artefactual University of Hull The University of York Information

ARCHIVES

Digital Preservation Requirements for Research Data Management

Jenny Mitcham, University of York
Chris Awre, University of Hull

18 November 2015

Jisc

DP and RDM

- To what extent should a research data management service address digital preservation?

- A lot! DP is an essential component of RDM
- Why?
  - The stick - Funder requirements/mandates (e.g., EPSRC, Horizon 2020)
  - But also...

Impact of no DP in RDM

...but we probably all have our own examples of data loss through lack of preservation planning or actions

artefactual University of Hull The University of York Information

It started with a Christmas wish…

Digital Archiving at the University of York

Home

WEDNESDAY, 11 DECEMBER 2013

My digital preservation Christmas wish list

All I want for Christmas is a digital archive.

Since I started at the Borthwick Institute for Archives I have been keen to adopt a digital preservation solution. Up until this point, exploratory work on the digital archive has been overtaken by other priorities, perhaps the most important of these being an audit of digital data held at the Borthwick and an audit of research data management practices across the University. The outcome is clear to me – we hold a lot of data and if we are to manage this data effectively over time, a digital archiving system is required.

By paparutzi on Flickr CC BY 2.0

In a talk at the SPRUCE end of project workshop a couple of weeks ago both Ed Fay and Chris Fryer spoke about the importance of the language that we use when we talk about digital archiving. This is a known problem for the digital preservation community and one I have myself come up against on a number of different levels.

In an institution relatively new to digital preservation the term 'digital archiving' can mean a variety of different things and on the most basic IT level it implies static storage, a conceptual box we can put data in, a place where we put data when we have finished using it, a place where data will be stored but no longer maintained.

ABOUT THIS BLOG

Some thoughts from a digital archivist working within a traditional archive at the University of York. Opinions expressed are my own.

CONTRIBUTORS

- Nathan Williams
- Julie Allinson
- Jenny Mitcham

BORTHWICK LOGO

Borthwick Institute For Archives

SUBSCRIBE

artefactual University of Hull The University of York Information

well actually a blog…

My requirements are listed below. Feedback is most welcome

#	Requirement
	INGEST
I1	The digital archive will enable us to record/store administrative information relating to the Submission Information Package (information and correspondence relating to receipt of the SIP)
I2	The digital archive will include a means for recording decisions regarding selection/retention/disposal of material from the Submission Information Package
I3	The digital archive will be able to identify and characterise data objects (where appropriate tools exist)
I4	The digital archive will be able to validate files (where appropriate tools exist)
I5	The digital archive will support automated extraction of metadata from files
I6	The digital archive will incorporate virus checking as part of the ingest process
I7	The digital archive will be able to record the presence and location of related physical material
	DATA MANAGEMENT
DM1	The digital archive will generate persistent, unique internal identifiers
DM2	The digital archive will ensure that preservation description information (PDI) is persistently associated with the relevant content information. The relationship between a file and its metadata/documentation should be permanent
DM3	The digital archive will support the PREMIS metadata schema and use it to store preservation metadata
DM4	The digital archive will enable us to describe data at different levels of granularity – for example metadata could be attached to a collection, a group of files or an individual file
DM5	The digital archive will accurately record and maintain relationships

- file extensions (1)
- file formats (4)
- file naming (1)
- Filling the Digital Preservation Gap (10)
- flickr (1)
- floppy disks (1)
- Foldermatch (1)
- google calendar (1)
- guidance (1)
- Hydra (1)
- iCalendar (1)
- ics (1)
- IDCC (1)
- iPRES (1)
- ISAD(G) (1)
- Jisc (6)
- mash up (1)
- media formats (3)
- metadata (2)
- migration (1)
- music (1)
- NDSA levels of digital preservation (1)
- Northern Collaboration (1)
- OAIS (3)
- obsolescence (1)
- open source (1)
- page turning (1)
- PDF/a (1)
- persistent identifiers (2)
- personal digital archives (1)
- PLANETS (1)

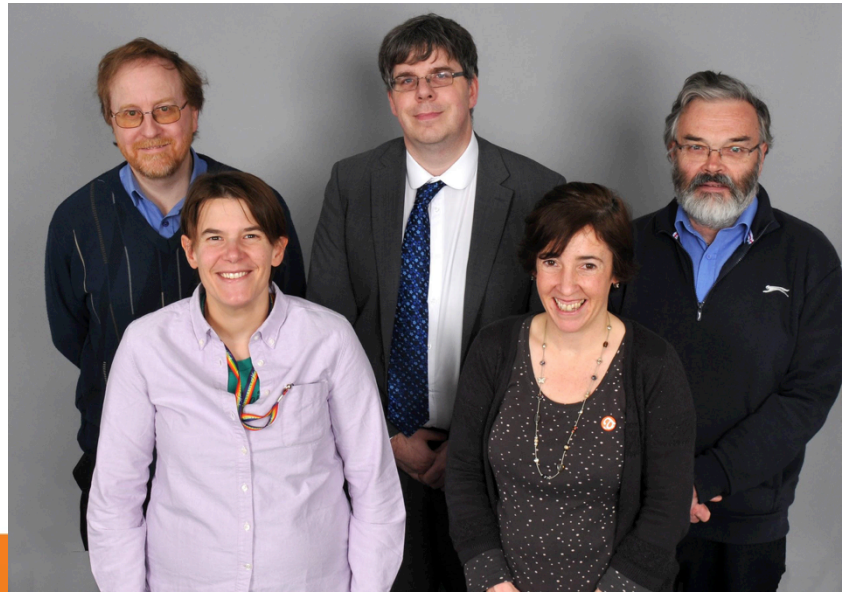
...but this was based on...

- The Open Archival Information System Reference Model (OAIS)
- ISO16363 (Audit and Certification of Trusted Digital Repositories)
- Work of others
  - Chris Fryer's business case work for the SPRUCE project
  - University of Leeds RoaDMaP project
- Discussion with colleagues and the wider community

artefactual University of Hull The University of York Information

...then "Filling the Digital Preservation Gap" came along

- Research Data Spring project funded by Jisc
- Aim: how can we fill the preservation gap for RDM using Archivematica
- We wanted to define our requirements for digital preservation
- Why start from scratch?

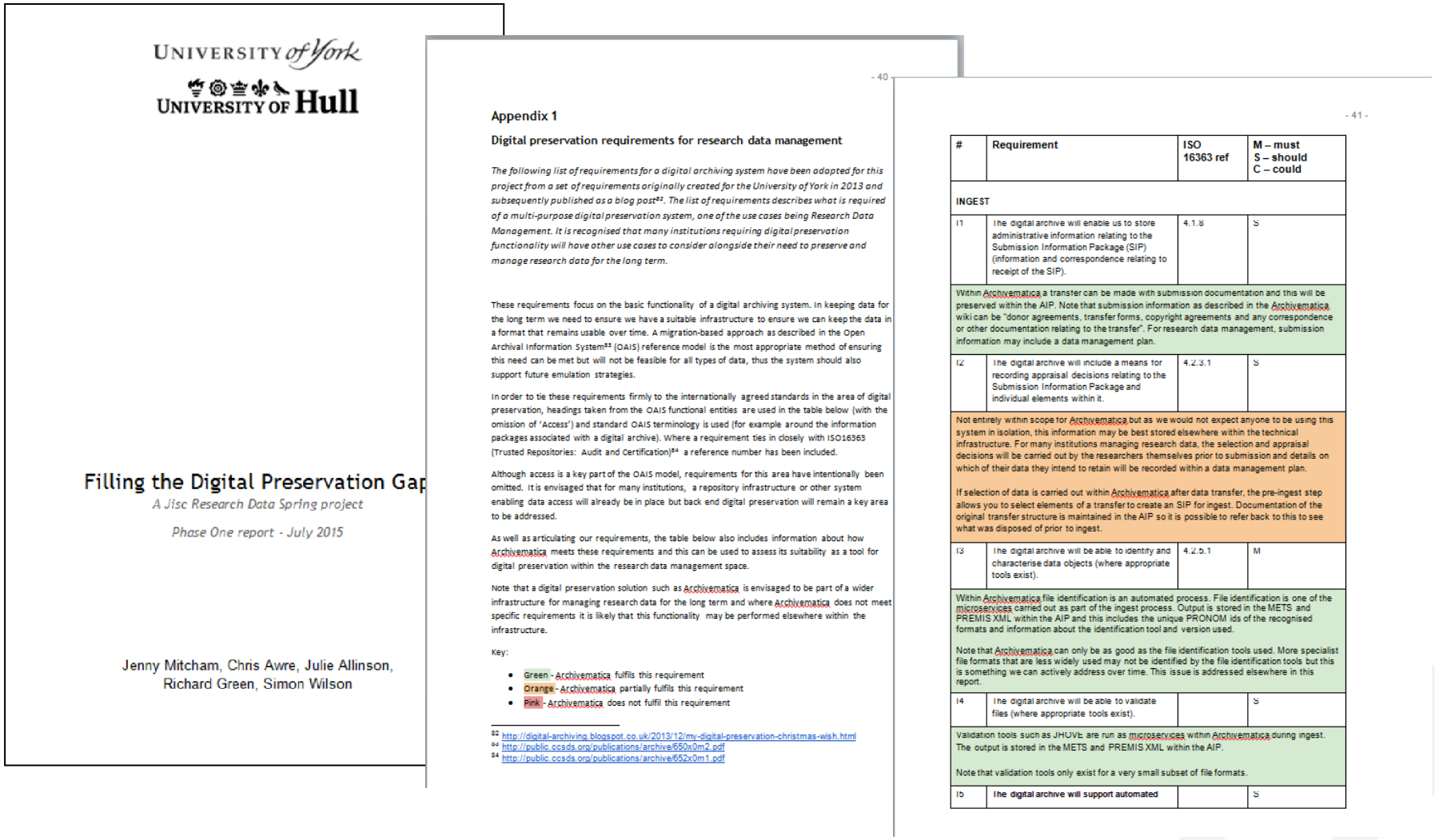


...and we made some additions

- "It will be possible to select and configure the required level of automation within the ingest workflow."
- "The digital archive will be able to process large numbers of files and files that are large in size."
- "A community of users will exist around the software or system to enable sharing of use cases, workflows and to promote developments in line with changes and innovations in the discipline of digital preservation."

artefactual | University of Hull | The University of York Information

Read about it in our project report



University of York
University of Hull

Filling the Digital Preservation Gap

A Jisc Research Data Spring project

Phase One report - July 2015

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

Appendix 1

Digital preservation requirements for research data management

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

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

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

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

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

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

Key:

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

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

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

artefactual | University of Hull | The University of York Information

Plugging in digital preservation

Archivematica as part of a Hydra preservation workflow in Hull

Where are we now?
Hull has a well established Hydra repository but we need to be able to preserve research data and other content for the long-term.

Why the need?
We have always intended that the repository should offer the option of long-term preservation but UK universities now have a mandate to preserve research data in particular.

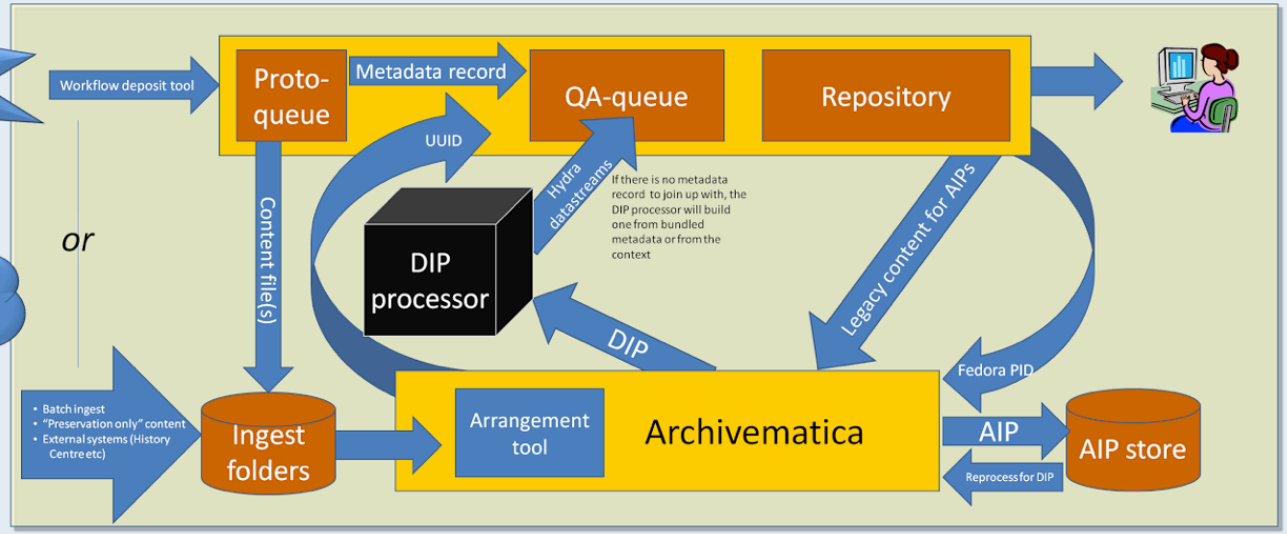
Why Archivematica?
Archivematica is a well-respected, open-source tool which seemed to offer much of the functionality that we needed. With the University of York we received a Jisc grant to test it out.

What are we doing now?
We and York now have a Phase 2 grant (ending November 2015) which is enabling us to work with Archivematica to improve its applicability to research data.

What do we hope to do?
We hope we shall be successful in bidding for a Phase 3 grant (January - June 2016) which would enable both Hull and York to build "proof-of-concept" systems.

Turns out that not only does Archivematica help with research data, it's good for other types of data too!

It looks something like this...



Your thoughts and questions are welcome!

c.awre @hull.ac.uk
r.green@hull.ac.uk
hydra.hull.ac.uk

hydra | University of Hull | University of York | archivematica | Jisc

Making DP work for RDM

- Digital preservation can be a component part of an RDM service infrastructure
  - It needs to be
- It needs to link into other components through open standards
  - Interfaces to other systems for passage of data
- Option to push data through a preservation workflow or not, as preferred
  - But encouraged...

Thanks for listening

Useful links:
Digital archiving blog: http://digital-archiving.blogspot.co.uk/
Archivematica: https://www.archivematica.org/en/
Report: http://dx.doi.org/10.6084/m9.figshare.1481170