# Exit Strategies and Techniques for Cloud-based Preservation Services

Matthew Addis
Arkivum
R21 Langley Park Way
Chippenham, UK, SN15 1GE
+44 (0) 1249 405060
matthew.addis@arkivum.com

## ABSTRACT

This poster presents an exit strategy for when organisations use cloud-based preservation services. We examine at a practical level what is involved in migrating to or from a cloud-hosted service, either to bring preservation in-house or to move to another service provider. Using work by Arkivum on providing Archivematica as a hosted service, we present how an organisation can use such a hosted service with assurance that they can exit without loss of data or preservation capability. Contractual agreements, data escrow, open source software licensing, use of independent third-party providers, and tested processes and procedures all come into play. These are necessary to mitigate the risks of a wide range of scenarios including vendor failure, service unavailability, change in customer preservation scale or budgets, and migration to or from an in-house approach. There is an existing body of work on how to trust and measure a service that a vendor might offer, for example using audit criteria for Trusted Digital Repositories or measuring service maturity using NDSA preservation levels. However, there has been far less work on how to quickly and safely exit providers of such services - despite this being an essential part of business continuity and disaster recovery. This poster presents some of the considerations and the practical approach taken by Arkivum to this problem including: use of open source software (Archivematica, and ownCloud), data escrow, contracts and handovers, use of vendor independent standards and interfaces (PREMIS, METS, Bagit) and technical migration support, e.g. exports of databases, configurations, software versions and updates. We believe the experience and approach that we have developed will be of use to others when considering either the construction or the use of cloud preservation services.

## Keywords

Exit strategy; exit plan; digital preservation; cloud services; software as a service; escrow; migration; hosted preservation.

## 1. MOTIVATION

Paul Wheatley, Head of Research and Practice at the Digital Preservation Coalition (DPC), presented some of the needs and challenges faced by the DPC membership as part of a talk at PASIG in March 2016 [1]. He articulated that whilst DPC members could see the value of cloud-based preservation services, there were also concerns and barriers to overcome. The top two issues are (a) the need for there to be some form of exit strategy when using a cloud preservation service, and (b) the need for customers of such services to be able to establish trust and perform checks on the quality of the service. Both prevent organisations from adopting preservation services and consequently from achieving the benefits that using these services can offer. This is a problem for the growing number of hosted preservation services, with examples including: Preservica[1], Council of Prairie and Pacific University Libraries (COPPUL)[2], Ontario Council of University Libraries (OCUL)[3], Archivematica hosting and integration with DuraCloud (ArchivesDirect)[4], and Archivematica hosting and integration with Arkivum's archive service in the UK (Arkivum/Perpetua)[5]. Work has been done on the benefits of such cloud services, how to compare and evaluate them, and why exit strategies are important [4][5][6]. These guidelines and comparisons are often based on criteria such as the Data Seal of Approval (DSA)[6], the Trusted Digital Repository standard (ISO16363)[7] or NDSA levels of digital preservation [3]. but don't go into detail on how to implement or verify an exit strategy.

## 2. APPROACH

Arkivum provides Archivematica[8] as a cloud hosted service, which is integrated with Arkivum's data archiving service. The service includes ownCloud[9] to provide easy upload and download of data. Our approach to providing a built-in exit-strategy for the service's users is to support migration from the Archivematica/Arkivum hosted solution to another Archivematica environment, which might be in-house or might be provided by another service provider. The concept of being able to migrate between preservation environments has been investigated by the SHAMAN [2] project amongst others, but we believe full support for migrating between preservation environments has yet to be implemented in a production preservation service. Given that Archivematica is already open source and supports open specifications (METS, PREMIS, Bagit) then we take the simple case of supporting migration between Archivematica instances rather than the general case of migrating to/from an arbitrary preservation environment. This allows the approach to be simpler and most importantly to be directly tested by users of the service. The approach consists of the following elements.

- All data produced by Archivematica (AIPs and DIPs) are stored in the Arkivum bit preservation service, which includes data escrow. Data escrow consists of a full copy of the data stored with an independent third-party without lock-in to Arkivum. If the user exits the service then there is a contractual agreement that allows them to retrieve the

[1] http://preservica.com/

[2] http://www.coppul.ca/archivematica

[3] http://www.ocul.on.ca/node/4316

[4] http://duracloud.org/archivematica

[5] http://arkivum.com/blog/perpetua-digital-preservation/

[6] http://datasealofapproval.org/en/

[7] http://public.ccsds.org/publications/archive/652x0m1.pdf

[8] https://www.archivematica.org/en/

[9] https://owncloud.org/

escrowed data directly from the escrow provider. Data is stored on LTO tape using LTFS as a file system. Data is contained within Bagit containers, which provide a manifest of all data files and their checksums. Each file is optionally encrypted using open standards (RSA and AES) and can be decrypted using open source tools if necessary, e.g. OpenSSL, by the user supplying the keys. Each data file is accompanied by an XML 'sidecar' file that contains metadata on the file, e.g. when it was originally ingested, which encryption keys were used, and the original file name, path and attributes. In this way, the user can retrieve their AIP and DIP files without lock-in.

- Archivematica databases and configuration are exportable from the service and can be downloaded by the user on a regular basis. For example, this includes Archivematica's internal database for storing processing state and AIP/DIP information, webserver configuration (nginx), indexes made of the files processed by Archivematica (elasticsearch). This export allows the user to in effect 'backup' their hosted Archivematica pipeline and storage service. The databases and configurations are snapshotted on a regular basis. This allows the ongoing 'state' of the service to be recorded and replicated into the users' environments.
- Log files are provided of the software versions and updates used in the hosted service, e.g. version of the Archivematica pipeline and storage service, underlying operating system, and peripheral services such as ownCloud. These logs are exported to allow the user to create their own record of the software versions used in the hosted service. This ensures that if the users try to recreate the service then they can do so using the same software versions and hence will be able to import/overlay the database and configuration backups.
- The database and configuration backups along with software version and update logs are all exported through ownCloud. This allows the user to automatically synchronise a local copy of these files into their environment without the need to explicitly remember to download them on a regular basis. Along with the AIP and DIPs stored in data escrow this means that the user has access to both their data and the information needed to take this data and rebuild a working Archivematica system around it.

We are currently working on a simple way for users to do a 'migration test' to verify that the information and data described above is complete and sufficient. Whilst it is easy to assert to a user that everything necessary has been done, the best way to validate this in practice is to perform an actual migration and demonstrate that a working Archivematica instance can be built from the supplied inputs. Arkivum already does this for data escrow through a 'USB key' based escrow test. When using the bit-preservation service the user can specify a test dataset that they want to use for an escrow test. This test data set is 'escrowed' to a USB key and delivered straight to the customer (or via the escrow provider if desired). The user can then validate that the escrowed data is recoverable and is identical to the test data set that they supplied. We are developing a similar approach for Archivematica. The user will be able to set up and use a 'test pipeline' in the hosted service and then ask for this to form the basis of a 'migration test'. The database and configuration etc. for this pipeline will be exported along with the test AIPs and DIPs that it generates. In a similar way to the escrow test, this will be delivered to the user in a self-contained form, e.g. USB key. We aim for this to include a working Archivematica instance configured using the test dataset and exports from the service, for example provided as a bootable drive. In this way, the user will be able to compare and validate that the migration successfully replicates the test pipeline in the hosted service. The test helps provide assurance that the full production pipelines can also be migrated if/when needed. This is important because the production pipelines may contain substantial amounts of data and hence doing actual migration tests of the whole service on a regular basis will typically not be practical.

## 3. CONCLUSION

Hosted preservation services offer many benefits but their adoption can be hampered by concerns over vendor lock-in and inability to migrate away from the service, i.e. lack of exit-plan. We have used Archivematica as a hosted service to investigate what is needed in practice to migrate from the service to an independent Archivematica instance. The approach includes data escrow, export of state information (e.g. databases and configuration), and most importantly a way for users to independently test and verify that migration is possible, i.e. the exit strategy can be successfully executed in practice.

## 4. REFERENCES


[1] Wheatley, Paul. 2016. The DPC Community: Growth, Progress and Future Challenges). Preservation and Archiving Special Interest Group meeting (PASIG), Prague. http://pasig.schk.sk/wordpress/agenda

[2] Watry, Paul. 2007. Digital Preservation Theory and Application: Transcontinental Persistent Archives Testbed Activity. The International Journal of Digital Curation. Issue 2, Volume 2. www.ijdc.net/index.php/ijdc/article/download/43/28

[3] Peltzman, Shira. 2016. Expanding NDSA Levels of Preservation. http://blogs.loc.gov/digitalpreservation/2016/04/expanding-ndsa-levels-of-preservation/

[4] Beagrie, Neil. 2015. Guidance on Cloud Storage and Digital Preservation. How Cloud Storage can address the needs of public archives in the UK. Report from The National Archives. http://www.nationalarchives.gov.uk/documents/CloudStorage-Guidance_March-2015.pdf

[5] AVPreserve. 2014. Cloud Storage Vendor Profiles. https://www.avpreserve.com/papers-and-presentations/cloud-storage-vendor-profiles/

[6] Digital Preservation Handbook, 2nd Edition, Digital Preservation Coalition © 2015 http://www.dpconline.org/advice/preservationhandbook