Preserving Research Data: Linking Repositories and Archivematica

Jenny Mitcham, Julie Allinson
University of York
Heslington, York, UK, YO10 5DD
+ 44 (0) 1904 321170
jenny.mitcham@york.ac.uk
julie.allinson@york.ac.uk

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

Christopher Awre, Richard Green, Simon Wilson
University of Hull
Hull, UK, HU6 7RX
+44 (0) 1482 465441
c.awre@hull.ac.uk
r.green@hull.ac.uk
s.wilson@hull.ac.uk

ABSTRACT

Applying digital preservation actions to research data is a practical step in carrying out research data management to its fullest extent and helps ensure this data remains usable in the future. This paper considers how repositories holding research data can link to an external third party tool, Archivematica, in order to carry out preservation actions as part of the data deposit workflow into the repository. We present experience from local use of Archivematica at the Universities of York and Hull in the Jisc Research Data Spring project "Filling the Digital Preservation Gap" as well as Archivematica as a shared service by Arkivum. A main focus across these use cases is a practical approach – parsimonious preservation – by using the Archivematica tools as they exist now whilst building a foundation for more comprehensive preservation strategies in the future. A key area of ongoing investigation covered by this presentation is dealing the with long tail of research data file formats, in particular how to best manage formats that are not immediately supported and need to be added to file registries such as PRONOM.

Keywords

Digital preservation; research data management; software as a service; repository integration; preservation workflows; file formats

1. AUDIENCE

This presentation is aimed at repository managers and related staff working to preserve digital content held within repositories. It is equally aimed at archivists and particularly digital archivists looking at ways to preserve both traditional archival material and other digital content collections, in particular research data.

2. BACKGROUND

Digital preservation should be seen as an integral part of Research Data Management (RDM). Research data is potentially very long lived, especially where it is irreplaceable and supports long running research studies, for example climate data, astronomy observations, and population surveys. This data will only remain usable if it undergoes active digital preservation to ensure that the applications of tomorrow can successfully find, retrieve, and understand the research data of today.

Digital Preservation is **"**the series of managed activities necessary to ensure continued access to digital materials for as long as necessary" where access is "continued, ongoing usability of a digital resource, retaining all qualities of authenticity, accuracy and functionality deemed to be essential for the purposes the digital material was created and/or acquired for" [1]. In the context of RDM, research data is kept to ensure that any research outputs based upon it are repeatable and verifiable[2] and also because research data has value through sharing so it can be reused and repurposed[3]. These underpin the ability to make research data openly available in a form that can be both used and trusted in the long-term.

Whilst digital preservation is clearly desirable, there can also be major challenges in its application, especially to diverse holdings such as University research outputs. This may come as a surprise given that there is no shortage of advice, guidelines and tools for digital preservation. There are dedicated organisations and resources available, including the Digital Preservation Coalition[4] and the Open Preservation Foundation[5]. There is a wide range of tools that can be used, for example as listed by COPTR[6]. There are increasingly well-defined processes for doing preservation, especially for data. Examples include workflows based on the functional model of the Open Archive Information System (OAIS)[7], which can be manifested in the policies/procedures of an organisation, for example the Archive Training Manual from the UK Data Archive[8] [9] Finally, there are also frameworks for assessing preservation maturity that provide a pathway for incremental progression along a preservation path, for example the NDSA Levels of Digital

[1] http://handbook.dpconline.org/glossary

[2] https://royalsociety.org/~/media/royal_society_content/policy/projects/sape/2012-06-20-saoe.pdf

[3] http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report,_Jan14_v1-04.pdf

[4] http://www.dpconline.org/

[5] http://openpreservation.org/

[6] http://coptr.digipres.org/Main_Page

[7] http://public.ccsds.org/publications/archive/650x0m2.pdf

[8] http://www.data-archive.ac.uk/curate/archive-training-manual

[9] http://www.dcc.ac.uk/sites/default/files/documents/RDMF11/HERVE.pdf

Preservation[10]. However, with this plethora of resources and tools comes complexity, cost, and a lot of choices that can lead to protracted timescales for getting preservation up and running within an institution.

3. SCOPE

This paper focuses on how Archivematica[11] can help institutions meet some of these challenges by delivering a practical solution to digital preservation. The paper focuses particularly on how this solution can be used for research data. This approach allows institutions to get started on the digital preservation ladder and then extend as their expertise and capabilities grow.

We show how Archivematica can provide a framework for digital preservation within an RDM infrastructure, including an example workflow for linking preservation with Institutional Repositories. The focus will be on the benefits of digital preservation and how it enables institutions to make their research data more accessible and useable over both the short and long terms.

The paper provides examples of how Archivematica is being applied in several contexts, and in particular will look at an ongoing project at the Universities of York and Hull which is actively investigating how Archivematica. This work is being undertaken as part of the Jisc "Filling the Digital Preservation Gap" project[12] on how Archivematica can be applied to research data [4][5], with a specific interest in workflow aspects[13] and how Archivematica can work with Institutional Repositories. Finally, the paper considers how institutions can lower the costs of adopting this solution thus enabling them to accelerate their preservation activities.

4. APPLYING ARCHIVEMATICA TO RESEARCH DATA PRESERVATION

Many of the benefits of using Archivematica stem from how it can be used to perform a technical audit that then underpins informed decisions on what to do about different types of research data. This is an essential part of 'parsimonious preservation'. This term was coined by Tim Gollins, Head of Digital Preservation at The National Archive in the UK [2],[3]. Being parsimonious means to 'get on and do' preservation in a simple and cost effective way that targets the immediate and real issues that digital content actually creates, rather than what the digital preservation community thinks might be problems in the future.

As University research data holdings diversify, digital content inexorably grows, budgets remain limited, and the benefits of easily accessible digital content become clear, there is never a more pressing time to apply the parsimonious approach. Archivematica provides a practical tool for parsimonious preservation, particularly in the areas of capturing and recording technical metadata within a preservation record (know what you have), and the safe storage (keep the bits safe) of data and metadata. Whilst focused on doing what can be done now, it also allows for additional tasks to be carried out in the future as required and as additional tools become available.

The approach of using Archivematica to inform decisions based on 'knowing what you have' can give an institution immediate visibility at time of deposit of whether the researcher's data is in a 'known' or 'unknown' format. For example, where a format is identified as 'unknown' (research data presents many uncommon formats) the institution can then work with the researcher on getting more information on the data format, securing software that can understand and render the data, or determining if the data format needs to be changed.

However, Archivematica is not purely about file format identification - there are a range of specific tools that could be used to identify files if this were the only requirement (for example, FIDO, FITS, Siegfried, JHOVE and DROID). File identification is just one of several micro-services initiated when a new dataset is transferred and ingested into the system[14]. Archivematica also performs other tasks such as checking the data for viruses (ClamAV), creating checksums (MD5, SHA1, SHA256 or SHA512), cleaning up file names, validating files (where appropriate tools are available) and carrying out format migrations (again where appropriate tools have been configured to do this). Currently, Archivematica identifies 720 file formats and has file format migration support for 123 of these using a wide range of tools (e.g. FFmpeg, ImageMagick, Ghostscript, Inkscape, and Convert). Whilst carrying out these tasks, Archivematica generates metadata describing the structure and technical characteristics of the dataset and this is packaged up (BagIt) and stored in the resulting Archival Information Package (AIP). These tools can all be used individually, but automation through Archivematica substantially reduces the time and effort involved in tool installation and then subsequent workflow automation.

5. PRESERVATION WORKFLOWS

The workflow shown below in Figure 1 is an example of how Archivematica could be integrated with an institutional repository and shows the use of Archivematica to assess and process content that has been deposited in the repository before it undergoes long-term archiving. Other options include using Archivematica to prepare content before deposit or using Archivematica to process content that has already been archived. These are discussed further in [1].

The workflow below shows how Archivematica can be used to ingest data created by a researcher as part of its deposit into the repository, this includes identifying which data is not in formats in Archivematica's Format Policy Register (FPR)[15], i.e. not in FIDO[16] or PRONOM[17], which can trigger an action to address this gap so that the data can be properly managed within the repository. In the first instance such management would allow the data to be presented correctly and links to relevant software made to enable engagement with it, adding value to the data in the repository over time rather than just holding it as a blob of uncharacterized data. In the longer term, knowledge of file formats within the repository also enables activities around Preservation Planning to take place, whether these consist of emulation or migration strategies.

[10] http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf

[11] https://www.archivematica.org

[12] http://www.york.ac.uk/borthwick/projects/archivematica/

[13] http://digital-archiving.blogspot.co.uk/2015/06/the-second-meeting-of-uk-archivematica.html

[14] https://wiki.archivematica.org/Micro-services

[15] https://wiki.archivematica.org/Administrator_manual_1.0#Format_Policy_Registry_.28FPR.29

[16] http://openpreservation.org/technology/products/fido/

[17] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

Archivematica in this scenario might be operated by a digital archivist, a subject specific librarian or some other form of data expert, for example as part of technical support within a given research group or department (the Editor within Figure 1).

The workflow has the following steps (these correspond to the numbers in Figure 1).

1. Researcher uploads data files to the Repository in the normal way. Alternatively, this might be the institution's CRIS system.

2. The Researcher adds descriptive metadata.

3.The Editor reviews the Researcher's dataset, e.g. against minimum repository requirements.

4. As part of the review process, the data files are uploaded to Archivematica

5. Metadata is added if necessary. Archivematica and the tools it applies is used to in effect perform quality control on the dataset, e.g. to flag any files that don't have identified file types or any files that don't conform to their file format specification.

6. Archivematica generates an AIP, which is returned to the repository and stored in Repository Storage.

7. The Editor reviews whether processing in Archivematica was successful and that the dataset is correctly represented by the AIP. The Editor then approves the Researcher's submission and the Researcher is notified.

8. The AIP is sent to Archive Storage for long-term retention.

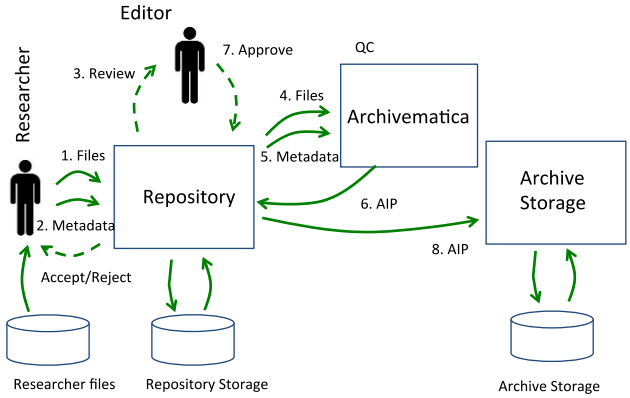

Figure 1 Example Archivematica workflow for research data preservation

Key to making Archivematica work with research data file formats is having a mechanism for reporting on unknown formats so that, in addition to local management, there is a way of adding value to the FPR and PRONOM by adding to the list of file format registered there and associated information about them. Work at York[18] (and reported in the Filling the Digital Preservation Gap Phase One report [4]) has highlighted the wide spectrum of research data file formats, and the long tail that preservation workflows will need to deal with over time. Though project work has started to address the problem through the addition of a small subset of research data formats[19] to PRONOM, this is clearly a problem that can only be addressed through wider collaboration and community engagement.

[18] http://digital-archiving.blogspot.co.uk/2016/05/research-data-what-does-it-really-look.html

[19] http://digital-archiving.blogspot.co.uk/2016/07/new-research-data-file-formats-now.html

6. LOWERING THE COST OF USING ARCHIVEMATICA

There are several tangible benefits to using preservation tools such as Archivematica for research data, however, these benefits will only be realized if the associated costs are low. Whilst Archivematica itself is open source and freely available, this does not mean it is cost neutral. Costs include the resources and infrastructure needed for executing digital preservation, for example the start-up and ongoing costs of installing and running Archivematica pipelines. The human costs associated with the time taken to learn how to apply and use a preservation system should also be taken into account.

Whilst setting up their own implementations of Archivematica for the preservation of research data as part of the "Filling the Digital Preservation Gap" project, the Universities of York and Hull will be reviewing the costs of having done so. Certainly there have been cost savings in being able to work together on this project. Some decisions can be taken jointly and technical solutions around the integration of Archivematica with our repository systems can be shared. There is also a clear benefit to being able to talk to other Archivematica users. The UK Archivematica group has been a helpful and supportive community to share ideas and discuss solutions and there are undoubtedly benefits to working in an open way that enables us to learn from other people's mistakes and replicate their successes. Doing so can lead to cost savings in the longer term.

Another way that costs can be reduced for institutions is through use of a hosting model whereby a third-party provider delivers tools such as Archivematica as a service. The service provider handles the issues of setting-up, running, managing and supporting pipelines which allows the institution to focus on the archival and business decisions on what to preserve, how to preserve it, and the business case on why it should be preserved. It also addresses a current and significant issue that institutions have finding staff with the necessary skills in the installation, operation and maintenance of preservation software as well as their institution having the capacity to host and run this software on appropriate IT servers and storage systems.

Examples of communities that have started to establish common digital preservation platforms around Archivematica include the Ontario Council of University Libraries (OCUL) which has integrated Archivematica as part of the Dataverse research data publishing platform resulting in the ability to ingest data files and metadata from Dataverse into Archivematica for digital preservation purposes[20]. Also relevant in this context is the Council of Prairie and Pacific University Libraries (COPPUL)[21], Archivematica hosting and integration with DuraCloud[22], and Archivematica hosting and integration with Arkivum[23].

In the UK, recent developments under the Jisc Research Data Shared Service[24] provide another example of how institutions can work together on a shared approach to the preservation and management of research data. Archivematica has been selected as one of the digital preservation systems under this shared service and work is underway to ensure that it can be integrated with a number of repository solutions. As this new service is

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

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

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

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

[24] https://www.jisc.ac.uk/rd/projects/research-data-shared-service

developed it provides a valuable opportunity for institutions to work together on their requirements and workflows and take advantage of a centrally hosted service.

Whilst using Archivematica as a hosted service from a third-party has many benefits, there are also several barriers to overcome. These include an assurance that an exit strategy is available in order to avoid lock-in to the hosting organization or to allow a continuity strategy that addresses the case where the service fails to be delivered. The use of open standards and data structures within Archivematica (for example PREMIS, METS, Bagit) is a key component of providing this assurance and allows migration to an alternative preservation service provider or in-house environment if needed.

7. CONCLUSION

Digital preservation of research data is an essential activity in ensuring that this data is accessible and usable in the future:

- Digital preservation has a valuable role to play in supporting the long-term availability and usability of research data, but it needs to be properly embedded into the research data management environment for these benefits to be realized.
- Digital preservation tools such as Archivematica can provide a quick way to get started with basic digital preservation whilst also providing a route for institutions to develop and apply more sophisticated techniques as their digital preservation maturity evolves.
- Doing preservation 'today' using Archivematica enables institutions to make practical parsimonious headway in the preservation of research data.
- Digital preservation tools are not currently able to recognize the range of file formats that researchers create. The digital preservation and research data community need to work together on improving the reach of file format registries and identification tools to help facilitate the preservation of research data.
- Archivematica is free but this does not mean implementation is cost neutral. There are ways of reducing these costs by sharing experiences and workflows, working together on integrations, and by taking advantage of the available hosting options.

8. REFERENCES

[1] Addis, M. 2015. RDM workflows and integrations for Higher Education institutions using hosted services, https://dx.doi.org/10.6084/m9.figshare.1476

[2] Gollins, T. 2009. Parsimonious preservation: preventing pointless processes!

http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf

[3] Gollins, T. (2012), Putting parsimonious preservation into practice,
http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation-in-practice.pdf

[4] Mitcham, J., Awre, C., Allinson, J., Green, R., Wilson, S. 2015. Filling the digital preservation gap. A Jisc Research Data Spring project: Phase One report - July 2015, http://dx.doi.org/10.6084/m9.figshare.1481170

[5] Mitcham, J., Awre, C., Allinson, J., Green, R., Wilson, S. 2016 Filling the digital preservation gap. A Jisc Research Data Spring project: Phase Two report – February 2016, https://dx.doi.org/10.6084/m9.figshare.207