UNIVERSITY of York

UNIVERSITY OF Hull

Filling the Digital Preservation Gap

A Jisc Research Data Spring project

Phase One report - July 2015

Jenny Mitcham, Chris Awre, Julie Allinson,
Richard Green, Simon Wilson

Authors:

Jenny Mitcham (jenny.mitcham@york.ac.uk) is Digital Archivist at the Borthwick Institute for Archives at the University of York
Chris Awre (c.awre@hull.ac.uk) is Head of Information Services, Library and Learning Innovation, at the University of Hull
Julie Allinson (julie.allinson@york.ac.uk) is the manager of Digital York at the University of York
Richard Green (r.green@hull.ac.uk) is an independent consultant working with the digital repository team at the University of Hull
Simon Wilson (s.wilson@hull.ac.uk) is University Archivist at the University of Hull

Acknowledgements

The authors of this report would like to thank the many organisations and individuals who contributed information for it. In particular we would like to thank staff at Artefactual Systems Inc, (especially Evelyn McLellan, Courtney Mumma and Justin Simpson), David Clipsham from The National Archives and members of the UK Archivematica group (particularly Matthew Addis) for their help, feedback and advice over the course of the project.

This report was funded by Jisc as part of its Research Data Spring initiative.

Preface

This report is divided into two distinct parts and it is the authors' hope that each of these parts can be read as a stand-alone text. Some readers will find Part A. "The rationale for RDM and for the use of Archivematica" a sufficient introduction to this field. Others, perhaps more familiar with the topic or having read Part A, will want a more detailed coverage of our investigations and this is provided in Part B "Archivematica as part of RDM: Detailed analysis".

As each part of the document is intended to be useful independent of the other, there is necessarily some overlap of the material covered but we have tried to keep this to a minimum.

Contents

Part A: The rationale for RDM and for the use of Archivematica

Why do we need a digital preservation system for research data?

Research data should be seen as a valuable institutional asset and treated accordingly. Research data is often unique and irreplaceable. It may need to be kept to validate or verify conclusions recorded in publications. Funder, publisher and often internal university requirements ask that research data is available for others to consult and is preserved in a usable form after the project that generated it is complete.

In order to facilitate future access to research data we need to actively manage and curate it. Digital preservation is not just about implementing a good archival storage system or 'preserving the bits' it is about working within the framework set out by international standards (for example the Open Archival Information System[1]) and taking steps to increase the chances of enabling meaningful re-use in the future.

What are the risks if we don't address digital preservation?

Digital preservation has been in the news this year (2015). An interview with Google CEO Vint Cerf in February grabbed the attention of the mainstream media with headlines about the fragility of a digital media and the onset of a digital dark age[2].

This is clearly already a problem for researchers with issues around format and media obsolescence already being encountered. In a 2013 Research Data Management (RDM) survey[3] just under a quarter of respondents to the question "Which data management issues have you come across in your research over the last five years?" selected the answer "Inability to read files in old software formats on old media or because of expired software licences". These are the sorts of problems that a digital preservation system is designed to address.

Due to its complexity digital preservation is very easy to put in the 'too difficult' box. There is no single perfect solution out there and it could be argued that we should sit it out and wait until a fuller set of tools emerges. A better approach is to join the existing community of practice and embrace some of the working and evolving solutions that are available.

Why are we interested in Archivematica?

Archivematica is an open source digital preservation system that is based on recognised standards in the field. Its functionality and the design of its interfaces were based on the Open Archival Information System and it uses standards such as PREMIS and METS to

---

[1] http://public.ccsds.org/publications/archive/650x0m2.pdf
[2] For example http://www.bbc.co.uk/news/science-environment-31450389 and http://www.theguardian.com/media-network/2015/may/29/googles-vint-cerf-prevent-digital-dark-age
[3] Research Data Management at the University of York: An audit of current practice, Jenny Mitcham, 2013

store metadata about the objects that are being preserved. Archivematica is flexible and configurable and can interface with a range of other systems.

A fully fledged RDM solution is likely to consist of a variety of different systems performing different functions within the workflow; Archivematica will fit well into this modular architecture and fills the digital preservation gap in the infrastructure.

The Archivematica website states that “The goal of the Archivematica project is to give archivists and librarians with limited technical and financial capacity the tools, methodology and confidence to begin preserving digital information today.” This vision appears to be a good fit with the needs and resources of those who are charged with managing an institution’s research data.

It should be noted that there are other digital preservation solutions available, both commercial and open source, but these were not assessed as part of this project.

Why do we recommend Archivematica to help preserve research data?

- It is flexible and can be configured in different ways for different institutional needs and workflows
- It allows many of the tasks around digital preservation to be carried out in an automated fashion
- It can be used alongside other existing systems as part of a wider workflow for research data
- It is a good digital preservation solution for those with limited resources
- It is an evolving solution that is continually driven and enhanced by and for the digital preservation community; it is responsive to developments in the field of digital preservation
- It gives institutions greater confidence that they will be able to continue to provide access to usable copies of research data over time.

What does Archivematica actually do?

Archivematica runs a series of microservices on the data and packages it up (with any metadata that has been extracted from it) in a standards compliant way for long term storage. Where a migration path exists, it will create preservation or dissemination versions of the data files to store alongside the originals and create metadata to record the preservation actions that have been carried out.

A more in depth discussion of what Archivematica does can be found in the report text. Full documentation for Archivematica is available online[4].

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

How could Archivematica be incorporated into a wider technical infrastructure for research data management?

Archivematica performs a very specific task within a wider infrastructure for research data management - that of preparing data for long term storage and access. It is also worth stating here what it doesn't do:

- It does not help with the transfer of data (and/or metadata) from researchers
- It does not provide storage
- It does not provide public access to data
- It does not allocate Digital Object Identifiers (DOIs)
- It does not provide statistics on when data was last accessed
- It does not manage retention periods and trigger disposal actions when that period has passed

These functions and activities will need to be established elsewhere within the infrastructure as appropriate.

What does research data look like?

Research data is hard to characterise, varying across institutions, disciplines and individual projects. A wide range of software applications are in use by researchers and the file formats generated are diverse and often specialist.

Higher education institutions typically have little control over the data types and file formats that their researchers are producing. We ask researchers to consider file formats as a part of their data management plan and can provide generic advice on preferred file formats if asked, but where many of the specialist data formats are concerned it is likely that there is no 'preservation-friendly' alternative that retains the significant properties of the data.

Research data can be large in size, and/or quantity. It often includes elements that are confidential or sensitive. Sensitivities are likely to vary across a dataset with some files being suitable for wider access and others being restricted. A one-size fits all approach to rights metadata is not appropriate. In some cases there will be different versions of the data that need to be preserved or different deposits of data for a single research project. Scenarios such as these are likely to come about where data is being used to support multiple publications over the course of a piece of research.

Research data may come with administrative data and documentation. These may be documents relating to ethical approval or grant funding, data management plans or documentation or metadata relating to particular files. The association between the research data and any associated administrative information should be maintained.

It can be difficult to ascertain the value of research data at the point of ingest. Some data will be widely used and should be preserved for the long term and other data will never be accessed and will be disposed of at the end of its retention period.

How would Archivematica handle research data?

Archivematica can handle any type of data but it should be noted that a richer level of preservation will only be available for some file formats. Archivematica (like other digital preservation systems) will recognise and identify a large number of research data formats but by no means the full range. For a smaller subset of these file formats (for example a range of raster and vector image and audio visual formats) it comes with normalisation rules and tools. It can be configured to normalise other file formats as required (where open source command line tools are available). Archivematica also allows for the flexibility of manual normalisations. This gives data curators the opportunity to migrate files in a more manual way and update the PREMIS metadata by hand accordingly.

For other data types (and this will include many of the file formats that are created by researchers), Archivematica may not be able to identify, characterise or normalise the files but will still be able to perform certain functions such as virus check, cleaning up file names, creating checksums and packaging the data and metadata up to create an archival information package.

Archivematica can handle large files (or large volumes of small files) but its abilities in this area are very much dependent on the processing power that has been allocated to it. Users of Archivematica should be aware of the capabilities of their own implementation and be prepared to establish a cut off point over which data files of a certain size may not be processed, or may need to be processed in a different way.

Archivematica uses the PREMIS metadata standard to record rights metadata. Rights metadata can be added for the Submission Information Package as a whole rather than in a granular fashion. This is not ideal for research data for which there are likely to be different levels of sensitivity for different elements within the final submitted dataset. The Archivematica manual suggests that fuller rights information would be added to the access system (outside of Archivematica)[5].

The use of Archival Information Collections (AICs) in Archivematica enables the loose association of groups of related Archival Information Packages (AIPs). This may be a useful feature for research data where different versions of a dataset or parts of a dataset are deposited at different times but are all associated with the same research project.

Archivematica is a suitable tool for preserving data of unknown value. Workflows within Archivematica and the processing of a transferred dataset from a Submission Information Package (SIP) to an Archival Information Package (AIP) can be automated. This means that some control over the data and a level of confidence that the data is being looked after adequately can be gained, without expending a large amount of staff time on curating the data in a manual fashion. If the value of the data is seen to increase (by frequent requests to access that data or as a result of assessment by curatorial staff) further efforts can be made to preserve the data using the AIP re-ingest feature and perhaps by carrying out a level of manual curation. The extent of automation within Archivematica can be configured so staff are able to treat datasets in different ways as appropriate. Institutions may have a range of approaches here, but the levels of automation that are possible provide a compelling argument for the adoption of Archivematica if few staff resources are available for manual preservation.

[5] https://www.archivematica.org/en/docs/archivematica-1.4/user-manual/ingest/ingest/#add-rights

What are the limitations of Archivematica for research data?

Archivematica should not be seen as a magic bullet. It does not guarantee that data will be preserved in a re-usable state into the future. It can only be as good as digital preservation theory and practice is currently and digital preservation itself is not a fully solved problem.

Research data is particularly challenging from a preservation point of view due to the range of data types and formats that are in existence, many of which are not formats that digital preservation tools and policies exist for, thus they will not receive as a high a level of curation when ingested into Archivematica.

As mentioned above, the rights metadata within Archivematica may not fit the granularity that would be required for research data. This information would need to be held elsewhere within the infrastructure.

One of Archivematica's strengths is its flexibility and the fact it can be configured to suit the needs of the institution or for a particular workflow. This however may also act as an initial barrier to use. It takes a bit of time to become familiar with Archivematica and to work out how you want to set it up[6]. It is also a tool that most people would not want to use in isolation, and considerable thought needs to go into how it needs to interact with other systems and what workflow may best suit your institution.

The user interface for Archivematica is not always intuitive and takes some time to fully understand. There is currently no indication within the GUI that Archivematica is processing or estimate of how long a particular microservice may have left to run. This is a limitation for large datasets if you are processing them through Archivematica's dashboard. For a more automated curation workflow this will not have any impact.

What costs are associated with using Archivematica?

Archivematica is a free and open source application but this does not mean it will not cost anything to run. As a minimum an organisation wishing to run Archivematica locally will need both technical and curatorial staff resource. A level of technical knowledge is required to install and troubleshoot Archivematica and perform necessary upgrades. Further technical knowledge is required to consider how Archivematica fits into a wider infrastructure and to get systems talking to each other. As Archivematica is open source, developer time could be devoted to enhancing it to suit institutional needs. Developer time can also be bought from Artefactual Systems, the lead developer of Archivematica, to fund specific enhancements or new functionality which will be made available to the wider user base. In order to make the most of the system, organisations may want to consider factoring in a budget for developments and enhancements.

It is essential to have at least one member of curatorial staff who can get to grips with the Archivematica interface, make administrative decisions about the workflow and edit the format policy registry where appropriate[7]. A level of knowledge of digital preservation is required for this, particularly where changes or additions to normalisation rules within the format policy registry are being considered. A greater number of curatorial staff working on

[6] Appendix 4 has been included within this report to help address this problem though it is recognised that there is no one-size-fits-all approach

[7] Note that changes to the format policy registry within Archivematica can be labour intensive and may also require technical staff resource to install new tools

Archivematica will be necessary the more manual steps there are within the workflow (for example if manual selection and arrangement, metadata entry or normalisations are carried out). This requirement for curatorial staff will increase in line with the volumes of data that are being processed.

Technical costs of establishing an Archivematica installation should also be considered. For a production system the following server configuration is recommended as a minimum:

- Processor: dual core i5 3rd generation CPU or better
- Memory: 8GB+
- Disk space: 20GB plus the disk space required for the collection

The software can be installed on a single machine or across a number of machines to share the workload. At the time of writing, the software requires a current version of Ubuntu LTS (14.04 or 12.04) as its operating system.

What other systems is Archivematica integrated with?

Archivematica provides various levels of integration with DSpace, CONTENTdm, AtoM, Islandora and Archivist's Toolkit for access and Arkivum, DuraCloud and LOCKSS for storage[8]. There are ongoing integrations underway with ArchivesSpace, Hydra, BitCurator and DataVerse[9].

In addition, Archivematica provides a transfer REST API that can be used to initiate transfers within the software, the first step of the preservation workflow. Archivematica's underlying Storage Service also provides RESTful APIs to facilitate the creation, retrieval and deletion of AIPs.

How can you use Archivematica?

There are 3 different ways that an institution might wish to use Archivematica:

1. Local - institutions may install and host Archivematica locally and link it to their preferred storage option
2. Arkivum[10] - a managed service from Arkivum will allow Archivematica to be hosted locally within the institution with upgrades and support available through Arkivum in partnership with Artefactual Systems. A remote hosting option is also available. Both include integration of Archivematica and Arkivum storage.
3. ArchivesDirect[11] - a hosted service from DuraSpace that combines Archivematica's preservation functionality with DuraCloud for storage

How could Archivematica be improved for research data?

It should be noted that Archivematica is an evolving tool that is under active development. During the short 3 months of phase 1 of our project "Filling the Digital Preservation Gap" we have been assessing a moving target. Version 1.3 was installed for initial testing. A month in,

[8] https://www.archivematica.org/en/
[9] As reported in an 'Introduction to Archivematica' webinar on the 9th April 2015
[10] http://arkivum.com/
[11] http://www.archivesdirect.org/

version 1.4 was released to the community. As we write this report, version 1.5 is under development and due for imminent release; a version of this has been made available to us for testing.

Archivematica is open source but much of the development work is carried out by Artefactual Systems the company that support it. They have their own development roadmap[12] for Archivematica but most new features that appear are directly sponsored by the user community. Users can pay to have new features, functionality or integrations built into Archivematica, and Artefactual Systems try to carry out this work in such a way to make the features useful to the wider user community and agree to continue to support and maintain the code base through subsequent versions of the software. This 'bounty model' for open source development seems to work well and keeps the software evolving in line with the priorities of its user base.

During the testing phase of this project we have highlighted several areas where Archivematica could be improved or enhanced to provide a better solution for research data and several of these features are already in development (sponsored by other Archivematica users). In phase 2 of the project we hope to be able to contribute to the continued development of Archivematica.

Who else is using Archivematica to do similar things?

Archivematica has been adopted by several institutions internationally but its key user base is in Canada and the United States. A list of a selection of Archivematica users can be found on their community wiki pages[13]. Some institutions are using Archivematica to preserve research data. Both the Zuse Institute Berlin[14] and Research Data Canada's Federated Pilot for Data Ingest and Preservation[15] are important to mention in this context.

Archivematica is not widely used in the UK but there are current implementations at the National Library of Wales and the University of Warwick's Modern Records Centre. Interest in Archivematica in the UK is growing. This is evidenced by the establishment this year of a UK Archivematica group which provides a local forum to share ideas and case studies. Representatives from 15 different organisations were present at the last meeting at Tate Britain in June 2015 and a further meeting is planned at the University of Leeds in the autumn.

[12] https://wiki.archivematica.org/Development_roadmap:_Archivematica
[13] https://wiki.archivematica.org/Community
[14] http://www.zib.de/features/shepherding-bits
[15] http://www.rdc-drc.ca/the-rdc-federated-pilot-for-data-ingest-and-preservation/

Part B: Archivematica as part of RDM: Detailed analysis

Executive summary

This document describes an analysis, undertaken over a twelve week period, of the possible benefits of using Archivematica as part of a research data management (RDM) workflow.

The report begins with a brief overview of Archivematica and explains that the authors' investigation seeks to explore how use of the product might provide digital preservation functionality within the context of RDM.

Early sections of the report briefly examine the requirements of research funding bodies in the UK and how these relate to digital research data preservation needs of institutions in the higher education community. The authors then go on to characterise research data and to explore the many forms that it might take.

The report notes that, using a "lightweight technology stack", Archivematica was tested to understand better how the product works and how it might fit, or be fitted, into potential local workflows. The testing was carried out with the multivarious forms of research data in mind and so using common and obscure file formats, large and small files, single and multiple files.

A brief, high-level, technical analysis of the software is followed by suggestions as to how Archivematica might be improved, particularly for use in an RDM context. Notwithstanding these suggestions, the conclusion of the report is that "Archivematica offers a practical and affordable solution for digital preservation in an RDM context, effectively 'filling the digital preservation gap' that was highlighted in the introduction of this report."

Introduction

In order to manage research data effectively for the long term we need to consider how we incorporate digital preservation functionality into our RDM infrastructures and workflows. The Engineering and Physical Sciences Research Council (EPSRC) requirements[16] state that research data should be retained for a minimum of ten years from last access. Though it is possible that over the course of ten years data will remain readable and re-usable, once longer timescales are required, this becomes less and less likely. Unless a digital preservation solution is in place it is unlikely that we will be able to continue to provide access to usable copies of research data over time.

Perhaps as a result of the long timescales at play here, universities have struggled to make the case to invest in digital preservation systems. Other (more immediate) elements of the RDM lifecycle and infrastructure have taken precedence, but there is an inherent risk in not

[16] https://www.epsrc.ac.uk/about/standards/researchdata/expectations/

fully exploring this element as there ultimately will be research data that universities will need to maintain and provide access to in perpetuity. Digital preservation systems should be very much considered as part of a wider RDM infrastructure but are currently a very real gap in existing provision.

Affordability is an issue here as there are few resources available for a issue which to many does not seem very immediate and where the range of data types and potentially huge size of datasets makes the problem seem insurmountable. This is why in this project we wish to explore the potential of a low cost solution, the open source digital preservation system Archivematica.

Archivematica[17] (developed and supported by Artefactual Systems[18]) is highly regarded in the digital preservation community, being the most advanced open source tool of its kind currently available[19]. One of its strengths is the fact it is based on internationally recognised open standards such as the Open Archival Information System Reference Model (OAIS)[20] and metadata standards such as PREMIS[21] and METS[22]. Archivematica is in continuous active development with many developments carried out by Artefactual Systems as a result of sponsorship by Archivematica users.

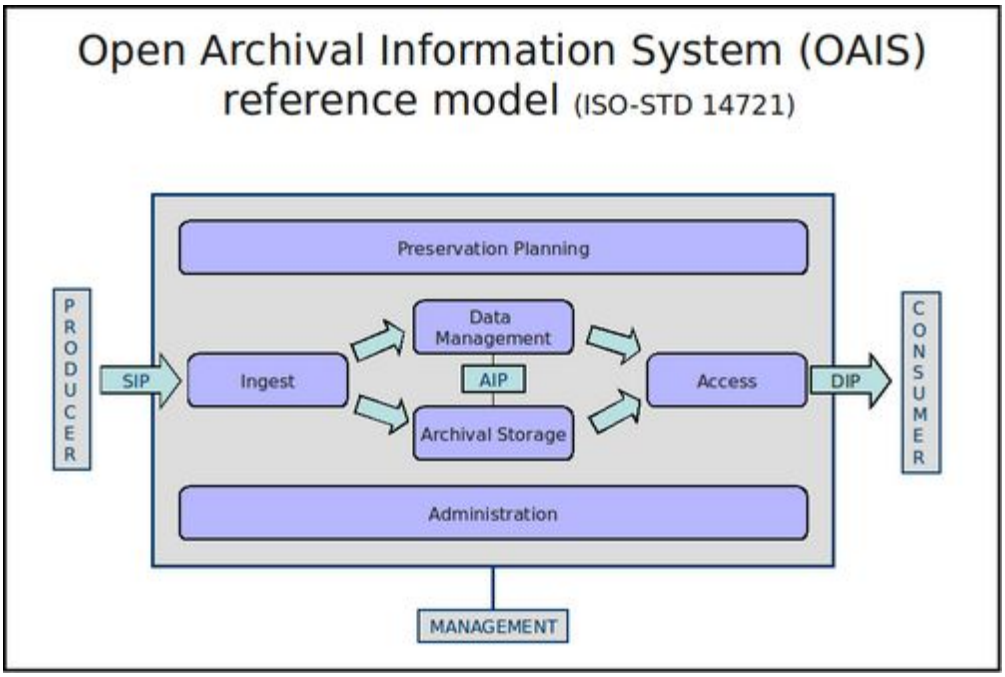


Figure 1 - OAIS reference model Source: Archivematica wiki (CC-BY-SA)

In very basic terms, Archivematica takes one or more files that need to be preserved, runs a set of microservices over them, generates from them a Submission Information Package (SIP), which is then processed to produce an Archival Information Package (AIP - the materials to be archived) and, optionally, a Dissemination Information Package (DIP - a user-oriented version of the materials for inclusion, for instance, in an institutional repository).

[17] https://www.archivematica.org/en/
[18] https://www.artefactual.com/
[19] There exists also the RODA repository system for preservation (http://www.roda-community.org/) but this is a different animal: rather than being suited to working with existing repository infrastructures, it is an integrated access and preservation system
[20] http://public.ccsds.org/publications/archive/650x0m2.pdf
[21] http://www.loc.gov/standards/premis/
[22] http://www.loc.gov/standards/mets/

There is a strong international community that has built up around Archivematica with a key user-base in the United States and Canada. It is regularly discussed at international digital preservation conferences such as iPRES[23]. A list of a selection of Archivematica users can be found on their community wiki pages[24]. Some institutions are using Archivematica to preserve research data. Both the Zuse Institute Berlin[25] and Research Data Canada's Federated Pilot for Data Ingest and Preservation[26] are important to mention in this context. Interest in Archivematica is now growing in the UK. This year the UK Archivematica group was formed and as of June 2015 have had two meetings to share experiences, knowledge and use cases.

The idea behind this project is to investigate Archivematica and explore how it might be used to provide digital preservation functionality within a wider infrastructure for Research Data Management. In phase one of the project we intend to find out what Archivematica has to offer RDM and what its main strengths and weaknesses are. We will explore where it might fit within an RDM infrastructure and workflow and highlight the benefits it could bring. Where weaknesses are highlighted, we will explore how these problems may be overcome, perhaps with a sponsored development which could be funded in phase two of this project. Sponsored developments may include extra functionality or enable better integration with other components of the RDM infrastructure. Phase three would see proof of concepts being established at Hull and York and their subsequent dissemination as case studies which would help inform the RDM community across the UK and beyond.

Funder requirements

As mentioned in the Introduction, a major driver in considering the need to preserve research data has been the EPSRC expectation that research data will be kept for 10 years after the last recorded request for access. This funder requirement necessitates specific planning for preservation, for which Archivematica may present a solution. In meeting the EPSRC expectations, the University of Hull took the view that a cross-disciplinary approach would be taken rather than simply focus on those specifically-funded by the EPSRC, as the expectations were themselves not disciplinary-specific per se. In doing so, this raised the question of what other funders were asking for re:preservation of research data. The Digital Curation Centre provides useful overviews of the funder requirements for research data generally[27]. This is mirrored in the SCONUL report 'Research Data Management: Briefing for Library Directors' in February 2015[28]. Following through on these, Table 1 lists major research funders and their specific requirements for preservation of research data.

Understanding these funder requirements will help to define the requirements for Archivematica in enabling preservation workflows for research data. Requirements can then be enshrined within local policies to facilitate local action and enforcement. Currently, for example, the University of York research data management policy has as one of its purposes the preservation of eligible data for re-use, through the capture of relevant preservation metadata and actions alongside the data itself: the time period required by the EPSRC is used as the guiding principle on length of preservation required. The University of

[23] http://www.ipres-conference.org/
[24] https://wiki.archivematica.org/Community
[25] http://www.zib.de/features/shepherding-bits
[26] http://www.rdc-drc.ca/the-rdc-federated-pilot-for-data-ingest-and-preservation/
[27] http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies
[28] http://www.sconul.ac.uk/page/research-data-management

Hull's draft policy states that the purpose of the policy is to meet the obligations of the University to funders (as well as establish research excellence), and that all data registered will be assessed for preservation in an appropriate repository.

Funder	Preservation requirement	Actions required	Length of time specified
AHRC[29]	"...the AHRC expects the digital outputs or technologies produced by projects to be preserved for an appropriate period after the end of project funding."	Technical Plan template has sections on Preserving your data. "You must consider preservation in four ways: what, where, how and for how long. You must also consider any institutional support needed in order to carry out these plans, whether from an individual, facility, organisation or service."	At least 3 years from the end of the project, though "... in many, if not most cases, a longer period will be appropriate."
BBSRC[30]	"Researchers are expected to ensure that data are maintained for a period of 10 years after the completion of the research project in suitable accessible formats using established standards."	Data management plan requirement has an emphasis on data sharing over time.	10 years after the completion of the research project.
EPSRC[31]	"Research organisations will ensure that EPSRC-funded research data is securely preserved for a minimum of 10 years from the date that any researcher 'privileged access' period expires or, if others have accessed the data, from last date on which access to the data was requested by a third party."	Local policy and plan for long-term management of research data should be in place institutionally.	10 years from date of last access requested.

[29] http://www.ahrc.ac.uk/SiteCollectionDocuments/Research-Funding-Guide.pdf
[30] http://www.bbsrc.ac.uk/documents/data-sharing-policy-pdf/
[31] https://www.epsrc.ac.uk/about/standards/researchdata/

ESRC[32]	"ESRC funds its data service providers to guarantee (curation and) long-term preservation of all research data deposited by grant holders."	All data should be offered to the UK Data Archive for preservation. A local copy may also be maintained if of value.	Not specified.
MRC[33]	"Data arising from MRC-funded research must be properly curated throughout its life-cycle and released with the appropriate high-quality metadata."	Complete data management plan, which includes a section on data preservation strategy and standards to be used.	10 years (taken from DCC information - the period is not apparent on the current data sharing webpages).
NERC[34]	"NERC considers that long-term, open access to the data that underpin research publications will help to ensure the integrity, transparency and robustness of the research record."	Complete data management plan for grant applications. Data produced should be assessed against a Data Value Checklist to inform its submission to a NERC data centre	A minimum of 10 years after completion of the research. However, for projects of major importance, this may need to be 20 years or longer. [35]
Cancer Research UK[36]	"This data sharing and preservation policy applies to all Cancer Research UK-funded research while recognising that different fields of study will require different approaches"	No prescribed approach is laid down, but "...requires [researchers] to make clear provision for [preserving and sharing data] when planning and executing their research."	Not specified.
Wellcome Trust [37]	"The Trust is happy to discuss issues relating to longer-term preservation and sustainability with researchers so as to help provide the support required to maximise the long-term value of key research datasets."	Complete data management and sharing plan, addressing issues of preservation within this.	A minimum of ten years, but research based on clinical samples or relating to public health might require longer storage to allow for long-term follow-up to occur.

[32] http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx
[33] http://www.mrc.ac.uk/research/research-policy-ethics/data-sharing/policy/
[34] http://www.nerc.ac.uk/research/sites/data/policy/
[35] http://www.nerc.ac.uk/research/sites/data/policy/datapolicy-guidance/
[36] http://www.cancerresearchuk.org/funding-for-researchers/applying-for-funding/policies-that-affect-your-grant/submission-of-a-data-sharing-and-preservation-strategy
[37] http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm

Horizon 2020 (EU)[38]	"All project proposals … include a section on research data management which is evaluated under the criterion 'Impact'. Where relevant, applicants must provide a short, general outline of their policy for: how will this data be curated and preserved?"	Complete data management submission.	Not specified.

Table 1: Analysis of funder requirements for data preservation

The information from funders suggests two current trends: that preservation and long-term access/re-use of data is an important part of requirements from all funders; but that the length of time that preservation should take place over is less consistently specified. Policies need to be realistic in their goals in order to give institutions the ability to fully comply, so it may be that the lack of detail about length of time reflects an uncertainty about what is feasible. A way of addressing this has been to specify in some cases the need for good quality metadata and the use of appropriate formats to assist with preservation: this reflects two key areas of preservation action that can be addressed using Archivematica.

It should be noted that funders are not the only external bodies making requirements for the long-term management of research data. Institutional policies have been mentioned already. Journals and publishers are also now starting to ask that data be preserved, in order to continue to underpin research publications and facilitate access to the data as part of research integrity. The Jisc-funded JoRD project[39] that completed in 2014 examined this emerging trend. This found more than 150 journals with a data sharing policy, but a lack of consistency across these and a focus on sharing rather than preservation per se. In describing a proposed standard framework for such policies the project included a section on preservation of the data, and emphasised the need for the use of open formats as a requirement. This area will undoubtedly evolve.

Community initiatives are also emerging as guidance and good practice for researchers. One such is the FAIR[40], which seeks to set out a document "...that is a general 'guide to FAIRness of data'" for community endorsement. A basic premise of such a document would be that data is both interoperable and re-usable, both areas that will be underpinned by preservation.

Requirements for a digital preservation system

As part of this project we have reviewed an existing set of requirements for digital preservation and demonstrated how Archivematica fits with these requirements. Full information about the requirements, how they were created and what was considered in and out of scope are available in Appendix 1 along with a brief assessment of how Archivematica

[38] http://www.eui.eu/Projects/EUIdatarepository/H2020datapilot.aspx
[39] https://jordproject.wordpress.com/
[40] https://www.force11.org/group/fairgroup

matches up to them. Requirements were seen to be the same for both Hull and York and were deliberately designed to be applicable to a range of digital preservation scenarios that Higher Education Institutions may have. Both Hull and York see research data as being just one element of a wider digital preservation need and any digital archive we put in place needs to be suitable for a range of data types, deposit scenarios and workflows.

Archivematica performs well against our requirements. Of the 33 requirements listed, Archivematica fully meets 23 of them. A further 9 requirements are partially met by Archivematica. Some of these (for example PP1 - allowing for preservation planning) will be met by new functionality in the pipeline for the next version of Archivematica. Others are in areas where we could reasonably expect other elements of a wider infrastructure for managing research data to fulfil these tasks (for example AS1 and 2 - integrity checking of files or I2 - recording selection and appraisal decisions). It should be noted that Archivematica does not in itself fulfil OAIS or provide a full solution for digital preservation, but instead should be seen as one piece of the jigsaw.

One requirement is not met by Archivematica. PP2 suggests that "Automated checking of significant properties of files[41] will be carried out post-migration to ensure these properties are adequately preserved" but acknowledges that this can only happen where appropriate tools exist, and places the weight of this requirement as a 'could' rather than a 'must' or a 'should'. There are few if any digital preservation solutions that would be able to fulfil this requirement currently though it is a realistic hope for the future. Where open source tools become available, Archivematica is well placed to incorporate these into its existing microservices workflow.

Developments and enhancements within Archivematica that we propose to fund during phase 2 of this project will directly help with moving it to a position that more closely aligns with our digital preservation requirements. By the end of phase 3 of this project we will be able to report on the impact that this has had.

The characteristics of research data

In order to assess the potential use of Archivematica for digital research data we need to understand something of the characteristics of research data, however, by its very nature, research data is difficult to characterise. It doesn't conform to a particular profile and isn't restricted to a simple set of file formats. It is not fixed and will change rapidly as data collection and analysis instruments and tools evolve. The nature of research that is carried out at academic institutions is often leading edge and thus moves on rapidly. Researchers are quick to explore and exploit new data collection and analysis techniques that become available and obsolescence of file formats is a very real issue.

Although it is difficult to get a definitive picture of research data, the University of York has some sources of information about the data that their researchers are producing and these have been consulted to help inform this project.

In 2013 at the University of York, the Research Data Management working group sent out an online questionnaire to research staff and students. The purpose of the questionnaire was to get a current picture of research data management at York during both the active research phase and after project completion. Information collected during this survey was valuable in

[41] http://www.significantproperties.org.uk/

giving the project team a broad overview of what types of research data are held by researchers.[42]

Types of data

As can be seen in figure 2 the most frequently held data types are documents or reports and spreadsheets. These are types of data that are well understood and we can have a high level of confidence that we can handle these effectively within a digital archive. There are however, many more types of data in the list, some of which that will present bigger challenges, for example those from 48 people who have 'Data automatically generated from or by computer programs', the 45 who have 'Data collected from sensors or instruments' and the 19 who have 'Software created for the project'.

Figure 2 - Answers to the question "Please select the main types of electronic research data you generate" from the 188 respondents to the 2013 online questionnaire at the University of York. From Research Data Management at the University of York: An audit of current practice, Jenny Mitcham, 2013

Size of data

Size of research datasets varies considerably across disciplines and individual research projects. Some researchers produce minimal quantities of data or none at all, the large body of their work being purely intellectual or publication-oriented, whereas other researchers produce large datasets on a daily basis. Again it is hard to generalise but one aspect of data size that is apparent from figure 3 is that researchers do find it difficult to estimate the size of the data they are likely to produce.

[42] Although the University of Hull has not conducted formal interviews or consultation on this precise area, there is nothing to suggest from the ongoing work and discussions centred around research data management plans that the situation at Hull is not the same.

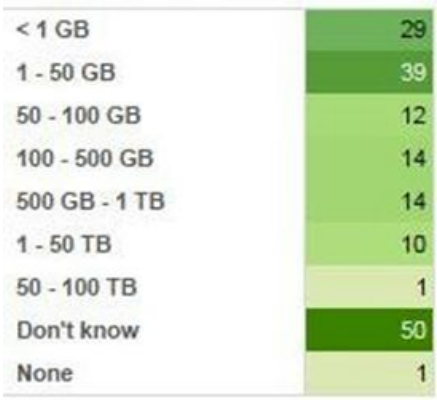


Figure 3 - Answers to the question "If your project is not yet complete, can you make an estimate of the 'final' size of your digital data?" from the 188 respondents to the 2013 online questionnaire at the University of York. From *Research Data Management at the University of York*: An audit of current practice, Jenny Mitcham, 2013

It is clear that some researchers are producing large quantities of data. There will be instances of individual files that are a large in size and instances where large numbers of smaller files are being created. We need to be aware of this when assessing solutions to manage and preserve research data.

Sensitivity of data

Many researchers hold data that is sensitive or confidential. In the RDM survey at the University of York 58% of researchers answered that they held some form of sensitive or confidential data for their current research project. This is an important issue to be aware of when planning for the management, preservation and re-use of research data. Institutions need to ensure that we are fully aware of the sensitivities and access restrictions that apply to each dataset and that we recognise that a broad brush approach is unlikely to be applicable - a single dataset may contain elements for which different levels of access should be applied.

Figure 4 - Answers to the question "Does the data that you hold for this project include any of the following categories of sensitive or confidential data?" from 188 respondents to the 2013 online questionnaire at the University of York. From Research Data Management at the University of York: An audit of current practice, Jenny Mitcham, 2013

Value of data

One of the key issues that can be hard to ascertain when thinking about the long term management of research data is that of value[43]. The value of research data is difficult to

[43] Although the focus of this project has been RDM the project team have actively sought to compare issues, scenarios, requirements and solutions with that of born-digital archives to determine commonality and the issue of value was felt to be especially true for RDM.

assess and can be very subjective. Some researchers would argue that their data has little value and limited re-use potential once their project has been completed. Other researchers may think their data is of primary importance and would be a key resource for future scholars. Some researchers won't see the potential for cross-disciplinary re-use. While there is some scope for defining different approaches to long term management based on perceived value, the bottom line is that in order to comply with funder mandates, publisher requirements and institutional policies, some data will need to be retained even if the researchers do not believe anyone will ever consult it.

The University of York RDM policy[44] mirrors the EPSRC expectations[45], assigning value based on actual re-use, or at least requests to access the data. The logic behind this is that those datasets that are seen to be in active use will be preserved and made available for as long as they are needed and those datasets that remain unused will gradually be disposed of, thus freeing up resources for more frequently accessed datasets. This strategy, although pragmatic, does risk the destruction of data whose value has not yet been discovered.

This characteristic of research data is one that is useful to note as it may have a direct impact on how data may be treated within a digital archive. Organisations may not wish to assign their limited resources to the task of preserving data for which the value is unknown but at the same time, there is a need to preserve 'valuable' datasets. There is certainly a balance to be struck here. It is clear that some processes and procedures around digital preservation would be best carried out up front at the point of data handover (for example those tasks that require a dialogue with the data creator to find out more about the file formats or request further documentation - it will become increasingly hard to get answers to these questions as time passes). However, preservation processes that are largely automated are an attractive proposition for research data.

Research data file formats and digital preservation systems

Having an understanding of the broad nature of research data is important for this project but it is also useful to explore the specifics of the file formats in use by researchers. These are files that are potentially going to need long term preservation so it is valuable to find out what they are and how they might be handled within a digital preservation system such as Archivematica.

A survey carried out at York in 2014 looked specifically at which software packages and applications researchers at the University of York use to carry out their research or create their research data. Though the focus of this particular survey was to assess software training needs for researchers, the information collected is very useful for this project in enabling us to see the software that is currently in use. From this information we can extrapolate information about the native file formats of these applications[46] and assess how they may be handled within a digital preservation system such as Archivematica.

---

[44] http://www.york.ac.uk/about/departments/support-and-admin/information-directorate/information-policy/index/research-data-management-policy/

[45] https://www.epsrc.ac.uk/about/standards/researchdata/expectations/

[46] Note disclaimer regarding this approach in Appendix 3

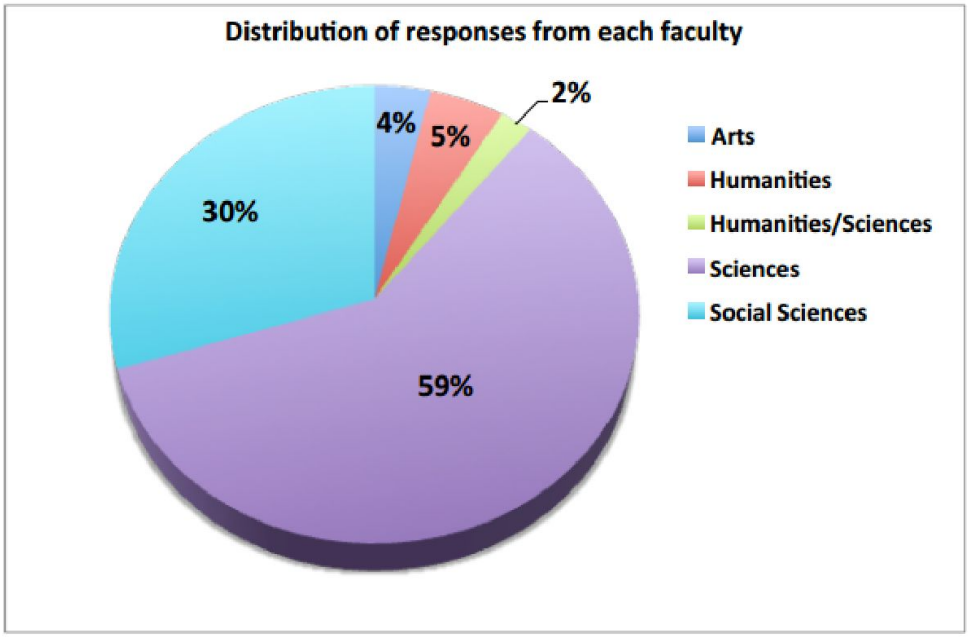


Figure 5 - Distribution of responses from each academic faculty (from 328 responses in total). From *Software and Training Questionnaire report* by Emma Barnes and Andrew Smith, 2014. Note the higher number of responses from the Sciences and low representation from Arts and Humanitites

This survey backs up the claim made above that research data file formats are likely to be very wide ranging. From the 328 responses to this questionnaire we have mentions of 260 different software applications. Figure 6 shows the top twenty applications but note what is not illustrated and perhaps most interesting is the very long tail of specialist formats that are used by only one or two respondents to the survey.

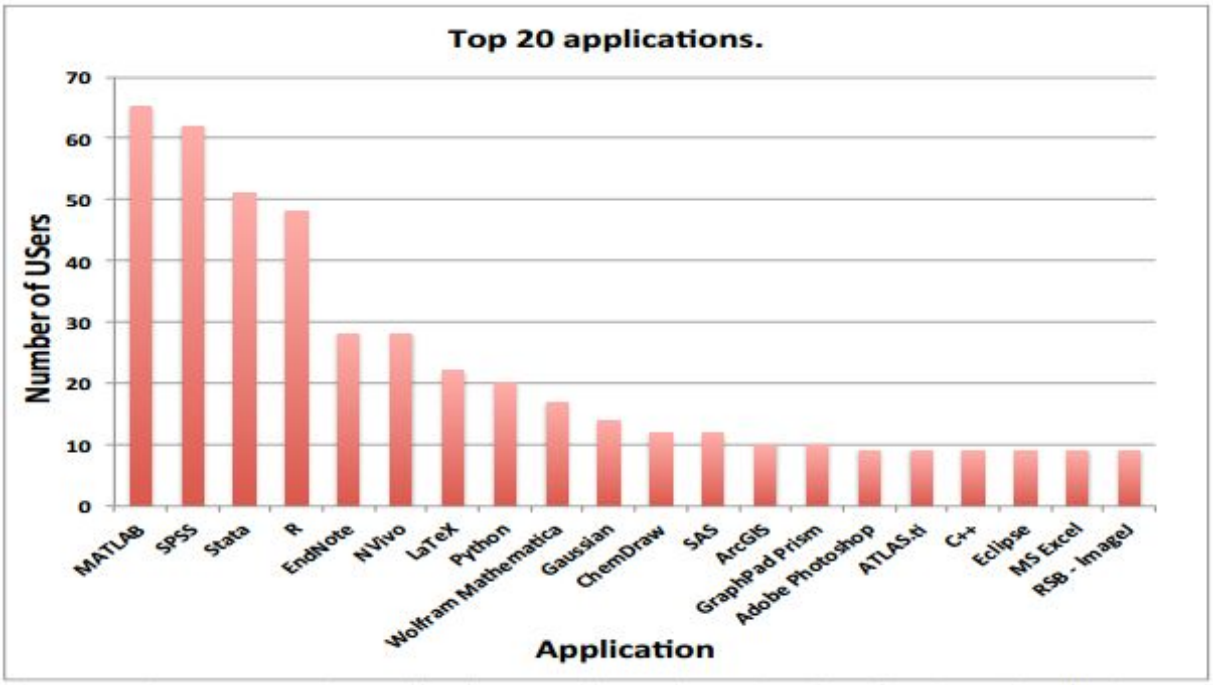


Figure 6 - The most popular applications used by at the University of York as described by the questionnaire responders. From *Software and Training Questionnaire report* by Emma Barnes and Andrew Smith, 2014.

One of the benefits to using a digital preservation system such as Archivematica is as a means to find out what types of data are held within the digital archive. Knowing what you've got is key to digital archiving and if this basic requirement isn't met, it can cause problems

further down the line. This is the case whether a migration-based strategy to preservation is used or a reliance on emulation in the future. If you don't know what type of file you have, how can you carry out preservation planning and establish a suitable format to migrate the file into? If you don't know what type of file you have, how can you find the right software to read it or locate a suitable emulation environment?

Key digital preservation standards back up this assertion. In the Open Archival Information System Reference Model[47] information about file formats should be a part of what is described as the 'Representation Information'. The Representation Information should include the information that is required in order for an end user to open and view a file and at a basic level this includes information about what file format the object is in.

In another key digital preservation standard, Audit and Certification of Trustworthy Digital Repositories (TRAC)[48], point 4.2.5.1 states that 'The repository shall have tools or methods to identify the file type of all submitted Data Objects' with supporting text that says that 'this is necessary in order to ensure that the repository's digital objects are understandable to the Designated Community'. So again this is seen as a necessary requirement.

The National Digital Stewardship Alliance (NDSA) "Levels of Digital Preservation"[49] is a helpful resource for those trying to build up digital preservation functionality from scratch. It is easier to digest than the key ISO standards mentioned above (OAIS and TRAC) and facilitates a more iterative step by step approach to setting up a digital preservation facility in bite sized chunks, acknowledging that setting up an OAIS is not something that is easily achievable in one step and TRAC compliance can be difficult to attain.

The NDSA model splits preservation activities into 4 levels, with each level moving you further towards best practice. Where file formats are concerned, coming in at level 2 is the need to have an inventory of the file formats in use within your digital archive. Tasks for level 3 and 4 build on this, asking for monitoring of file format obsolescence and performing file migrations and emulation where required. It is clear from this model that if you don't know what formats you hold, you can not move forward to the higher levels of digital preservation. Accompanying documentation stresses that the priority for organisations should be to 'get one's house in order' before moving on to migration and emulation and that the basic knowledge of what file formats you hold is key.

As part of this project we have done some initial work looking at the top 20 applications used by researchers at the University of York in order to understand how well they may be automatically identified within a digital preservation system such as Archivematica. Key to file identification in many digital preservation systems is PRONOM[50] and Archivematica is no exception. PRONOM is managed and hosted by The National Archives. On the surface it is a database of file formats, but it is of primary importance for those of us who are managing digital assets for the long term. Each file format and version that PRONOM knows about is assigned a unique identifier (PUID) and these identifiers give us the means to search and reference files of a particular type and version across different platforms and domains. File identification tools such as FIDO (which is used within Archivematica) use PRONOM signatures to try and identify the format of each file and if the signature is advanced enough,

---

[47] http://public.ccsds.org/publications/archive/650x0m2.pdf
[48] http://public.ccsds.org/publications/archive/652x0m1.pdf
[49] http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf
[50] http://www.nationalarchives.gov.uk/pronom/

more detailed information about the version of that format. Once an identification has been made, the relevant PUID will be associated with that file in the metadata held within the AIP and by Archivematica. This PUID can then be used within preservation plans to reference files of a particular type that need to treated in a particular way. PRONOM identifiers can be seen alongside formats in the Preservation Planning tab of Archivematica and the Archivematica documentation strongly encourages you to use these if creating new formats within Archivematica[51].

Assessing how well our top 20 research data applications are represented in PRONOM is a good indicator of how well they will be identified as part of the Archivematica ingest process. From the details of the applications, a list of native file formats from these applications was compiled and searches were carried out within the PRONOM website. This exercise wasn't always straightforward. Though some applications had an obvious native format that we could assume researchers were using, others had such a wide range of formats that could be input or output, it was hard to do a proper assessment. For those researchers using Adobe Photoshop for example it is impossible to know which image format they hold their data in. Though Photoshop has its own native file types, it is possible to open and save images in a wide range of other formats.

It was encouraging to see that over half of the software applications were represented in some way within PRONOM. Coverage was partial for many of the software applications, see for example Wolfram Mathematica for which PRONOM holds an entry for the native notebook (.nb) format but not for other types of file which internet research suggests can be output from the software (for example .nbp which appears to be a legacy notebook format). However, perhaps in this instance we can assume that for current research data we are less likely to encounter legacy formats.

Of the range of formats that can be created by ArcGIS, 12 of these were represented in PRONOM with .mxd being one notable gap in the coverage. Mxd files are project files for ArcGIS and as detailed by the Archaeology Data Service (ADS) Guide to Good Practice[52] are not well suited to long term preservation, however it is still likely that researchers will deposit files such as these in their final datasets so we do still need to be able to identify them in an automated fashion at the point of ingest.

The only application in the top 20 list that appeared to be fully represented by PRONOM was Microsoft Excel with 9 different versions of the native spreadsheet format listed.

Fuller details of the assessment of PRONOM against our top 20 research data applications is available in Appendix 3.

Early on in the project it was decided that we should feed into the ongoing development of PRONOM to ensure that more research data formats will be automatically identified. As described on the PRONOM website, they are keen to accept new information to help enhance the available resource and to allow for the identification of a wider range of formats [53]. To help feed into this process we have been collecting sample data from researchers focusing on our top 20 applications and sharing these files with The National Archives to aid the creation of file signatures. We are grateful to those researchers who volunteered to

[51] https://www.archivematica.org/en/docs/fpr/#updating
[52] http://guides.archaeologydataservice.ac.uk/g2gp/Gis_3-1
[53] http://www.nationalarchives.gov.uk/aboutapps/PRONOM/default.htm#new_info

share sample data with us and also to staff at The National Archives who have agreed to spend some time working with this data.

Clearly this work is valuable as a one off, but far more so if it can be seen as an ongoing concern. Assessment has not been carried out of the 'long tail' of file formats (240 of which were not illustrated in figure 6) and as mentioned in previously file formats for research data are likely to be subject to frequent change. It is clear that as we start to ingest research data into a digital archive we will encounter many files that will not be automatically identified. As we move towards a proof of concept for archiving research data we should continue to engage with the team who maintain PRONOM and promote discussions within the wider digital preservation community about a sustainable or more automated way to address this problem.

Archivematica testing

The "Archivematica process" has been tested, using a "lightweight", rather than production-capable, technology stack (see appendix 2) against a number of different scenarios and with a range of file types, both common and unusual. The scenarios include processing:

- single files in a wide variety of file types and sizes
- multiple files of a single file type
- multiple files of mixed file type
- multiple, mixed files in a directory structure
- multiple, mixed files in a directory structure which was rearranged and edited during ingest

Single files in a wide variety of file types and sizes

A variety of files of differing file type and size were processed through Archivematica. The file types included a number of common varieties and a number of obscure ones. Archivematica coped with these well even correctly identifying some files from a specialist simulation package.

Archivematica was tested with a range of file sizes from a few hundred bytes up to approximately 2GB. Problems started to emerge when the files reached 500MB; processing an mpeg file of this size Archivematica produced a DIP successfully but failed in the process of creating an AIP although a 1GB wav file completed successfully. Correspondence on the Archivematica user forum[54] suggests this may be due to a time-out problem and that the time-out period can be adjusted. Some contributors have suggested that Archivematica's use of the checksum standard SHA-512 is very time consuming and that less thorough, but 'good enough', standards such as MD5 might help the situation[55]. (In our 'Future work' section we suggest that the ability for a user to choose their checksum standard might make a good, new feature.) Testing with a 2GB mpeg file Archivematica would not even start

54 https://groups.google.com/forum/#!searchin/archivematica/large$20files/archivematica/NLD5o-n4pQw/nc9hLF4lOggJ

55 https://groups.google.com/forum/?fromgroups#!searchin/archivematica/md5/archivematica/NLD5o-n4pQw/wQMezSSjN0sJ

the ingest process. At this early stage in the project, with a lightweight technology stack and limited technical support, these issues were not further investigated and illustrate that Archivematica is not 'out-of-the-box' but requires fine tuning. It is known, however, that there are Archivematica users successfully processing files of up to 500GB.

Multiple files of a single file type

With file sizes up to a few megabytes this process worked well.

Multiple files of mixed file type

With file sizes up to a few megabytes this process worked well.

Multiple, mixed files in a directory structure

With file sizes up to a few megabytes this process worked well. The directory structure is not reflected in the 'objects' directory of the DIP merely in the METS file.

Large numbers of files

We did not test Archivematica with a very large number of files in a single ingest; it was felt that we could not carry out a meaningful experiment on a server stack designed only for lightweight testing. Correspondence on the Archivematica community mailing list suggests that Archivematica should, in principle, deal with an ingest of any size[56].

Multiple, mixed files in a directory structure which was rearranged and edited during ingest

With file sizes up to a few megabytes this process worked well. Again, the directory structure is not reflected in the 'objects' directory of the DIP merely in the METS file.

Metadata and rights

Dublin Core (DC) metadata and rights information can be added during the early stages of Archivematica's workflow although the documentation makes plain that this is not intended for use "as-is" in an access environment. The rights information and DC elements are added to the METS file in both AIPs and DIPs. Any DC metadata is additionally added to the AIP serialised as json.

RDM workflows and Archivematica use

Whilst the main function of Archivematica is to aid in the preservation of data, the preservation functions need to be seen in the context of wider workflows that may also involve making the data accessible to users, perhaps via an institutional repository. There seem to be three essential scenarios for using Archivematica in conjunction with such a delivery system.

[56] https://groups.google.com/forum/?fromgroups#!topic/archivematica/ZXa66P_yYIY

The first scenario would have Archivematica taking lead place in the workflow and being used to generate AIPs as the basis for preservation and possibly DIPs as the basis for any delivery surrogates that an institution might wish to make available through its repository.[57] This might be illustrated as follows:

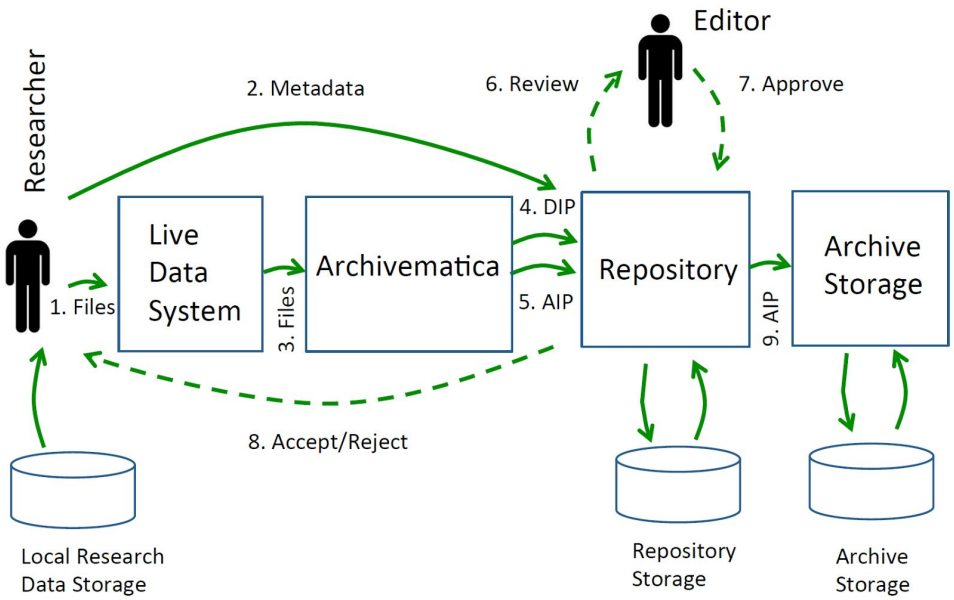
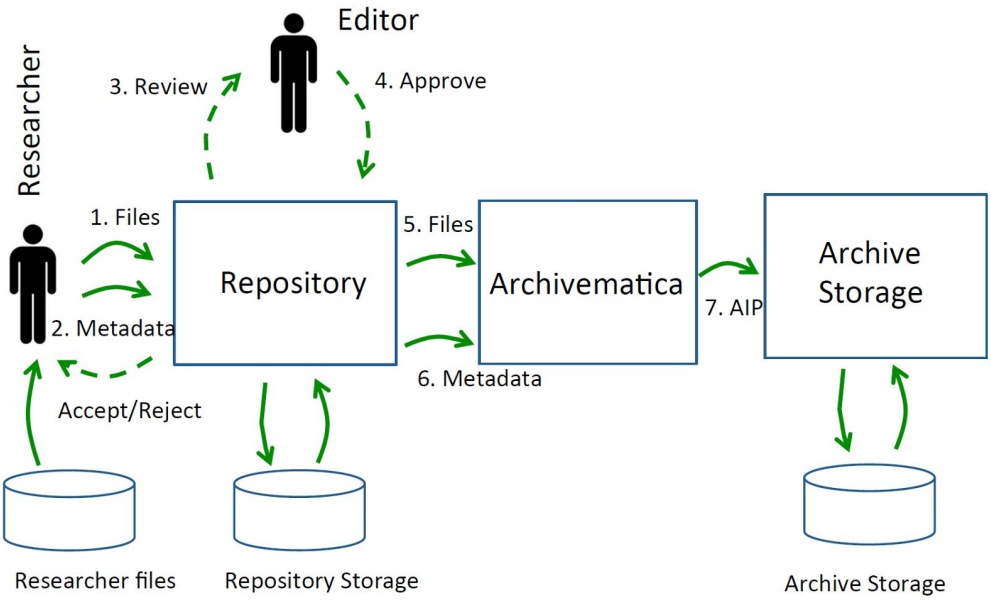


Figure 7 - A possible "Archivematica first" repository-integrated workflow. Source: Matthew Addis[58]

The second scenario has the research data first placed in an institutional repository and then, in some sense exported into, or imported by, Archivematica with the intention of generating AIPs for safe-keeping - in effect, a dark archive. In this case DIPs would have no apparent use as the material for Archivematica processing has come from a dissemination platform.



[57] It is worth mentioning that this is the approach used when Archivematica is used in conjunction with AtoM, Artefactual's own archival description software

[58] Matthew Addis's diagrams are taken from Addis, M. *Archivematica RDM workflows* http://dx.doi.org/10.6084/m9.figshare.1476856

Figure 8 - A possible “repository first” integrated workflow. Source: Matthew Addis

A third scenario is intermediate between these two and envisages a situation where some wider workflow makes a call out to Archivematica as an intermediate stage in processing. For example:

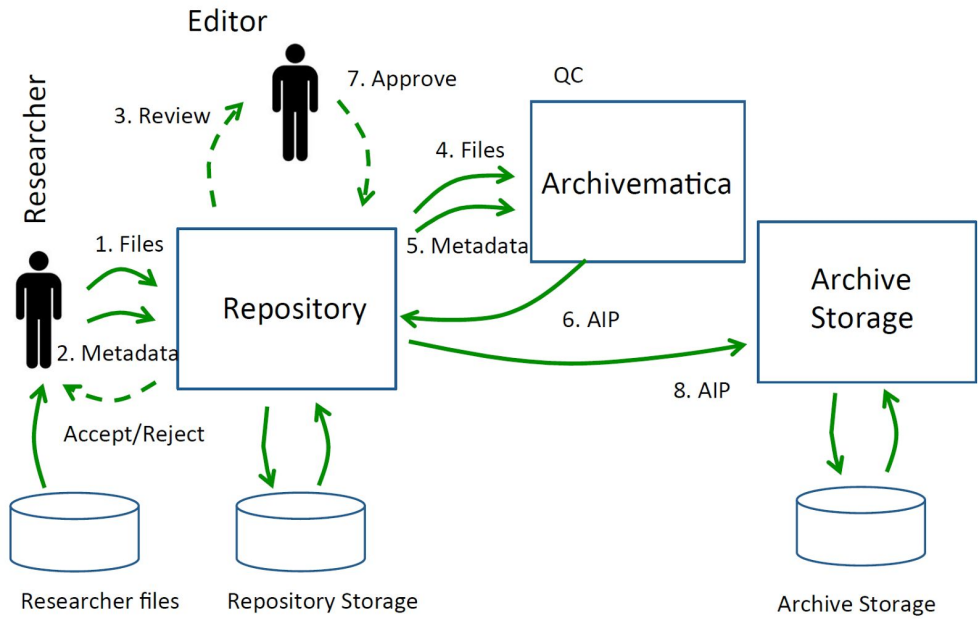


Figure 9 - A possible repository-integrated workflow in which Archivematica is employed as an intermediate stage. Source: Matthew Addis

Of course, there will be a number of subtle variations on these three, rather crude, characterisations. Note, in particular, that in the first of Matthew Addis’s diagrams above it is not a necessity that the AIPs be routed through the repository on their way to archive storage.

Specifically in respect of the second scenario above, Archivematica offers a number of options by which the data for processing may be provided; one of these is to provide for a “DSpace export”[59]. The Archivematica manual notes that “Archivematica can act to facilitate a “dark archive” for a DSpace repository - i.e. providing back-end preservation functionality while DSpace remains the user deposit and access system.” In this scenario, a user exports data from DSpace via its inbuilt export function and the package thus created is imported to Archivematica. We understand that there is some work being carried out in the US to provide similar functionality in respect of a Hydra repository. It is worth a note of caution about a ‘standard’ Hydra integration as whilst DSpace ships as a package and therefore there is some level of uniformity between one installation and another, Hydra is essentially a toolkit which users have employed to build a repository to their own design, on top of the Fedora repository.

More generally, it would be useful to further explore options for integration of Archivematica with other repository platforms (Hydra, Fedora, Eprints etc), including options for both custom integrations and API-based solutions.

Archivematica’s workflow

[59] https://www.archivematica.org/en/docs/archivematica-1.4/user-manual/transfer/dspace/

The first stage in the Archivematica workflow, called “transfer”, requires the user to select a directory containing the files or files to be processed; the directory chosen can be the top level of a tree. The process undertakes a number of preliminary actions including a virus check and initial file format verification before creating a SIP and placing it in the ingest queue. A universally unique identifier (UUID) is created for this transfer but it does not follow the data through the system; however any AIP created from it is given a further UUID which is reflected in the content of any associated DIP - allowing them to be linked.

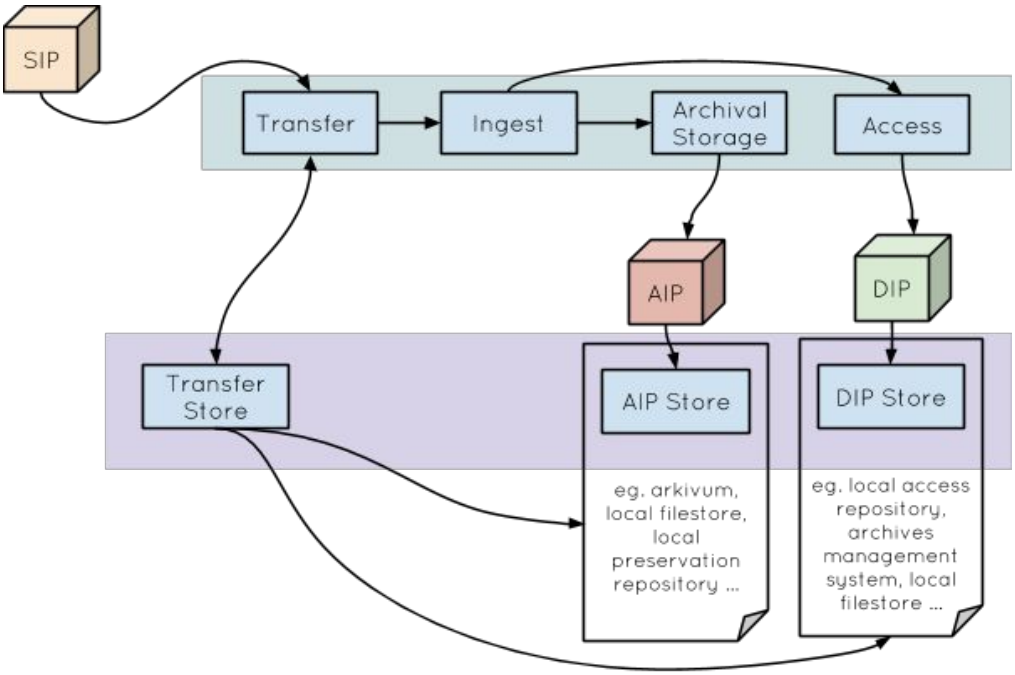


Figure 10 - Archivematica’s standard transfer. Source: Julie Allinson

Optionally, Archivematica version 1.3.x, as tested in Hull and York, allows the contents of a SIP to be edited and rearranged prior to ingest. The files are first transferred in whatever directory structure they may have but then sent to the “transfer backlog” instead of forward for processing into an AIP and/or DIP. The tree structure can then be retrieved by the user and individual files can be chosen and passed into a new directory structure of his/her design. Effectively this allows some archival arrangement of the content. Once this process is complete, the new tree structure is passed forward for processing. The structure in the transfer backlog is retained (and so further work could be done on it) with files that have been sent forward to processing greyed out.

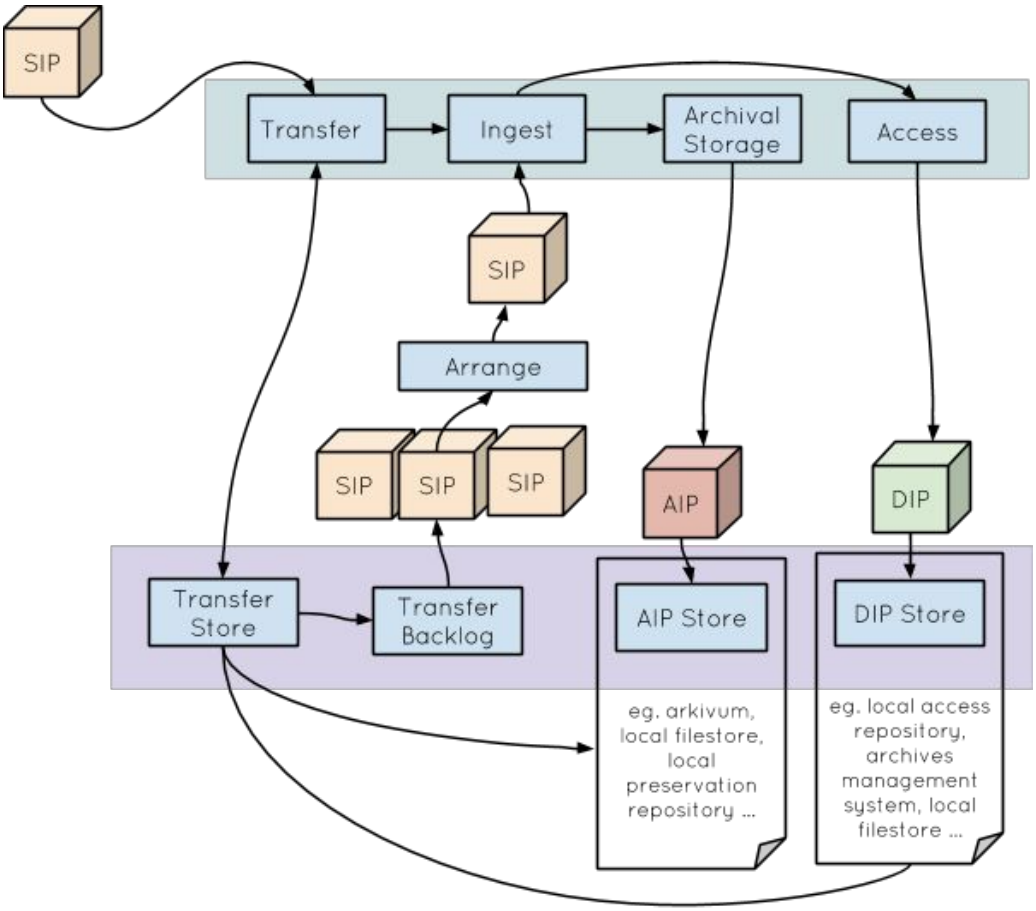


Figure 11 - Archivematica's SIP arrangement. Source: Julie Allinson

Also optionally, DC metadata and/or rights information can be added to the SIP using templates provided; any such metadata is passed into the AIP and DIP METS files. DC is separately represented as a json file in the AIP.

The next stage, the "ingest" process offers a number of options for normalizing (sic) the files for preservation or for preservation and access. The end result of using the automated processes should be an AIP and, if requested, a DIP. The AIP is structured in conformance with the widely-adopted BagIt[60] file packaging format.

Archivematica AIPs consist of a number of elements contained (by default) in a 7-zip file.

- in a "data" directory are logs, the object(s) for preservation, matching thumbnails and a METS file
- two files related to the BagIt process
- two manifests; one containing checksums for the preservation objects and one containing checksums for the bagit process files

Archivematica DIPs consist of three elements bundled together in a .tar file:

- a folder containing the delivery surrogates
- a folder containing thumbnail images matched to each of the delivery surrogates
- a METS file

[60] http://www.dcc.ac.uk/resources/external/bagit-library

If any of the delivery surrogates is an image, the thumbnail is a reduced version of this, otherwise it is an icon. It would be possible to take a DIP from Archivematica and from it build a repository object for dissemination. The focus of interest for the Universities of Hull and York would be to build Hydra objects, but the essential building blocks present would allow content for other repository systems to be built too. It seems likely that the most productive approach would be to run scripts against the Archivematica DIP which employ Hydra gems to construct an object at an appropriate point in the local repository workflow and that detailed descriptive metadata should be added as necessary before the object is made available for dissemination. (Whilst Archivematica can record simple Dublin Core metadata and make this available via the METS file in the AIP and any DIP, such metadata may well be insufficient in a repository context.) If a Phase 2 should be funded for this project, that is certainly the approach that team members in Hull would wish to follow.

Archivematica is natively capable of preparing DIPs for upload to AtoM, CONTENTdm, and of transferring metadata to Archivists' Toolkit for use with a web based access system. As Archivematica's users are, currently, primarily to be found in the US and Canada it is perhaps not surprising that there is no provision for use with the EPrints software; this potential functionality would not seem a major challenge should a demand arise.

A forthcoming version of Archivematica (1.5) offers the facility to reingest an AIP from store and to reprocess it. This may be useful in an RDM context if particular research data was not processed for a dissemination system (for example, a repository) during its initial ingest but is later found to be required. Its AIP can be retrieved from the preservation store and a DIP created from it. We understand that Artefactual Systems have code which may assist in unpacking the accompanying METS file to help populate metadata for the dissemination materials subsequently created.

How would Archivematica handle research data?

Archivematica can handle any type of data but it should be noted that the level to which it will carry out automated preservation actions are dependent on the file formats ingested. There are three different levels as highlighted below:

Level of curation within Archivematica	**Description of how files are handled within Archivematica**	**Example file formats**
Full	Archivematica identifies the file format and has a normalisation strategy for preservation as defined in the format policy registry. It records information about the file format and carries out the necessary normalisations for preservation, creating PREMIS metadata and storing that alongside the files in the final AIP. And other preservation actions as	● JPEG image (would be normalised to a TIFF format) ● PDF file (would be normalised to a PDF/A) ● MPEG (would be normalised to a WAV file) ● EPS files (would be normalised to SVG)

	described below.	
Mid	Archivematica identifies the file format but there is no normalisation strategy defined in the format policy registry. No normalisation is carried out but the file (along with metadata including the information about what it is) is packaged up into an AIP And other preservation actions as described below. The institution may choose to supplement this level of curation with manual normalisation actions targeted at particular types of files perceived to be at risk.	● Microsoft Word document ● Excel spreadsheet ● SAS files ● ArcGIS files ● LaTeX files ● Websites
Basic	Archivematica doesn't recognise the file format but is still able to carry out the base level of preservation actions (for example virus checking, cleaning up file names, creating checksums) and package the data and any metadata generated into an AIP. The institution may choose to supplement this level of curation by carrying out further investigation into those files that could not be identified. This may involve checking the data management plan, speaking to the researcher, actively enhancing the tools for file identification by submitting information to PRONOM or adding support for unknown formats into Archivematica directly.	● MATLAB files ● STATA files ● Gaussian files

Table 2: The different levels of support within Archivematica depending on file format

Archivematica can handle large files (or large volumes of small files) but its abilities in this area are very much dependent on the processing power that has been allocated to it. Users of Archivematica should be aware of the capabilities of their own implementation and be prepared to establish a cut off point over which data files of a certain size may not be processed, or may need to be processed in a different way.

Archivematica uses the PREMIS metadata standard to record rights metadata. Rights metadata can be added for the Submission Information Package as a whole rather than in a granular fashion. This is not ideal for research data for which there are likely to be different levels of sensitivity for different elements within the final submitted dataset. The Archivematica manual suggests that fuller rights information would be added to the access system (outside of Archivematica)[61].

The use of Archival Information Collections (AICs) in Archivematica enables the loose association of groups of related Archival Information Packages (AIPs). This may be a useful feature for research data where different versions of a dataset or parts of a dataset are deposited at different times but are all associated with the same research project.

Archivematica is a suitable tool for preserving data of unknown value. Workflows within Archivematica and the processing of a transferred dataset from a Submission Information Package (SIP) to an Archival Information Package (AIP) can be automated. This means that some control over the data and a level of confidence that the data is being looked after adequately can be gained, without expending a large amount of staff time on curating the data in a manual fashion. If the value of the data is seen to increase (by a request to access that data for example) further efforts can be made to preserve the data using the forthcoming AIP re-ingest feature[62] which would enable the retrieval of the AIP for further processing, for example to create a Dissemination Information Package (DIP). Manual curation can also be carried out during this re-ingest process if deemed necessary.

The level of automation within Archivematica can be configured so staff are able to treat datasets in different ways as appropriate. Different institutions may have varying approaches here, but the levels of automation that are possible provide a compelling argument for the adoption of Archivematica if few staff resources are available for more manual preservation interventions.

Technical analysis

Archivematica is a free and open source application. It comprises two components: the storage service (latest version 0.7) and Archivematica (latest version 1.4). There are at least three different approaches to deploying Archivematica:

1. Full, local installation - institutions may install and host Archivematica locally and link it to their preferred storage option
2. Arkivum[63] - a managed service from Arkivum will allow Archivematica to be hosted locally within the institution with upgrades and support available through Arkivum in partnership with Artefactual. A remote hosting option is also available. Both include integration of Archivematica and Arkivum storage.
3. ArchivesDirect[64] - a hosted service from DuraSpace that combines Archivematica's preservation functionality with DuraCloud for storage

In the case of a locally installed and maintained production system, Archivematica recommend a minimum server configuration of

- Processor: dual core i5 3rd generation CPU or better

[61] https://www.archivematica.org/en/docs/archivematica-1.4/user-manual/ingest/ingest/#add-rights
[62] Scheduled for version 1.5 of Archivematica and due for release in July 2015
[63] http://arkivum.com/
[64] http://www.archivesdirect.org/

- Memory: 8GB+
- Disk space: 20GB plus the disk space required for the collection

A less well specified machine can be used for testing and evaluation. The software can be installed on a single machine or across a number of machines to share the workload. At the time of writing, the software requires a current version of Ubuntu LTS (14.04 or 12.04) as its operating system.

Storage Service

The Storage Service[65] is separate to the core Archivematica software and is used by Archivematica to manage storage locations. The Storage Service does not need to reside on the same server as Archivematica and, indeed, can serve multiple Archivematica installations (called Pipelines).

Each Pipeline has a number of Spaces. Each Space has an 'access protocol' for where data is stored. These include LOCKSS and DuraCloud, along with local and NFS filestore. Spaces contain Locations for the different types of package (Transfers, DIPs, AIPs etc.).

Another job of the Storage Service is to maintain a list of all of the packages in the system and to handle deletion requests.

The Storage Service has an in-built API, documented within the code[66]:

```
List (api/v1/file/) supports:
GET: List of files
POST: Create new Package

Detail (api/v1/file/<uuid>/) supports:
GET: Get details on a specific file

Download package (/api/v1/file/<uuid>/download/) supports:
GET: Get package as download

Extract file (/api/v1/file/<uuid>/extract_file/) supports:
GET: Extract file from package (param "relative_path_to_file" specifies which file)

api/v1/file/<uuid>/delete_aip/ supports:
POST: Create a delete request for that AIP.

Validate fixity (api/v1/file/<uuid>/check_fixity/) supports:
GET: Scan package for fixity
```

Archivematica

Archivematica is the core software. It's functionality is largely driven by its web-based dashboard, with the background digital preservation activities co-ordinated by a suite of

[65] https://ww.archivematica.org/en/docs/storage-service-0.7

[66] https://github.com/artefactual/archivematica-storage-service/blob/stable/0.7.x/storage_service/locations/api/resources.py#L390

pluggable micro-services. The format policy registry (fpr) allows users to define define format policies for handling different file formats. It also has a rest API which can do two things:

- Approve a transfer in a specified 'watch' directory.
- List unapproved transfers[67].

This enables the automation and monitoring of the initial transfer into Archivematica's workflow. Metadata can also be supplied with this transfer.

It has been acknowledged by the community that there is a lack of developer documentation for Archivematica and that it would be very helpful, especially to new users, if this could be addressed. (This matter is mentioned again towards the end of the next section, 'Future work'.

Future work

During phase one of this project we have been testing Archivematica and assessing its capabilities for preserving research data. In our view, Archivematica represents a good solution, the more so because through community input and support, it is continually evolving. In the specific context of RDM needs, there are still areas where improvements could be made and we have discussed these with the wider UK Archivematica user community[68], with other international Archivematica users and with Artefactual Systems. For several of these areas, developments are already in progress (having been prioritised and sponsored by other institutions). One area of particular interest to us is the AIP re-ingest functionality[69] currently being sponsored by the Zuse Institute Berlin. This new functionality (available in version 1.5 and due for release soon) is key to some of the workflows that we envisage putting in place for the preservation of research data as it will enable an AIP to be re-processed by Archivematica for example to create a DIP on request, we have highlighted some further areas of development to enable this to be carried out in a more automated fashion.

Another area we saw as a high priority for development work was the addition of reporting functionality within Archivematica. Archivematica is very much a back-end system for processing and packaging data for preservation. It does not currently give you statistics about what file formats you hold or how many AIPs you have ingested in a certain month for example[70]. We have found several ongoing developments in this area which may help fulfil this need.

- Binder[71], a new open source system created by Artefactual Systems and the Museum of Modern Art (MOMA) was released in May 2015[72] and is designed to be

---

67 https://www.archivematica.org/en/docs/archivematica-1.4/user-manual/administer/dashboard-admin/#rest-api

68 In a presentation given to the UK Archivematica group in June 2015 we discussed our ideas for sponsored developments and asked the group to complete a feedback form to tell us how useful these developments would be to them for their own proposed Archivematica implementations.

69 https://wiki.archivematica.org/AIP_re-ingest

70 Though it is possible to extract this data by querying Elasticsearch

71 http://binder.readthedocs.org/en/latest/

72 https://groups.google.com/forum/?fromgroups#!searchin/archivematica/binder/archivematica/FXsRgbrBywE/WGTHfkDB-rgJ

used alongside Archivematica and AtoM. The project team has not had the opportunity to explore this in any real detail, but a short overview of this system[73] shows some powerful reporting capability on the datasets and individual files that have been ingested into Archivematica. This system has great potential but in its current form is tailored to MOMA's own infrastructure and also requires integration with a specific museum collections system in order to function correctly.

- The Bentley Historical Library at the University of Michigan are currently working on a Mellon funded project through which they are sponsoring an Appraisal and Arrangement tab[74] within Archivematica which will be made available in the first half of 2016. This will provide more detailed and graphical reporting on the contents of a particular transfer prior to ingest. It will be possible for example to see the contents of the transfer broken down by MIME type, file format, file size etc and this information can be used to help inform decisions about the arrangement of SIP.
- DMAonline[75] is a Jisc funded Research Data Spring project led by Lancaster University. The project team is developing an administrative dashboard for research data management to give an overview of how an institution is performing in terms of its compliance with policies, and other important metrics. This dashboard view will pull statistics from a number of different systems related to research data management and could be used to give managerial, research support or curatorial staff an overview of performance and progress.

We wish to explore these new features and developments in more detail in phases two and three of the project.

There remain some high priority issues that we would like to work with Artefactual Systems to solve during phase two of the project and these are listed in the table below:

Development priority	**Benefits for research data management**
Automation of DIP generation on request - building on the AIP re-ingest functionality within Archivematica which allows an AIP to be re-processed and allows for a delay in the generation of a DIP until such a time as it is requested, this feature will enable further automation of this process.	This feature is of particular benefit to those situations where the value of data is not fully understood. It is unnecessary to create an access copy of all research datasets as some of them will never be requested. In our workflows for long term management of research data we would like to trigger the creation of a copy of the data for dissemination and re-use on request rather than create one by default.
METS parsing tools - creating a Python library which could be used by third party applications to understand the METS file that is contained within an Archivematica DIP. Additionally an HTTP REST service would be developed to allow third party applications to interact with the library in a programming language agnostic fashion.	This is key to being able to work with the DIP that is created by Archivematica within other repository or access systems that aren't integrated with Archivematica. Both York and Hull have repositories built with Fedora and Hydra and this feature will allow the repositories to better understand the DIP that Archivematica creates. This development is in no way specific to a Fedora/Hydra repository and will equally

[73] https://www.youtube.com/watch?v=TelwvLkt-84&feature=youtu.be
[74] https://wiki.archivematica.org/Appraisal_Arrangement_tab
[75] http://www.dmao.info/

	benefit other repository platforms in use for RDM.
Improved file identification - an initial feature will enable Archivematica to report on any unidentified files within a transfer alongside access to the file identification tool output. Further enhancements could help curatorial staff to submit information to PRONOM by partially automating this process.	It has been highlighted in this report that the identification of research data file formats is a key area of concern when managing research data for the longer term. This feature will help users of Archivematica see which files haven't been identified and thus enable them to take action to establish what they hold. This feature will also encourage communication with PRONOM to enhance the database of file formats for the future, thus enabling a more sustainable community approach to addressing this problem.
Generic Search API - the development of a proof of concept Search REST API for Archivematica allowing third party applications to query Archivematica for information about the objects in archival storage.	There is a need to be able to produce statistics or reports on RDM processes in order to obtain a clear picture of what data has been archived. This development will enable these statistics to be generated more easily and sustainably. For example this would enable tools such as the DMAonline dashboard in development at Lancaster University to pull out summary statistics from Archivematica.
Support for multiple checksum algorithms - Currently Archivematica generates sha256 checksums for all files, and inserts those into premis fixity tags in the METS file. In addition, 2 premis:events are generated for each file. All 3 of these entries are currently hardcoded to assume SHA256. This development would include support for other hash algorithms such as MD5, SHA1 and SHA512	Research data files can be large in size and/or quantity and may take some time to process through the Archivematica pipeline. One of the potential bottlenecks highlighted in the pipeline is checksums which are created at more than one point in the process. SHA256 checksums can take a long time to create and it has been highlighted that having the option to alter the checksum algorithm within Archivematica could speed things up. Having additional configuration options within Archivematica will help institutions to refine their pipelines and reduce bottlenecks as appropriate.
Documentation - a specific piece of work to improve and enhance Archivematica documentation for developers, particularly with regard to their automation tools project.	The automation tools project has been used by a handful of Archivematica users to fully automate an Archivematica pipeline. As mentioned in this report, the ability to automate processes relating to preservation are a key consideration where few resources are available to manually process data of unknown value. Fuller documentation of how an automated workflow can be configured within Archivematica using the APIs that exist would be very helpful for those considering

	using Archivematica for RDM.

Table 3: Areas highlighted for development work and rationale

Developments will be carried out by Artefactual Systems using an agile methodology and they will work closely with the Universities of Hull and York to ensure developments are meeting the requirements of the project. Artefactual Systems and the project team are particularly keen to work on features that have a greater benefit to the wider Archivematica community and development plans are created with this in mind. Costs for development work include a community support fee of 10% which supports the addition of new features to a subsequent public release of the Archivematica. It helps cover the costs of merging the code into a public release branch, ensuring that the features are maintained over time and is also used to providing public documentation and user support for the features.

Whilst Artefactual Systems carries out the necessary development work during phase two, ongoing work at Hull and York will further consider how Archivematica might fit into our own workflows for research data and will include detailed planning for our proof of concept installations in phase three.

Conclusion

Archivematica offers a practical and affordable solution for digital preservation in an RDM context, effectively “filling the digital preservation gap” that was highlighted in the introduction of this report. Archivematica is based on international digital preservation standards. Its flexibility enables it to be configured for many different workflows, increasing its suitability for use by many institutions with varying workflows and requirements. Workflows can be largely automated or a more manual approach tailored to the dataset in question can be taken. Archivematica is an evolving and continually developing system with a growing international user base. It can be installed locally or hosted externally and it can be integrated with a variety of different access and storage systems.

Archivematica is just one piece of the wider digital preservation landscape. It is not an isolated solution and much of its appeal comes from the fact that it includes many other open source tools that perform specific data curation functions within it. Archivematica is greater than the sum of its parts and will grow and develop as these tools evolve and new tools emerge and are incorporated into it.

Archivematica users should engage with the wider digital preservation community to contribute to and enhance the tools and resources that are available. Submitting information about file formats to PRONOM is one way of doing so and has been highlighted as a key factor that we should address to enable the identification of research data file formats, many of which are currently not recognised by file identification tools. Sharing workflows and digital preservation use cases is equally valuable and this is something we hope to do more of as we continue our work with Archivematica and establish a proof of concept implementation for the preservation of research data.

Glossary

ADS: The Archaeology Data Service

AIP: Archival Information Package - processed information sent to the archival store for preservation

Archivists' Toolkit: An archival data management system[76]

AtoM: AtoM (or Access to Memory) is Artefactual Systems' own archival description software which can be used to put archival holdings online

CONTENTdm: A software solution from the Online Computer Library Center (OCLC) allowing digital collections to be made available across the web[77]

Dark archive: In reference to data storage, an archive that cannot be accessed by any users. Access to the data is either limited to a set few individuals or completely restricted to all. (Webopedia 2015-05-19)

Delivery surrogate: Generally a copy of a file provided for delivery purposes so that the original file need not be accessed

DC: (in the context of this report) Dublin Core metadata

DIP: Dissemination Information Package - information created from the material being archived intended for sending to a user

DSpace: A widely adopted, community open source, institutional repository solution now stewarded by the DuraSpace organisation[78]

DuraSpace: A not-for-profit organisation in the USA which stewards, amongst other products, Fedora and DSpace

EPrints: A well-adopted institutional repository solution in use mainly within the UK and Western Europe more generally [79]

EPSRC: Engineering and Physical Sciences Research Council

Fedora: (in the context of this report) An open-source digital repository platform[80]

FIDO: Format Identification for Digital Objects - A command-line tool to identify the file formats of digital objects.

---

[76] http://www.archiviststoolkit.org/
[77] http://www.contentdm.org/
[78] http://www.dspace.org/
[79] http://www.eprints.org/uk/
[80] http://www.fedora-commons.org/

FPR: Format Policy Registry - within Archivematica the FPR defines the actions, tools and settings to apply to a file of a particular file format (e.g. conversion to preservation format, conversion to access format)

Hydra: A repository solution based on a number of “best-of-breed” open-source components, including Fedora[81]

METS: The METS metadata schema is a widely adopted standard for encoding descriptive, administrative, and structural metadata

NDSA: The National Digital Stewardship Alliance

Normalisation: The process of converting ingested objects into a small number of pre-selected formats in order to make them more suitable for preservation or access

OAIS: Open Archival Information System

PREMIS: The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability

PRONOM: A resource provided by the National Archives in the UK providing definitive information about file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value.

PUID: A persistent unique identifier. In the context of this report used by PRONOM to identify unique file format signatures

RDM: Research Data Management

SIP: Submission Information Package - information sent from its producer for archiving

Sufia: A Hydra Ruby gem which provides basic functionality for a form of institutional repository

TRAC: *Trustworthy Repositories Audit & Certification: Criteria and Checklist*. Version 1.0. Chicago: CRL, February 2007.

UUID: a universally unique identifier

[81] http://projecthydra.org/

Appendix 1

Digital preservation requirements for research data management

The following list of requirements for a digital archiving system have been adapted for this project from a set of requirements originally created for the University of York in 2013 and subsequently published as a blog post[82]. The list of requirements describes what is required of a multi-purpose digital preservation system, one of the use cases being Research Data Management. It is recognised that many institutions requiring digital preservation functionality will have other use cases to consider alongside their need to preserve and manage research data for the long term.

These requirements focus on the basic functionality of a digital archiving system. In keeping data for the long term we need to ensure we have a suitable infrastructure to ensure we can keep the data in a format that remains usable over time. A migration-based approach as described in the Open Archival Information System[83] (OAIS) reference model is the most appropriate method of ensuring this need can be met but will not be feasible for all types of data, thus the system should also support future emulation strategies.

In order to tie these requirements firmly to the internationally agreed standards in the area of digital preservation, headings taken from the OAIS functional entities are used in the table below (with the omission of 'Access') and standard OAIS terminology is used (for example around the information packages associated with a digital archive). Where a requirement ties in closely with ISO16363 (Trusted Repositories: Audit and Certification)[84] a reference number has been included.

Although access is a key part of the OAIS model, requirements for this area have intentionally been omitted. It is envisaged that for many institutions, a repository infrastructure or other system enabling data access will already be in place but back end digital preservation will remain a key area to be addressed.

As well as articulating our requirements, the table below also includes information about how Archivematica meets these requirements and this can be used to assess its suitability as a tool for digital preservation within the research data management space.

Note that a digital preservation solution such as Archivematica is envisaged to be part of a wider infrastructure for managing research data for the long term and where Archivematica does not meet specific requirements it is likely that this functionality may be performed elsewhere within the infrastructure.

Key:

- Green - Archivematica fulfils this requirement
- Orange - Archivematica partially fulfils this requirement
- Pink - Archivematica does not fulfil this requirement

[82] http://digital-archiving.blogspot.co.uk/2013/12/my-digital-preservation-christmas-wish.html
[83] http://public.ccsds.org/publications/archive/650x0m2.pdf
[84] http://public.ccsds.org/publications/archive/652x0m1.pdf

#	Requirement	ISO 16363 ref	M – must S – should C – could
INGEST			
I1	The digital archive will enable us to store administrative information relating to the Submission Information Package (SIP) (information and correspondence relating to receipt of the SIP).	4.1.8	S
Within Archivematica a transfer can be made with submission documentation and this will be preserved within the AIP. Note that submission information as described in the Archivematica wiki can be “donor agreements, transfer forms, copyright agreements and any correspondence or other documentation relating to the transfer”. For research data management, submission information may include a data management plan.			
I2	The digital archive will include a means for recording appraisal decisions relating to the Submission Information Package and individual elements within it.	4.2.3.1	S
Not entirely within scope for Archivematica but as we would not expect anyone to be using this system in isolation, this information may be best stored elsewhere within the technical infrastructure. For many institutions managing research data, the selection and appraisal decisions will be carried out by the researchers themselves prior to submission and details on which of their data they intend to retain will be recorded within a data management plan. If selection of data is carried out within Archivematica after data transfer, the pre-ingest step allows you to select elements of a transfer to create an SIP for ingest. Documentation of the original transfer structure is maintained in the AIP so it is possible to refer back to this to see what was disposed of prior to ingest.			
I3	The digital archive will be able to identify and characterise data objects (where appropriate tools exist).	4.2.5.1	M
Within Archivematica file identification is an automated process. File identification is one of the microservices carried out as part of the ingest process. Output is stored in the METS and PREMIS XML within the AIP and this includes the unique PRONOM ids of the recognised formats and information about the identification tool and version used. Note that Archivematica can only be as good as the file identification tools used. More specialist file formats that are less widely used may not be identified by the file identification tools but this is something we can actively address over time. This issue is addressed elsewhere in this report.			
I4	The digital archive will be able to validate files (where appropriate tools exist).		S
Validation tools such as JHOVE are run as microservices within Archivematica during ingest. The output is stored in the METS and PREMIS XML within the AIP. Note that validation tools only exist for a very small subset of file formats.			
I5	The digital archive will support automated extraction of metadata from files.		S

Apache Tika is run as a microservice during the ingest of files into Archivematica and output from this tool is stored in the METS and PREMIS XML within the AIP. Note that tools for automatic extraction of metadata only exist for a very small subset of file formats.			
I6	The digital archive will virus check files on ingest.		S
Archivematica runs virus checks as part of the transfer (and ingest?) process. Information about virus checking is included within the PREMIS and METS XML. If a virus is detected within a file it will be sent to the 'failed' directory and all processing on that SIP will stop until the problem is resolved by an administrator.			
I7	The digital archive will be able to record the presence and location of related physical material.		C
In theory this information could be recorded as a related resource under 'relation' within the Dublin Core metadata for an Archivematica AIP. However we would argue that this information is best stored elsewhere within the wider infrastructure.			
I8	It will be possible to select and configure the required level of automation within the ingest workflow.		S
The transfer and ingest process within Archivematica can be configured to allow for a high or low level of automation by pre-selecting choices at many of the decision points within the administration tab. Though not all the decision points can be fully automated using the options available within the GUI, there are reports from users who have managed to fully automate the process by alternative means.			
I9	The digital archive will be able to process large numbers of files and files that are large in size.		M
This requirement is hard to accurately define as we do not have a firm idea of the volumes of data that the system will need to manage. Also, it is difficult to separate out the capabilities of Archivematica from the server specifications of a particular installation. Many problems encountered when processing large volumes of material with Archivematica could be solved by adding more processing power. Recommended minimum requirements for installation of a production version of Archivematica are available from Archivematica's administrator documentation.[85] Discussions on the Archivematica mailing list show that Archivematica has been successfully used to process 500GB files[86] and discussions at the June 2015 UK Archivematica meeting included an example of 10,000 tiny files being processed by Archivematica as one transfer. Similar tests with 100,000 tiny files on the same installation have caused Archivematica to fail.			
DATA MANAGEMENT			

85 https://ww.archivematica.org/en/docs/archivematica-1.4/admin-manual/installation/installation/#installation

86 https://groups.google.com/forum/#!searchin/archivematica/large$20files/archivematica/NLD5o-n4pQw/nc9hLF4lOggJ

DM1	The digital archive will generate persistent, unique internal identifiers.	4.2.4	M
A unique internal identifier is generated at both AIP and file levels and incorporated into AIP and filenames			
DM2	The digital archive will ensure that preservation description information (PDI) is persistently associated with the relevant content information. The relationship between a file and its metadata/documentation must be permanent.	4.2.6.3	M
Any documentation that is included in the SIP will be included in the AIP. All technical and preservation metadata generated by Archivematica will also be wrapped up in the AIP			
DM3	The digital archive will support the PREMIS metadata schema and use it to store preservation metadata.		M
Archivematica creates and stores PREMIS metadata as part of the ingest process and as preservation actions (normalisations) are carried out. This PREMIS metadata is stored within a METS XML wrapper as part of the AIP			
DM4	The digital archive will enable us to describe data at different levels of granularity – for example metadata may be attached to a collection, a group of files or an individual file.		S
This requirement is fulfilled in a partial way by Archivematica. Preservation and technical metadata are generated at file level. Descriptive (Dublin Core) and rights metadata can be created only at project/collection level. Although for research data we do not envisage describing data at any level of detail beyond project level, it is possible that information about rights may need to be more granular. This area needs further thought but it may be possible to implement this elsewhere within the wider infrastructure.			
DM5	The digital archive will accurately record and maintain relationships between different representations of a file (for example, from submitted originals to dissemination and preservation versions that will be created over time).		M
This functionality is very much a part of Archivematica. This is achieved using a unique identifier which is allocated to a submitted file, and included in any subsequent representations that are created.			
DM6	The digital archive will store technical metadata extracted from files (for example that which is created as part of the ingest process).		M
Archivematica stores comprehensive technical metadata including details of all of the tools used as part of the AIP			
PRESERVATION PLANNING			

PP1	The digital archive will allow preservation plans (such as file migration/normalisation) to be enacted on individual or groups of files.		M
Within Archivematica, rules are in place to normalise files (migrate them) to different formats as appropriate for preservation/dissemination and this is enacted as part of the ingest process. The current version of Archivematica (1.4) only allows these rules to be run as a one off process during ingest but AIP re-ingest functionality due for release in July 2015 (in version 1.5) will enable Archivematica to push an AIP back through the ingest process at a later date and should allow further migrations to occur over time. The full functionality of this feature is still to be explored.			
PP2	Automated checking of significant properties of files will be carried out post-migration to ensure these properties are adequately preserved (where appropriate tools exist).		C
This is not actually carried out within Archivematica but it is up to the user of the system to carry out thorough checks of the outputs of the tools and migration pathways chosen within the format policy registry. Once normalisation has been carried out of a file within Archivematica the user has the option of manually checking and approving the normalisations. Default format policy choices within Archivematica are based on a comprehensive analysis of the significant properties of the samples as well as tests of many tools, and the results of these tests are publicly available on the Archivematica wiki.			
PP3	The digital archive will record actions, migrations and administrative processes that occur whilst the digital objects are contained within the digital archive.	4.2.10	M
Detailed information of any automated migrations and administrative processes are stored within the AIP created by Archivematica. Where migrations are carried out manually, PREMIS metadata can also be added for inclusion in the AIP. Note, this feature does assume a one-to-one relationship between an original and migrated file which may not always be the case.			
ADMINISTRATION			
A1	The digital archive will allow for disposal of data where appropriate.		M
It is possible to delete an AIP within Archivematica and record the reason for this deletion. This functionality is essential for research data management for which clear retention and disposal policies will need to be adhered to. Note that file level deletions within an AIP are not supported within Archivematica.			
A2	A record must be kept of data disposal including what was disposed of, when it was disposed of and reasons for disposal.		S
When triggering the deletion of an AIP within Archivematica it is possible to record the reason for deletion and this will be visible to the storage service administrator when actioning that deletion. Disposal decisions may be best recorded elsewhere within the wider infrastructure.			
A3	The digital archive will have reporting capabilities so statistics can be collated. For example it would be useful to be able to report		S

	on numbers of files, types of files, size of files, preservation actions carried out.		
Reporting within Archivematica is limited but there may be other ways we can achieve this and these should be more fully explored in later phases of this project. We are investigating the overlap with another Jisc Research Data Spring project, 'Data Management Administration Online (DMAOnline)'[87] at Lancaster University with its proposed dashboard view of research data in various systems. We also need to investigate whether the recently launched Binder[88] system may fill this reporting gap. Alternatively, reporting may be something we can set up ourselves using the MySQL data that sits behind the system.			
ARCHIVAL STORAGE			
AS1	The digital archive will actively monitor the integrity of digital objects on a regular and automated schedule with the use of checksums.	4.4.1.2	M
This is not a standard feature of Archivematica though a command line fixity checker tool exists which can be installed and integrated with the Archivematica Storage Service[89]. As a matter of course checksums are generated by Archivematica and stored as part of the AIP but routine integrity checking is not performed. This functionality may be addressed elsewhere within the the wider infrastructure and is already a feature of some archival storage services (see for example Arkivum[90]).			
AS2	Where problems of data loss or corruption occur, The digital archive will have a reporting/notification system to prompt appropriate action.	5.1.1.3.1	M
As above.			
AS3	The digital archive will be able to connect to, and support a range of storage systems		S
A number of different storage options can be configured within Archivematica and it is possible to have several different storage options allowing for different types of storage and levels of security depending on the nature of the data, its perceived value or how frequently you envisage it will be accessed. Initial work to build a link between Archivematica and the Arkivum storage service was sponsored by the University of York in 2014[91]			
GENERAL			
G1	The digital archive will be compliant with the Open Archival Information System (OAIS) reference model.		S
Archivematica was created with OAIS in mind. The GUI leads you through the relevant OAIS functional entities and the language used throughout the application is consistent with that used within the OAIS reference model.			

[87] http://www.dmao.info/
[88] https://binder.readthedocs.org/en/latest/
[89] https://github.com/artefactual/fixity
[90] http://arkivum.com/
[91] http://digital-archiving.blogspot.co.uk/2014/12/making-connections-linking-arkivum-with.html

G2	The digital archive will integrate with a range of repository or content management systems		M
Archivematica integrates with dSpace, CONTENTdm and Islandora. The Universities of Hull and York both have existing Fedora repositories (and are also using or planning to use Hydra). The flexible nature of Fedora and Hydra mean that local implementations can differ greatly and therefore a direct integration may be difficult. But this does not mean Fedora users can not work with it. Options include, the existing SWORD integration or using some form of monitoring and messaging service to identify new additions to Archivematica storage locations. There is some ongoing development work involving Archivematica and Hydra by the Bentley Historical Library but this is not likely to be complete within the timescale of this Research Data Spring project and may not meet the needs of our own Hydra implementations.			
G3	The digital archive will integrate with our archival management systems.		S
Many institutions interested in using Archivematica for research data will also have additional use cases for a digital archive, for example born digital archives coming in through a more traditional archives route. This is certainly the case for the University of Hull (Hull History Centre) and the University of York (Borthwick Institute for Archives). A digital preservation system that can integrate and work alongside existing archival management systems to enable more wider use would be of benefit to these organisations. The Hull History Centre uses Calm[92] to manage information about its archives. There is no current link between Archivematica and Calm but this is something that could be explored as a sponsored development. The Borthwick Institute for Archives has recently adopted AtoM[93] to manage information about its archives. Archivematica and AtoM are both supported by Artefactual systems and are designed to complement each other. AtoM is the recommended access front end to Archivematica. Archivematica also integrates with Archivists Toolkit[94]			
G4	The digital archive will have APIs or other services for integrating with other systems.		S
Archivematica has a REST API that offers some functionality around transfers and APIs for interacting with the storage service, and a SWORD API already exists. There may be other APIs required to get Archivematica working effectively within a wider RDM infrastructure			
G5	The digital archive will be able to incorporate new digital preservation tools (for migration, file validation, characterisation etc) as they become available.		S
Within Archivematica there is an interface for adding new migration tools or commands and setting up new rules. Archivematica's microservices architecture has been designed in such a way to enable new tools to be plugged in as required. The microservices have been named to reflect the functions they perform rather than the exact tool that carries out this function. Developers contribute to the development of tools such as FITS to make them better and more scalable. As of			

[92] http://www.axiell.co.uk/calm-for-archives
[93] https://www.accesstomemory.org/en/
[94] http://www.archiviststoolkit.org/

July 2015 Artefactual Systems has become a member of the Open Preservation Foundation[95] and their contribution will focus on support and development of the FIDO tool[96]			
G6	The digital archive will include functionality for extracting and exporting the data and associated metadata in standards compliant formats.		M
Archivematica uses open standards where possible. Metadata is in XML format, uses recognised standards and is packaged with the AIP. Archivematica packages its AIPs using BagIt[97] which is an open standard for storage and transfer of files and metadata. Archival storage is separate to Archivematica so exporting the package from storage should be a requirement of the storage system in use.			
G7	The software or system chosen for the digital archive will be supported and technical help should be available.		S
Archivematica is open source but supported by Artefactual Systems. An active mailing list exists for technical support and Artefactual Systems (and other community members) are quick to respond to any queries.			
G8	The software or system chosen for the digital archive will be under active development.		S
Archivematica is very much in active development with several new versions being made available over the course of a year. Archivematica's development roadmap is published online[98]. Specific developments happen quicker if users of the system are able to fund the necessary work. Alternatively, as Archivematica is open source it is possible for anyone with the necessary skills to help develop the system to meet their needs.			
G9	A community of users will exist around the software or system to enable sharing of use cases, workflows and to promote developments in line with changes and innovations in the discipline of digital preservation.		S
Archivematica has an active international user forum[99] for users to share resources and discuss issues relevant to the software. In January 2015 a UK Archivematica group was established to provide another means for sharing experiences with Archivematica. This UK group has a mailing list and holds a meeting in the UK two or three times a year to catch up with other local users and share experiences.			

[95] http://openpreservation.org/

[96] http://openpreservation.org/news/artefactual-systems-becomes-the-newest-member-of-the-open-preservation-foundation/

[97] http://www.dcc.ac.uk/resources/external/bagit-library

[98] https://www.archivematica.org/wiki/Development_roadmap:_Archivematica

[99] https://groups.google.com/forum/?fromgroups#!forum/archivematica

Appendix 2

Server configurations for testing

Archivematica has been set up at the Universities of Hull and York to allow the project team to push sample data through the system and test the capabilities of Archivematica during phase one of the project. Server configurations and Archivematica versions are described below.

Hull:

- Archivematica v1.3.0
- Ubuntu 12.04.5 LTS 64bit (VM)
- 2 GB ram
- Intel(R) Xeon(R) CPU E7- 4870 @ 2.40GHz
- 50GB+ storage

York:

- Archivematica v1.3.2
- Ubuntu 14.04.1 LTS 64bit (VM)
- 4 GB ram
- 2 CPUs
- 25GB storage

Note that the recommended specifications for a test and production version of Archivematica are available via the Archivematica documentation[100]

While phase one of the project was underway a new version of Archivematica was released. Version 1.4 was made available to the community on 27th May 2015[101]. The project team decided not to upgrade at this point but agreed to maintain current versions for testing. The new features available in version 1.4 are of interest to the project team and have been noted but do not have a significant impact on our intended use of Archivematica or the way it might handle research data. Similarly, at the very end of the phase one work version 1.5 was nearing release and a hosted version of the release-candidate software was made available to the team. We are grateful to Archivematica for allowing us access in order to test the AIP re-ingest facility which was seen as a potentially important requirement for RDM work.

---

100 https://ww.archivematica.org/en/docs/archivematica-1.4/admin-manual/installation/installation/#installation

101 https://www.archivematica.org/wiki/Archivematica_Release_Notes#Archivematica_1.4

Appendix 3

Coverage of top 20 file formats within PRONOM

The table below records the top 20 research data file formats at the University of York and indicates how they are represented in PRONOM as of June 2015. File format representation was checked in PRONOM using the web interface[102], primarily using the file format search to search for a particular application name, but some other types of search (for example by file extension) were also employed to double check initial findings.

It should be noted that making assumptions about file formats on the basis of a list of software applications in use does not necessarily give a true picture of the actual range of file formats that would be deposited within a research data archive. For the purposes of this exercise an assumption has been made that the native file format of an application will be the format that research data is output in, but it is recognised that this will not always be the case. Many packages support a range of formats and data can be imported and exported in a variety of ways.

Several applications have a selection of native file formats, some of which are more widely used than others. It may be that the full range of these formats is not represented in PRONOM but that the formats that we are most likely to receive in a research data deposit are covered. Until we start receiving data from researchers that has been selected for deposit we will not know whether this is the case or not.

Not all of these packages have a native format as such, for example Eclipse[103] is an integrated development environment that can be used to program in a variety of different languages. The nature of the files that will be output will differ depending on the programming language used. Programming languages (several of which are represented individually in this list) can themselves be challenging to identify.

NVivo is another interesting application. It is used as a tool to analyse qualitative data and supports a wide range of different data types, for example “audio files, videos, digital photos, Word, PDF, spreadsheets, rich text, plain text and web and social media data”[104]. Its native formats are .nvp and .npvx but it is not clear whether researchers would consider these files to be their ‘research data’ or prefer just to deposit the qualitative data that they analysed alongside their findings as described within an associated publication.

Note that at the time of writing, discussions with The National Archives were underway regarding increasing the range of PRONOM to include some of these research data file formats. Samples of a selection of these formats have been supplied to The National Archives to help inform this process

Software/application	Coverage in PRONOM	Notes
MATLAB	No coverage	

[102] https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=new
[103] https://en.wikipedia.org/wiki/Eclipse_%28software%29
[104] https://en.wikipedia.org/wiki/NVivo

SPSS	Partial coverage	2 SPSS files represented: fmt/638 (SPSS Data File .sav) and fmt/274 (SPSS Output File .spv).
Stata	No coverage	
R	No coverage	
EndNote	Near complete coverage	5 EndNote files represented: fmt/326 (EndNote Connection File .enz), fmt/327 (EndNote filter File .enf), fmt/328 (EndNote Import File . enw/.enr), fmt325 (EndNote Library .enl), fmt/324 (EndNote Style File .ens).
NVivo	No coverage	
LaTeX	Partial coverage	3 LaTeX files represented: fmt/280 (LaTeX Master document), fmt/281 (LaTeX Subdocument), fmt/160 (TeX/LaTeX Device Independent Document).
Python	No coverage	
Wolfram Mathematica	Partial coverage	1 Wolfram Mathematica file represented: fmt/201 (Mathematica Notebook .nb)
Gaussian	No coverage	
ChemDraw	Partial coverage	1 ChemDraw file represented: fmt/378 (Chemical Draw Exchange Format .cdx)
SAS	Near complete coverage	11 SAS files represented: x-fmt/192, x-fmt/355, x-fmt/356 and x-fmt/601-608 but this doesn't represent the full range as described here: http://support.sas.com/documentation/cdl/en/hostwin/63285/HTML/default/viewer.htm#introsasfiles.htm
ArcGIS	Near complete coverage	12 ESRI ArcView/ArcInfo formats represented - missing .mxd file
GraphPad Prism	Near complete coverage	2 GraphPad Prism file formats represented: fmt/575 (GraphPad Prism .pzm), fmt/576 (GraphPad Prism .pzf)
Adobe Photoshop	Partial coverage	2 Adobe Photoshop file formats represented: x-fmt/92 (Adobe Photoshop .psd), fmt/667 (Photoshop curve file .acv, .atf). Some native Photoshop formats not represented but it should be noted that many users of Photoshop may not be using its native formats and may be exporting their research data in a variety of other formats (many of which will be recognised by PRONOM)
ATLAS.ti	No coverage	
C++	No coverage	
Eclipse	No coverage	No native file formats as such - this application is a development environment
MS Excel	Complete coverage	36 MS Excel file formats represented - no omissions noted
RSB - ImageJ	Partial coverage	No RSB - ImageJ native files listed directly but it saves images in TIFF format by default and works with a number of other image formats that are represented in PRONOM

Appendix 4

Processing configuration used within Archivematica

One of the strengths of Archivematica is that it is configurable and customisable. This gives the user the power to be able to set it up in a way that is appropriate for their needs and their workflows, rather than making do with a one-size-fits-all approach. Whilst this should be seen as a good thing it can also serve as a barrier to those who are new to the field of digital preservation or unsure how to get started with Archivematica.

The following information about the processing configuration as it has been set up for testing a research data workflow at York has been included in this report for those who wish to understand the decision process that we have gone through to configure Archivematica for our purposes. This configuration will not suit all users of Archivematica but defines a largely automated workflow that could be used for research data.

Decision point	Setting	Rationale
Send transfer to quarantine	No	A quarantine stage would be useful perhaps if data was coming from external sources. The purpose of this is to reduce the risk posed by viruses, allowing a suitable time to pass before ingest and allowing virus definitions to update before carrying out the virus scan. This step wasn't seen to be necessary for research data which will be coming from internal sources and subjected to network virus checks before transfer.
Approve normalization	Yes	In other workflows we would take a less automated approach to curation and having the opportunity to see the normalisations would be beneficial. For research data we do not have the human resources to carry out manual curation and prefer to trust that the migration tools are doing their job properly, thus we automatically accept the normalisation that has occurred
Store AIP	Yes	In all circumstances we wish the AIP to be stored (in a location as defined in the AIP storage location).
Transcribe files (OCR)	Yes	This microservice uses the Tesseract tool[105] to transcribe text from images (but not from pdf files). Although this is not likely to be a useful feature for much of the research data that we ingest, it is worth having this microservice enabled by default for those instances where this is useful (unless this results in any noticeable drop in performance).

[105] https://code.google.com/p/tesseract-ocr/

Generate transfer structure report	No	This would be a useful feature if arrangement[106] of the data was carried out prior to ingesting the transfer as the transfer structure report would remain a permanent record of the structure of the original dataset (original order). In our implementation of Archivematica for research data we do not envisage doing any SIP arrangement therefore do not need a record of the transfer structure. On ingest the structure of the SIP will be recorded in the structMap within the METS file and this will be a permanent record of the directory structure of the data.
Remove from quarantine after ? days	N/A	Not applicable for our research data implementation as we do not intend to use the quarantine feature
Create SIP(s)	Create single SIP and continue processing	This option should be selected to automate the step of moving the transferred data into the ingest tab in order to create an SIP. We may however choose to halt processing after the transfer stage of Archivematica if further work on file identification is required at this point.
Extract packages	Yes	This is useful where zip files are included within the SIP - the extract packages microservice unzips the files so they can be processed individually
Normalize	Normalize for preservation	Whilst we acknowledge that for the vast majority of data types that researchers produce there will be no migration pathway, we are keen to take advantage of any automated normalisations (or file migrations) that exist within the Format Policy Registry. For example for documents, images, audio and video files - this will increase our levels of confidence that the data will be kept in a usable format for the future.
Reminder: add metadata if desired	Continue	We do not envisage any manual descriptive or rights metadata entry at this stage of the process for research data, preferring instead to explore a more automated approach to importing metadata[107]
Examine contents	Examine contents	It is acknowledged that research data may contain sensitive and personal information. Enabling the Examine contents microservice will cause a tool called bulk_extractor[108] to be run on the transferred data. This tool is capable of identifying and outputting text reports about personal information contained

106 https://www.archivematica.org/en/docs/archivematica-1.4/user-manual/ingest/ingest/#arrange-a-sip-from-backlog

107 https://www.archivematica.org/en/docs/archivematica-1.4/user-manual/transfer/import-metadata/#import-metadata

108 http://forensicswiki.org/wiki/Bulk_extractor

		in the set of objects. Obviously it can not report on all personal information but can highlight strings of characters that conform to a set pattern such as telephone numbers, e-mail addresses and credit card numbers.
Select file format identification command (Transfer)	Fido version 1 PUID runs Identify using Fido	It is important for our workflows that the most advanced file format identification tools are used at the Transfer stage of processing in Archivematica. Fido[109] has been chosen for this reason rather than identification via File Extension or no file identification at all. Note that in version 1.4 Siegfried has been added as an additional file identification tool[110] and discussion about the pros and cons of using this tool instead of Fido have begun on the Archivematica mailing list[111]. We are interested in exploring the potential of highlighting which files have not been identified at this stage and finding ways to educate the file identification tools so that a wider range of research data file types can be identified in the future.
Select file format identification command (Ingest)	Fido version 1 PUID runs Identify using Fido	See above
Select file format identification command (Submission documentation & metadata)	Fido version 1 PUID runs Identify using Fido	See above
Delete packages after extraction	No	It may be worth keeping the packages in case the package itself is a file format. For example problems have been reported on the Archivematica mailing list[112] around false identifications with Fido in which a Microsoft Excel file (.xlsx) was wrongly identified as a zip file and extracted. Although in reality the recent Microsoft Office formats are container formats that contains the xml files and other content, from a preservation point of view we would not want to lose the original container format.
Select compression algorithm	Uncompressed	There is a discussion about this decision on the Archivematica mailing list[113]. There is a balance here between the benefits of reducing the size of the AIP for storage and

[109] http://wiki.opf-labs.org/display/KB/FIDO+usage+guide
[110] https://wiki.archivematica.org/Archivematica_Release_Notes#Archivematica_1.4
[111] https://groups.google.com/forum/?fromgroups#!searchin/archivematica/siegfried/archivematica/SBfILdoxMNs/WSiqJJbK-egJ
[112] https://groups.google.com/forum/?fromgroups#!searchin/archivematica/extract$20packages/archivematica/dtsFz--8g_8/kMkryCNIYR0J
[113] https://groups.google.com/forum/?fromgroups#!searchin/archivematica/uncompressed/archivematica/f_1mvAOlotQ/iqrSkiwlGkYJ

		the knock on effects this will have on processing time.
Select compression level	N/A	This setting is not required when 'Uncompressed' is selected above. However, if compression is selected the compression level should be decided based on the perceived trade off between the amount of time it takes to compress, and the resulting benefits in the reduction of AIP size. This issue was discussed in more detail in a recent thread on the Archivematica mailing list[114] and it was recommended that levels 1, 3 or 5 were selected on this basis.
Store AIP location	Store AIP in standard Archivematica Directory	For our test implementations this is adequate. Additional locations have not been set up within Archivematica's Storage Service. A decision on where we want AIPs to be safely stored will be made as we move towards the proof of concept phase of the project. Factors to consider when selecting additional storage locations will be the number and nature of backups and whether regular integrity checks are run.
Store DIP location	Store AIP in standard Archivematica Directory	See above. Again this is fine for testing. For the York proof of concept we envisage that DIPs will not be created by default but only on request. A workflow will need to be defined for moving DIPs to an appropriate location if and when access to the data is requested.

114 https://groups.google.com/forum/?fromgroups#!searchin/archivematica/compression%7Csort:date/archivematica/f_1mvAOlotQ/iqrSkiwlGkYJ