UNIVERSITY of York

UNIVERSITY OF Hull

Filling the Digital Preservation Gap

A Jisc Research Data Spring project

Phase Three report - October 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 our funders and several organisations and individuals who contributed information for it. In particular we would like to thank staff at Artefactual Systems for their advice and support throughout the project and The National Archives for their work on new research data file signatures in PRONOM and advice as we created our own. Also Fergus McGlynn and Cottage Labs for their hard work on our proof of concept implementations. Thanks also to other members of the digital preservation community that have engaged with our work on file formats; both those who have profiled the files in their digital archive using DROID (Max Eckart from Bentley Historical Library, Pawel Jaskulski a former trainee at Norfolk Record Office and Rachel MacGregor from Lancaster University) and those who have blogged about their attempts to create file format signatures for PRONOM (Andrea Byrne of Archives New Zealand and David Heelas a former trainee at Hull History Centre). Carl Wilson from the Open Preservation Foundation has also been very helpful in discussing the file formats problem with us.

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

Contents

Executive summary

This report describes work carried out at the Universities of Hull and York on phase 3 of the Filling the Digital Preservation Gap project. The work described here built on the work that was carried out in phases 1 and 2 of the project. The report is in three parts.

The first section of the report describes proof of concept implementations of Archivematica for the preservation of research data at Hull and York. These implementations were integrated with other systems and services in use for managing research data at each institution. Though both implementations integrated Archivematica with a Fedora based repository, the implementations were not identical due to differing research information and storage systems as well as institutional requirements. Both institutions have successfully created largely automated systems for the longer term preservation of research data using Archivematica as a key element of the infrastructure.

The second section describes another strand of the project which looked specifically at the issue of research data file formats and the challenges involved in identifying these files automatically with current tools and registries. As well as exploring the nature of the problem (with the comparison of file format profiles and identification rates and methods) further research focused on how we might improve on this result by enhancing the available registries. A targeted piece of work in this area has increased the number of research data types within the PRONOM registry but there is a recognition that further community effort in this area is necessary. The report concludes with a set of recommendations for further work in this area.

The final section of the report details the outreach work that the project team carried out during phase 3 through presentations, publications and blog posts.

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 Filling the Digital Preservation Gap was 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]

Phase 2 of the project built on the phase 1 work, using the findings of the feasibility study to take practical steps to enhance Archivematica for use for research data preservation. We worked with Artefactual Systems on five discrete developments, and recognising that improving documentation for Archivematica lifts one of the barriers to uptake, we also funded a small piece of documentation work. The development work was designed to address some of the features of research data (trying to reduce the bottlenecks around creating SHA256 checksums for large datasets and implementing a way to highlight unidentified files) and the system integrations and workflows required for the research data infrastructure (allowing better integration with repository systems, automating the generation of a Dissemination Information Package (DIP) as part of the re-ingest process and enabling third party tools to access information for reporting purposes). Additionally in phase 2 the project team worked on their own implementation plans, mapping out how Archivematica would be implemented at each institution. These, along with full details of the development and dissemination work are available in our phase 2 report:

> *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[2]

As each phase of this project builds upon the previous phases it is suggested that readers familiarise themselves with the phase 1 and 2 reports in order to fully understand the context of the project. This report, for example, references the implementation plans in the phase 2 report which provide details of the workflows we were trying to establish.

---

[1] http://dx.doi.org/10.6084/m9.figshare.1481170
[2] http://dx.doi.org/10.6084/m9.figshare.2073220

This report describes the work that has been carried out during phase 3 of Filling the Digital Preservation Gap. Phase 3 ran for a period of six months from 14th March to the 14th September 2016. Work in phase 3 had the following aims:

- To establish proof of concept implementations of Archivematica at the Universities of Hull and York, integrated with other research data systems at each institution
- To investigate the problem of unidentified research data file formats and consider practical steps for increasing the representation of research data formats in PRONOM[3]
- To continue to disseminate the outcomes of the project both nationally and internationally and to a variety of different audiences

Our work in these areas will be discussed in detail in this report.

1. Implementation

Given the available budget and timescales at play, the purpose of our implementation work at Hull and York Universities was to establish a proof of concept rather than a production installation of Archivematica for research data. Our priority was to demonstrate that the workflows and integrations described in our phase 2 report were possible. Although Hull and York have some common infrastructure (namely Fedora and Hydra repositories), there are also differences in systems and facilities (for example CRIS and data storage) so our Archivematica implementations are not identical but instead fit with priorities and workflows at an institutional level. Whilst being able to benefit from discussions across the project team on how best to implement Archivematica we also feel it is useful for demonstrate how Archivematica can be established in two different institutional contexts.

To aid comparison between the two implementations, their key features are summarised in the table below.

Infrastructure element or function	**University of Hull implementation**	**University of York implementation**
Metadata deposit	Via Box, in a file associated with the data file(s)	Via PURE
Data deposit	Via Box, though a shared folder	Bespoke web form developed in Ruby/Rails
Transfer via ...	Box Watcher (a script that watches the shared Box folder and automatically initiates the transfer to Archivematica)	Watched directory (when something is placed in this directory it initiates a transfer to Archivematica)
SIP packaging	BagIt	SIP organised into a folder structure understandable by Archivematica. No

[3] http://www.nationalarchives.gov.uk/PRONOM/

		packaging (e.g., BagIt) applied.
Ingest into ...	Archivematica	Archivematica
Archivematica workflow via ...	Automation Tools (fully automated)	Automation Tools (fully automated)
DIP created?	Always	On request
DIP sent to ...	Hydra repository (via DIP processor to unpack DIP and generate Hydra objects)	Hydra application backed with Fedora 4 repository
DIP discovered via ...	Hydra repository	Data catalogue (to be defined)
AIP sent to	University of Hull Research Storage Service	University of York filestore

Table 1: A comparison of the proof of concept implementations at Hull and York, highlighting the similarities and differences

The following section of the report describes each implementation in turn.

Implementation at Hull

As noted in our phase 2 report, “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.” With this in mind, the Hull implementation uses the Box collaborative cloud storage system[4] as its point of ingest. Box is available to all Hull staff and students through an institutional subscription, and is integrated with the campus single sign on system, so can be used to store files from many different workflows. Thus, Box takes the place of the Ingest folders from the original architecture described. We have also focused on how we use this ingest route, as it is more relevant to our use case over direct ingesting to Hydra, which we will re-visit at a later date.

[4] https://www.box.com

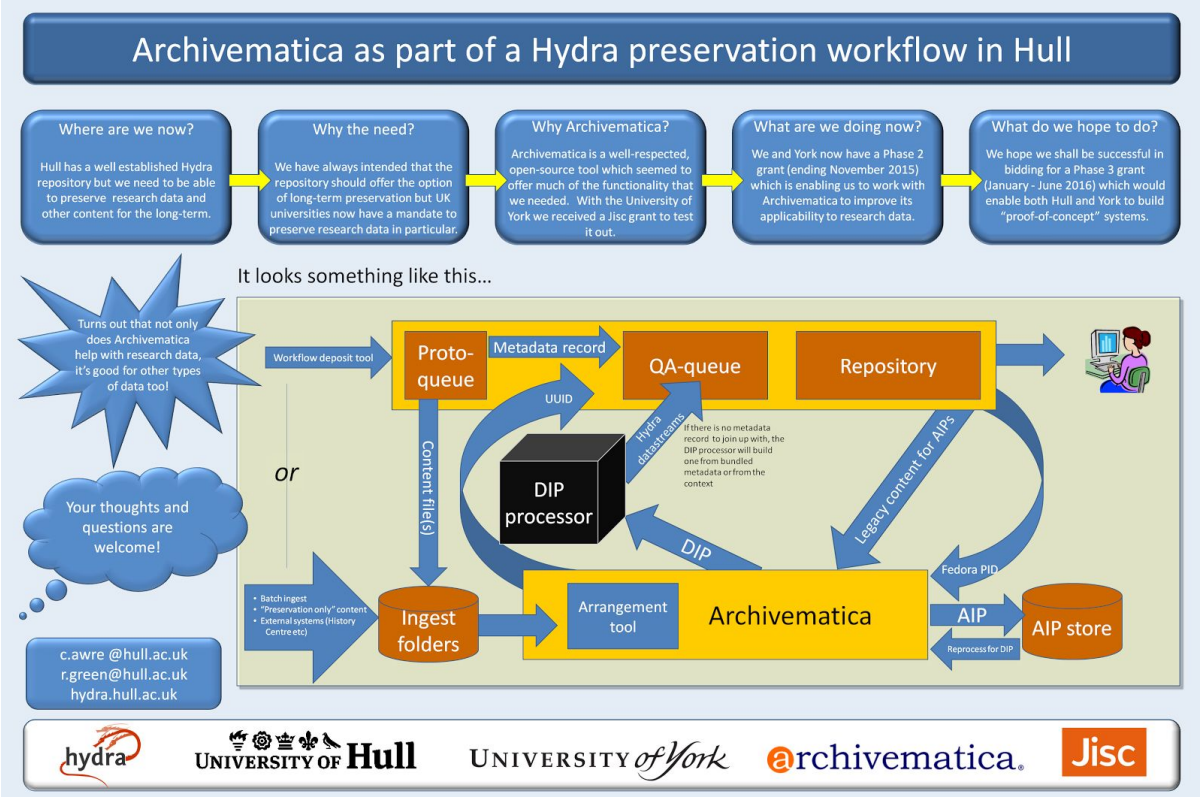


Figure 1: Hull RDM preservation architecture

How it works

Depositors assemble the material that they wish to deposit in a Box folder within their own Box space. This material may be a single file or multiple files, it may also be a folder structure containing multiple files in an organised hierarchy; the resulting repository content may be a single object or multiple objects. These objects describe the file(s) that have been placed in the University's Research Storage Service and may optionally contain copies of the data for download. Users will be provided with instructions for preparing up to four simple files (depending on their needs) for inclusion in their folder and these will determine the way in which their data will be handled (see full instructions in Appendix 1). In practice we anticipate that many users will find they need only to generate one of these files, that which provides descriptive metadata to go with the content (the additional files deal with more complex deposit requirements). This descriptive file is a simple text file with a number of defined fields (including all the Dublin Core fields) of which only "title" and "author(s)" are mandatory.

When the user has assembled their data they use the facility in Box to share the relevant folder with a user called "Archivematica". The proof of concept system keeps a watch for new shares of this type and will process any that appear. Our "Box watcher" service will check the contents of the shared folder and, if it is valid, will create a "bag" (as in the BagIt standard[5]) which is passed to Archivematica. Archivematica then processes this bag of information to create an Archival Information Package (AIP), which is passed to preservation storage in the Research Storage Service, and a Dissemination Information Package (DIP) which is passed on for further processing. The final stage of Hull's proof of concept system sees the contents of this DIP transformed into one or more Hydra objects, which are then

[5] https://en.wikipedia.org/wiki/BagIt

passed into the quality assurance queue of the Hydra repository for checking prior to provision of access as appropriate.

The additional coding for this proof of concept system was carried out by Cottage Labs[6] and is available from Hull's github repository[7].

Future work

The next steps for the University of Hull are twofold:

- To test the proof of concept with a variety of real life examples of data, giving specific attention to the different use cases, from single files through to a directory with hierarchy. The output from these tests will be disseminated separately, and inform the definition of a production instance of the proof of concept that we can provide to researchers wishing or required to archive their data.
- To use the same system as the basis for a digital archive for the City of Culture 2017, encompassing collections of business materials, materials from artists, and materials generated by those attending events. This will use the proof of concept as a kernel for a broader system, adding additional functionality to assist with the processing of the materials for archival purposes.

The project has also made us aware of the value of getting individual tools to do what they do best and combining these as required to achieve a solution that is greater than the sum of its parts. We will also be continuing to liaise with York, as well as other members of the Hydra and Archivematica UK communities on ongoing development of the tools to facilitate their future development.

Implementation at York

York's proof of concept implementation is very close to what was envisaged during our phase 2 project: "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 term 'RDMonitor' has been dropped in favour of 'Research Data York' to mirror the name of our local RDM service. At the time of writing the phase 2 report we had not investigated using Automation Tools[8], nor were we aware of the Puree gem[9] from Lancaster University Library, both of these have saved much local development effort and demonstrate our desire to use community solutions over local ones wherever practical.

[6] http://cottagelabs.com
[7] https://github.com/organizations/uohull
[8] https://github.com/artefactual/automation-tools
[9] https://github.com/lulibrary/puree

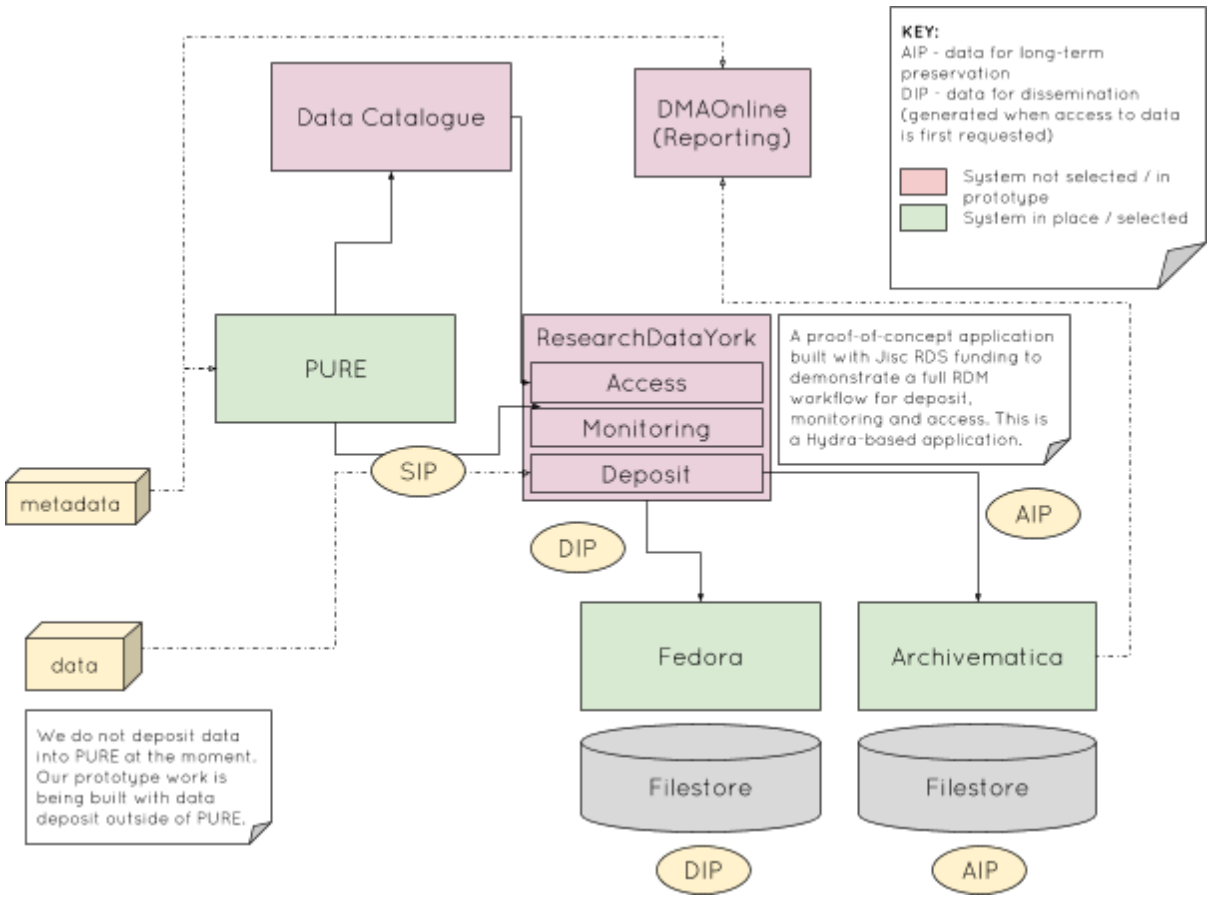


Figure 2: York RDM Architecture Diagram

How it works

At York, researchers are required to add metadata about datasets they produce into PURE, our research information system. Research data staff in the library then review the metadata and liaise with the researcher on the deposit of the data. 'Research Data York' (RDYork) is designed to streamline this process. RDYork is a Hydra application that provides an admin view on datasets in PURE. It pulls new or updated records in on a nightly basis, or on demand by PURE id. Research data staff can then record information about the status of the dataset - do we need further info from the researcher? do we need to store the data locally or has it been deposited in an external data archive?

For data that we need to store locally, RDYork then generates two URLs for each dataset we need to store locally - one for the data deposit and one for data access. The former is sent to the researcher for them to make the deposit. Currently we support file, folder and Google Drive uploads. We would also be keen to explore Hull's approach for sharing the data and see if we can also offer upload by sharing with a specific Google account via the Drive API.

On receiving a data deposit, upload to Archivematica and subsequent processing of information packages is fully automated and the RDYork application is updated to show the status of the data. Once the AIP is stored by Archivematica, we can delete the locally deposited copy.

One feature of York's implementation that our phase 2 development work was designed to facilitate was our DIP creation workflow. It was agreed that we didn't want to create dissemination copies of the data by default. The rationale for this is that we don't know the reuse value of data deposited with the RDM service and, especially where datasets are large, we would rather create dissemination copies when the data access is requested.

The second URL generated by RDYork (mentioned above) is a data access URL. This URL will be made available alongside metadata about the dataset within our data catalogue. When a user reaches this page, one of two things will happen:

- If this is the first time the data has been requested, the user will be asked to supply a valid email address and will then be alerted when the data is available.
- If a prior request has been made, the data will be automatically available for download at the URL.

Research data staff are alerted by email to the initial request for access, so they can make sure there are no restrictions on the data. They then approve the request (if appropriate) through the RDYork application and another automatic process is initiated to request the creation of a Dissemination Information Package (DIP) by Archivematica. Files from the DIP are ingested into the Fedora repository sitting behind the RDYork application.

The code for York's proof of concept system is available from the Library & Archives Technical Team github repository[10]. A screencast showing the working application is also available[11].

Future work

Almost in parallel with the Research Data Spring projects Jisc were planning a Research Data Shared Service[12]. The programme to build this service has now started and the resulting system will be managed and hosted by Jisc, and will offer three core modules : repository, preservation and reporting. The Phase 1 and 2 reports from Hull and York have been influential for scoping the preservation module and demonstrating how repository and preservation systems could be integrated. After a tendering exercise there are commercial and open source offerings for each module, including Archivematica (for preservation) and Hydra (for the repository). Over 20 pilot institutions have been recruited (including York) and all have identified preservation as a priority.

York plan to continue the implementation work we have started but ensure it continues to align with the Jisc Shared Service. With longer timescales at play for the establishment of the Shared Service we are also keen to move the prototype we have developed into production in the meantime so that our research data team have a system through which they can manage deposits to RDYork.

There are several areas we would like to address in order to refine and polish our implementation work and move it into production:

- Complete work on the fully automated DIP re-ingest process.
- Carry out further testing and implement improvements as required.
- Finalise the data model (see appendix 2) and align this with broader Hydra activity at York.
- Deploy code to a production architecture.

[10] https://github.com/digital-york/researchdatayork
[11] https://www.youtube.com/watch?v=3cl5W_7gYvM
[12] https://www.jisc.ac.uk/rd/projects/research-data-shared-service

- Investigate the University of Lancaster's 'preservation' gem[13] as a standard means of interacting with Artefactual's Automation Tools.
- Consider extracting some functionality from the RDYork application into separate gems to better able reuse of the code-base. The code to extract the file structure from the Archivematica METS file has already elicited external interest, for example
- Investigate Hull's 'Box Watcher' approach with Google Drive.
- Extend functionality to support data deposits that aren't made via PURE, in particular those from PhD students (who don't have PURE accounts at York).
- Consider how to marry the automated workflows defined here with the need to solve the file format identification problem (described in a later section of this report).

In addition, we will be testing out Lancaster's DMAOnline[14] reporting tool with data from PURE and Archivematica.

Lessons Learned

At the start of phase 1 we set out to answer the question: 'is Archivematica suitable for preserving research data?', and at the end of phase 3 our conclusion remains that it is. Developments in the software during the course of phases 2 and 3 (both those that the project funded and others sponsored by other institutions), make Archivematica better able to deal with larger datasets, easier to automate, and introduce options for querying Archivematica to assess RDM readiness. The proof of concept implementations at both Hull and York successfully demonstrate fully automated ingest workflows and both make use of Artefactual Systems' Automation Tools in order to do this.

Some lessons learnt from the project include:

- It can be hard to balance tight project deadlines (ours) with product release cycles (Archivematica) where an externally developed system is being used.
- Translating the phase 2 implementation plans into development plans benefitted from conscious and detailed discussion on how ideas translated into required code.
- Establishing the server infrastructure on which we needed to implement the proof of concept required negotiation and regular communication with local IT staff, particularly around firewall and security settings. Making the clear distinction between a test proof of concept and a production system at an early stage will help define the level of work required.
- When you don't have much time, work with good developers. If this resource is not available (with free development cycles to match the project timeframe) in house, have external resource available to tap into.
- Archivematica is reasonably easy to install in a standard configuration but there are documentation gotchas, particularly if you want to diverge or delve further into the flexibility of the system. Improving developer documentation remains a challenge for Archivematica.
- Developing out in the open, seeking peer review and discussion work is definitely the right approach.

[13] https://github.com/lulibrary/preservation
[14] http://www.dmao.info/

2. Research data file formats

In our phase 1 report we discussed the nature of research data and highlighted the many and varied software applications and file formats that are in use for research. In the report it was noted that “It is clear that as we start to ingest research data into a digital archive we will encounter many files that will not be automatically identified”[15] but this hypothesis was not formally tested.

Our work during phase 2 partially addressed this problem by funding a development within Archivematica which would enable users of the system to more easily locate those files with a format that can not be identified automatically. Highlighting these files would enable the digital curator to carry out follow up actions. Further development work to enhance and improve this functionality was also proposed and discussed.

In phase 3 we continued this theme alongside our implementation work. Preserving digital data isn’t solely reliant on the implementation of a digital preservation system, it is also necessary to think about related challenges that will be encountered and how they may be addressed.

Profiling research data

Since May 2015 the University of York has been accepting datasets from researchers into its institutional repository. Many of these deposits came as a result of the EPSRC Expectations [16] thus the majority of the data was scientific in nature, much of it originating from the Department of Chemistry. We used this collection of research data to test the hypothesis from our phase 1 report regarding the variety of formats and difficulties that would be encountered when trying to automatically identify formats.

We ran DROID[17] (a file identification tool from The National Archives) over the files in the collection to establish which file formats within the collection could be automatically identified and the results of this exercise were published as a blog post[18].

Our initial sample size wasn’t large, and actual analysis was carried out on a sample smaller than anticipated. When the analysis was underway it became clear that DROID does not look inside all of the zip files that we hold[19]. Datasets submitted to Research Data York typically arrive zipped up and we do not extract the files as a matter of course. The sample included 8 .rar files which contained another 1291 digital objects which were not included in

[15] A fuller discussion on the nature of research data and the common research data software applications in use at the University of York can be found in our phase 1 report
[16] https://www.epsrc.ac.uk/about/standards/researchdata/expectations/
[17] http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/
[18] http://digital-archiving.blogspot.co.uk/2016/05/research-data-what-does-it-really-look.html
[19] A response received from The National Archives on 19th May 2016 states that “DROID can only probe within the following archival container formats: zip, gzip, tar, plus the web archival formats .arc and .warc. It is on our development roadmap for DROID to add the ability to probe further archival containers, including .rar, 7z, .bz, .iso, which would provide the requested functionality, but we don't currently have a target release date for this”

the statistics (though a quick check showed that the percentage of identified files in this additional sample was strikingly similar to that reported).

The headlines were quite startling. Of the 3752 files analysed, only 1382 (37%) were assigned a file format identification by DROID. This success rate appeared to be very low but we were keen to compare this figure with similar profiles from research data and other born digital collections, so, within the blog post, a question was posed:

> *Is identification of 37% of files a particularly bad result or is it similar to what others have experienced?*

As part of this project, a comparable test was carried out on the born digital holdings of Hull University Archives and the results were very different with an impressive 98% of files identified[20]. Bentley Historical Library, Norfolk Record Office also engaged in this exercise and shared their results[21]. This was a great example of how the digital preservation community can work together and learn from each other.

These profiling exercises certainly seemed to demonstrate that 37% of files identified at York *was* a particularly low result, but considering the small sample size for the York research data it was decided that it would be beneficial to look at DROID profiles of other research datasets. The University of Hull carried out a DROID profiling exercise on the research data held on their research storage area network[22], which currently holds work in progress as well as final outputs, and the results were quite different to those for York research data with 89% of files assigned an identification by DROID. The sample size for this study was very large (leading to issues when trying to analyse the results) and a look at the top ten identified formats showed that 70% of the files on this storage area were TIFF images (from research groups carrying out medical imagery for the most part). This is perhaps not typical of research datasets and this will certainly have contributed to the high format identification rate.

Lancaster University also carried out a similar study, looking specifically at research datasets that had been deposited in their institutional repository[23]. A file format identification rate of 46% was not too far off initial findings from York, though as can be seen in the table below, the method of identification differed.

[20] The results of this exercise were published as a blog post: http://digital-archiving.blogspot.co.uk/2016/08/research-data-is-different.html
[21] Results from Bentley Historical Library: http://archival-integration.blogspot.co.uk/2016/06/born-digital-data-what-does-it-really.html; results from Norfolk Record Office: http://digital-archiving.blogspot.co.uk/2016/09/file-format-identification-at-norfolk.html
[22] The results of this exercise are not published elsewhere
[23] These findings from Lancaster University were shared directly with project team and are not yet published

Institution and test data	No of files in sample	% of files identified	% identified by signature	% identified by container	% identified by extension
RESEARCH DATA					
University of York Research Data	3,752	37%	48%	5%	47%
University of Hull Research Data	10,174,380	89%	data not collected		
University of Lancaster Research Data	24,069	46%	90%	1%	9%
OTHER DIGITAL ARCHIVES					
Hull University Archives Born Digital Holdings	270,867	98%	data not collected		
Bentley Historical Library Born Digital Holdings	731,949	90%	88%	10%	2%
Norwich Record Office Born Digital Holdings	49,117	96%	83%	15%	2%

Table 2: A summary of DROID analysis of results for several institutions. Note that for ease of comparison, figures quoted are rounded up or down to nearest whole number.

As well as the low file format identification rate for York research data, another finding reported in the blog post related to the method of identification for those research data files that were identified. Only 53% of the identified files in York's research dataset were identified by signature or container (methods which suggest more reliable and accurate identification) and this contrasts to 98% at both Bentley Historical Library and Norwich Record Office and 91% for research data at Lancaster University. The remaining identifications were carried out by file extension. File extension is not the most reliable identification method, given that unrelated files coming from different software applications can share the same extension (for example .dat files as discussed below). In the York results 47% of identified files were identified in this way which does suggest that further human intervention may be required to validate those identifications.

Another interesting finding from the work to identify data deposited with Research Data York was the large number files with no file extension. Of the files that were not automatically identified, files with no extension made up 26% of the total. Given the reliance on identification by extension within this sample, this does rather reduce our chances of identifying these files. The second largest group of unidentified files (12%) were files with a .dat extension. This is a fairly common file extension and represents a whole range of files

produced by a variety of different software applications, hardware or operating systems[24] so this isn't a problem that could be solved by the creation of a single file format signature.

Of course there may be other things at play here, for example it would be interesting to explore whether the date of the sample data is a factor in the results. The York research data sample was all deposited within the last 18 months and been modified relatively recently. As stated in the blog post "The data is mostly fairly recent, as suggested by the last modified dates on these files, which range from 2006 to 2016 with the vast majority having been modified in the last five years"[25]. As illustrated in the graph below there is only a small quantity of data from 2006 and the next oldest dates from 2009. Not all of the other studies quoted included details of file dates, but the Bentley example demonstrates that the data was collected over a longer period of time and the range of dates present is wider: "The vast majority of the data was last modified in the past 15 years, and our peaks are in in 2006 and 2008."[26]

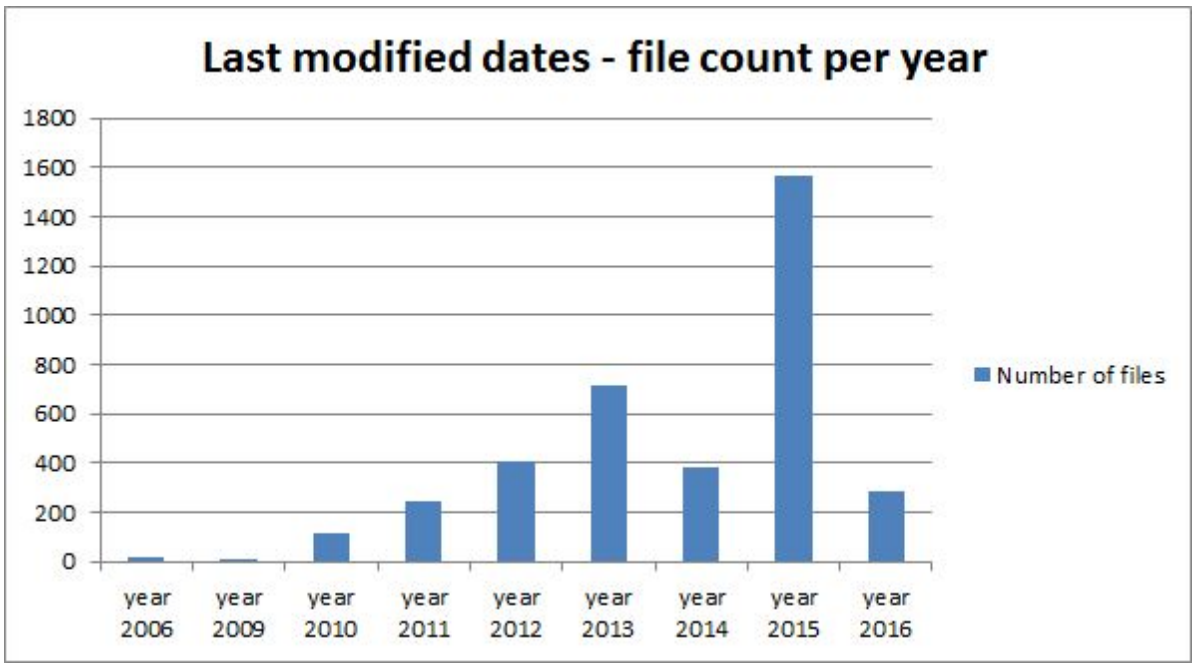


Figure 3: Last modified dates in York's research data sample

We could speculate that given the pace of technological change, it is possible that data that has been worked on more recently may be more likely to have been saved in newer file formats and those newer formats may not yet have made their way into the PRONOM registry.

Our initial blog post was a fairly quick and informal snapshot of research data at York but as others have contributed to the picture it becomes clear that the landscape is a varied one. It would be useful to carry out a more formal piece of work in this area with a larger sample of contributing institutions and guidelines in place regarding the PRONOM signature file version and internal settings in use within DROID. This would allow for easier comparison of results across the institutions and more solid conclusions to be reached.

[24] See for example the list at Just Solve the File Format Problem (by no means exhaustive): http://fileformats.archiveteam.org/wiki/Category:File_formats_with_extension_.dat
[25] http://digital-archiving.blogspot.co.uk/2016/05/research-data-what-does-it-really-look.html
[26] http://archival-integration.blogspot.co.uk/2016/06/born-digital-data-what-does-it-really.html

However, it is clear from work carried out in this area so far that there is scope for increasing the number of research data formats within the PRONOM database. This issue is addressed in the next section of the report.

Identifying Research Data File Formats

As mentioned in our phase 1 report "As we move towards a proof of concept for archiving research data we should continue to engage with the team who maintain PRONOM and promote discussions within the wider digital preservation community about a sustainable or more automated way to address this problem."

One of the goals of this phase of the project was therefore to try and increase the representation of research data file formats within the PRONOM database whilst encouraging wider engagement in the digital preservation and research data community.

Signature creation at The National Archives

The majority of file format signature development work for PRONOM is undertaken at The National Archives (TNA) in the UK. They are understandably driven by their own priorities - so that they can manage the formats that they hold - but they also carry out work for the community on request. Anyone may submit information and sample files to the PRONOM team so that signature development work can be carried out but the timescales for this unfunded community work will vary greatly (and will also depend on the quality and accuracy of the information submitted)[27].

Due to the short timescales at play for this project, we directly funded some signature development work to ensure that it would be completed within the project time frame.

At York, the research data profiling work described above and the top 20 research data applications at York as mentioned in our phase 1 report helped inform our selection of file formats for submission to PRONOM. We chose Gaussian input files[28] and JEOL NMR Spectroscopy files[29]. A longer discussion about these formats and the submission to PRONOM is available as a blog post[30]. The two new signatures were released on 29th June 2016 as part of DROID signature file version 85[31].

Priorities at Hull were informed through a survey of research data file formats generated, and subsequent discussions with researchers self-identifying as having more unusual formats. This led to the creation of six new signatures. Firstly the addition of an ESRI ArcMap Document[32]. We had noted in our phase 1 report that ESRI files are well represented in PRONOM but that the .mxd file was not included so this work was designed to fill that gap.

---

[27] A blog post from Paul Young provides a good summary of the community engagement around PRONOM: http://blog.nationalarchives.gov.uk/blog/identifying-digital-file-formats-collaborative-effort/
[28] http://www.nationalarchives.gov.uk/PRONOM/fmt/894
[29] http://www.nationalarchives.gov.uk/PRONOM/fmt/895
[30] http://digital-archiving.blogspot.co.uk/2016/07/new-research-data-file-formats-now.html
[31] http://www.nationalarchives.gov.uk/aboutapps/pronom/release-notes.xml
[32] http://www.nationalarchives.gov.uk/PRONOM/fmt/916

Five further signatures were also developed for AmiraMesh file formats[33]. These new signatures were released on 27th July 2016 in version 86.

Creating our own PRONOM signatures

It was interesting working directly with TNA on signature development work and discussions with the PRONOM team led to a deeper understanding of how new entries in PRONOM are created. Following on from this, the project team decided to try to create their own file format signatures for PRONOM. The file format identification problem is a large and complex one and it was recognised that it could be solved more quickly with wider and more active community engagement. As noted in Paul Young's blog post[34], the PRONOM team commit to releasing 100 new PRONOM records per year. One way we could increase this figure and prioritise research data formats would be through more direct engagement with this problem. Clearly we are not the first to engage with this task, but we wanted to establish whether it was practical for 'an average digital archivist' to attempt file signature creation.

York's experiences at signature development are discussed in full in a blog post[35]. The resulting signature for an OMNIC Spectral Data File[36] was made available in DROID signature file version 88 on the 27th September 2016.

Hull were also able to create their own PRONOM signature and this work by Transforming Archives trainee Dave Heelas has again been documented in a blog post[37]. This work led to the creation of a signature for Final Draft Document 5-7[38] again in signature version 88.

Inspired by the work of this project, Andrea Byrne from Archives New Zealand also took on the challenge and again blogged about her progress[39]. Her detailed post describes some of the detective work necessary to understand the files she chose to investigate as well as some of the benefits of working in an open and community-focused way. The resulting file signatures are for AppleSingle 1 and 2[40] and are now available in signature version 88 along with a revised entry for AppleDouble Resource Fork 1[41].

The key points to note from this recent community signature creation work are as follows:

- It is possible for data curators and digital archivists to actively contribute to signature development (and it is not a task only suited to those who are more technically minded).

---

[33] http://www.nationalarchives.gov.uk/PRONOM/fmt/917, http://www.nationalarchives.gov.uk/PRONOM/fmt/918, http://www.nationalarchives.gov.uk/PRONOM/fmt/919, http://www.nationalarchives.gov.uk/PRONOM/fmt/920, http://www.nationalarchives.gov.uk/PRONOM/fmt/921

[34] http://blog.nationalarchives.gov.uk/blog/identifying-digital-file-formats-collaborative-effort/

[35] http://digital-archiving.blogspot.co.uk/2016/08/my-first-file-format-signature.html

[36] http://www.nationalarchives.gov.uk/PRONOM/fmt/963

[37] http://hullhistorycentre.blogspot.co.uk/2016/08/from-plans-to-digital-content-daves.html

[38] http://www.nationalarchives.gov.uk/PRONOM/fmt/964

[39] http://openpreservation.org/blog/2016/09/08/making-the-switch-from-user-to-user-and-contributor-my-first-file-format-signature/

[40] http://www.nationalarchives.gov.uk/PRONOM/fmt/967, http://www.nationalarchives.gov.uk/PRONOM/fmt/968

[41] http://www.nationalarchives.gov.uk/PRONOM/fmt/966

- Start with the low-hanging fruit - binary files are much easier to develop signatures for than ASCII. Starting with files represented in the institution's own collections is also a more meaningful starting point and it is often possible to locate additional sample files to test the identification work.
- Signature development is not always straightforward - input from the PRONOM team at TNA is required to check, refine and test the signatures before incorporating them into a PRONOM signature release[42].
- There are benefits to be had from sharing experiences around signature development with the wider digital preservation community (for example via the PRONOM google group[43]).
- Enhanced and improved documentation[44] around signature development would help the community engage more as would moving the signature development utility[45] from prototype to production.

As mentioned above, this flurry of work specifically targeted binary formats, however it should be noted that many of the research data formats in the York sample were ASCII files. File format signatures can be developed for ASCII files (if the files conform to a set structure which can be defined) but these are typically more complex to create than signatures for binary files, using regular expressions to describe their identifying characteristics[46]. It is also helpful to think beyond file identification of single digital objects and consider recognising groups of files that might together make up a more complex digital object, for example a website or software application.

Discussion

Why PRONOM?

The file format problem is a not a new one for the digital preservation community. A recent attempt to address this issue as a community, Just Solve the File Format Problem[47], has produced a useful resource but does not currently include many of the research data file formats highlighted in our profiling work. Furthermore, the real need as highlighted by our project is for a resource or registry which provides a means of automatically identifying file formats and allocating a unique identifier. Just Solve the File Format Problem was a project to bring together disparate sources of information about file formats, not specifically to identify them.

PRONOM, first released in 2002, is designed both to store information to aid automatic identification of files and to provide a means of uniquely identifying the format (via the Persistent Unique Identifier or PUID) and has demonstrated a longevity that other format

---

[42] In a blog post entitled 'A week of file format research' David Clipsham describes the work he undertook in the course of a week to help refine submitted signatures before incorporating them into PRONOM: http://openpreservation.org/blog/2016/08/31/a-week-of-file-format-research/

[43] https://groups.google.com/forum/#!forum/pronom

[44] Helpful documentation does exist but it does not describe the full process of signature development (including use of the signature development utility and testing using DROID): http://www.nationalarchives.gov.uk/documents/information-management/pronom-file-signature-research.pdf

[45] http://www.nationalarchives.gov.uk/pronom/sigdev/index.htm

[46] Andy Jackson discusses further identification methods for text-based formats in a blog post: http://anjackson.net/2016/06/08/frontiers-in-format-identification

[47] http://fileformats.archiveteam.org/

registries have not been able to match[48]. PRONOM is designed to work alongside DROID, a file identification tool also developed by TNA. However, there are many other tools that can be used to identify files - see for example those listed in the tool registry COPTR[49]. Archivematica gives the operator the choice of the PRONOM based tools FIDO[50]or Siegfried [51] to carry out format identifications, or there is the option of identification by file extension[52].

One of the benefits of using a PRONOM based tool for identification is that a PUID will be assigned to the file to match the identifier of the format in PRONOM. In theory this should allow reporting and search and retrieval based on file format[53] and would help facilitate the sharing of information about file formats with others, for example with other digital repositories about strategies for the preservation of a particular collection of files. Within Archivematica the PUID is stored within the PREMIS metadata for the digital object and is used by the Format Policy Registry (FPR)[54] which defines which preservation or dissemination actions should occur to a particular type of file. If a file is not identified, using the tools provided within Archivematica the system will not store an identifier for the format and further automated preservation actions through the FPR will be limited and non-specific.

What to accept?

One possible solution to the file format problem as described would be to limit the types of files that would be accepted within the digital repository. This is a tried and tested approach for certain disciplines and data archives[55] and is one that is also consistently supported by the digital preservation literature. For example The National Digital Stewardship Alliance's Levels of Digital Preservation[56] recommends at level one that an organisation carrying out digital preservation activities should "... encourage use of a limited set of known open formats ...". This is problematic for institutions facing the task of preserving research data. Researchers within an institution will use such a wide range of specialist hardware and software and it will be hard for the repository and research support staff to provide appropriate advice on suitable formats. For much of the data there will be no obvious preservation format for that data.

We also wish to encourage the innovative and cutting-edge research that is going on within our institutions so limiting the range of formats deposited would be counter-productive. At the University of York we encourage researchers (through both our RDM training sessions and RDM webpages[57]) to consider file formats throughout their project and think about the longevity and accessibility of the formats they select, but this is only guidance and ultimately we leave it up to the researcher to decide what formats to deposit their data in. We accept

---

[48] See for example the Unified Digital Format Registry which announced it would no longer continue in April 2016: http://udfr.org/

[49] http://coptr.digipres.org/Category:File_Format_Identification

[50] http://coptr.digipres.org/FIDO_(Format_Identification_for_Digital_Objects)

[51] http://coptr.digipres.org/Siegfried

[52] Identification by file extension is carried out with a basic python script in Archivematica.

[53] This functionality is not yet available in Archivematica

[54] https://www.archivematica.org/en/docs/archivematica-1.5/user-manual/preservation/preservation-planning/#fpr

[55] See for example depositor guidelines for the UK Data Archive: https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats and Archaeology Data Service: http://archaeologydataservice.ac.uk/advice/FileFormatTable

[56] http://ndsa.org/activities/levels-of-digital-preservation/

[57] http://www.york.ac.uk/library/info-for/researchers/data/organising/#tab-1

these formats and will preserve them on a best efforts basis. Understanding the file format moves us one step closer to preservation and reuse over the longer term.

Automation v. human intervention

It is perhaps worth highlighting the slight paradox in our project approach. Our proof of concept implementation work set out to establish a pragmatic and parsimonious[58] approach to digital archiving involving freely available tools and a high degree of automation in order to limit staff intervention in the process. However, if the file format identification problem is to be solved, this will of course require a substantial amount of staff time to highlight areas where work is required, prioritise and research the formats and submit them for signature development. This problem could be addressed at a higher level with targeted project funding to improve identification rates for research data, but given the rate of technological change it will always be an ongoing issue that institutions will need to engage with periodically.

Future work

Our project has highlighted the challenges around file format identification for research data and this information will feed into the ongoing Jisc Research Data Shared Service project[59] through which more resource is available to address the problem. Through the Shared Service project a piece of work will be carried out by the Open Preservation Foundation[60] to scope this issue and assess solutions to facilitate the preservation of research data. This piece of work will include an investigation of identification methods that are not based on PRONOM (for example Apache Tika[61] and the Unix File utility[62]) and consider how information about formats that are identified by these other methods can be submitted to PRONOM in order that we can uniquely identify them and share information across preservation systems. Some prior work to compare and aggregate the contents of different format registries already exists[63] and it is anticipated that there will benefits to revisiting this problem.

Recommendations

As discussed above, we have further investigated the extent of the file format problem for research data and carried out some work to increase the representation of research data file formats within PRONOM.

We recognise that the research data formats that this project has added to PRONOM are not a solution in themselves but should be seen as just the start of a bigger piece of community work in this area. We are keen to see this work continued by the digital preservation and

---

[58] The term Parsimonious Preservation was coined by Tim Gollins and refers to the simple and affordable steps you can take to start preserving digital material using freely available tools: http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf

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

[60] http://openpreservation.org/

[61] http://coptr.digipres.org/Apache_Tika

[62] http://coptr.digipres.org/Fine_Free_File_Command

[63] See for example Andy Jackson's format registry aggregator: http://www.digipres.org/formats/ and recent work by Tim Allison: http://openpreservation.org/blog/2016/10/04/apache-tikas-regression-corpus-tika-1302/

research data community in order to enable better research data identification nationally and internationally. The following recommendations aimed at a number of stakeholders suggest a way forward to help us achieve this as a community:

For data curators

- Actively engage with issues around file format identification. This could include:
    - Sharing file format profiles
    - Noting and sharing information about files that are wrongly identified by the available tools[64]
    - Noting and sharing information about files that are not identified by the available tools
    - Engaging in signature development work
    - Making sample files available to test corpuses (if appropriate permissions can be sought)
- Greater engagement with researchers on the value and necessity of recognising and recording the file formats they will use/generate to inform effective data curation.

For TNA

- Provide enhanced documentation describing how to engage in the process of signature creation.
- Greater active engagement with the community about what is currently being worked on and what the priorities are (transparency of process).
- Greater active engagement with the community to seek example file formats for testing and development purposes.
- Facilitate better integration between PRONOM and third party digital preservation tools and systems - for example automatic notifications which allow the tools to pick up new signatures as they are released[65].
- Facilitation of crowdsourcing efforts to suit need.

For digital preservation tool providers

- Digital preservation tools should incorporate better methods for engaging with the file format identification problem. For example:
    - Highlighting unidentified files as an integrated element of the ingest process and to increase awareness for data curators of the implications of this for future workflow and preservation actions
    - Re-running file identification tools
    - Sharing file identification profiles and reports
    - Efficient methods of incorporating up-to-date PRONOM sigs

For educators

- Those working in digital preservation should have a basic understanding of how file identification works and how they can contribute to the available registries and tools. The mechanics of file identification and how to contribute to the tools and registries should be taught in digital preservation training courses.

---

[64] A good example of this level of sharing was the winning poster at iPRES 2016 “To Act or not To Act - Handling File Format Identification Issues in Practice” from ETH Zurich: https://twitter.com/ipres2016/status/783217695835783168

[65] This notification currently works for DROID but not other PRONOM-based tools

For funders

- Time and money are barriers to engagement with this problem - particularly given the fact that it is too big for one institution to solve alone. There is often an interest and willingness to engage, but people struggle to make the case for this community work against local priorities. Funding for this work would reduce the barriers to engagement.

For digital preservation membership organisations

- Encourage community effort in this area - for example by organising training days, workshops and hackathons[66] around signature development.
- Facilitate the sharing of DROID profiles to establish what the priorities should be for signature development work.

For researchers

- Supply adequate metadata about submitted datasets. Clear and accurate metadata about file formats and hardware/software dependencies will aid file format identification and future preservation work.
- Be open to sharing sample files for test data corpuses and to aid signature development where appropriate.

[66] See for example the recent and successful JHOVE hack day organised by the OPF: http://openpreservation.org/blog/2016/10/19/jhove-online-hack-day-report/

3. Outreach

During the six months of phase 3 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 and 2 project reports. We attended a wide range of different events, engaging with audiences from the UK and beyond. As in previous phases of this project, we have been encouraged by the level of interest the project has generated and have received positive feedback on the work we are doing.

Outreach channels consisted largely of presentations and posters at organised events and blog posts published on the University of York's Digital Archiving blog[67]; our project was also featured in other publications. Outreach work continued beyond the active project period to include dissemination of our phase 3 work.

Events

International Digital Curation Conference (IDCC16) - Amsterdam (22-24 February 2016)

Jisc Research Data Management Shared Service Workshop: An institutional perspective - Jenny Mitcham (presented in the Jisc Research Data Management Shared Service Pilot workshop)

"Filling the digital preservation gap" for Research Data - Jenny Mitcham, Julie Allinson, Chris Awre, Richard Green, Simon Wilson[68]

'Digital Preservation: Strategic Issues' - National Library of Wales (25 February 2016)

A collaborative approach to "filling the digital preservation gap" for Research Data Management - Julie Allinson

UK Archives Discovery Forum - Kew (17 March 2016)

Poster: *Filling the Digital Preservation Gap* - Jenny Mitcham and Simon Wilson

[67] http://digital-archiving.blogspot.co.uk/

[68] http://www.dcc.ac.uk/sites/default/files/documents/IDCC16/Parallel%20B/Session%203/Jen%20Mitcham.pdf

Figure 4: UK Archives Discovery poster

UK Archivematica group meeting - York (22 March 2016)

"Filling the digital preservation gap" for research data: Results from phase 2 and plans for phase 3 - Jenny Mitcham[69]

Research Data, Records and Archives: Breaking the Boundaries - Edinburgh (18 April 2016)

A collaborative approach to "filling the digital preservation gap" for Research Data Management - Chris Awre[70]

Open Repositories (OR16) - Dublin (13-16 June 2016)

Prototyping a digital preservation pipeline with Archivematica, Fedora 4 and Hydra - Julie Allinson and Justin Simpson (Artefactual Systems)

Jisc and CNI conference - Oxford (6 July 2016)

Addressing the preservation gap at the University of York - Jenny Mitcham

---

69 https://docs.google.com/presentation/d/1NbzpxB27yNWRQP2oUr-8qeApFq5lSUc2jiAMqD8sWOo/edit?usp=sharing

70 https://media.ed.ac.uk/media/Chris+Awre+%28Head+of+Information+Services%2C+University+of+Hull%29/1_ip395t3t/41567111

Hydra Virtual Connect (7 July 2016)

Hydra For Research Data - Julie Allinson and Matthew Phillips (University of Durham)

TNA Digital Transformation Day - Kew (25 July 2016)

Going digital: a case study from the Borthwick Institute for Archives - Jenny Mitcham

Jisc Research Data Network meeting - Cambridge (6 September 2016)

Implementing Archivematica for Research Data Preservation at York and Hull - Jenny Mitcham

UK Archivematica group meeting - Lancaster (14 September 2016)

Filling the Digital Preservation Gap: update on phase 3 work - Julie Allinson and Jenny Mitcham[71]

iPRES conference - Bern (3-6 October 2016)

Preserving Research Data: Linking Repositories and Archivematica - Jenny Mitcham, Matthew Addis (Arkivum), Julie Allinson, Chris Awre, Richard Green, Simon Wilson[72]

Hydra Connect - Boston (3-6 October 2016)

Hydra, research data and Archivematica - Julie Allinson, Richard Green

Poster: *Filling the Digital Preservation Gap: integrating Archivematica and Hydra for Research Data Management*

Research Data Spring Showcase - Birmingham (20 October 2016)

Lightning talk and demo: *Filling the Digital Preservation Gap* - Julie Allinson, Chris Awre, Jenny Mitcham

---

71 https://docs.google.com/presentation/d/1b-yF36iD3lIqB1WxVbOyiwgSqmvUWk0TE3k97ejFuuM/edit?usp=sharing

72 http://www.ipres2016.ch/frontend/organizers/media/iPRES2016/_PDF/IPR16.Proceedings_4_Web_Brochuere_Link.pdf

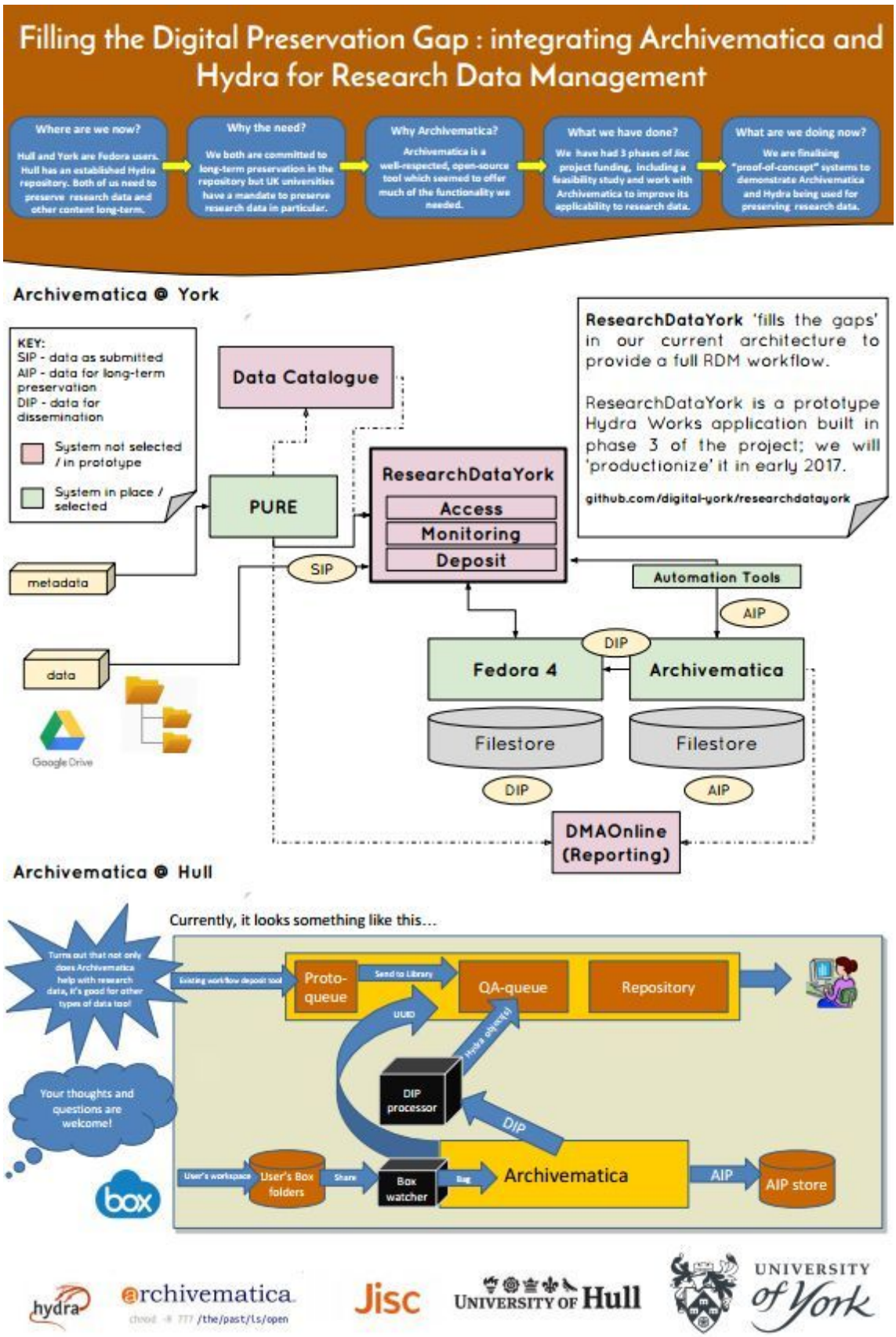


Figure 5: Poster at Hydra Connect

Other publications

The National Archives

Details of our project were featured as a case study within
Consultation on a new Strategic Vision for the Archives Sector
http://www.nationalarchives.gov.uk/documents/archives/consultation-strategic-vision-for-archives.pdf

Nestor

An edited version of one of our project blog posts "My first file format signature" was translated into German for inclusion in a new nestor publication series. Nestor is the German competence network for digital preservation[73]:
http://files.dnb.de/nestor/kurzartikel/thema_03-Meine_erste_Dateiformatsignatur.pdf

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 3 work) are listed below:

Title of blog post	Date of release (2016)	No of views[74]
New "Filling the Digital Preservation Gap" report released http://digital-archiving.blogspot.co.uk/2016/02/new-filling-digital-preservation-gap.html -	5th February 2016	513
Kicking off phase 3 of "Filling the Digital Preservation Gap" http://digital-archiving.blogspot.co.uk/2016/04/kicking-off-phase-3-of-filling-digital.html	1st April 2016	338
Research data - what does it *really* look like? http://digital-archiving.blogspot.co.uk/2016/05/research-data-what-does-it-really-look.html	31st May 2016	1302
Modelling Research Data with PCDM http://digital-archiving.blogspot.co.uk/2016/07/modelling-research-data-with-pcdm.html	4th July 2016	971

[73] http://www.langzeitarchivierung.de/Subsites/nestor/EN/Home/home_node.html
[74] as of 17th October 2016

New research data file formats now available in PRONOM http://digital-archiving.blogspot.co.uk/2016/07/new-research-data-file-formats-now.html	4th July 2016	597
Research data is different http://digital-archiving.blogspot.co.uk/2016/08/research-data-is-different.html	5th August 2016	510
My first file format signature http://digital-archiving.blogspot.co.uk/2016/08/my-first-file-format-signature.html	19th August 2016	614
Filling the Digital Preservation Gap - a brief update http://digital-archiving.blogspot.co.uk/2016/08/filling-digital-preservation-gap-brief.html	30th August 2016	309
UK Archivematica group at Lancaster http://digital-archiving.blogspot.co.uk/2016/09/uk-archivematica-group-at-lancaster.html	16th September 2016	265
File format identification at Norfolk Record Office (a guest post) http://digital-archiving.blogspot.co.uk/2016/09/file-format-identification-at-norfolk.html	21st September 2016	363
Some highlights from iPRES 2016 http://digital-archiving.blogspot.co.uk/2016/10/some-highlights-from-ipres-2016.html	11th October 2016	307

Table 3: Project blog posts released during Phase 3 and number of page views recorded

Project website

The project website is hosted at the Borthwick Institute for Archives at the University of York. This is available at http://www.york.ac.uk/borthwick/projects/archivematica/.

In the period from the 9th December 2015 (when these stats were last reported in the Phase 2 report) to the 16th October 2016, there were 559 pageviews representing 480 unique visits to the page. Average time spent on the page was 4 minutes and 11 seconds.

Figure 6: The project website

Project reports

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 3113 times in the intervening period and downloaded 620 times[75].

In February 2016 our phase 2 project report was made available via Figshare (https://dx.doi.org/10.6084/m9.figshare.2073220). This report has been viewed 3265 times in the intervening period and downloaded 465 times[76].

These reports are also available from the University of Hull repository.

[75] Statistics collected on 17th October 2016
[76] Statistics collected on 17th October 2016

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

Automation Tools: A set of python scripts, that are designed to automate the processing of transfers in an Archivematica pipeline[77]

Box: A service used at the University of Hull for secure file sharing, storage and collaboration[78]

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[79]

DROID: (Digital Record Object Identification) An automatic file format identification tool from The National Archives

EPSRC: Engineering and Physical Sciences Research Council

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

Figshare: A repository where researchers, institutions and publishers can share research outputs[81]

Hydra: A repository solution based on a number of “best-of-breed” open-source components, including Fedora[82]

JHOVE: JSTOR/Harvard Object Validation Environment is an extensible software framework for performing format identification, validation, and characterization of digital objects[83]

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

---

[77] https://github.com/artefactual/automation-tools
[78] https://www.box.com/en-gb/home
[79] http://www.dmao.info
[80] http://www.fedora-commons.org/
[81] http://www.figshare.com
[82] http://projecthydra.org/
[83] http://openpreservation.org/technology/products/jhove/

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

PCDM: Portland Common Data Model is a flexible, extensible domain model that is intended to underlie a wide array of repository and DAMS applications. Currently being implemented by the Hydra and Islandora communities.

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.[84]

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.[85]

PUID: PRONOM Unique Identifier[86] - an extensible scheme for providing persistent unique identifiers for records in the PRONOM registry

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

[84] http://www.loc.gov/standards/premis/
[85] https://www.nationalarchives.gov.uk/PRONOM/Default.aspx
[86] http://www.nationalarchives.gov.uk/aboutapps/pronom/puid.htm

Appendix 1: Draft instructions for use of Hull implementation

These draft instructions are aimed at researchers at the University of Hull and are intended to give the reader a sense of how a Hull production system would be used. Whilst most of the functionality referred to has been implemented in our proof of concept system, some has not: in particular, only a subset of the Dublin Core metadata terms are currently supported and only dataset, journal article, book, book chapter and photograph are functional from the list of possible content models.

Depositing digital content for preservation and discovery using Box folders

Overview

The process described here allows a user to deposit materials for long term digital preservation and to have a record of them placed in the University's digital repository, Hydra. Optionally, copies of the files can be made available through Hydra for download.

Deposit options

Material to be deposited may be a single file or multiple files, it may also be a folder structure containing multiple files in an organised hierarchy.

The resulting repository content may be a single object or multiple objects. These objects describe the file(s) that have been placed in the University's preservation store and may optionally contain copies of the data for download.

As a depositor, in addition to your files for deposit, you must provide one or more simple text files, as described below, that indicate how you wish your content to be dealt with.

One or more files to be represented in a single object

This first option allows you to deposit one or more files as a group. The files will be kept together and there will be a single object in the repository that describes them all. Optionally, any or all of the files can be copied to the repository for download.

1. Create a folder in your Box space in which to assemble your content.

2. Copy into this folder the files that you wish to deposit.

3. You must now add to the Box folder one or more simple text files that will determine how your content is processed through our automated system. Any simple text editor can be used to create the files, for instance Notepad in Windows or TextEdit on an Apple.

   Create a file called “description.txt”
   - It **must** contain a title that describes the materials deposited
   - It **must** contain the name(s) of the creator(s) of the material in the form “Lastname, firstname” or “Lastname, Firstname Intital(s).” Multiple authors should be separated by semicolons.
   - It **must** contain a list of files (“visibleFiles”) that are to be made available for download in Hydra. The list can be blank (no visible files) or may contain the word “all” (all files to be visible). Multiple filenames must be separated by semicolons.
   - It may contain a “contentModel” field that describes the primary type of material being deposited. If this field is omitted, the default is “dataset”. 1) (See list in Table
   - It may contain additional Dublin Core metadata fields (see list in Table 2), for instance a description of the materials deposited or a specific citation that you would like used in references. **Not all DC fields are currently implemented.**

     The following are valid examples of “description.txt”. Note that the field labels begin with a lower case letter and are followed by a colon.

```
title: Data package with all files visible in Hydra

creator: Mouse, Mickey; Mouse, Minnie; Duck, Donald A.

description: Specimen data package with a small number of files.

contentModel: dataset

visibleFiles: all
```

```
title: Data package with all files visible in Hydra and with additional
metadata

creator: Mouse, Mickey; Mouse, Minnie; Duck, Donald A.

description: Specimen data package with a number of files, all to be
downloadable through Hydra.

visibleFiles: all

citation: Mouse, Mickey; Mouse, Minnie; Duck, Donald A. (2016) "Research
data from the Disney Project" University of Hull

subject: Cartoons; Disney, Walt; Animation
```

```
title: Data package with selected files visible in Hydra

creator: Mouse, Mickey; Mouse, Minnie; Duck, Donald A.

description: Specimen data package with a small number of files. Some
```

```
files for preservation only – not to be made available in Hydra, three
copy files available for download.

visibleFiles: file1.jpg; file2.pdf; file3.log
```

```
title: Data package with no files visible in Hydra

creator: Mouse, Mickey; Mouse, Minnie; Duck, Donald A.

description: Specimen data package with a small number of files. Files
for preservation only – copies not to be made available in Hydra, just a
metadata record.

contentModel: dataset

visibleFiles:
```

artwork	drawing	letter	policyOrProcedure
book	event	licence	presentation
bookChapter	genericContent	map	regulation
conferencePaper	guidance	meetingPapers	report
conferenceAbstract	handbook	movingImage	software
conferencePoster	internetPublication	musicScore	sound
dataset	journalArticle	newsletterArticle	
diagram	learningMaterials	photograph	

Table 1: Acceptable content model types

abstract	coverage	hasFormat	isVersionOf	requires
accessRights	created	hasPart	language	rights
accrualMethod	creator	hasVersion	license	rightsHolder
accrualPeriodicity	date	identifier	mediator	source
accrualPolicy	dateAccepted	instructionalMethod	medium	spatial
alternative	dateCopyrighted	isFormatOf	modified	subject
audience	dateSubmitted	isPartOf	provenance	tableOfContents
available	description	isReferencedBy	publisher	temporal
bibliographicCitation	educationLevel	isReplacedBy	references	title
conformsTo	extent	isRequiredBy	relation	type
contributor	format	issued	replaces	valid

Table 2: Dublin Core metadata acceptable in description files

4. If you have citations of a paper or papers related to data that you are depositing, for instance in Bibtex format, you may add this/these to your folder in a file called "citations.bib". **Other formats are not currently supported.**

5. If there is information you wish to pass to the repository staff in the University Library, create another simple text file called "readme.txt" in your Box folder. The text in this folder will be visible to them when they come to check the repository object that is created and they can then contact you if necessary.

6. If you are satisfied that everything is as you want it to be, use the Box "Share" option for your folder to share it with the email address archivematica@hull.ac.uk.

7. Essentially, that's it! If you check back from time to time you will see that your folder name has the status of your material appended. So, for instance, a folder called "Mydata" will go through a number of stages starting with "Mydata – processing" and ending with "Mydata – all processing complete". Once you reach this last stage you can, should you wish, delete the folder; do not delete it until then!

More than one file, each to be represented in its own object

This second option allows you to create multiple objects from the files that you deposit. Optionally, any or all of the files in these objects can be copied to the repository for download.

1. Create a folder in your Box space in which to assemble your content.

2. Copy into this folder the files that you wish to deposit.

3. You must now add to the Box folder a comma-separated-values (csv) file that will determine how your content is processed through our automated system. A csv file is most easily created in a spreadsheet application

   Create a set of headings in row 1 corresponding to the metadata fields that you intend passing into the system and then a row of metadata for each file to be deposited.
   - You **must** first pass the filename that each row of metadata applies to
   - You **must** pass a title that describes the materials deposited in that object
   - You **must** pass the name(s) of the creator(s) of the material in each object in the form “Lastname, firstname” or “Lastname, Firstname Intital(s).” Multiple authors for the same material should be separated by semicolons.
   - Each file row **must** contain a “visibleFiles” entry if the file is to be available for download in Hydra. The visibleFiles entry should be set to “all” if a copy of the file is to be made available for download. If it is left blank, no copy file will be available.
   - You may pass a “contentModel” field that describes the type of material being deposited. If this field is omitted, the default is “dataset”. (See list in Table1)
   - It may contain additional Dublin Core metadata fields (see list in Table 2), for instance a description of the materials deposited or a specific citation that you would like used in references. **Not all DC fields are currently implemented.**

   The following is an example of “description.csv” as laid out in a spreadsheet app. Note that the field labels begin with a lower case letter.

filename	title	creator	description	subject	contentModel	visibleFiles
picture.jpg	My picture	Mouse, Mickey	Darling Minnie	Minnie Mouse	photograph	
story.ppt	My presentation	Duck, Donald A.	My research work	Nautical history	presentation	all
data.xls	My spreadsheet	Mouse, Minnie; Mouse, Mickey	My data			all

   Assuming that the named files are present in your folder, processing this description.csv would result in three objects in Hydra with corresponding data in the preservation store. The first object would contain the metadata for a preserved image file, but the image would not be available for download. The second object would describe the presentation and make a copy available for download. The final object would describe a spreadsheet of data and make a copy available for download; this would be recognised as a dataset within the repository even though the contentModel was not specified.

4. If there is information you wish to pass to the repository staff in the University Library, create another simple text file called “readme.txt” in your Box folder. The text in this folder will be visible to them when they come to check the repository object that is created and they can then contact you if necessary.

5. If you are satisfied that everything is as you want it to be, use the Box “Share” option for your folder to share it with the email address archivematica@hull.ac.uk.

6. Essentially, that’s it! If you check back from time to time you will see that your folder name has the status of your material appended. So, for instance, a folder called “Mydata” will go through a number of stages starting with “Mydata – processing” and ending with “Mydata – all processing complete”. Once you reach this last stage you can, should you wish, delete the folder; do not delete it until then!

A hierarchy of files to be represented in a single object

This final option allows you to deposit a hierarchy (folder structure) of files. The files will be kept together in their folder structure which is turned into a zip file. There will be a single object in the repository that describes them and, optionally, the zip file can be copied to the repository for download.

1. Create a folder in your Box space in which to assemble your content.

2. Copy into this folder the folder that contains the hierarchy so that you get a folder within the Box folder

3. You must now add to the Box folder one or more simple text files that will determine how your content is processed through our automated system. Any simple text editor can be used to create the files, for instance Notepad in Windows or TextEdit on an Apple.

Create a file called “description.txt”

- It **must** contain a title that describes the materials deposited
- It **must** contain the name(s) of the creator(s) of the material in the form “Lastname, firstname” or “Lastname, Firstname Intital(s).” Multiple authors should be separated by semicolons.
- It **must** contain a “visibleFiles” entry to show whether a copy of the zip file should be made available for download in Hydra. The entry can be blank (file not available) or may contain the word “all” (file to be downloadable).
- It may contain a “contentModel” field that describes the primary type of material being deposited. If this field is omitted, the default is “dataset”. (See list in Table 1)
- It may contain additional Dublin Core metadata fields (see list in Table 2), for instance a description of the materials deposited or a specific citation that you would like used in references. **Not all DC fields are currently implemented.**

The following are valid examples of “description.txt”. Note that the field labels begin with a lower case letter and are followed by a colon.

```
title: Hierarchical data package with zip file visible in Hydra and with
additional metadata

creator: Mouse, Mickey; Mouse, Minnie; Duck, Donald A.

description: Specimen data package with a tree of files. Files for preservation
only – not to be made available in Hydra.

visibleFiles: all

citation: Mouse, Mickey; Mouse, Minnie; Duck, Donald A. (2016) "Research data
from the Disney Project" University of Hull

subject: Cartoons; Disney, Walt; Animation

contentModel: dataset
```

```
title: Hierarchical data package with additional metadata – zip file for
preservation only – no copy downloadable through Hydra

creator: Mouse, Mickey; Mouse, Minnie; Duck, Donald A.

description: Specimen data package with a tree of files. Files for preservation
only – not to be made available in Hydra.

visibleFiles:

citation: Mouse, Mickey; Mouse, Minnie; Duck, Donald A. (2016) "Research data
from the Disney Project" University of Hull

subject: Cartoons; Disney, Walt; Animation
```

4. If you have citations of a paper or papers related to data that you are depositing, for instance in Bibtex format, you may add this/these to your folder in a file called "citations.bib".

5. If there is information you wish to pass to the repository staff in the University Library, create another simple text file called "readme.txt" in your Box folder. The text in this folder will be visible to them when they come to check the repository object that is created and they can then contact you if necessary.

6. If you are satisfied that everything is as you want it to be, use the Box "Share" option for your folder to share the Box folder (not the folder contained within it) with the email address archivematica@hull.ac.uk.

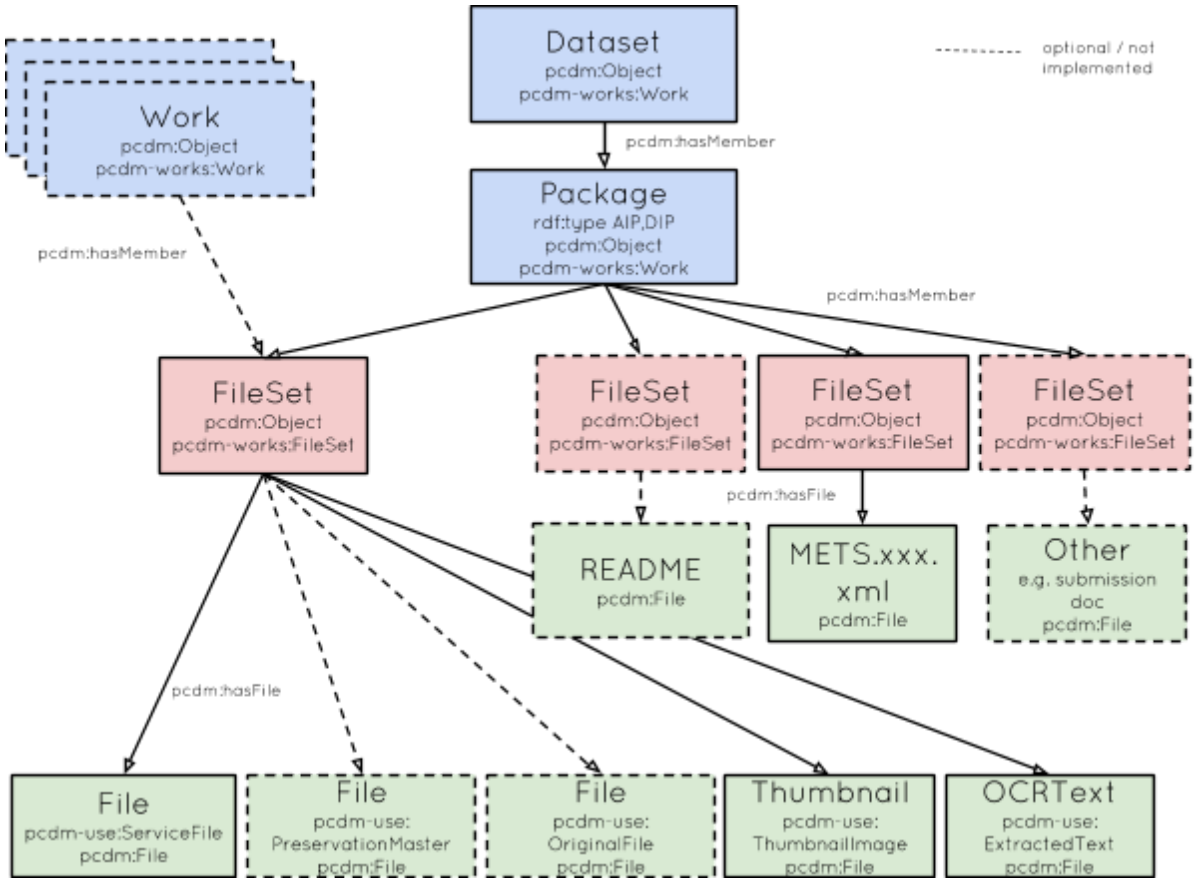
7. Essentially, that's it! If you check back from time to time you will see that your folder name has the status of your material appended. So, for instance, a folder called "Mydata" will go through a number of stages starting with "Mydata – processing" and ending with "Mydata – all processing complete". Once you reach this last stage you can, should you wish, delete the folder; do not delete it until then!

Appendix 2: A Draft PCDM-based Data Model for Datasets

The Portland Common Data Model (PCDM) is "a flexible, extensible domain model that is intended to underlie a wide array of repository and DAMS applications" [87]. It is informing the latest developments of Hydra and Islandora in their implementations on top of Fedora 4. Aligning our model with PCDM means that we are designing with interoperability in mind, and adding to the growing list of example PCDM models.

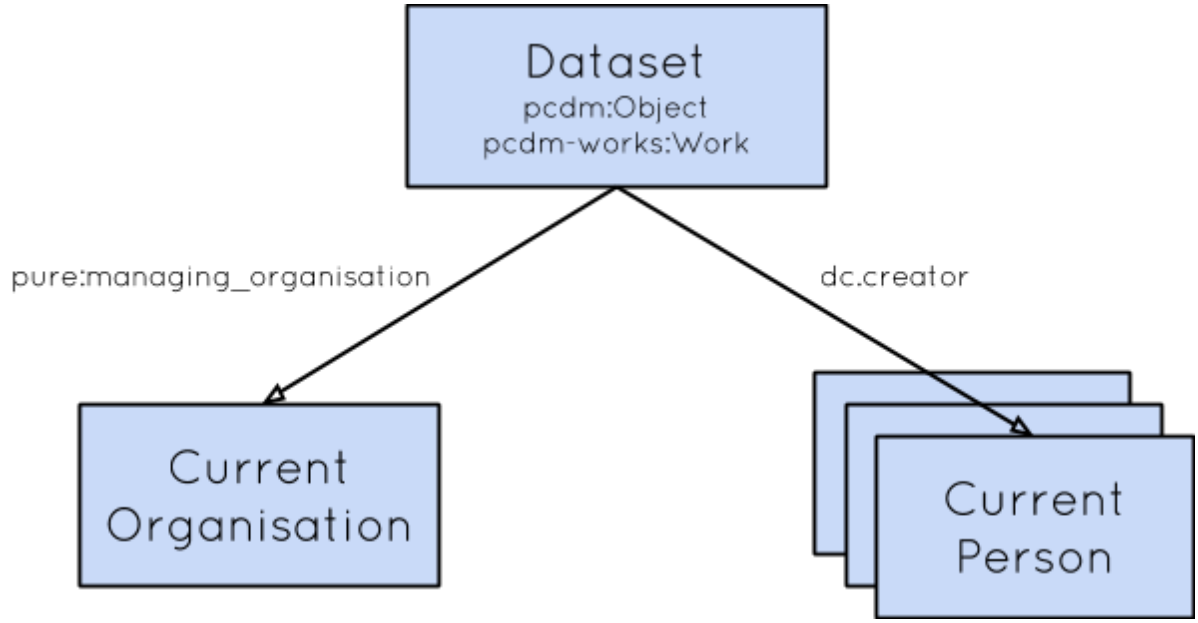
For the 'live' document, please see:
https://docs.google.com/document/d/1QPw9kLqRFzI5aStRr3nlBqqj5BzkI3eFZY4zNjOmo8w/edit#heading=h.1jggb1ac1v2v

PCDM Model



[87] https://github.com/duraspace/pcdm

People and Organisations Model



Namespaces

am: <http://dlib.york.ac.uk/ontologies/oais-archivematica#>
dlib: <http://dlib.york.ac.uk/ontologies/generic#>
dc: <http://purl.org/dc/terms/>
identifiers: <http://id.loc.gov/vocabulary/identifiers/>
sorg: <http://schema.org/>
foaf: <http://xmlns.com/foaf/0.1/>

Models

Dataset

A Dataset is a discrete set of related data files. The files can be of any type. In PURE a single Dataset metadata record has a one-to-one relationship to a Dataset as described here. A Dataset does not, however, *require* an accompanying PURE record. Local descriptive metadata can be added with the exception of pureUuid which *must* be derived from PURE.

Code:
https://github.com/digital-york/dlibhydra/blob/master/lib/dlibhydra/models/works/dataset.rb

Class	Property	Expected Object Type	Usage
Dataset	rdf:type	URI	dcat:dataset
	dc:title	Literal (String)	1 (datacite mandatory) From PURE

	dc.creator	Current Person	1..n (datacite mandatory) From PURE
	identifiers:doi	URI	0..1 (datacite mandatory) From PURE
	dc.publisher	Literal (String)	1 (datacite mandatory) From PURE
	dc.available	Literal (String)	1 (datacite mandatory) From PURE
	dlib:pureUuid	Literal (String)	1 From PURE
	pure:managingUnit	Current Organisation	1 From PURE
	pure:pureLink	URI	0..n From PURE
	dc:accessRights	Literal (String)	1 From PURE
	dlib:embargoRelease Date	Literal (String)	1
	dlib:lastAccess	Literal (String)	0..1
	dlib:retentionPolicy	URI/Object	0..1
	dlib:for_indexing	Literal (String)	1 From PURE
	dlib:restriction_note s	Literal (String)	0..1

	pcdm:hasMember	Package	0..n

Data available from PURE but not currently added:

- Publications and projects
- People with roles other than 'Creator'
- Description
- Geographical
- Temporal
- Date of production
- Workflow / Visibility

Data not available from PURE Web Services:

- Embargo
- Contact person
- Relations to other datasets
- Restrictions narrative

Class	Property	Expected Object Type	Usage
CurrentPerson	rdf:type	URI	sorg:Person pure:PurePerson
	skos:prefLabel	Literal (String)	1
	foaf:familyName	Literal (String)	1
	foaf:givenName	Literal (String)	1
	pure:pureUuid	Literal (String)	1
	pure:puretype	Literal (String)	1

Class	Property	Expected Object Type	Usage
CurrentOrganisation	rdf:type	URI	sorg:Organisation pure:PureOrganisation
	skos:prefLabel	Literal (String)	1

	foaf:name	Literal (String)	1
	pure:pureUuid	Literal (String)	1
	pure:puretype	Literal (String)	1

Package

A Package represents a deposit of data files that form part of a larger Dataset. A Package with the rdf:type am:ArchivalInformtionPackage (AIP) will contain sufficient metadata to identify the location of the stored AIP (aipUuid as a minimum). A Package with the rdf:type am:DisseminationInformtionPackage (DIP) with contain references to the files comprising the DIP and the dipUuid as a minimum). A single Package can be both AIP and DIP. Folder structure for the original deposit shall be retained in such a way that it can be re-created to the end user, for example in an Archivematica METS file. Other files may be included in the Package, for example submission documentation or a readme file. These should be distinguished from the data files themselves.

Code:
https://github.com/digital-york/dlibhydra/blob/master/lib/dlibhydra/models/works/package.rb

Class	Property	Expected Object Type	Usage
Package	rdf:type	URI	dlib:Package am: ArchivalInformtionPackage am: DisseminationInformtionPackage
	skos:prefLabel	Literal (String)	1
	am:aipUuid	Literal (String)	0..1 from Archivematica
	am:aipStatus	Literal (String)	1 from Archivematica
	am:aipSize	Literal (String)	0..1 from Archivematica

	am:aipCurrentPath	Literal (String)	0..1 from Archivematica
	am:aipCurrentLocation	Literal (String)	0..1 from Archivematica
	am:dipUuid	Literal (String)	0..1 from Archivematica
	am:dipStatus	Literal (String)	0..1 from Archivematica
	am:dipSize	Literal (String)	0..1 from Archivematica
	am:dipCurrentPath	Literal (String)	0..1 from Archivematica
	am:originPipeline	Literal (String)	0..1 from Archivematica (dip)
	am:dipCurrentLocation	Literal (String)	0..1 from Archivematica
	am:aipResourceUri	Literal (String)	0..1 from Archivematica
	am:dipResourceUri	Literal (String)	0..1 from Archivematica
	dlib:requestorEmail	Literal (String)	0..n