide optional installation, customization, help-desk, hosting and service level agreements for those institutions that lack the capacity to implement or support their own digital curation 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. Additional service providers are encouraged to collaborate and contribute to the ongoing development of the Archivematica platform.

## 6. GET INVOLVED

Like any newly launched open-source project, Archivematica is growing its network of implementation institutions, end-users, developers, solution providers, and funding sponsors. If you think that the Archivematica open-source technology, agile development methodology and micro-services conceptual framework is a good fit for your institution, then we encourage you to get involved in the project and help to define its future. 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.