UNIVERSITY of York

UNIVERSITY OF Hull

Filling the Digital Preservation Gap

A Jisc Research Data Spring project

Phase Two report - February 2016

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 (Evelyn McLellan, Sarah Romkey, Justin Simpson and Nick Wilkinson) for their hard work, advice and commitment to working with us on this project, John Krug (Lancaster University) for his input into our work on the generic search API and The National Archives for their work on new research data file signatures in PRONOM during the course of this project. Thanks also to Michael Shallcross and Max Eckard (Bentley Historical Library), Andrew Berger (Computer History Museum), Laura Peaurt (University of Sheffield), Chris Fryer (Parliamentary Archives) and members of the UK Archivematica group who submitted valuable advice and feedback on our proposed file identification work.

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

Contents

- Contents
- Introduction
- Enhancements to Archivematica
  - Deliverable 1: Automated DIP regeneration
    - The problem
    - The proposed solution
    - The end result
  - Deliverable 2: METS parsing tools
    - The problem
    - The proposed solution
    - The end result
  - Deliverable 3: Generic search REST API (proof-of-concept)
    - The problem
    - The proposed solution
    - The end result
  - Deliverable 4: Support multiple checksum algorithms
    - The problem
    - The proposed solution
    - The end result
  - Deliverable 5: Enhance PRONOM integration
    - The problem
    - The proposed solution
    - The end result
  - Deliverable 6: Automation tools documentation
    - The problem
    - The proposed solution
    - The end result
- Implementation plans
  - University of Hull
    - The Problem
    - The proposed solution
  - University of York
    - The Problem
    - The Proposed Solution
- Exploration of an ‘above campus’ option for Archivematica
- Linking Archivematica to local institutional repositories
- Outreach
  - Events
    - Hydra Preservation Interest Group - via Skype (13th August 2015)
    - Northern Collaboration Conference - Leeds (10th September 2015)
    - Hydra Connect (Minneapolis, 21st-24th September 2015)
    - UK Archivematica group meeting - Leeds (6th November 2015)
    - iPres conference - Chapel Hill, North Carolina (6th November 2015)
    - RDMF14 (Research Data Management Forum) - York (9th November 2015)
    - Jisc Research Data Management Shared Service Requirements workshop - Birmingham (18th November 2015)
    - DPC members webinar - webex (25th November 2015)

IDCC (International Digital Curation Conference) - Amsterdam (February 2016)
Podcast
Blogs
Project website
Phase 1 project report
Glossary
Appendix 1: Hydra in Hull - Preservation workflows
Author's note:
Table of contents
Background
The document concludes with details of a proposed proof-of-concept implementation corresponding to Phase 3 of the JISC project mentioned above.
Archivematica
Why do we recommend Archivematica to help preserve research data?
What does Archivematica actually do?
How could Archivematica be incorporated into a wider technical infrastructure for research data management?
Types of repository content
Metadata-only records with no associated local content to manage
Content for dissemination but with no apparent requirement for long-term preservation
Content for long-term preservation but with no apparent need for dissemination
Content with a need both for dissemination and for long-term preservation
Ingest methods
Proposed, revised, workflow
Transitioning to the new workflow
A proof-of-concept implementation for Jisc
Appendix 2 : Implementation Plan for Archivematica and RDMonitor at York
Overview
Workflows
Deposit and Preservation Workflow
Before the transfer to Archivematica:
Transfer to Archivematica for storage and preservation:
After the transfer to Archivematica:
Discovery and Access Workflow
Data Access DIP Creation Workflow
Management, reporting and administration workflows
Requirements and Specification
RDMonitor
Data Uploader
SIP Processor
Archivematica Transfer (Submission Information Package - SIP)
DIP Processor
Dissemination Information Package (DIP)
Data Model for Datasets and DIPs
Dataset
DIP
Project Scope

Introduction

In order to manage research data effectively for the long term we need to consider how we incorporate digital preservation functionality into our Research Data Management (RDM) workflows. The idea behind the "Filling the Digital Preservation Gap" 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.

Phase 1 of the project investigated the need for digital preservation as part of a wider infrastructure for research data management and looked specifically at how the open source digital preservation system Archivematica could fulfil this function. Archivematica was installed and tested locally and the project team assessed how it would handle research data of various types. Areas for improvement were highlighted and a plan put in place for enhancing Archivematica to make it more suitable for incorporating into an infrastructure for research data management. The details of this work have been fully documented in a report that was produced at the end of phase 1.

> *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[1]

The phase 1 report is referenced frequently in this document and as the content within this phase 2 report builds heavily on previous work, it is suggested that readers familiarise themselves with the first report in order to fully understand the context of the project.

This report describes the work that has been carried out during phase 2 of the "Filling the Digital Preservation Gap" project. Phase 2 ran from 27th July to the 27th November 2015 with a focus on developing Archivematica further and preparing for local implementation in phase 3.

Work in phase 2 had the following aims:

- Work with Artefactual Systems to develop Archivematica in a number of areas (highlighted in our phase 1 report) in order to make the system more suitable for fitting into our infrastructures for research data management
- Develop our own detailed implementation plans for Hull and York to establish how Archivematica will be incorporated into our local infrastructures for research data
- Consider how Archivematica could work as an above campus installation
- Consider how digital preservation is addressed by the projects in phase 2 of Research Data Spring[2]
- Continue to spread the word, both nationally and internationally, about the ongoing work of our project

It was agreed that the development work carried out by Artefactual Systems could run beyond the phase 2 project dates in order to fit with their own timetable and other areas of work, thus testing of the final deliverables and publication of this report was delayed until early 2016.

---

[1] http://dx.doi.org/10.6084/m9.figshare.1481170

[2] Note that a commentary on this has been published as a blog post: http://digital-archiving.blogspot.co.uk/2015/12/the-research-data-spring-projects.html

Enhancements to Archivematica

At the end of phase 1 of our project we concluded that Archivematica could be used for preserving research data, and highlighted several areas where improvements to Archivematica would be beneficial. During phase 2 we have funded work by Artefactual Systems to develop Archivematica in these areas. For each of our deliverables, the problem we were trying to solve, the proposed solution and the result is documented in the following sections of this report.

Several of these pieces of work do not represent discrete developments that could be completed within the timeline or resources of this project. For a number of our areas of development, work carried out is just the start of a solution and there is potential for others to take these ideas further. By scoping the problems and starting the process of working towards a solution we feel that we have helped move Archivematica into a better place for research data. Other institutions will be able to benefit from this work as well as building on it in the future.

In order to facilitate future development in these areas, we have been open in the way that we have carried out the work, blogging about our plans and highlighting progress in the Archivematica mailing lists (both UK and internationally) and in our other outreach activities (see Outreach section of this report for further details). Artefactual Systems have also made documentation about this work available on the Archivematica wiki so that others can easily locate information about our plans.

The development work that has been carried out thus far has already been referenced and picked up by other Archivematica development projects and has generated considerable interest in the community. Our work on the search API for Archivematica (deliverable 3) is being examined as part of an extensive Mellon funded project[3] currently underway at the Bentley Historical Library (University of Michigan). Our work on DIP generation (deliverable 1) is likely to be of immediate use for staff working at Simon Fraser University in Vancouver and the fact that we have added functionality for working with uncompressed AIPs will also be used and possibly extended by the Museum of Modern Art (MoMA) in New York. There are also plans to enhance some of the work we have sponsored around METS parsing (deliverable 2) through a continuation of the AIP reingest work that the Zuse Institute, Berlin have initiated.

Additionally, the work we have sponsored (particularly deliverables 1 and 2) and the implementation that we hope to carry out in phase 3 of this project will be of particular interest to a number of other institutions that are looking at ways to integrate Archivematica with their Fedora based repositories.

Thinking more specifically about the use of Archivematica for research data management, the Ontario Council of University Libraries (OCUL) has sponsored an integration between Archivematica and Dataverse (a system very much geared towards archiving, sharing and citing research data). They are also interested in the development we have begun in deliverable 5 and our investigations into research data file formats. They too have recognised this as a problem area and are keen to see a greater range of research data file formats represented in PRONOM.

[3] http://archival-integration.blogspot.co.uk/

All sponsored developments described below will be made available in version 1.5 or 1.6 of Archivematica. Version 1.5 is due for release in February 2016 and it is anticipated that version 1.6 will be released in the Spring.

Deliverable 1: Automated DIP regeneration

The problem

As noted in the report from phase 1 of this project, research datasets can be large, of mixed formats and their value may not be fully understood. Creating access copies may be unnecessary as some datasets will never be requested for re-use. Currently, there is no way of automating the process of asking Archivematica to create a DIP 'on demand' after the AIP has been processed.

The proposed solution

In our workflows for long term management of research data we would like the option to initiate the creation of a copy of the data for dissemination and re-use on request rather than create one by default. We were keen that this process could be triggered automatically in order to fit within the automated workflows we would like to put in place for the archiving of research data. Some relevant and useful work on AIP re-ingest[4] has recently been funded by the Zuse Institute Berlin and we proposed to build on this in order to further automate this process.

The end result

There are three aspects to the completion of this deliverable. The first two are complete. These allow the following actions:

1) Programmatically find out whether a DIP exists for a given API by requesting information about the API.
2) Trigger an AIP re-ingest and thereby start the DIP creation process via a REST API call.

At present, there is still a manual step of approving the re-ingest, so the final piece is to approve without manual intervention and fully automate the DIP regeneration. This work will be completed within the next few months.

Deliverable 2: METS parsing tools

The problem

Out of the box Archivematica integrates with a number of third party access systems (for example AtoM, DSpace, CONTENTdm) but not with Fedora and several other repository systems. Hull and York both run repositories built on Fedora and Hydra and need their systems to be able to make sense of the access copy of the data or DIP that Archivematica creates. This can be problematic as the METS files generated by Archivematica can be large and complex and difficult for other applications to interpret.

[4] https://wiki.archivematica.org/AIP_re-ingest

The ability to use an Archivematica DIP as the basis for a digital object in Fedora was key to the needs of both institutions. This process will be at the heart of implementing their proof of concept systems in a subsequent phase of the project.

The proposed solution

Rather than develop a solution that is specific to our own Fedora repositories we wanted to create something with potential for wider application by other third party access systems in use for RDM. Another use case that emerged was the search API developed in deliverable 3 (described below).

The proposed solution is to create a METS reader/writer library that will provide an API for working with the METS files produced for both AIPs and DIPs.

The end result

A first version of the library is available from Artefactual's labs repository[5]. This version will take an archivematica METS file and convert it to json for more efficient processing by the requesting application. The METS file can be accessed within the DIP and contains information about the files in original SIP and preservation AIP.

Although, for the purposes of this project, the work on this deliverable is done, the current library could be extended to, for example, service requests for specific parts of the METS file, such as elements in the Dublin Core metadata.

Deliverable 3: Generic search REST API (proof-of-concept)

The problem

There is a need to be able to produce statistics or reports on RDM in order to obtain a clear picture of what data has been archived. In appendix 1 of our phase 1 project report it was noted that Archivematica does not currently meet our requirements in the area of reporting[6]. It is invaluable for RDM administrators and data curators to be able to view summary statistics about what data is held within the digital archive in order to monitor compliance, assess risks and analyse the take up of the service. In order to use Archivematica for preserving research data we wanted to ensure that we could report on the Archival Information Packages (AIP)s and Dissemination Information Packages (DIP)s that Archivematica had created.

These are the types of questions that we would like to be able to answer:

- How many files are in archival storage?
- What is the total volume of files in archival storage? (in terms of file size)
- How many AIPs are there in total?
- How many files have been identified (ie: have a PRONOM id or similar)?
- How many files are unidentified?
- How many files have been normalised for preservation?
- How many files were not normalised for preservation?
- How many AIPs have DIPs?
- How many AIPs do not have DIPs?

---

[5] https://github.com/artefactual-labs/mets-reader-writer

[6] See requirement A3 in Appendix 1 of phase 1 report: http://dx.doi.org/10.6084/m9.figshare.1481170

- How many files ingested are invalid/not well formed?

The proposed solution

The proposed solution is to create a REST API to facilitate external applications querying the contents of Archivematica. This development will enable statistics to be generated more easily and openly. For example this would enable tools such as the DMAOnline[7] dashboard in development at Lancaster University (also with Jisc Research Data Spring funding) to pull out summary statistics from Archivematica. We intend to test this proposal in phase 3 of our projects.

In developing this solution we wanted to keep an open mind about how the REST API might be used in future. Our development therefore was not limited to just enabling communication with DMAOnline. We are keen that the feature can be used in different ways and by other third party systems.

The end result

A first version of the search API has been developed and demonstrated. It allows queries to be made for information about storage service locations, packages and files. Preliminary search REST API documentation is available[8].

Prototype of the Generic Search REST API endpoint

For example, the search API would allow the following queries to be easily made:

1. Find the package with a given UUID
2. Find packages in a given location
3. Find the status (eg. PENDING or DELETED) or size of a given package
4. Find all files for a given package

[7] http://www.dmao.info/

[8] https://github.com/artefactual/archivematica-storage-service/blob/dev/issue-8895-search-api/docs/search.rst and https://github.com/artefactual/archivematica-storage-service/blob/6593e03f3a576f2a706458339abf18397b1c1f84/docs/search.rst

5. Find all files over a given size for all packages

Beyond research data management, there are many Archivematica use cases which could make use of the search API. Digital preservation managers have a need to analyze their AIP storage to assess their archives for risk related to file formats and make sound decisions for normalization or migration in the future. Because the REST API has been developed in a generic way, it can be used by applications like DMAOnline, but also opens up possibilities for new applications with different functionality or purposes.

Deliverable 4: Support multiple checksum algorithms

The problem

As highlighted in our phase 1 project report, research data files can be large in size and/or quantity and may take some time to process through the Archivematica pipeline. We need to ensure that our workflows for preserving research data are scalable and suitable both for large individual files and large quantities of smaller files. One of the potential bottlenecks in the current Archivematica pipeline is checksum generation - this occurs at more than one point in the process and can be time consuming[9].

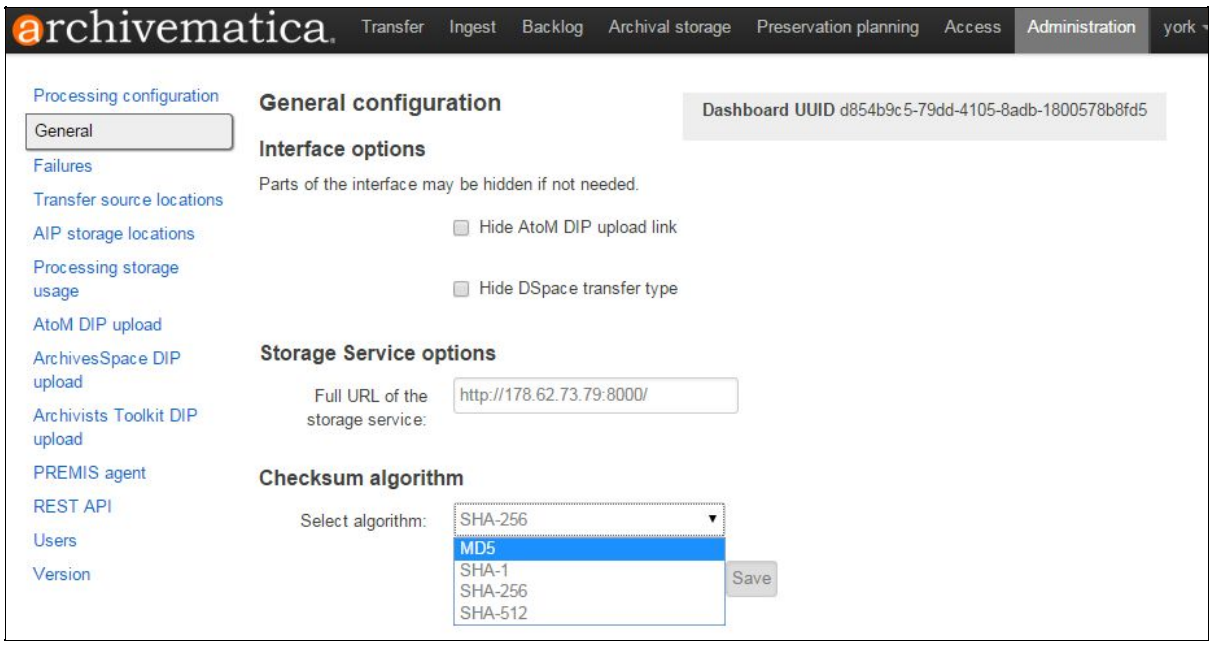
The proposed solution

Archivematica is hardcoded to use the SHA256 algorithm to create checksums and these can take a long time to generate. Other checksum algorithms (e.g. MD5, SHA1 and SHA512) are widely used and may be acceptable alternatives in some circumstances. For instance, if an institution is concerned about file integrity rather than file authenticity the MD5 algorithm (although less robust than SHA256) may well be adequate and faster to compute. Updating Archivematica to allow the administrator to choose the hash algorithm and ensure that this information is also recorded in the PREMIS metadata could help to solve this problem and give institutions the ability to make their own decisions based on local priorities.

The end result

Artefactual extended the capabilities of Archivematica to allow the user to choose a checksum algorithm prior to ingest. Choices are MD5, SHA1, SHA256 or SHA512.

[9] There are some interesting discussions on the Archivematica mailing list about the use of checksums within Archivematica and the time taken to generate them using different hash algorithms: https://groups.google.com/forum/?fromgroups#!searchin/archivematica/checksum$20md5/archivematica/NLD5o-n4pQw/_c3yyg1yDMIJ



Screenshot of the new feature within Archivematica. In the administration tab an institution can now select the checksum algorithm for Archivematica to use

```
<?xml version="1.0" encoding="ASCII"?>
- <mets:mets xsi:schemaLocation="http://www.loc.gov/METS/
http://www.loc.gov/standards/mets/version18/mets.xsd" xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:mets="http://www.loc.gov/METS/">
    <mets:metsHdr CREATEDATE="2015-12-02T09:43:40"/>
  - <mets:amdSec ID="amdSec_1">
      - <mets:techMD ID="techMD_1">
          - <mets:mdWrap MDTYPE="PREMIS:OBJECT">
              - <mets:xmlData>
                  - <premis:object xsi:schemaLocation="info:lc/xmlns/premis-v2
                    http://www.loc.gov/standards/premis/v2/premis-v2-2.xsd" version="2.2"
                    xsi:type="premis:file" xmlns:premis="info:lc/xmlns/premis-v2">
                      - <premis:objectIdentifier>
                          <premis:objectIdentifierType>UUID</premis:objectIdentifierType>
                          <premis:objectIdentifierValue>85ae79cc-cde8-49e5-88ce-
                              2751b0d94b8a</premis:objectIdentifierValue>
                        </premis:objectIdentifier>
                      - <premis:objectCharacteristics>
                          <premis:compositionLevel>0</premis:compositionLevel>
                        - <premis:fixity>
                            <premis:messageDigestAlgorithm>md5</premis:messageDigestAlgorithm>
                            <premis:messageDigest>ef0bcd5e8152527dd4b3d4b3db9cab06</premis:messageDigest>
                          </premis:fixity>
                          <premis:size>2108579844</premis:size>
                        - <premis:format>
```

An extract from the METS file for one of the AIPs created whilst testing this new feature. Note that the PREMIS metadata has been updated to store information about the checksum algorithm used (MD5 in this instance)

Comparative ingest tests were run on a specially commissioned virtual machine at Digital Ocean in London using eight CPU cores and 16GB of RAM. Storage was limited and so the tests were run with a 1.96GB MPEG-2 video file using each of the checksum alternatives in turn.

	MD5	**SHA1**	**SHA256**	**SHA512**
Overall ingest time (mm:ss)	12:55 2nd run 11:50	11:40	10:55 2nd run 12:57	12:29

It might have been anticipated, from the experience of others, surveyed during Phase 1 of this project, that the increasing complexity of checksum would have a significant influence on the overall ingest time. In fact one might conclude from the figures above that, within a likely margin of natural variation, that the checksum algorithm has little impact on ingest time. It would be useful to re-run these tests several times again, and then with other types of file of similar size, and with much bigger files (say 20GB or 100GB) but the facilities available to us for Phase 2 did not permit this.

Although an interesting guide, the crude comparison above clearly has significant failings. We have since been given access to the actual processor times rather than looking at the ingest time as a whole. For the MD5 and SHA256 checksum calculations above, these were 8 and 17 seconds respectively (to the nearest second) demonstrating the sort of efficiency we were hoping for. In addition tests were run with a 25GB file giving a comparison of some 84 seconds for MD5 against 160 seconds for SHA256. On this very limited evidence, using MD5 would save some 50 minutes of processing time per terabyte of data. In terms of Archivematica's overall processing, this one microservice (generate checksums) represents about 2-3% of the total processing time. This may vary with different workloads. Changing to MD5 would likely save at most 1-1.5% of the total processing time. However, checksums are calculated two other times during ingest, currently as SHA256; if we were to extend the work here to cover both of those cases, then the time savings would be tripled. Further investigation is clearly called for and we hope to be able to revisit this during phase 3 of our project using our own local implementations.

On available evidence, it seems that the choice of checksum algorithm did not make a significant impact on overall processing time in Archivematica for files up to 2GB, but we envisage that for institutions dealing with larger files, or collections of files totalling in the tens of gigabytes and more, this new feature should prove to be more significant. This development work has also given users a choice of checksum for the first time; some institutions may have a preference for one type over another and they should now be able to employ their algorithm of choice.

Deliverable 5: Enhance PRONOM integration

The problem

It was highlighted in our phase 1 project report that the identification of research data file formats is a key area when managing research data for the longer term. Research data comes in a wide and varied range of file formats, many of which will not currently be recognised by file format identification tools. 'Knowing what you've got' is an issue of primary concern to those who are charged with managing digital data for the long term.

In our testing of Archivematica in phase 1 of this project we noted how Archivematica handled files that were not identified by the file identification tools it utilises. It is possible within Archivematica to view the file identification report and see which files were not identified, but this information was not presented in a way that was particularly easy for a digital curator to work with and there were no further options to enable or to encourage additional work with these non-identified files, either to identify them by other means or to enable them to be submitted to the file format registry (PRONOM[10]).

[10] https://www.nationalarchives.gov.uk/PRONOM/Default.aspx

During phase 1 of the project we were able to supply some samples of research data files to The National Archives to enable them to assess whether they would be able to create file signatures for PRONOM. It is encouraging to see that a signature for MATLAB (highlighted in our phase 1 report as the University of York's most popular research data application) has been incorporated into the next signature release. The National Archives are also currently working on other research data file signatures using samples submitted by this project. While this is encouraging news, it is worth noting that populating PRONOM is not a one-off exercise. We need to find ways to continue to engage and submit samples in order that new file signatures can be created as the need arises.

The proposed solution

This is a large and complex problem and not one that can be solved quickly and easily. With the resources available we were keen to carry out some initial work on this in the hope that we could start to work towards a solution and that this could be developed and enhanced further by other institutions at a later date given the fundamental nature of file identification to digital preservation.

We planned for an initial feature that would enable Archivematica to report on unidentified files within a transfer alongside access to the file identification tool output. This feature will help users of Archivematica see which files haven't been identified and thus allow them to take further action if they wish to do so.

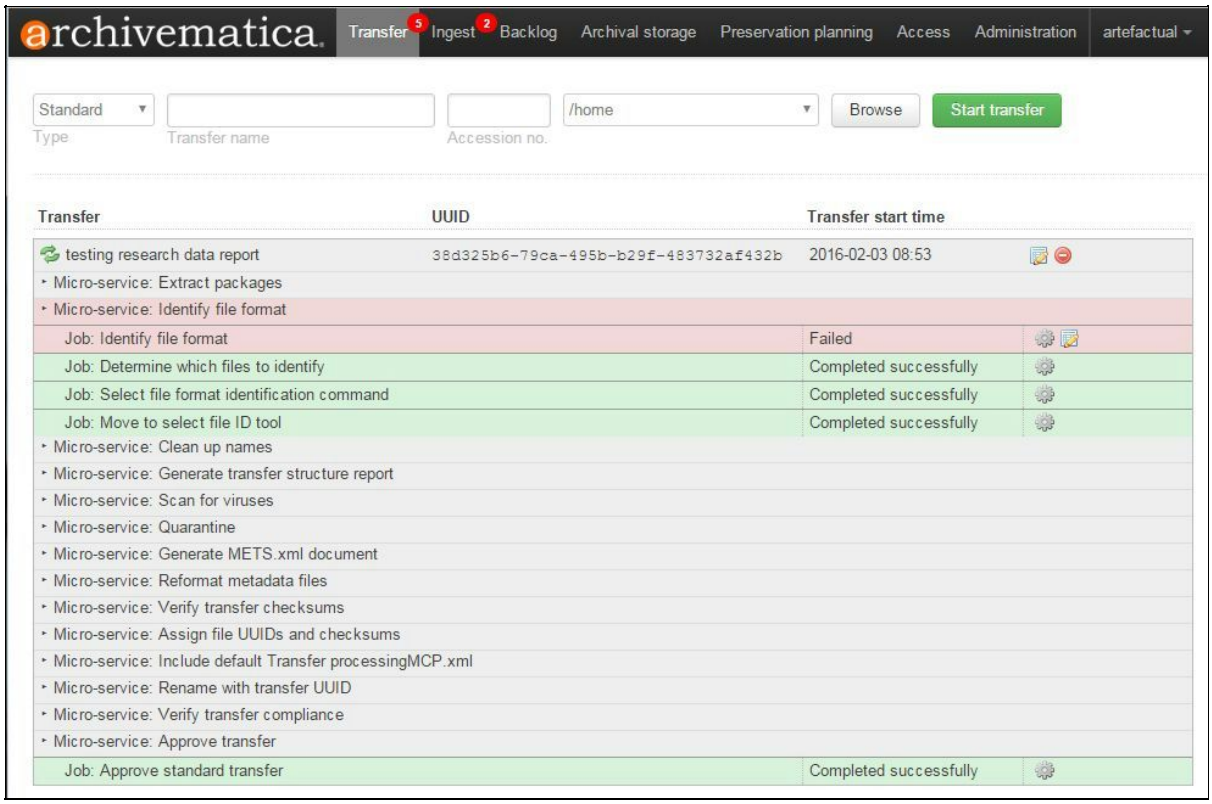
We had hoped to take this development further to enable curatorial staff to submit information about unidentified files directly to PRONOM or to carry out a number of actions on the unidentified files but this additional work was not thought to be feasible in the time frame available (and would have warranted proper scoping with the involvement of key stakeholders). What we proposed to do here was to focus on the initial report and establish what else could be done with the available time. Allowing the operator to view a report of unidentified files provides the foundation for future work in this area.

The end result

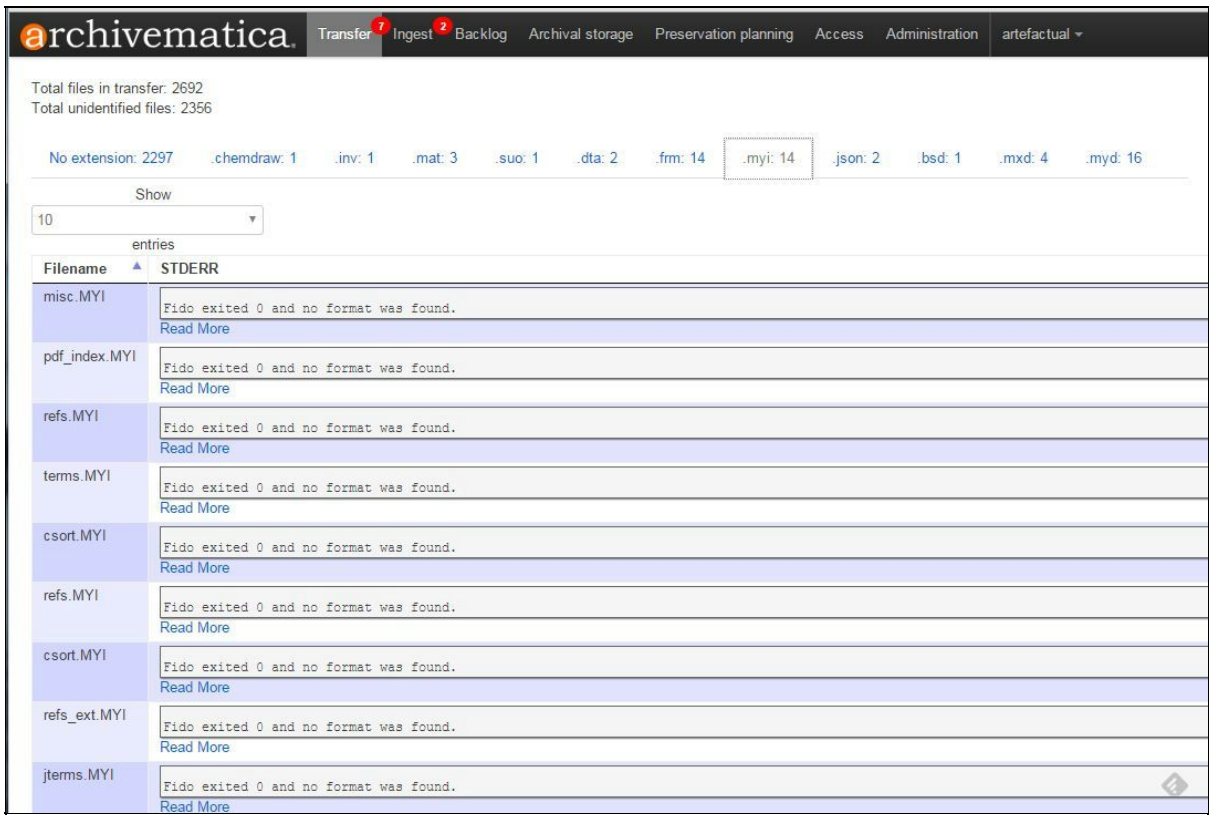
Development work carried out under this deliverable now enables Archivematica users to view a more user-friendly report of unidentified files during the transfer process. As can be seen in the screenshot below, unidentified files are flagged within the dashboard when the failure of the microservice is reported on.

Clicking on the report icon for this microservice takes you to the new report feature that has been developed for this deliverable. Whereas with previous versions of Archivematica it was difficult to see which files had not been identified, the new report (pictured below) will allow the operator to clearly see which files have not been recognised.

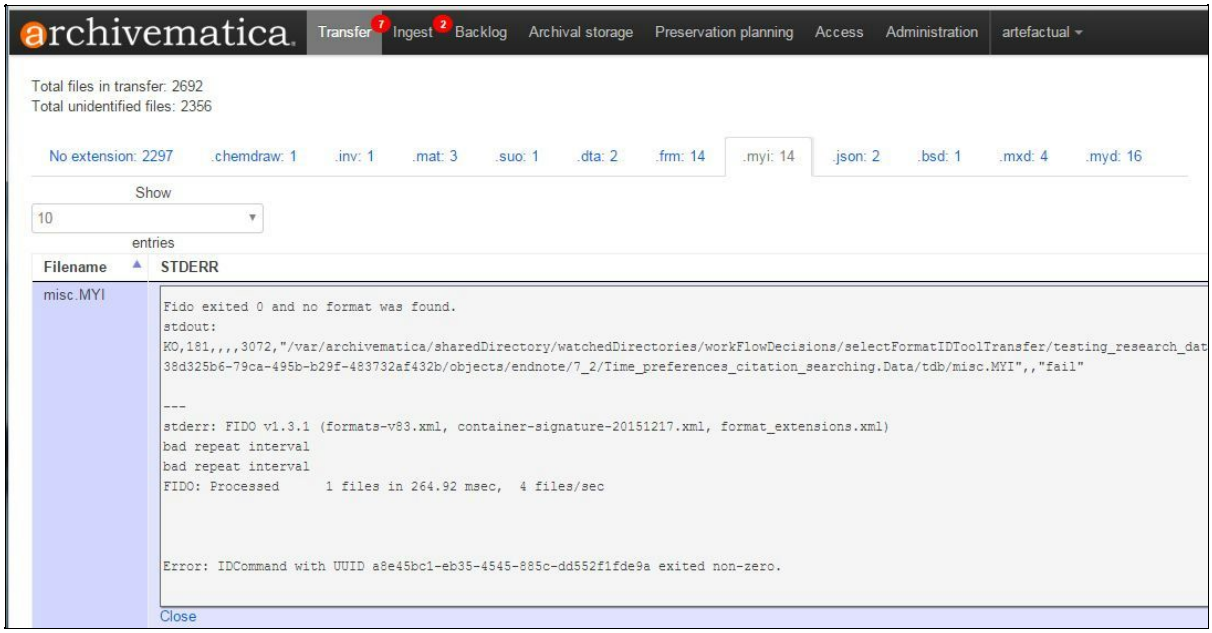
For a ease of viewing (particularly where large numbers of unidentified files are present), the report splits the unidentified files into tabs by file extension and the number of files that can be viewed per tab can be configured by the user to enable up to 100 files to be viewed on one page.



The Identify file format microservice at the Transfer stage displays a report icon where unidentified files have been encountered (the report icon can be seen next to the cog on the Failed 'Identify file format' microservice)



The new unidentified file report in Archivematica. Different file extensions are organised into separate tabs with file counts displayed



Clicking the 'Read More' button displays more detailed output from the file identification tool - for example the FIDO tool includes information about the directory where the unidentified file is located

During the course of phase 2, The project team discussed with Artefactual Systems, Archivematica users and digital preservation professionals how we could develop Archivematica further to make this feature more useful. Several ideas came out of these discussions but at the most basic level a report of unknown files was seen to be the most useful starting point for many of the other suggested enhancements[11].

In the future this reporting feature could be developed further to include:

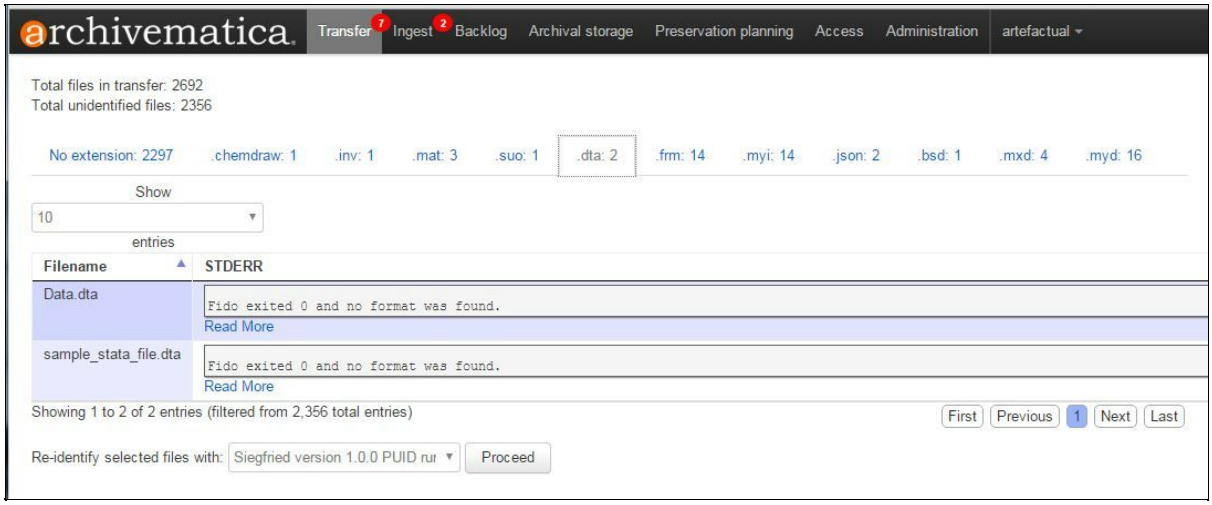
1. The ability to re-run file identification using a different tool
2. The ability to enter a PRONOM ID (PUID) manually
3. The ability to enter a description of the file manually (for example, in the scenario illustrated above I may know through discussion with the content creator what these .myi files are, and in the absence of an identification by the selected tool I could make a note of this information within Archivematica for future reference)
4. The ability to resolve conflicts, for example where different identification tools produce different results, or indeed where the same tool produces a range of results
5. The ability to correct identification errors, for example where a file from the late 1980's with a .MOV extension has been recognised (by file extension) as a QuickTime file. The operator may be aware that QuickTime was not released until 1991[12] so would want to override this identification manually[13]
6. The ability to interact in a more direct way with PRONOM, submitting sample files and other information about formats as appropriate

Artefactual Systems intend to make the first of these options available in a future release of Archivematica (as illustrated below) but further sponsorship will be required to take other features forward.

[11] Some of these thoughts are discussed in a blog post: http://digital-archiving.blogspot.co.uk/2015/11/file-identification-lets-talk-about.html

[12] https://en.wikipedia.org/wiki/QuickTime

[13] Thanks to Andrew Berger (Computer History Museum) for sharing this use case



This screenshot shows the beginnings of the file re-identification feature, which will be available in a future release of Archivematica. With a user able to select or re-run a file identification tool to work on those files that are not previously identified

Deliverable 6: Automation tools documentation

The problem

One of the potential barriers to institutions that may be considering adopting Archivematica are the difficulties of installing and configuring the system. Archivematica comes with an online user manual[14] which is updated with each new version, however there are still inevitably some areas where documentation could be improved in order to enable users to more quickly get to grips with the system.

When an institution sponsors a development within Archivematica through Artefactual Systems, a 10% community support fee is added - among over things this covers the cost of documenting the new features within the user manual. This is a good approach to documenting an open source system but inevitably there are gaps. Whilst individual new features may be well documented, users just getting started with Archivematica would benefit from other more generic documentation such as a user-friendly overview of the installation process and an introduction to the available APIs.

Artefactual Systems has also been working on a project of relevance to the RDM community called the automation tools project. This has been utilised by a handful of Archivematica users to fully automate an Archivematica pipeline. The ability to automate processes relating to preservation is of obvious benefit where few resources are available to manually process data of unknown value, such as is the case for many institutions tackling the preservation of research data. 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.

The proposed solution

We were keen to kickstart the development of some updated documentation for Archivematica, that would give an introductory overview of Archivematica's technical architecture and describe the process of installing, configuring and executing the full Archivematica stack. The series would cover many of the things a developer needs to know

[14] https://www.archivematica.org/en/docs/

to get Archivematica up and running and set up an automated workflow through the available API's.

The end result

During the course of phase 2, Artefactual Systems has been working on some new documentation on Archivematica's technical architecture and the first installment of this will be available with the release of version 1.5 later this month. It is hoped that this new documentation will enable new Archivematica users to more quickly get an installation up and running. As with many of the other deliverables, this is not a discrete and finite piece of work, but one which will grow and develop with the system itself. It is envisaged that other topics and themes will be introduced as and when resources allow.

Implementation plans

Part of the work carried out during phase 2 of the project was to develop implementation plans for prototypes to be built during the final phase of the project (if funded). The prototypes are intended to demonstrate Archivematica in use as part of an RDM workflow. Developing our implementation plans has allowed us to establish the extent and feasibility of the work required.

Both Hull and York are users of the Fedora Commons repository software and so both prototypes will have Fedora as their access repository component. But the work in phase 3 is intended to demonstrate lightweight, re-usable and modular approaches that could be applied to workflows with different components, for example a different research data upload process or a different repository such as EPrints, DSpace or Figshare.

York's prototype will focus on RDM compliance, processing datasets submitted by our academics via York's research information system, PURE. To keep the prototype achievable in a short time-scale, York will focus only on open datasets, not those with restrictions.

Hull's prototype will focus on the fit with existing workflows and will look to develop an approach which is fairly generic and therefore easily adaptable to the local circumstances of other adopters.

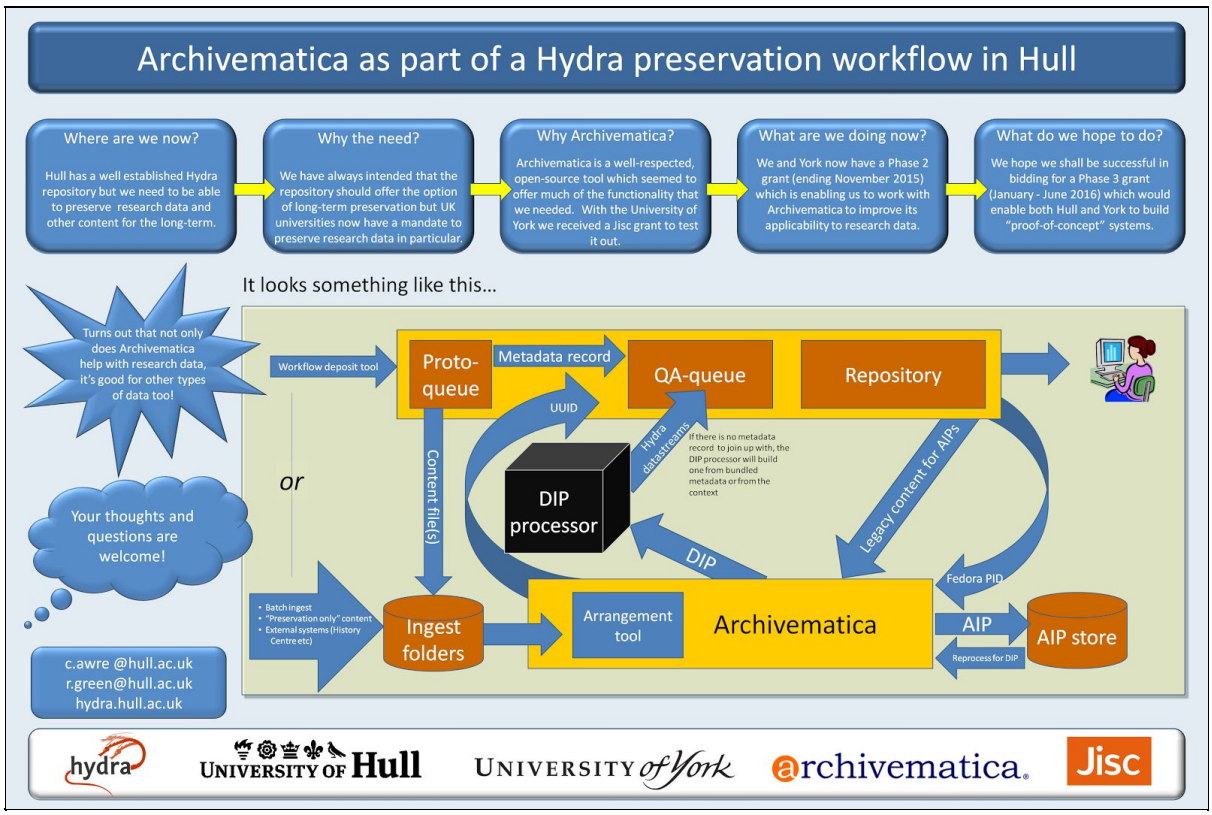
Central to both approaches will be a software element that we have termed the "DIP processor" and this will be developed jointly by Hull and York. Although the end result will be somewhat different at the two institutions, to deal with different requirements and workflows, we propose to develop common code blocks where this is an appropriate methodology.

University of Hull

The Problem

The University of Hull has been running a digital repository for more than seven years. It has always been part of the thinking that this repository system should form part of an information architecture capable of preserving content as well as disseminating it but this second element has not yet been developed. The UK's current mandates around the preservation of research data have provided a timely incentive to add the preservation

capability. In the wider scheme of things, Hull needs functionality that is capable of providing "preservation on request" for other types of digital content in addition to research data and so the proof-of-concept implementation for phase 3 needs to be a pathway through the workflows which address this bigger picture. A poster created for the Hydra Connect conference[15] in autumn 2015 gives some indication of this wider context:



Poster prepared for Hydra Connect 2015 showing Hull's proposed overall workflow

The proposed solution

Hull has a number of use cases which require the batch ingest of repository content (here, research data) and its subsequent processing to produce and store a preservation package (an AIP in Archivematica terms). Hull's approach will be always to produce also a DIP although the dissemination file(s) contained within it may not be used in the digital repository. The DIP will provide the necessary information with which to produce a repository object capable of disseminating the data or else a metadata-only record where re-use of the data is deemed unlikely at the time of ingest. The proposed solution employs a system of "watch folders" which should be easily adapted to workflows in other institutions, being agnostic of any "front-end" software that might have prepared the ingest in some way. Ultimately the system should be capable of dealing with the ingest of a single or multiple files, with or without metadata:

- One or more files with no, or minimal, accompanying descriptive metadata, each requiring a repository record
- One or more files with accompanying metadata which, excepting the filename is essentially the same in all cases, each requiring a repository record
- One or more files with an accompanying metadata manifest providing detailed descriptive metadata for each file, each requiring a repository record

[15] See https://wiki.duraspace.org/display/hydra/Hydra+Connect+2015

- One or more files with no, or minimal, accompanying descriptive metadata, requiring a common repository record
- One or more files with accompanying metadata, requiring a common repository record

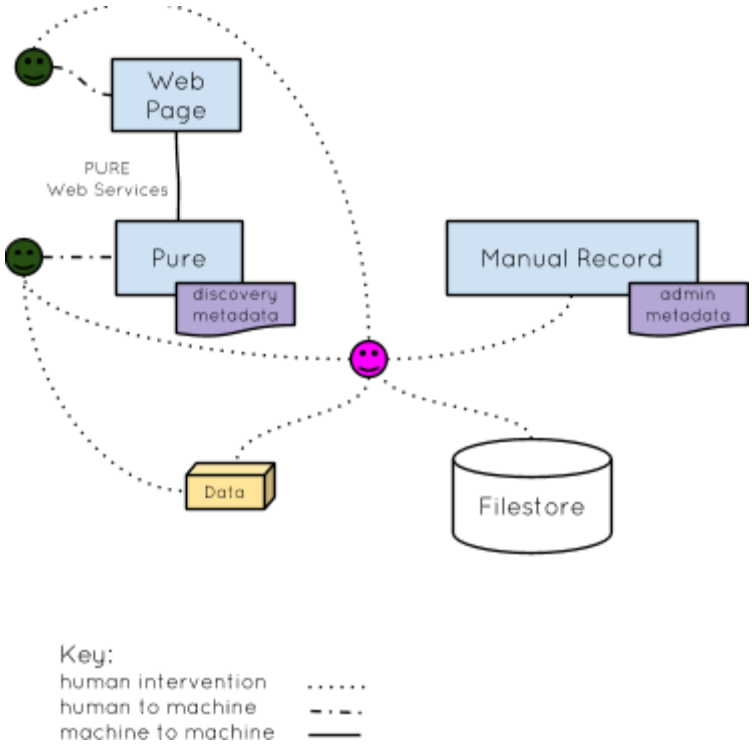
In all cases there may or may not be a need for the repository object to disseminate the data file. Development will start from a single file with moderate metadata and capability will be extended as far as time and resources will allow.

Further information is available in Appendix 1: Hydra in Hull - Preservation workflows.

University of York

The Problem

The expectations of the EPSRC around research data management[16] have required us, along with all other EPSRC-funded institutions, to put in place policies and workflows around managing research data produced by academics at York.



Current York RDM workflow

Our existing workflow was put together with limited resources to meet the immediate need to address EPSRC expectations, however the way we are doing things currently is far from ideal and not a sustainable solution for managing research data. Some of the issues are as follows:

- There are lots of manual steps - this means the process is time consuming and there are risks that inconsistencies will be introduced or mistakes will be made
- Important data around compliance is stored in manually maintained systems, with manual reporting

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

- Access requests are made ad hoc by email with a potential delays in giving access to data
- There is no checking of the data itself - this means there is risk that the data contains viruses, may be corrupted or encrypted.
- No file format identification is carried out to try and understand the types of data we are managing and to help with the future preservation of that data
- No preservation planning is in place - without knowing what types of files we have, it is difficult to then take steps to preserve and provide access to that content in a meaningful way
- There are no systems in place to ensure data is unchanged over time - this should be a key feature of any system to manage data for the long term

York has had, since 2008, a Fedora Commons repository (YODL[17]) for storing multimedia resources produced by the University. This repository is not currently integrated into our RDM workflow, although public datasets will be uploaded to YODL by Library staff to provide public access. YODL is not currently preservation-oriented - the primary focus is to provide access to data not to preserve it for the long term, so this is a gap we are keen to fill.

The Proposed Solution

The York proof of concept will utilise existing systems and resources (YODL, PURE[18], university filestore) alongside the digital preservation system Archivematica to complete our data management workflow in line with University of York RDM policy and funder requirements.

As already noted, phase 2 of this project has enhanced Archivematica for use as part of an RDM infrastructure and will enable Archivematica to be more easily integrated with other systems.

The solution we are implementing focuses on the following use cases:

A member of academic staff submits a dataset:

1) to support a publication
2) as a stand alone ‘data archive’ for a research project

A PhD student submits a dataset:

3) to support a publication
4) to support their examined thesis
5) that is their examined thesis[19]

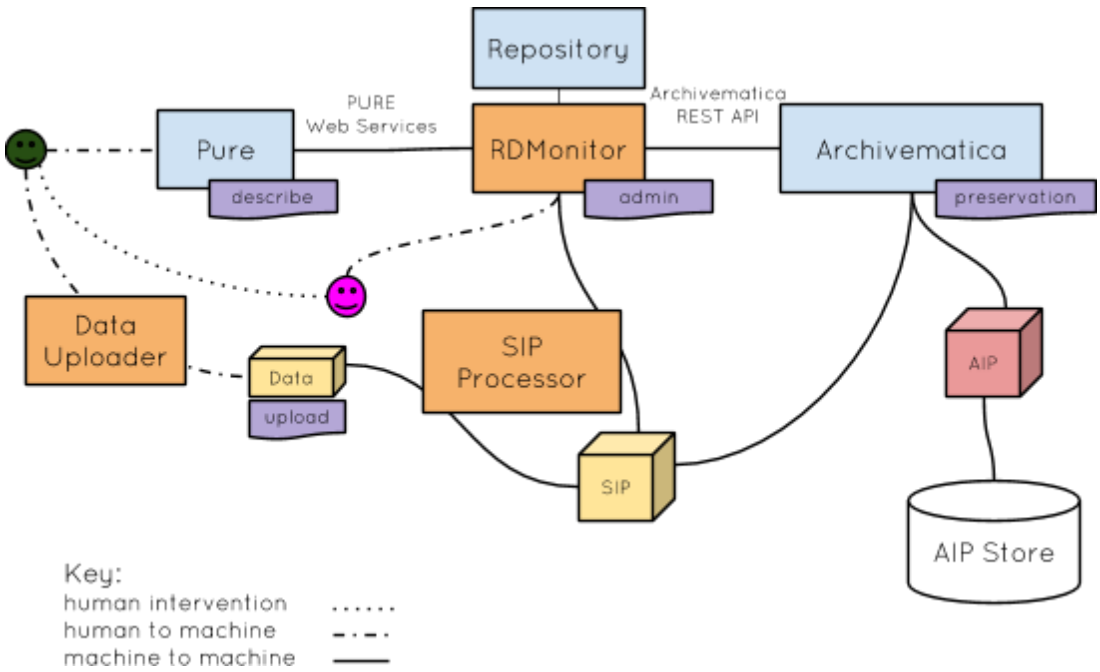
We propose to build a lightweight application to fill the gaps in our current workflow and provide the necessary ‘glue’ to facilitate interoperability between systems. In our case this will comprise the following components:

- **SIP Processor**: this component will be supplied with data and other files (eg. metadata, data management plan) and will build an Archivematica-ready submission package (SIP).

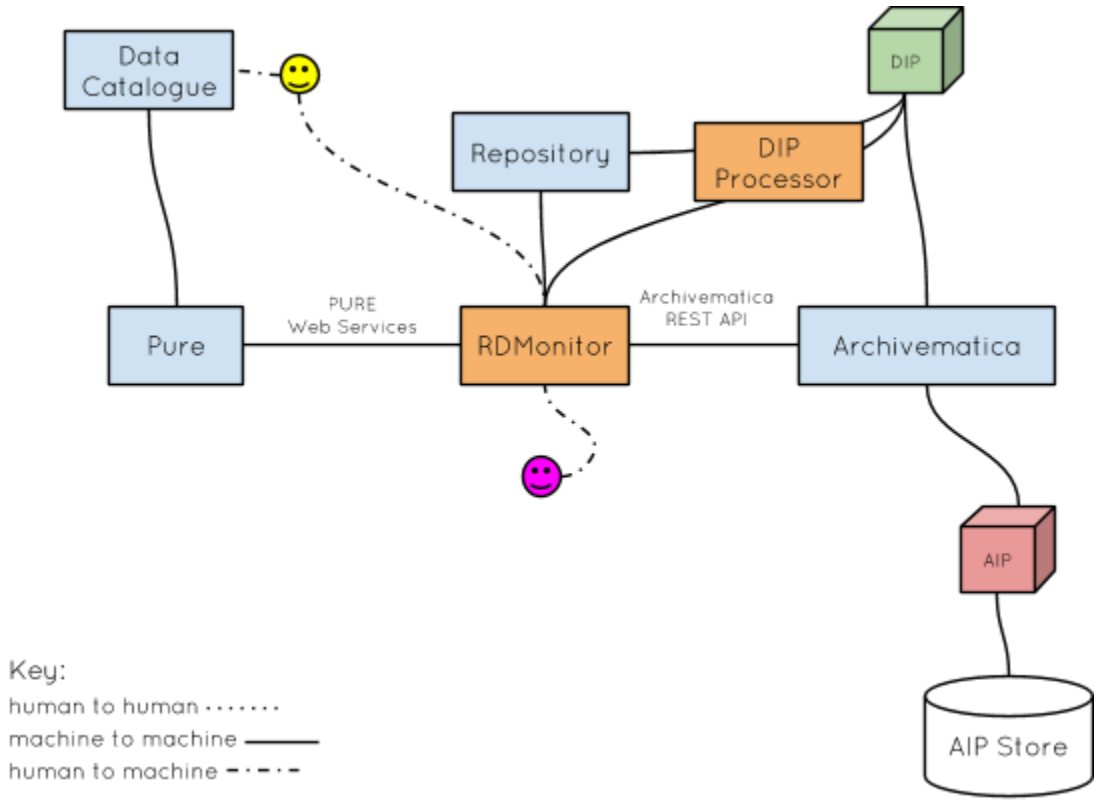
[17] https://dlib.york.ac.uk/yodl/app/home/index
[18] https://www.elsevier.com/solutions/pure
[19] for example if the thesis itself is submitted in the form of a website with associated multimedia

- **DIP Processor**: this component will access the Archivematica dissemination package (DIP), then build and ingest Fedora objects to our local specification
- **Data Upload**: this component will allow us to collect data and associated information from the depositor, outside of PURE; we also plan to investigate collecting data from within PURE itself
- **Research Data Monitor (RDMonitor)**: this component will be a web application for our data management staff, which will provide them with information about the status of a dataset (eg. metadata created, dataset not yet uploaded, date of last access) and also enable them to action events, such as creating the DIP.

The two diagrams below illustrate that, although seemingly more complex than our existing workflow, the proposed development will eliminate all but the most necessary manual steps.



York's proposed deposit and preservation workflow



York's proposed access workflow

All components will be developed in Rails and will be designed to be re-usable for other use cases and repository platforms. Code for all of these will be freely available in Github.

Intended to be an integrated solution, the RDMonitor tool will use information from the PURE Web Services about datasets described in PURE; from Archivematica about the storage of the datasets themselves and from our Fedora repository about the access copies of data.

The roles of different systems within the proposed solution are are as follows:

- Metadata about datasets, including information about how to access them, will reside in **PURE**.
- Data itself will be processed in **Archivematica** before being moved into Archival Storage alongside metadata about preservation actions and file formats.
- Access copies of requested data will be made available through **York Digital Library** (Fedora4[20] repository with Hydra-based[21] front-end).
- Discovery will be via the **Data Catalogue**[22].
- **DMAOnline**[23] will be used to view administrative data about our RDM systems
- We will continue to encourage use of **DMPonline**[24] by researchers and will encourage the deposit of data management plans in our RDM workflows

The overarching rationale for the proof of concept is:

- It **automates** various manual steps in the current workflow

[20] http://www.fedora-commons.org/
[21] http://projecthydra.org/
[22] The software for this is yet to be decided at York.
[23] DMAOnline is a tool currently in production at the University of Lancaster also funded by Jisc as part of their Research Data Spring initiative. See http://www.dmao.info/
[24] https://dmponline.dcc.ac.uk/

- It **integrates** with existing systems to access and store information
- It provides **digital preservation functionality** to ensure we have a better idea about what we have and help us identify future risks
- It **plugs recognised gaps** in our current RDM infrastructure
- It **offers a model that could be adopted by other institutions** and/or shared services

Further information is available in Appendix 2: Implementation Plan for Archivematica and RDMonitor.

Exploration of an 'above campus' option for Archivematica

The work carried out to date within the project has focused on having a local implementation of Archivematica. Recognising that not every site will have the capacity to work on this basis, the project was interested to investigate to what extent Archivematica could be used in an 'above campus' situation, where an institution makes use of a remote implementation of Archivematica.

A review of such scenarios highlighted four instances of where Archivematica is being used in this way:

A. The Council of Prairie and Pacific University Libraries (COPPUL) Archivematica-as-a-Service (AaaS) provision
B. The implementation of Archivematica by Archives Wales (ARCW)
C. The ArchivesDirect service provided by DuraSpace
D. Integration between Archivematica and Arkivum

The COPPUL service was initiated as a way of enabling member institutions to take advantage of digital preservation services beyond simple preservation storage. Collaboration with Artefactual led to an implementation that makes use of the University of British Columbia's EduCloud cloud hosting service[25]. This combination of commercial company and university-based cloud hosting provided a degree of accountability amongst the involved institutions that might not have been present if the cloud hosting had been with, for example, Amazon.

The desire to facilitate access to preservation functionality compatible with the OAIS Reference Model, was at the heart of the decision by ARCW to look at a range of potential solutions that could be used across a number of Local Authority Archives, including DAITSS [26], Xena[27], and Archivematica, so that members could gain access to this functionality without local IT requirement[28]: Preservica in the Cloud[29] has also been examined. Notwithstanding that no local IT was required to run the applications, there were IT issues in establishing effective links to the shared service that they are conscious of needing to address in time, particularly around network speed for file upload; central storage was a core component of the shared service approach being tested. A more detailed exploration of

[25] http://www.coppul.ca/archivematica
[26] https://daitss.fcla.edu/
[27] http://xena.sourceforge.net/
[28] http://www.archiveswales.org.uk/projects/digital-preservation/
[29] http://preservica.com/edition/cloud-edition/

Archivematica highlighted the good logical flow of the system and the advantages of different pipelines to manage different materials. ARCW is currently looking at options for turning this testing into a service.

The two other services listed are more commercial offerings. Interestingly, they are both driven by organisations that have storage as their main service, which they are then looking to enhance through the processing of content for preservation prior to ingest to the store. DuraSpace has been offering its DuraCloud managed cloud storage service for some time now. Adding Archivematica enhances and adds value to the storage being offered[30]. Arkivum was initially integrated with Archivematica in a project funded by the University of York in 2014[31] and the two services have been working together on other projects subsequently. Arkivum offers flexibility in its offering according to need and whilst the other solutions have been content agnostic or focused on digital archives, a specific area of work for Arkivum has been to address the digital preservation needs of research data.

Key to all of these above campus solutions is that they (as evidenced clearly by the last two examples) make use of storage that is closely linked to the Archivematica instance. In the case of COPPUL it was the availability of EduCloud that enabled the AaaS service to go ahead, whilst ARCW is working in close liaison with the National Library of Wales to ensure that archived materials are stored appropriately. The storage isn't always directly connected: ArchivesDirect makes use of Amazon S3 and Glacier, but the service takes responsibility for linking the two so that institutions locally do not need to. Also common across most of the services is the existence of a cost model that provides different levels of both storage and computing resource for the Archivematica instance. For example, ArchivesDirect[32] and COPPUL[33] both offer three levels of engagement at different prices, dependent on the level of service.

Central to making an above campus option for Archivematica work is how the content gets picked up for processing by Archivematica in the first place. This requires the content to be placed in watched folders that Archivematica can trawl, or it can come straight from a local repository or store as appropriate (e.g., COPPUL has worked with local DSpace repositories to capture content directly to Archivematica from there). ARCW found both good (different watch folders/pipelines) and bad (network speeds) factors here. In the provision and take-up of any service, clarity on how content can be best uploaded, and the limits of how this works should be taken into account.

When we proposed looking at this area, we were interested in doing so from the perspective of understanding how the work of the project could be taken forward in different ways. The advent during this Phase 2 of the Jisc Research Data Management Shared Service initiative [34] has highlighted that there is a real need to provide some form of above campus digital preservation capability, in this case focused on research data. The experience elsewhere of using Archivematica as a shared service highlights how the system could apply in this context.

---

[30] http://archivesdirect.org/
[31] http://digital-archiving.blogspot.co.uk/2014/12/making-connections-linking-arkivum-with.html
[32] http://archivesdirect.org/pricing
[33] http://www.coppul.ca/archivematica
[34] http://researchdata.jiscinvolve.org/wp/2015/10/07/jisc-rdm-shared-service-pilot/

Linking Archivematica to local institutional repositories

When Hull and York decided to work together on this project, it was partly on the basis that we had a common need to identify how our Fedora/Hydra repositories could work with Archivematica so we could effectively manage research data within these. The implementation plans described highlight how this can be enabled at the two sites. Whilst different they apply the same principles for working with Fedora/Hydra, and respect the local infrastructure and systems that each repository needs to integrate with.

Looking beyond this, we were also keen to ensure that the solutions we identified were not solely those that would work in a Fedora/Hydra environment. Ultimately, Archivematica outputs an AIP and optionally a DIP; work within this project has enhanced the facilities of Archivematica so that the DIP can be more easily used as the basis for a repository object which will be ingested for discovery and delivery. This was considered vital for research data where there is a need from funders to make the data and its metadata available in many cases, alongside the need to preserve it for longer term access.

In order to move data from Archivematica into Fedora/Hydra, both Hull and York will be implementing a go-between, a tool we have titled the DIP Processor. This will take the METS file within an Archivematica DIP and parse this for the information needed to build the relevant Fedora object so that this can be ingested into the repository using repository tools and workflows. The project's work to develop METS parsing tools will provide a more generic way of achieving much of this by parsing the DIP into json, easier to work with than METS itself. The proof of concepts proposed for Phase 3 would put this into action.

The aim of the METS parsing tools is also to enable connections to be made with other repository systems. To what degree these are used will depend on the scenario for a repository. There are two basic courses of action that can be followed given the DIP output from Archivematica:

- The DIP can be provided as a zip file and can be ingested into a repository and stored in that zipped file format. This scenario does not anticipate the file being made accessible as any user unzipping the file would not easily be able to navigate it or interpret it for the content it held.
- The DIP can be unpacked and processed and the components used to build objects for ingest to the repository.

The latter scenario is the one being tackled by the project. There is work ongoing by Artefactual to add the capability for the first scenario for ingest to DSpace, and equivalent work could be undertaken to do the same for EPrints. The dilemma presented by this is that the files are not then directly accessible, which may impact on the access requirements for the content and the repository. See, for example, the use case at Edinburgh below, based on conversation with staff there.

> **DSpace use case**[35]
> The University of Edinburgh is considering the role that Archivematica may play in assisting their processing of digital materials. Their use cases require both

[35] Many thanks to Claire Knowles, Library Digital Development Manager, and Kirsty Lee, Digital Preservation Curator, for their time in describing this.

preservation and access, and they are seeking to move the processed files into DSpace for this latter purpose. They would also like the files to be themselves indexed and searchable. The focus is on archival materials rather than research data at this time, though the DataShare service, also based on DSpace, is also interested in what might be possible. Given that focus, there is also a need to link in ArchivesSpace, the archives cataloguing system. This will be the main point of access, referencing materials in DSpace.

Edinburgh has been following the work at the Bentley Historical Library at the University of Michigan, where they are tackling a similar scenario. There is work ongoing here on moving content to DSpace, and an exploration of whether to use SWORD 2 or ResourceSync as the mechanism. More will become apparent in the first part of 2016 as the Bentley work completes.

The METS parsing tools are designed such that it should be feasible to write the equivalent of the project's DIP Processor for either DSpace or EPrints. This would then enable the appropriate information to be extracted from the DIP. Once extracted, the file(s) need to be pushed or pulled (see discussion at Bentley above) into the repository, and ingested using normal repository workflows as for Fedora/Hydra.

A converse link to a repository is where the repository is the source of the content for processing into Archivematica, with the intention of creating an AIP that can be placed in a dark archive AIP store. There is functionality in Archivematica 1.4 to enable this for DSpace [36], which has been applied as part of the COPPUL AaaS service (see previous section), and equivalent functionality for EPrints could be added to complement this. The architecture for Hull's implementation of Archivematica also includes this loop for content already in the repository that will not have been passed through Archivematica, so that we can retrospectively enhance existing records with additional preservation metadata..

Outreach

During the four months of phase 2 of this project, substantial efforts were made to ensure that we kept people informed about what we were doing. As well as promoting the existence of our phase 1 project report, we also wanted to use our outreach opportunities as a means of making new contacts, finding out who else is working in this area now, and who may be thinking of establishing similar solutions in the future. These goals were met largely because of the wide range of different events we attended, both UK and international. The project team have had numerous conversations about Archivematica and our work with individuals from different institutions throughout phase 2 of the project and have been encouraged by the level of interest the project has generated.

Outreach channels consisted largely of presentations at organised events and continuing to report on project progress using the University of York's Digital Archiving blog[37].

[36] https://www.archivematica.org/en/docs/archivematica-1.4/user-manual/transfer/dspace/
[37] http://digital-archiving.blogspot.co.uk/

Events

Hydra Preservation Interest Group - via Skype (13th August 2015)

"Filling the digital preservation gap" an update from the Jisc Research Data Spring project at York and Hull - Jenny Mitcham and Julie Allinson
http://www.slideshare.net/JennyMitcham/filling-the-digital-preservation-gapan-update-from-the-jisc-research-data-spring-project-at-york-and-hull

Northern Collaboration Conference - Leeds (10th September 2015)

A collaborative approach to "filling the digital preservation gap" for Research Data Management - Jenny Mitcham
http://www.slideshare.net/JennyMitcham/a-collaborative-approach-to-filling-the-digital-preservation-gap-for-research-data-management

Hydra Connect (Minneapolis, 21st-24th September 2015)

Poster: *Archivematica as part of a Hydra preservation workflow in Hull* - Richard Green and Chris Awre
https://hydra.hull.ac.uk/resources/hull:11580

UK Archivematica group meeting - Leeds (6th November 2015)

Project update: A collaborative approach to "filling the digital preservation gap" for Research Data Management - Julie Allinson
http://www.slideshare.net/JennyMitcham/project-update-a-collaborative-approach-to-filling-the-digital-preservation-gap-for-research-data-management

iPres conference - Chapel Hill, North Carolina (6th November 2015)

"Filling the Digital Preservation Gap" with Archivematica - Jenny Mitcham
http://www.slideshare.net/JennyMitcham/filling-the-digital-preservation-gap-with-archivematica

RDMF14 (Research Data Management Forum) - York (9th November 2015)

"Filling the Digital Preservation Gap" for RDM with Archivematica - Chris Awre, Jenny Mitcham and Sarah Romkey
http://www.slideshare.net/JennyMitcham/a-collaborative-approach-to-filling-the-digital-preservation-gap-for-research-data-management-55738944

Jisc Research Data Management Shared Service Requirements workshop - Birmingham (18th November 2015)

Digital preservation requirements for research data management - Chris Awre and Jenny Mitcham
http://www.slideshare.net/JennyMitcham/jisc-shared-service-requirements-presentation-18th-november-2015

DPC members webinar - webex (25th November 2015)

"Filling the Digital Preservation Gap" with Archivematica - Jenny Mitcham and Simon Wilson

http://www.slideshare.net/JennyMitcham/filling-the-digital-preservation-gap-with-archivematica-55738788

IDCC (International Digital Curation Conference) - Amsterdam (February 2016)

An abstract for a practice paper has been accepted

Podcast

A Jisc podcast interview with Chris Awre and Jenny Mitcham was recorded 9th September 2015 and was conducted by Jisc Press Office.
https://www.jisc.ac.uk/podcasts/research-data-spring-automatically-preserving-research-data-24-sep-2015

Blogs

The project team have been blogging about the project on the University of York's Digital Archiving blog: http://digital-archiving.blogspot.co.uk/

Blog posts relating to the project (either describing or informing our phase 2 work) are listed below:

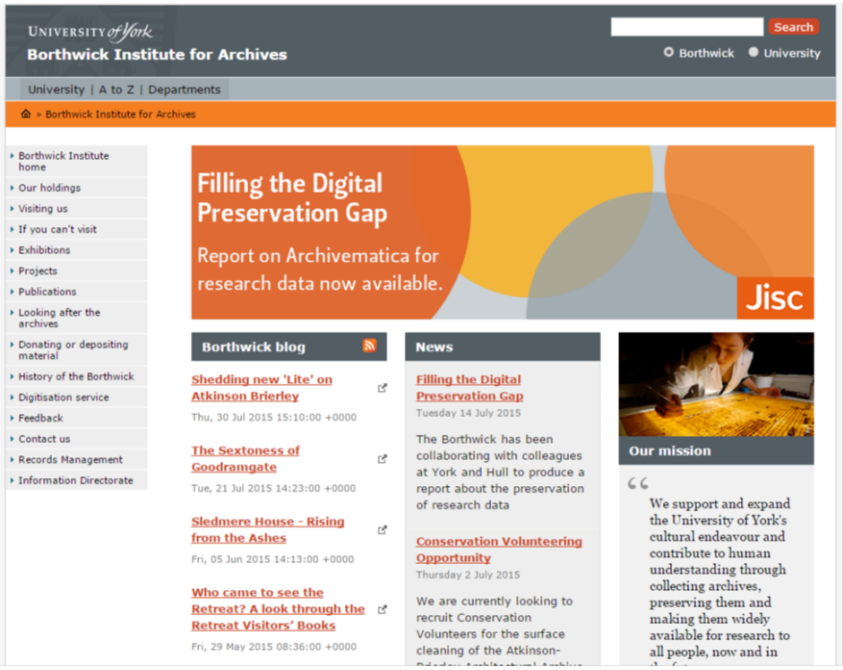
Title of blog post	Date of release	No of views[38]
Archivematica for research data? The FAQs http://digital-archiving.blogspot.co.uk/2015/07/archivematica-for-research-data-faqs.html - *a copy of the first section of our phase 1 report - published as a blog to enable more people to discover it*	24th July	627
Enhancing Archivematica for Research Data Management http://digital-archiving.blogspot.co.uk/2015/08/enhancing-archivematica-for-research.html - *a description of the developments we are proposing to fund during phase 2 of our project*	28th August	371
Spreading the word at the Northern Collaboration Conference http://digital-archiving.blogspot.co.uk/2015/09/spreading-word-at-northern.html - *a report on a presentation we gave on the project at this conference*	18th September	97
Spreading the word on the "other side of the pond" http://digital-archiving.blogspot.co.uk/2015/10/spreading-word-on-other-side-of-pond.html - *a report from the Hydra Connect conference and the poster that we presented there*	29th October	193

[38] as of 8th December 2015

iPRES workshop report: Using Open-Source Tools to Fulfill Digital Preservation Requirements http://digital-archiving.blogspot.co.uk/2015/11/ipres-workshop-report-using-open-source.html - *a report on a workshop at the iPRES conference which included a presentation about our project*	12th November	421
The third UK Archivematica user group meeting http://digital-archiving.blogspot.co.uk/2015/11/the-third-uk-archivematica-user-group.html - *a report on the UK Archivematica user group meeting which included an update from our project*	16th November	163
Sharing the load: Jisc RDM Shared Services events http://digital-archiving.blogspot.co.uk/2015/11/sharing-load-jisc-rdm-shared-services.html - *a report on this event, attended by members of our project team*	25th November	169
File identification ...let's talk about the workflows http://digital-archiving.blogspot.co.uk/2015/11/file-identification-lets-talk-about.html - *a discussion of some of the things we have been considering as we try and improve the way Archivematica handles unidentified files*	27th November	291
Research Data Spring - a case study for collaboration http://digital-archiving.blogspot.co.uk/2015/12/research-data-spring-case-study-for.html - *a case study written during the DPC Digital Preservation Handbook booksprint for the chapter on collaboration*	2nd December	113
Addressing digital preservation challenges through Research Data Spring http://digital-archiving.blogspot.co.uk/2015/12/the-research-data-spring-projects.html - *a synthesis of the phase 2 Research Data Spring projects and how together they help to solve some of our digital preservation problems*	8th December	108

Project website

A project web page has been established on the website of the Borthwick Institute for Archives at the University of York. This is available at http://www.york.ac.uk/borthwick/projects/archivematica/. For most of the period covering phase 2 the project has been further publicised with a banner image on the front page. In the period from early July (when the page was established) to 8th December 2015, there were 228 pageviews representing 188 unique visits to the page.



The Borthwick website showing the front page banner image promoting the Archivematica project

Phase 1 project report

In mid July at the second sandpit workshop our phase 1 project report was made available via Figshare (http://dx.doi.org/10.6084/m9.figshare.1481170). This report has been viewed 1698 times in the intervening period[39].

The report is also available from the University of Hull repository; currently we have no statistics for this version.

[39] Statistics collected on 8th December 2015

Glossary

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

API: Application Programming Interface - protocol that allows integration between software for example to allow third-party developers to create additional functionality for a piece of software

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

CONTENTdm: A software solution from the Online Computer Library Center (OCLC) allowing digital collections to be made available across the web[40]

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)

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

DMAOnline: Data Management Administration Online - a Data Spring Project based at Lancaster University it seeks to provide a single dashboard view of its RDM activities[41]

DSpace: A widely adopted, community open source, institutional repository solution now stewarded by the DuraSpace organisation[42]

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 [43]

EPSRC: Engineering and Physical Sciences Research Council

Fedora: (in the context of this report) An open-source digital repository platform[44]

Figshare: a repository where researchers, institutions and publishers can share research outputs[45]

---

[40] http://www.contentdm.org/
[41] http://www.dmao.info
[42] http://www.dspace.org/
[43] http://www.eprints.org/uk/
[44] http://www.fedora-commons.org/
[45] http://www.figshare.com

Hydra: A repository solution based on a number of “best-of-breed” open-source components, including Fedora[46]

json: - JavaScript Object Notation - lightweight data-interchange format that is easily read by humans and parsed by machines and is supported by all modern browsers

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

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. Version 3 of the standard has just been released.[47]

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.[48]

PURE: Research information system from Elsevier used at York.

Rails: Ruby on Rails - web application framework that provides structures for a database, web service or web pages. It uses json or XML for data transfer and html for display.

RDM: Research Data Management

SHA256: (and SHA-512, md5) hash algorithms that create the unique digital signature or checksum that can be used to prove a file has not changed over time. A single change to a file would produce a different hash value using the same algorithm.

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

UUID: a universally unique identifier

YODL: York Digital Library - a Fedora Commons repository at the University of York for storing multimedia content

[46] http://projecthydra.org/
[47] http://www.loc.gov/standards/premis/
[48] https://www.nationalarchives.gov.uk/PRONOM/Default.aspx

Appendix 1: Hydra in Hull - Preservation workflows

A framing document considering the use of Archivematica to provide preservation functionality for Hydra, the University's repository

v2.0

Richard Green

December 2015

Author's note:

Version 1 of this document was dated August 2015 and comprised essentially the content up to and including the section titled "Transitioning to the new workflow". (Minor amendments to the original text have been made in the light of subsequent thinking.) This version 2 of the document extends the proposal to cover the construction of a proof-of-concept implementation for Phase 3 of the Jisc project "Filling the digital preservation gap".

Table of contents

- Background
- Archivematica
  - Why do we recommend Archivematica to help preserve research data?
  - What does Archivematica actually do?
  - How could Archivematica be incorporated into a wider technical infrastructure for research data management?
- Types of repository content
  - Metadata-only records with no associated local content to manage
  - Content for dissemination but with no apparent requirement for long-term preservation
  - Content for long-term preservation but with no apparent need for dissemination
  - Content with a need both for dissemination and for long-term preservation
- Ingest methods
- Proposed, revised, workflow
- Transitioning to the new workflow

Background

Whilst the Hydra repository at the University of Hull was built with long-term content preservation in mind, this aspect of its functionality has never yet been implemented. Rather, long-term digital preservation has always been one of the things on the "wish list", the priority being to keep things "safe for the time being" – in other words, the short- to medium-term.

In spring 2015 the University of York, with the University of Hull as a partner, was awarded a Jisc grant for the first phase of a potentially three phase project, "Filling the preservation gap", to investigate the possible role of the open source application Archivematica in the long-term preservation of research data. This first phase of initial investigation resulted in a substantial report[49] and the team was successful in gaining funding for a second phase. Amongst other things, this second phase includes funding for Artefactual, the primary developers of Archivematica, to implement a number of enhancements to better equip the software for preservation of research data and also for the two universities to plan a possible phase three which would see proof-of-concept implementations of the proposed systems at each site.

It became apparent during the phase one work that the use of Archivematica would, indeed, be beneficial in any repository workflow designed to facilitate the long-term preservation of research data. Further, though, it became apparent that essentially the same approach could be used for any repository content deemed worthy of long-term management. This relatively brief document is an attempt to re-imagine the workflows around Hydra in Hull such that the option of long-term preservation is an integral part of all ingest routines regardless of the content type (document, image, multimedia, data, etc) or its origin (postgraduate students, research staff, administrators, archivists, external donors, etc). Given the basis of the Jisc grant, it is a *sine qua non* that these suggested workflows deal appropriately with the specific case of research data. Any redeveloped system should be essentially compliant with the widely accepted OAIS model for preservation systems.

[49] Mitcham, Jenny; Awre, Christopher L.; Allinson, Julie; Green, Richard A.; Wilson, Simon P. (2015) *Filling the digital preservation gap : A JISC Data Spring project : Phase One report - July 2015* See https://hydra.hull.ac.uk/resources/hull:11243

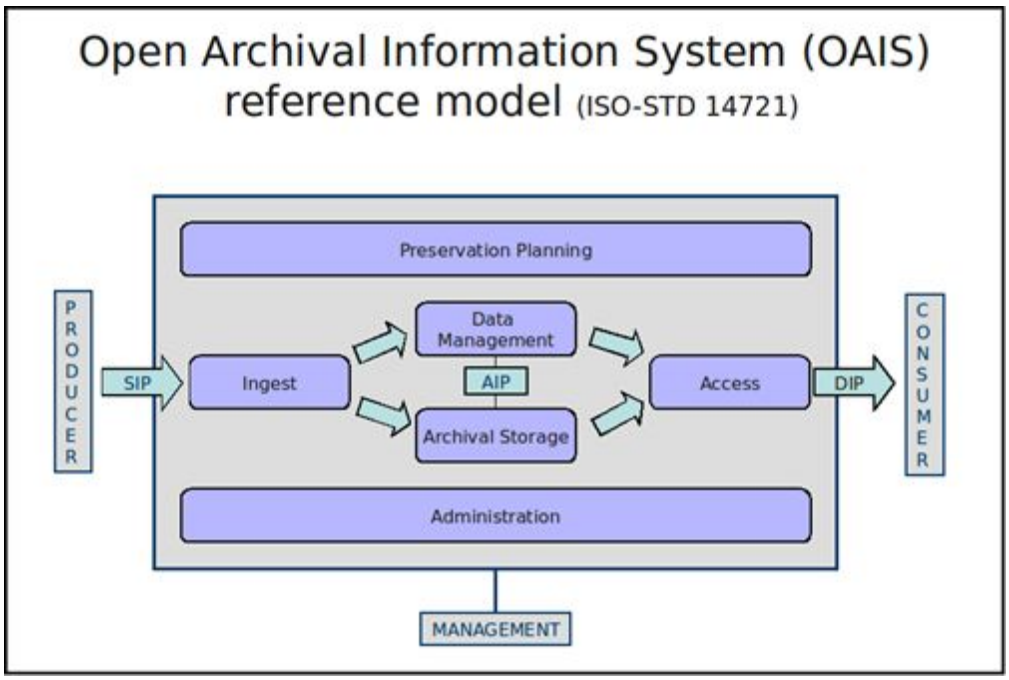


Figure 1: OAIS reference model Source: Archivematica wiki (CC-BY-SA)

The document concludes with details of a proposed proof-of-concept implementation corresponding to Phase 3 of the JISC project mentioned above.

Archivematica

For readers unfamiliar with Archivematica, it will be useful to quote here from the Phase One report referenced above[50] (but bear in mind this document was written to report on Archivematica's potential use with research data specifically):

> 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 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.

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

[50] Op. cit. pp 5-6

- 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 micro-services 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.[51]

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.

The first sentence in last quoted section, "How could Archivematica be incorporated...", is important. Hull sees its repository as an element in an overall infrastructure supporting learning, teaching and research. Other parts of this wider infrastructure are in place to provide the services in the bullet list.

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

Types of repository content

The University repository system overall will contain a range of different content types with a range of different long-term preservation needs. It will be helpful to characterise these at the outset:

- Metadata-only records with no associated local content to manage
- Content for dissemination but with no apparent requirement for long-term preservation
- Content for long-term preservation but with no apparent need for dissemination
- Content with a need both for dissemination and for long-term preservation

Metadata-only records with no associated local content to manage

Metadata-only records have no local content associated with them and there is thus no binary file to preserve. The totality of the new workflows that will be proposed here will need to include provision for such records.

Content for dissemination but with no apparent requirement for long-term preservation

It is likely that there will be content added to the repository that is seen as useful in the short- to medium-term but for which there is no perceived long-term preservation need. It may be appropriate to use an ingest workflow for this material similar to that which exists at the present time, one that makes no use of Archivematica. However, it may be that it will prove more straightforward to pass this material through Archivematica anyway against the possibility that long-term preservation becomes desirable, or simply to take advantage of some of Archivematica's micro-services - for example to generate technical metadata.

Content for long-term preservation but with no apparent need for dissemination

In the case of research data, in particular, there may be a requirement from funding bodies for a repository to preserve it for a period of time even though the likelihood of anyone requesting access to it is minimal. In this case it may be appropriate to provide for its long-term preservation but not to make it available for dissemination and to provide only a metadata record for discovery purposes. This course of action may be particularly useful in the case that the data content is large, in terms of storage requirements, and so the prospect of a dissemination copy in addition to an archival copy is unattractive. The workflows that will be proposed here should enable a dissemination copy to be created retrospectively if access is later required.

Content with a need both for dissemination and for long-term preservation

Perhaps the largest body of material provided to the repository will fall into this category; there is a need for access to it and for its long-term preservation. This implies that there will be two copies of the content in the repository's overall infrastructure: one for dissemination and one for preservation.

Ingest methods

In order to develop workflows to ingest the different types of content described above, we need also to consider the forms in which such content might be presented to the repository (examples are intended to be illustrative only).

- A metadata-only record
- Content generated in the proto-queue stage of the existing repository workflow
- One or more files with no, or minimal, accompanying descriptive metadata, each requiring a repository record
    - Self-deposited collection of research data
    - Some archival materials
- One or more files with accompanying metadata which, excepting the filename, is essentially the same in all cases; each file requires a repository record, for example
    - EAD record with multiple associated files
    - Sequential page images of a book
- One or more files with an accompanying metadata manifest providing detailed descriptive metadata for each file, each file requiring a repository record, for example
    - Greenhill image project
- One or more files with no, or minimal, accompanying descriptive metadata, requiring a common repository record (components of a compound or complex object)
    - Hopefully this would be a rare occurrence, but it should be catered for
- One or more files with accompanying metadata, requiring a common repository record (components of a compound or complex object)
    - Larkin Centre multimedia files

In the case of multiple files, it is possible that they are presented in a tree structure and/or a form of zip file.

The workflows proposed here should be able to deal with all these possibilities in an appropriate fashion. Ideally, they will be a logical development of existing workflows rather than something completely new.

Proposed, revised, workflow

What follows in this section should be treated as an overview. Much more detailed work would be required to specify the detailed workflows.

The following diagram represents the proposed, revised workflow:

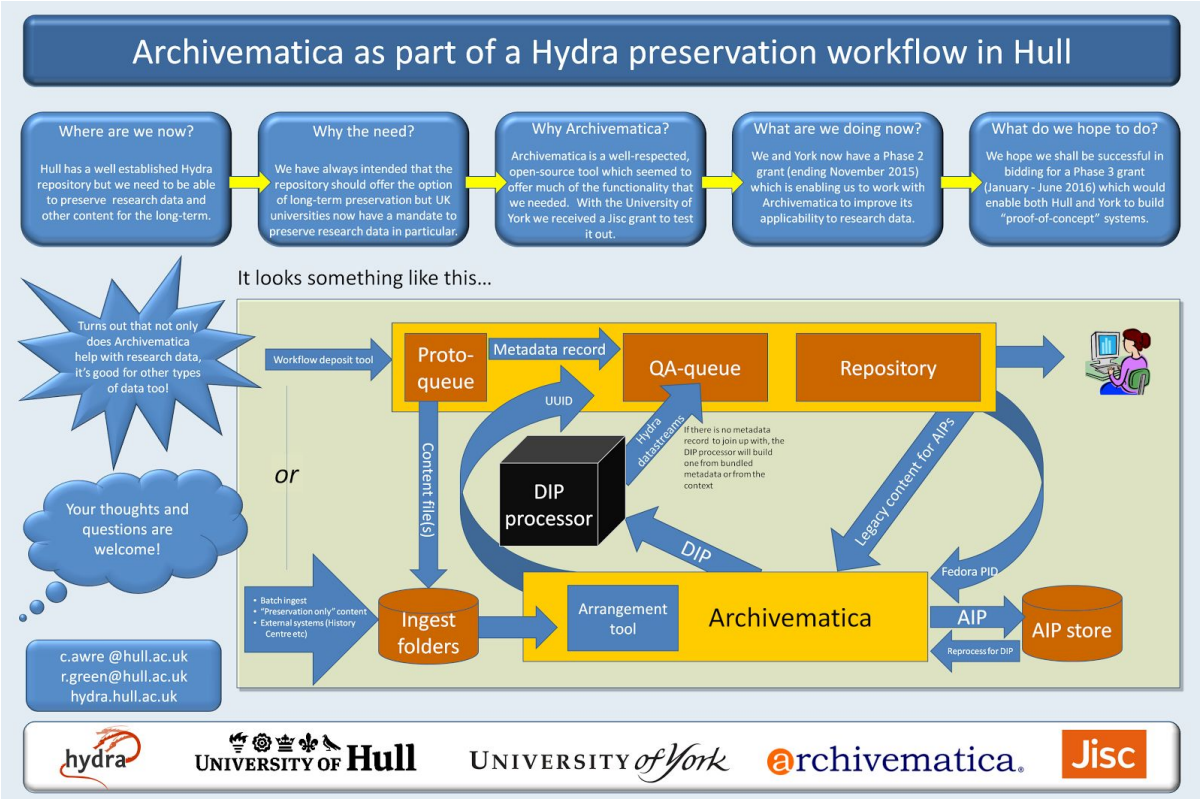


Figure 2: Proposed Hydra/Archivematica workflow

The upper part of the diagram largely corresponds to the existing workflow for Hydra in Hull. In this, content is generated in the proto-queue, passed to the QA-queue and thence to the repository proper. This would remain the case for human-generated, metadata-only, records. For records with associated binary content, the proto-queue interface would be adapted so that, as content files are added to a record they are passed first to Archivematica; subsequently the DIP processor (see below) will insert dissemination copies of these files and any additional metadata into the relevant Hydra object and promote it to the QA-queue.

In addition to the facilities offered by the repository at present, the system will provide ingest folders which would be the collecting point for at least three further categories of content:

- content that was not deemed to require a dissemination copy (and therefore not worthy of time spent generating a detailed descriptive metadata record via the proto-queue),[52]
- content sent to the repository from other systems (for instance, files and a corresponding EAD record from the University Archives), and
- "batches" of content, for instance an image collection

Each deposit into the ingest folders (be that the materials for a single record having an associated single file, or some larger grouping) would be held in its own sub-folder. It may be useful to have these sub-folders contained in a number of different top-level folders, each one collecting content from a specific source and/or of a specific type.

The ingest folders will be "watched" and when a new subfolder is located the content will be passed to an appropriate Archivematica pipeline. Archivematica will then undertake appropriate

[52] This is a subject for debate. The proto-queue tool could be used to generate a very basic descriptive metadata record; otherwise a very simple deposit tool might be needed. Even then, someone with appropriate access to university systems might make a deposit "by hand" – by creating a sub-directory and placing content within it.

processing, depending on the form of content, and produce in all cases a dissemination information package (DIP) and archival information package (AIP).

The DIP will contain all the data required to produce a basic Hydra object within the repository and a DIP processor (probably based largely on the code currently used for proto-queue processing) will be written to deal with it. Where there is a pre-existing metadata record in the proto-queue, the dissemination files and any additional metadata will be injected into it and it will be passed to the QA-queue. Where material has been introduced into the repository system via the ingest folders the DIP tool will build appropriate Hydra objects and place them in the QA-queue. In this second case, in addition to the content for the Hydra object (metadata and binary file(s)), the DIP processor will ideally require two further pieces of information:

- the content model that should be associated with the Hydra object (ETD, data file, etc) in the absence of which it will default to "generic content", and
- information as to whether the Hydra object created from the DIP should have content for dissemination or be created without this, ie metadata-only. The default would likely be to create a content-bearing object; metadata-only would be appropriate to research data that is not likely to be called upon

The AIP, in all cases, will be passed to the AIP store. Each AIP has a unique, universal identifier (UUID) which is available in the DIP and which can therefore be included in the metadata of the Hydra object.

This workflow differs in two significant respects from possible approaches that we have previously discussed:

- all binary content for the repository is processed using Archivematica. This allows us to take advantage of a number of micro-services used by Archivematica that are of potential benefit even if long-term preservation is not seen to be a priority (amongst which the use of FITS, or similar, to extract technical metadata; and a file characterisation service)
- all content-bearing items for the repository will acquire an AIP. We would need to create functionality within the Hydra stack to delete an AIP from the AIP store should the associated Hydra object be fully deleted (as opposed to hidden)

It follows, from the workflow proposed above that the DIP sent from an ingest folder for processing should normally contain descriptive metadata in addition to just the binary payload for an object. This might be in a number of forms, as set out earlier in this document, and the DIP processor would need some "intelligence" to distinguish between them and to perform accordingly. Amongst the possibilities:

- a file containing a single MODS record corresponding to the binary payload(s) (this may be a structured XML file or in the form of a spreadsheet)
- a file containing multiple MODS records, each one corresponding to a file within the deposit (this would likely be in spreadsheet form)
- a file containing an EAD record corresponding to one or more files within the deposit
- no metadata record

In this last case, no metadata record, a very basic record could be built using the subfolder name as a title, elements of the technical metadata, and the UUID for the AIP. This scenario will

probably be confined to collections of research data that are deemed likely to remain unused whilst in the repository system.

As in previous discussions, and as represented at the right-hand side of figure 2, material already in the Hydra repository worthy of long-term preservation can be exported via Fedora and an AIP generated using Archivematica. In such cases the Object and the AIP could be linked via the Fedora PID.

Transitioning to the new workflow

Should the decision be taken to go ahead with the changes proposed here, the transition can easily be phased. It will be perfectly possible to leave the current operation of the repository unchanged whilst the new functionality for the use of Archivematica with ingest folders is developed. Once that is working satisfactorily, then work could begin to modify the workflow between the repository proto-queue and QA-queue.

A proof-of-concept implementation for Jisc

A proof-of-concept implementation of part of this overall proposal for the Jisc "Filling the preservation gap" project should be targeted specifically at workflows for research data. Logically it will start with one of the most straightforward scenarios and then be extended to other situations as time and resources allow (ultimately, and beyond the project, to cover all the needs described above). Thus, as noted in the project's Phase 2 report, development will start from a single file with moderate metadata and from there capability will be extended as far as time and resources will allow.

The general process can be characterised in very crude terms as follows:

	Process	**Notes**
	Content found in watch folder	
	Call transfer API to start transfer automatically	There will need to be some trap to ensure that the content in the watch folder is stable; i.e. that all files have been deposited
	Archivematica creates AIP and DIP AIP to store, DIP to DIP processor	Always create a DIP even if we require a metadata only record - that way we get the technical metadata, file characterisation, etc. The dissemination copies can just be discarded. We shall need a policy on how long DIPs are kept after use
	DIP processor receives DIP	Maybe need to burst the DIP's .tar file if we can't extract direct from its compressed format
	Check DIP for a [metadata.xls] file	Or some reserved name. May be a single

		line file - same basic descriptive metadata for each file. May be multiline - different metadata for each file. May be no file.
	Determine number of files to process	
⇓	For each file	
	Is DIP requested by the human workflow? Yes: then there is already a Fedora object, no need to create one No: create a basic Fedora object	
	If there is no metadata file create descriptive metadata from context If there is a metadata file: match file with metadata entry and get descriptive metadata, add to object	"Context" yet to be defined. Could be associated with the watch folder, derived from SIP title, or...
	Extract tech metadata etc from DIP and add to object	Need to decide how the tech metadata would be held in a compound Fedora 3 object. Probably sufficient for p-o-c to take the entire block and create a techMetadata datastream
	If the Metadata only flag is not set add dissemination file (and PREMIS?) Finish off object	
⇑	More files? Loop.	Only applicable to batch ingest. If from the human workflow it will be a SIP for a single file to go back into the object under construction.
	DIP from human workflow? Send "completed" message back.	

The methodology used around the watch folders must be kept as generic as possible so that it is applicable to a wide range of circumstances.

Objects created by our DIP processor will be Fedora 3 objects in line with our existing repository content. Where possible we will employ common code for the processor with York. Not only will this give us a shared workload, but as York will not be using Fedora 3 (rather, we understand, Fedora 4), this should go part way to demonstrating that the idea is relatively easily adapted between repository systems. We would hope that the idea might be taken up in the context of DSpace and EPrints repositories.

Appendix 2 : Implementation Plan for Archivematica and RDMonitor at York

Julie Allinson and Jen Mitcham, December 2015
University of York

Implementation Plan for Archivematica and RDMonitor
Workflows
Deposit and Preservation Workflow
Before the transfer to Archivematica:
Transfer to Archivematica for storage and preservation:
After the transfer to Archivematica:
Discovery and Access Workflow
Data Access DIP Creation Workflow
Management, reporting and administration workflows
Requirements and Specification
RDMonitor
Data Uploader
SIP Processor
Archivematica Transfer (Submission Information Package - SIP)
DIP Processor
Dissemination Information Package (DIP)
Data Model for Datasets and DIPs
Dataset
DIP
Project Scope
Workplan for Phase 3
Summary

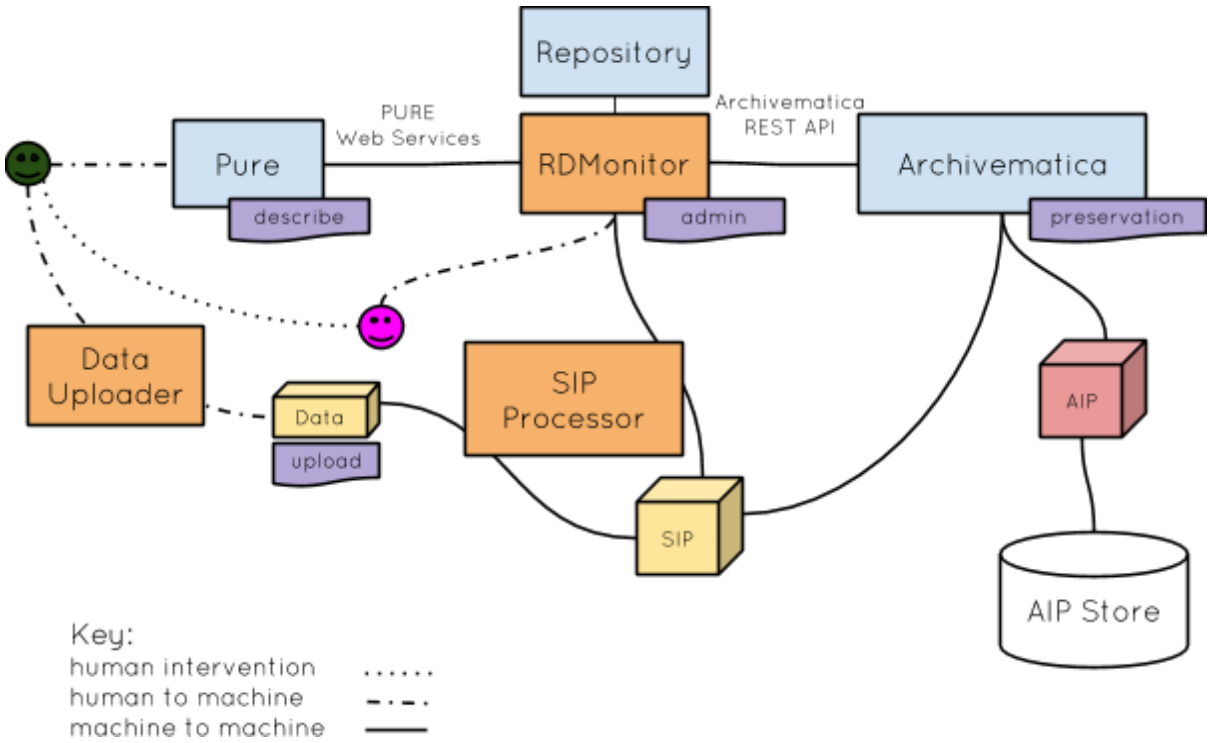
Overview

This implementation plan is intended to be read in conjunction with the corresponding section in the final report for phase two of the Filling the Digital Preservation Gap project. It's aim is to provide some low-level detail on the implementation proposed at York, and has been used during phase two to help us test and validate the feasibility of the proposed proof of concept.

Workflows

What follows is a detailed description of how research data would be captured and processed within our proof of concept implementation:

Deposit and Preservation Workflow



The Proposed Deposit and Preservation Workflow

Before the transfer to Archivematica:

Step	Description	Comments
1.	Researcher enters metadata about a dataset in PURE	
2.	Library RDM staff review the metadata and contact the researcher with any queries to ensure the metadata record is complete. They also establish the nature, size and access requirements for the data, including pointing the researcher to an external data repository as appropriate.	This is a manual step. It is necessary to ensure that data is appropriately described. It is due diligence for ensuring we do not release data that is sensitive or close data that should be open.
3.	If it is established that the data will be stored locally, Library RDM staff organise the transfer of data to centrally managed RDM filestore. The RDMonitor will generate a custom upload link for the dataset.	Need to establish cut off point for 'small' and 'large' data
3a.	If the data is small in size, the researcher is sent an upload link which leads to a short form for uploading the data. An option to supply the DMP or a link to the DMP will be available here, along with other supporting files, such as data agreements. The form would also act as a data transfer agreement between the depositor and the University. The depositor would need to confirm that they have the right to deposit the data and to agree to our standard open licence.	The proposal here is to build a very simple upload form pre-populated as far as possible with existing info. This would make use of our our new Hydra infrastructure for the Digital Library. See below for further discussion of this, and other options. The DMPOnline[53] service may develop an API to make retrieving the DMP more

[53] https://dmponline.dcc.ac.uk/

		automated. This would be a fire and forget form, once record has been created, the researcher cannot edit it.
3b.	If the data is large, the form would be used as the data transfer agreement, but the data would not be uploaded. The Library RDM staff will organise with IT colleagues for a copy of the data to be transferred.	It is anticipated that all data will be on University filestore, so moving/copying it around will be straightforward.
4.	The RDMonitor tool will provide Library RDM staff with info about the status of the dataset, eg. that it been uploaded by the researcher. Once it has been uploaded it can be sent to archival storage at the 'click of a button'.	See notes below for specification for the Submission Information Package (SIP). It is also possible to add extra files at this point, for example a copy of the data transfer agreement.
9.	A SIP is created (with structure as defined later on in this document) and placed in a directory that is watched by Archivematica and a call is made to the Archivematica transfer API to start the transfer.	see Archivematica documentation[54] re creating a structured SIP before transfer

Transfer to Archivematica for storage and preservation:

Step	Description	Comments
1.	Once transfer API has been called Archivematica picks up the SIP and adds it to the ingest queue.	
2.	The ingest begins automatically.	
3.	Archivematica processes and checks the data and creates an Archival Information Package (AIP) (extracting metadata and carrying out normalisations where appropriate) and stores it in it's Archival Store	The archival store will be filestore provided by IT Services. Multiple copies will be kept and checksums will be monitored.

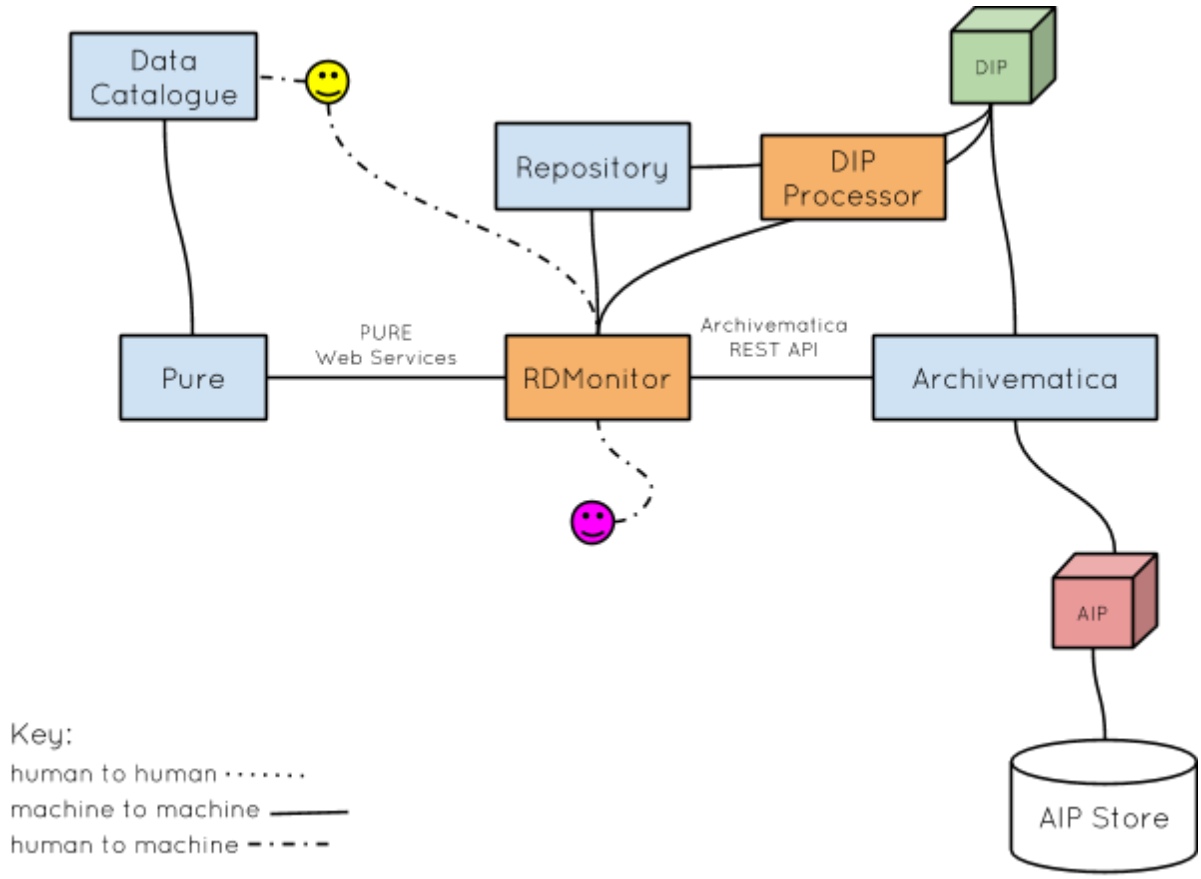
After the transfer to Archivematica:

1.	The RDMonitor gets back the archivematica UUID for the AIP and stores it. Library RDM staff are alerted that this process is complete.	

54 https://www.archivematica.org/en/docs/archivematica-1.4/user-manual/transfer/transfer/#create-submission

2.	Once the item is stored, Library RDM staff can generate the DOI in PURE.	Contingent on data catalogue decision. The PURE DOI will point to the PURE portal. If we do not use the PURE portal then DOI creation will be carried out by the RDMonitor. It will be possible to create the DOI in advance of the data being stored, if necessary.
3.	Researcher is informed of the DOI for their dataset by automated email.	
4.	The RDMonitor generates a URL for the data that the Library RDM staff add into PURE	This URL will either take the user to the downloads for the dataset (if the Dissemination Information Package) DIP has been created) OR will allow them to request access (when the DIP has not yet been created).
5.	As part of the discussion with the researcher, Library RDM staff will have established whether we want to immediately create a DIP (essentially an access copy of the data), or leave that to happen on data access request. If the former, then the DIP can be generated by Archivematica in the next stage of the process.	

Discovery and Access Workflow



The Proposed Discovery and Access Workflow

Step	Description	Comments
1.	User discovers data through the data catalogue and requests the data via the data access URL (noted above)	Each dataset will have a URL even if not currently available to download
2a.	If the DIP has already been generated the user will be presented with a description of the data and a download link. The dataset will be displayed according to it's original file structure, with an option to download the whole dataset as a zip file.	
2b.	If the DIP has not yet been created the user will be presented with a request button and asked for their email address	Potential for other elements in the form if for example we need to collect further information about intended use.
3.	Library RDM staff are alerted to the request, review information about the request and the data from within the RDMonitor and initiate the creation of a Dissemination Information Package (DIP) by Archivematica. They may need to check with the depositing department and/or researcher. If for any reason data access cannot be granted they will communicate with the requestor.	Manual step necessary here again for due diligence to ensure we are not releasing data when we shouldn't. This is intended for Open access requests.

4.	The DIP is transferred into the Digital Library for access.	
5.	An automated email is sent to the requestor providing the download link.	Note, this is the same link that was added to PURE initially, now the DIP is available that link leads to a download page. No further updating in PURE is needed.
6.	On accessing the downloads, stats are captured by the RDMonitor, including date of last access. If the depositor has requested it, they can be automatically alerted to the download of data.	These stats may be simply drawn from server access logs.

Data Access DIP Creation Workflow

Step	Description	Comments
1.	Archivematica creates a DIP and sends a callback on completion with information about the DIP location.	This functionality has been developed by Artefactual Systems in phase 2 of the project.
2.	RDMonitor picks up this callback and initiates the creation of a DIP Object in Fedora, structured as per our specification below.	

Management, reporting and administration workflows

1.	Reporting and analytics on RDM - We will use DMAOnline[55] to provide high level statistics on usage of our RDM services (for example data volumes, access statistics etc). Summary information will be pulled from PURE, DMPOnline, Fedora and Archivematica to enable reports to be generated.	This will help us measure compliance with University and funder policies
2.	Disposal of data - notifications after 10 years from last access will be generated and sent to RDM staff in line with University of York RDM policy. Data will not be automatically deleted, there will be a human step to double check that a deletion is appropriate.	This will be carried out by RDMonitor - storing statistics about last access stats and the generation of notifications when data has not been accessed for 10 years

Requirements and Specification

[55] http://www.dmao.info/

RDMonitor

We are proposing to build a lightweight interface for RDM staff to perform actions related to datasets.

The RDMonitor will:

1. Provide an interface for RDM staff to review the status of datasets
2. Retrieve information about datasets from PURE via its Web Services
3. Store sufficient metadata to associate dataset records in PURE with the data object stored in Archivematica (for example the Archivematica and PURE UUIDs)
4. Track the data deposit and preservation process so that a user can easily see whether data has been stored and whether an access copy has been created
5. Initiate the creation of a DIP by Archivematica (on request)
6. Send automated emails to the requestor to notify them of the availability of the data and to depositor that their data has been requested (if needed).
7. Record information about embargoes and automate email alerts when an embargo is ending
8. Access information about downloads to calculate date of last access
9. Initiate and record data deletion (eg. after 10 years) and automate sending alerts to RDM staff and depositors, provide a decision point for RDM staff to review whether to delete, keep or otherwise archive the dataset.

Data Uploader

We are proposing to build a custom data upload interface as part of the RDMonitor, based on existing work in our Hydra project.

- Create a custom deposit URL which would grab metadata from PURE to speed deposit and avoid re-keying of data
- Capture additional information which we may need from the depositor, for example a README or other documentation on how to understand the data
- Allow upload of or link to DMP
- Act as the data transfer agreement and licence acceptance point
- Provide simple drag and drop for data, including whole folders (retaining folder structure)

SIP Processor

We are proposing to build a component (or ‘gem’) that will build an archivematica-ready package structure from the data it is passed:

1. Create required directory structure (see below for details of SIP structure)
2. Construct a metadata.csv file
3. Add the additional submission documents
4. Add the data
5. Optionally create checksum
6. Add any additional submission documentation

7. Sanitise data and metadata and set permissions to ensure data/metadata is in an Archivematica-ready state
8. Ensure file structure can be adequately passed in the DIP
9. Move to the Archivematica watched directory

Archivematica Transfer (Submission Information Package - SIP)

The SIP should be structured in the following way before transfer to Archivematica - see Archivematica documentation for further information:

- **/logs** - this directory would be empty
- **/metadata** - this directory will include DC metadata, rights and licence information and the PURE UUID in a csv file (metadata.csv). It may additionally include checksums for the dataset.
    - **/submissionDocumentation** - this directory will include the licence/deposit agreement, the DMP and any other submission documentation as appropriate (for example any relevant correspondence)
- **/objects** - this directory will contain the dataset itself (including the full directory structure of that dataset)

DIP Processor

We are proposing to build a reusable component (or 'gem') that will facilitate the retrieval of the DIP and it's repackaging for the repository.

The DIP Processor will:

- Request creation of DIP
- Receive callback to say the DIP creation has completed
- Retrieve and unpack the DIP
- Build Fedora object(s) and create relationships
- Save to Fedora

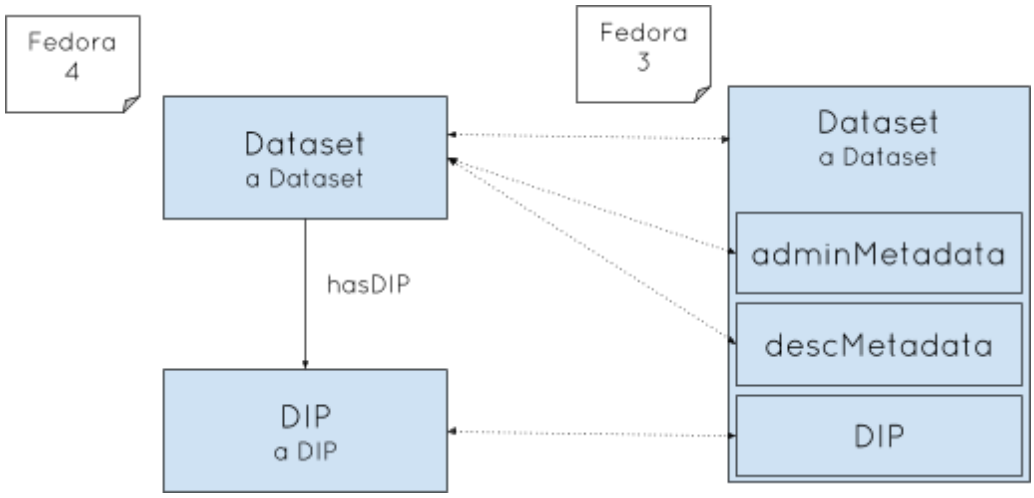
Dissemination Information Package (DIP)

The Archivematica DIP is structured as follows.

- **/objects** - the dataset files, normalized for access
- **/thumbnails** - thumbnail images
- **METS** - the METS file for the DIP
- **processingMCP.xml** - details of the processing archivematica has done

Data Model for Datasets and DIPs

York will implement the prototype in Fedora 4. Two Fedora objects will be created, one containing information about the Dataset, and the other containing the DIP itself. The diagram below shows how the same model might be implemented in Fedora 3, using a single object and object datastreams.



An outline metadata model is in the tables below to indicate what information will be captured and in what format.

Class	Property	Expected Object Type	Usage
Dataset	rdf:type	URI	
	pure_uuid	Literal(String)	1
	sip_uuid	Literal(String)	1
	aip_uuid	Literal(String)	1
	data_status	Literal(String)	1 .. n not yet uploaded | uploaded | stored | store_failed | access | access_failed
	access_copy	URI	link to DIP object
	embargo	Literal(String)	1
	date of last access	Literal(String)	1

Class	Property	Expected Object Type	Usage
DIP	rdf:type	URI	

	skos:prefLabel	Literal(String)	1
	dataset	URI	1
	additionalFiles	URI	0..n

Project Scope

In phase 3 of our Research Data Spring project we want to focus on the core functionality of the system using reasonably straightforward test datasets. We recognise however that this will not fulfill all expected scenarios. In the future there is potential to expand the functionality of RDMonitor in the following ways, but we do not envisage these possibilities will be addressed during phase 3 of our project:

1. Include information about physical data (requests for access etc)
2. Show how the same tool might be used for other (non RDM) workflows, eg. digitization
3. Integrate with DMPOnline
4. Do more with the contents of the dataset, unpacking into different object types, dealing with mixed permissions
5. Support different workflows for different types of access restriction, eg. requestor must complete an access request form, requestor must have an academic referee.
6. Support confidential and restricted data, eg. with different filestore and new workflows for access.