

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.



This report is licensed under a Creative Commons CC-BY-NC-SA 2.0 UK licence.

Contents

[Contents](#)

[Executive summary](#)

[Introduction](#)

[1. Implementation](#)

[Implementation at Hull](#)

[How it works](#)

[Future work](#)

[Implementation at York](#)

[How it works](#)

[Future work](#)

[Lessons Learned](#)

[2. Research data file formats](#)

[Profiling research data](#)

[Identifying Research Data File Formats](#)

[Signature creation at The National Archives](#)

[Creating our own PRONOM signatures](#)

[Discussion](#)

[Why PRONOM?](#)

[What to accept?](#)

[Automation v. human intervention](#)

[Future work](#)

[Recommendations](#)

[For data curators](#)

[For TNA](#)

[For digital preservation tool providers](#)

[For educators](#)

[For funders](#)

[For digital preservation membership organisations](#)

[For researchers](#)

[3. Outreach](#)

[Events](#)

[International Digital Curation Conference \(IDCC16\) - Amsterdam \(22-24 February 2016\)](#)

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

[UK Archives Discovery Forum - Kew \(17 March 2016\)](#)

[UK Archivemata group meeting - York \(22 March 2016\)](#)

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

[Open Repositories \(OR16\) - Dublin \(13-16 June 2016\)](#)

[Jisc and CNI conference - Oxford \(6 July 2016\)](#)

[Hydra Virtual Connect \(7 July 2016\)](#)

[TNA Digital Transformation Day - Kew \(25 July 2016\)](#)

[Jisc Research Data Network meeting - Cambridge \(6 September 2016\)](#)

[UK Archivemata group meeting - Lancaster \(14 September 2016\)](#)

[iPRES conference - Bern \(3-6 October 2016\)](#)

[Hydra Connect - Boston \(3-6 October 2016\)](#)

[Research Data Spring Showcase - Birmingham \(20 October 2016\)](#)

[Other publications](#)

[The National Archives](#)

[Nestor](#)

[Blogs](#)

[Project website](#)

[Project reports](#)

[Glossary](#)

[Appendix 1: Draft instructions for use of Hull implementation](#)

[Depositing digital content for preservation and discovery using Box folders](#)

[Overview](#)

[Deposit options](#)

[One or more files to be represented in a single object](#)

[More than one file, each to be represented in its own object](#)

[A hierarchy of files to be represented in a single object](#)

[Appendix 2: A Draft PCDM-based Data Model for Datasets](#)

[PCDM Model](#)

[People and Organisations Model](#)

[Namespaces](#)

[Models](#)

[Dataset](#)

[CurrentPerson](#)

[CurrentOrganisation](#)

[Package](#)

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:

<http://dx.doi.org/10.6084/m9.figshare.1481170> } RDM@archivematica.org | www.archivematica.org | July 2015. Jenny Mitcham, Chris Awre, Julie Allinson, Richard Green, Simon Wilson¹

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:

<http://dx.doi.org/10.6084/m9.figshare.2073220> } RDM@archivematica.org | www.archivematica.org | February 2016. Jenny Mitcham, Chris Awre, Julie Allinson, Richard Green, Simon Wilson²

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 Archivemata 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 PRONOM³
- 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 Archivemata 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 Archivemata 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 Archivemata we also feel it is useful to demonstrate how Archivemata 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 Archivemata)	Watched directory (when something is placed in this directory it initiates a transfer to Archivemata)
SIP packaging	BagIt	SIP organised into a folder structure understandable by Archivemata. 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 system⁴ 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>

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 Labs⁶ and is available from Hull's github repository⁷.

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 Archivemata 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 Archivemata 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 Tools⁸, nor were we aware of the Puree gem⁹ 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>

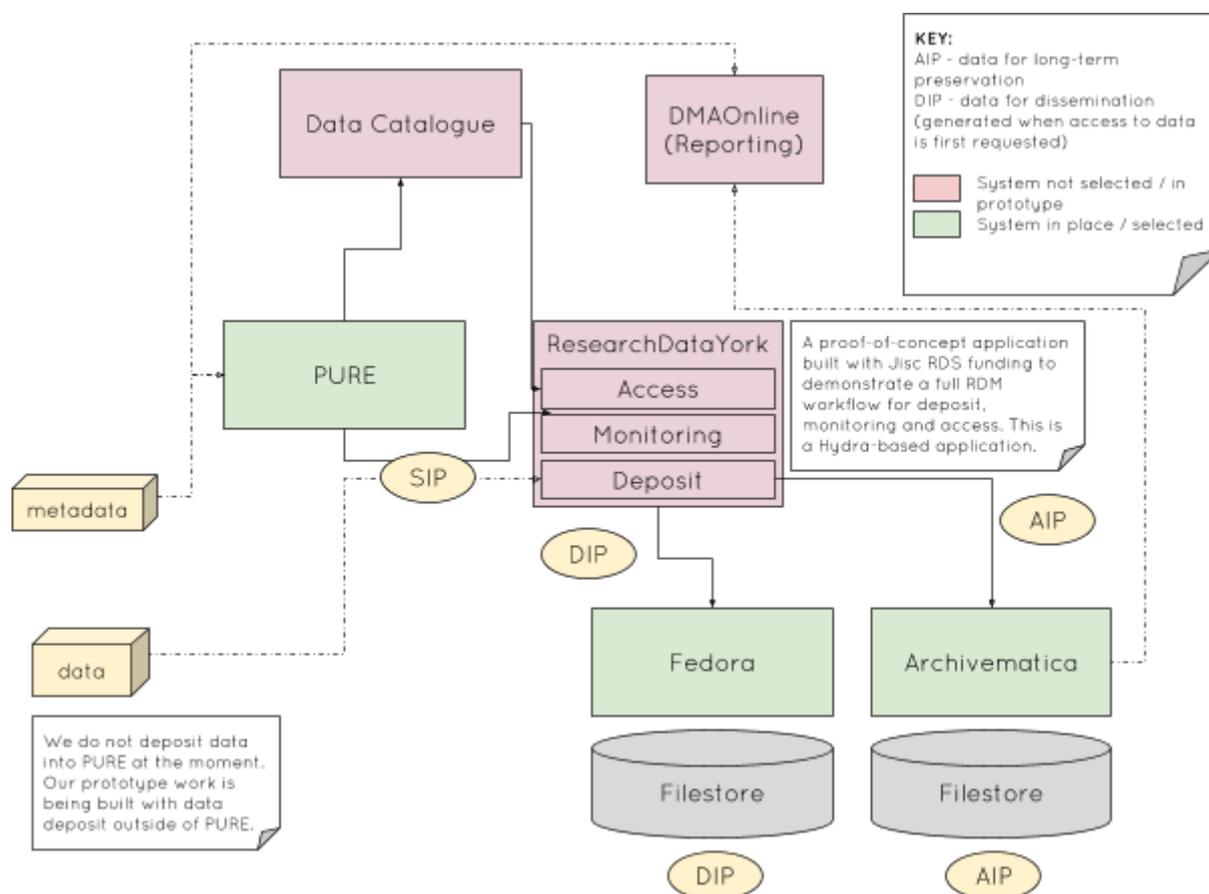


Figure 2: York RDM Architecture Diagram

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 R DYork (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 R DYork 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 R DYork application.

The code for York's proof of concept system is available from the Library & Archives Technical Team github repository¹⁰. A screencast showing the working application is also available¹¹.

Future work

Almost in parallel with the Research Data Spring projects Jisc were planning a Research Data Shared Service¹². 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 R DYork.

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=Uz^Vnk_u_MVa&list=PLjA_H_1a`Jf2_

¹² <https://www.jisc.ac.uk/rd/projects/research-data-shared-service>

- Investigate the University of Lancaster's 'preservation' gem¹³ 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 DMAOnline¹⁴ 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 Tools¹⁵ 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 Archivemata 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 Expectations¹⁶ 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 DROID¹⁷ (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 post¹⁸.

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 hold¹⁹. 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:

Q. How many files in the sample were identified by DROID? What percentage of files were identified by DROID? What file formats were identified by DROID? What file formats were not identified by DROID? What file formats were identified by DROID but not by the other methods? What file formats were identified by the other methods but not by DROID? What file formats were identified by both DROID and the other methods? What file formats were identified by neither DROID nor the other methods?

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 identified²⁰. Bentley Historical Library, Norfolk Record Office also engaged in this exercise and shared their results²¹. 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, 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 network²², 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 repository²³. 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 systems²⁴ 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

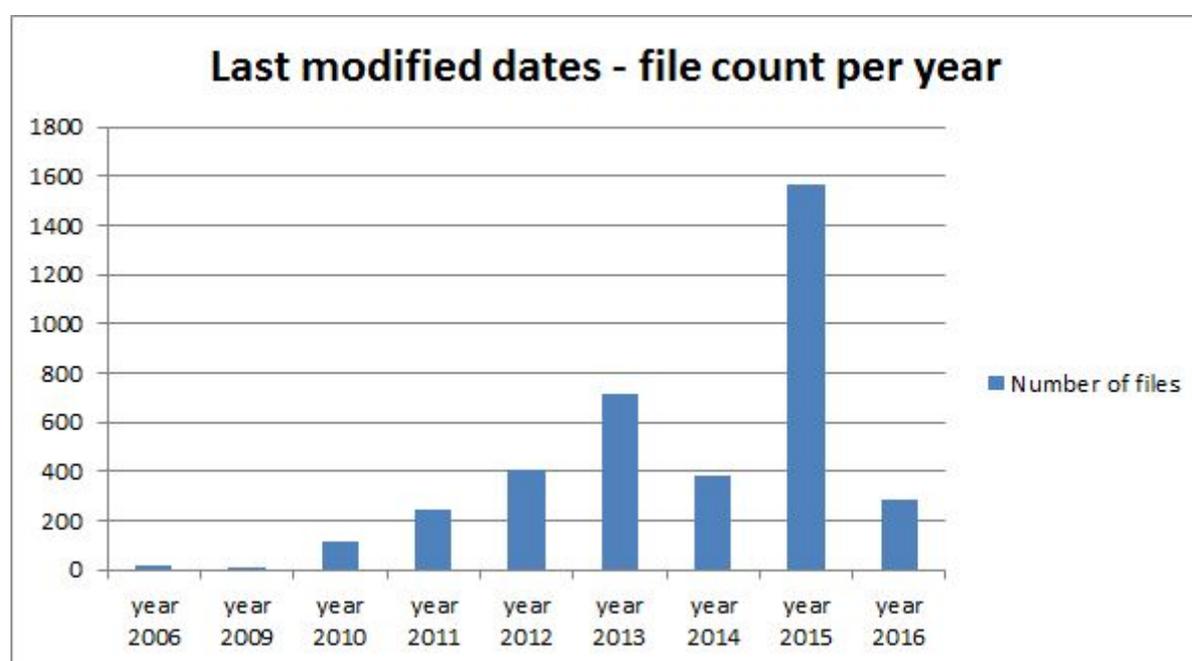


Figure 3: Last modified dates in York's research data sample

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 files²⁸ and JEOL NMR Spectroscopy files²⁹. A longer discussion about these formats and the submission to PRONOM is available as a blog post³⁰. The two new signatures were released on 29th June 2016 as part of DROID signature file version 85³¹.

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 Document³². 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 formats³³. 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 post³⁴, 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 post³⁵. The resulting signature for an OMNIC Spectral Data File³⁶ 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 post³⁷. This work led to the creation of a signature for Final Draft Document 5-7³⁸ 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 progress³⁹. 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 2⁴⁰ and are now available in signature version 88 along with a revised entry for AppleDouble Resource Fork 1⁴¹.

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 release⁴².
- There are benefits to be had from sharing experiences around signature development with the wider digital preservation community (for example via the PRONOM google group⁴³).
- Enhanced and improved documentation⁴⁴ around signature development would help the community engage more as would moving the signature development utility⁴⁵ 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 characteristics⁴⁶. 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 not a new one for the digital preservation community. A recent attempt to address this issue as a community, Just Solve the File Format Problem⁴⁷, 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 match⁴⁸. 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 COPTR⁴⁹.

Archivematica gives the operator the choice of the PRONOM based tools FIDO⁵⁰ or Siegfried⁵¹ to carry out format identifications, or there is the option of identification by file extension⁵².

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 format⁵³ 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 archives⁵⁵ and is one that is also consistently supported by the digital preservation literature. For example The National Digital Stewardship Alliance's Levels of Digital Preservation⁵⁶ 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 webpages⁵⁷) 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_\(Format_Identification_for_Digital_Objects\)](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 parsimonious⁵⁸ 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 project⁵⁹ 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 Foundation⁶⁰ 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 Tika⁶¹ and the Unix File utility⁶²) 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 exists⁶³ and it is anticipated that there will be 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 tools⁶⁴
 - 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 released⁶⁵.
- 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 hackathons⁶⁶ 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 blog⁶⁷; 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)

Jenny Mitcham (presented in the Jisc Research Data Management Shared Service Pilot workshop)

Jenny Mitcham, Julie Allinson, Chris Awre, Richard Green, Simon Wilson⁶⁸

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

Julie Allinson

UK Archives Discovery Forum - Kew (17 March 2016)

Poster: Jenny Mitcham and Simon Wilson

⁶⁷ <http://digital-archiving.blogspot.co.uk/>

⁶⁸



Figure 4: UK Archives Discovery poster

UK Archivematica group meeting - York (22 March 2016)

Chris Awre, Jenny Mitcham, Julie Allinson, Justin Simpson, Richard Green, Simon Wilson, Chris Awre, Richard Green & Simon Wilson (Hull) | Jenny Mitcham⁶⁹

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

Chris Awre, Jenny Mitcham, Julie Allinson, Justin Simpson, Richard Green, Simon Wilson, Chris Awre, Richard Green & Simon Wilson (Hull) | Chris Awre⁷⁰

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

Julie Allinson, Justin Simpson, Julie Allinson and Justin Simpson (Artefactual Systems)

Jisc and CN1 conference - Oxford (6 July 2016)

Jenny Mitcham, Jenny Mitcham

69

<https://docs.google.com/presentation/d/1NbzpxB27yNWRQP2oUr-8qeApFq5ISUc2jiAMqD8sWOO/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)

Poster: j.allinson@durham.ac.uk & m.phillips@durham.ac.uk - Julie Allinson and Matthew Phillips (University of Durham)

TNA Digital Transformation Day - Kew (25 July 2016)

Poster: j.mitcham@kew.ac.uk - Jenny Mitcham

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

Poster: j.mitcham@kew.ac.uk & j.allinson@durham.ac.uk - Jenny Mitcham

UK Archivematica group meeting - Lancaster (14 September 2016)

Poster: j.allinson@durham.ac.uk & j.mitcham@kew.ac.uk - Julie Allinson and Jenny Mitcham⁷¹

iPRES conference - Bern (3-6 October 2016)

Poster: j.mitcham@kew.ac.uk & m.addis@arkivum.com | j.allinson@durham.ac.uk & c.awre@durham.ac.uk & r.green@durham.ac.uk & s.wilson@durham.ac.uk - Jenny Mitcham, Matthew Addis (Arkivum), Julie Allinson, Chris Awre, Richard Green, Simon Wilson⁷²

Hydra Connect - Boston (3-6 October 2016)

Poster: j.allinson@durham.ac.uk & r.green@durham.ac.uk - Julie Allinson, Richard Green

Poster: j.allinson@durham.ac.uk & j.mitcham@kew.ac.uk - Julie Allinson & Jenny Mitcham

Research Data Spring Showcase - Birmingham (20 October 2016)

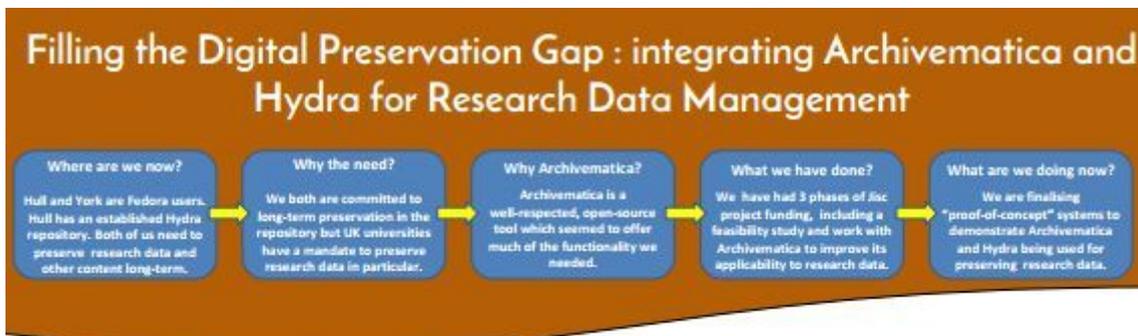
Lightning talk and demo: j.allinson@durham.ac.uk - Julie Allinson, Chris Awre, Jenny Mitcham

⁷¹

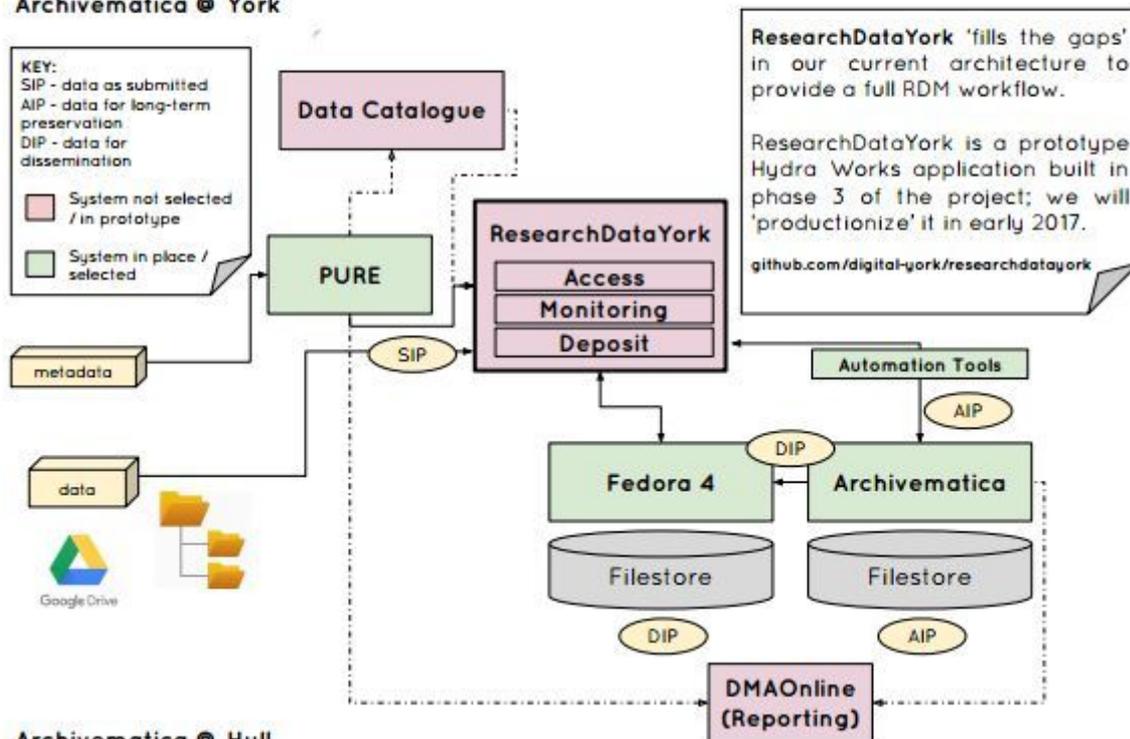
<https://docs.google.com/presentation/d/1b-yF36iD3llqB1WxVbOyiwgSqmvUWk0TE3k97ejFuuM/edit?usp=sharing>

⁷²

http://www.ipres2016.ch/frontend/organizers/media/iPRES2016/PDF/IPR16.Proceedings_4_Web_Broschuere_Link.pdf



Archivemata @ York



Archivemata @ Hull

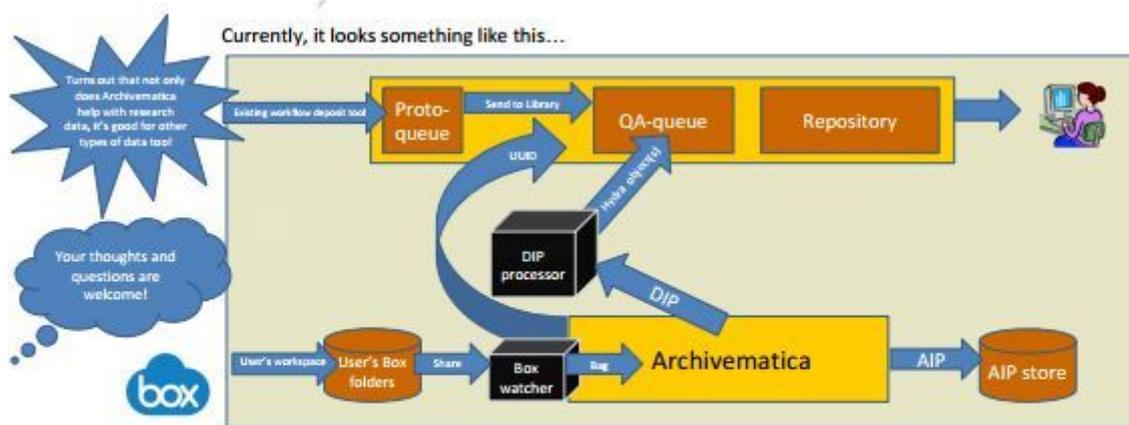


Figure 5: Poster at Hydra Connect

Other publications

The National Archives

Details of our project were featured as a case study within

<http://www.nationalarchives.gov.uk/documents/archives/consultation-strategic-vision-for-archives.pdf>

<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 preservation⁷³:

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 Ë	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 Archivemata group at Lancaster http://digital-archiving.blogspot.co.uk/2016/09/uk-archivemata-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/archivemata/>.

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.

The screenshot shows the website for the Borthwick Institute for Archives at the University of York. The header includes the university logo, the institute name, a search bar, and navigation links for 'Borthwick' and 'University'. A breadcrumb trail indicates the current page is 'Projects > Archivematica'. A left-hand navigation menu lists various site sections, with 'Projects' expanded to show 'Planning your funded research project'. The main content area features a banner titled 'Filling the Digital Preservation Gap' with the subtitle 'Report on Archivematica for research data now available.' and the Jisc logo. Below the banner is a paragraph of text describing the project's start in March 2015, its goal to explore digital preservation solutions like Archivematica, and its funding from Jisc's Research Data Spring program in collaboration with the University of Hull.

The diagram below the text illustrates the digital preservation workflow. It starts with a 'SIP' (Source Information Package) box on the left. An arrow points to a 'Transfer' box, which leads to an 'Ingest' box, then to 'Archival Storage', and finally to 'Access'. Below this top row, there are three storage components: 'Transfer Store', 'AIP Store', and 'DIP Store'. Arrows show data flow from 'Transfer' to 'Transfer Store', from 'Ingest' to 'AIP Store', and from 'Access' to 'DIP Store'. The 'AIP Store' and 'DIP Store' boxes contain examples of storage systems: 'eg. arkivum, local filestore, local preservation repository ...' for AIP and 'eg. local access repository, archives management system, local filestore ...' for DIP. A feedback arrow points from the 'Transfer Store' back to the 'SIP' box.

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 times⁷⁵.

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 times⁷⁶.

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 pipeline⁷⁷

Box: A service used at the University of Hull for secure file sharing, storage and collaboration⁷⁸

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 activities⁷⁹

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 platform⁸⁰

Figshare: A repository where researchers, institutions and publishers can share research outputs⁸¹

Hydra: A repository solution based on a number of “best-of-breed” open-source components, including Fedora⁸²

JHOVE: JSTOR/Harvard Object Validation Environment is an extensible software framework for performing format identification, validation, and characterization of digital objects⁸³

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 Identifier⁸⁶ - 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

..

AaZ'uxnVZyyXZYwWUZx'aZxZ'Njn, yM|yZx{n'XZunyq'k MZXdy'nxjnl`{Zxk'Xb{N|uxZyZx'fMl'nl`
MIX{n'aM'Z'IMZ'Vhx'n_{aZk'ujWZX'b_{aZ'BI'bfZxyq't-y'Xb{N|uxZunyq'nxtS)†XxM'5u{nl'Nj}†SVhulZy'
n_{aZ'bz'y'VM'UZ'k'NXZ'M'N'N'UJZ'axn|`a')†XxM'nx'Xn, IjnN'a'

..

Deposit options

..

2 MZXdy{n'UZ'XZunyq'ZX'k M'UZ'Myb`jZ'bz'nxk |j{bzjZ'bz'yS'k M'Njn'UZ'M'njXZx'y{x|V| xZ'
Vhl {N'b`k |j{bzjZ'bz'y'b'M'nx'M'byZX'at'X'Va't'a'

..

AaZ'xZy|j{b`'xZunyq'nxt'Vhl {Zl {k M'UZ'Myb`jZ'nUHZV'nxk |j{bzjZ'nUHZV'y'aAaZyZ'nUHZV'y'XZYwWUZ'
{aZ'bz'y'A{aM'aM'Z'UZZI'ujWZX'b_{aZ'BI'bfZxyq't-y'uxZyZx'fMl'nl'y{nZ'MIX'k M'nu{nl'Nj}†Vhl {N'b`
VhulZy'n_{aZ'X'M'nx'Xn, IjnN'a'

..

°y'IMXZunyq'nxs'b'IMX'q'nl' {n'tn|x'bz'y'nx'XZunyq'Stn|'k |y{uxn'fXZ'nl'Z'nxk'nxZ'ytk'ujZ'Z'bz'yS'
Nj'XZYwWUZx'UZjn, S{aM'b'Xb'MZ'an, 'tn|', bya'tn|x'Vhl {Zl {n'UZ'XZNY'q', k'a'

..

One or more files to be represented in a single object

..

Aaby'by{n'u{nl'Nj'n, y'tn| {n'XZunyq'nl'Z'nxk'nxZ'bz'y'M'y'M'xn|u'aAaZ'bz'y', bzj'UZ'i'Zu{ {n'Z{aZx'
MIX{aZxZ', bzj'UZ'Myb`jZ'nUHZV'`b_{aZ'xZunyq'nxt {aM'XZYwWUZy{aZk'Nj'a'5u{nl'Nj}†SM'†nx'Nj'n_
{aZ'bz'y'VM'UZ'VhulZx' {n'XZunyq'nxt'nx'Xn, IjnN'a'

..

Üa'† xZMZ'M'njXZx'b'†n|x'En..yuWZ'b', ab'a {n'NjyZk'UjZ'tn|x'Vhl {Zl {a'

Ýa'† nut'b {n {aby'njXZx' {aZ'bz'y' {aM'†n|', bya' {n'XZunyq'a'

Pa...Jn| k | y{ l n, 'NXX{ n { aZ Ln...njXZxnl Z' n xk nxZ yk ujZ {Z. { hZy{ aM', hJ XZ{ Zxk b Z'an, ' tn| xVhl {Zl { by'uxnVZyZX{ axn| ` a n| xMl{ nk MZX yty{ Zk a' I t ytk ujZ {Z. { ZXk nx'VMl 'UZ | yZX{ n VxMZ { aZ hZyS_ nx' b y{ MVZ ' 3n{ ZuM' b H b Xn, y' nx' AZ. { ' Xk' nl ' Ml ' uujZ'

† xMZ MhZ VjZ XZ -XZYWtu{ tnl a'. {®

- o 'k | y{ Vhl { Nl ' Ml hZ { aM' XZYWtuZy { aZ' k MZX yty{ XZuny{ ZX'
- o 'k | y{ Vhl { Nl { aZ l Nl Z y' An_ { aZ VxM' nxy' An_ { aZ' k MZX yty{ { aZ' n xk -1M' l Nl ZS_ by{ l Nl Z' n x-1M' l Nl ZS' by{ l Nl Z' t { hZy' A' 2 | j{ hZj' Ml { anxy' yan| jX' UZ' yZuM' MZX' Ut yZk b' hJnl y'
- o 'k | y{ Vhl { Nl ' Ml by{ n_ hZy' - f y' hZj' hZy' A' { aM' MZ { n' UZ' k NXX' M' M' M' jZ' n x' Xn, l jnM' b) t XxM' AaZ j by{ VMI UZ' UjMl i j n f y' hZj' hZy' Anxk M' Vhl { Nl { aZ', n xk -Mj' ® j' hZy' { n' UZ' f y' hZj' A' 2 | j{ hZj' hZj' Nl Zyk | y{ UZ' yZuM' MZX' Ut yZk b' hJnl y'
- o 'k M' Vhl { Nl ' M' Vhl { Zl { 2 nXZj' ® hZjX' { aM' XZYWtuZy { aZ' u xk M' t { tuZ' n_ k MZX yty{ UZb' ` XZuny{ ZX' a' _ { aby' hZjX' by' nk hZ{ XZS{ aZ' XZ' Ml j{ by' - X M' M' yZ { ® UÄ; @Z j by{ b' ANUjZ'
- o 'k M' Vhl { Nl ' NXX' hZ tnl M' z | Uj b † n xZ' k Z{ NXX' M' M' hZjXy' yZ Z j by{ b' ANUjZ' YÄS_ n x' b y{ M' VZ' M' XZYWtu{ tnl ' n_ { aZ' k MZX yty{ XZuny{ ZX' n x' M' yZ Vb' V' M' tnl { aM' tn| ' n| jX' j b Z | yZ X' b' xZ' Z xZl VZ y' 3n' Mj' ž t hZjXy' MZ' V' xZl j t tk ujZk Zl { ZX' a'

AaZ' n j n, b ` MZ' f y' XZ. Nl ujZ y' n_ -XZYWtu{ tnl a'. {® 3n{ Z { aM' { aZ' hZjX' j M' UZjy' UZ' b' , hZa' M' n, Zx' V yZ j Z { Zx' M' X' MZ' n j n, ZX' Ut ' M' hJnl a'

```

\↔→æİÁĖá\áÁ*á'←á&æÁ}↔\áÁá→Áâ↔→æbÁ{↔b↔â→æÁ↔^ÁÔ}ääá Á
ÁÁ
'ääá\~äíÁR~|bæÉÁR↔'←æ]İÁR~|bæÉÁR↔^^↔æİÁĖ|'←ÉÁĖ~^á→ääÁÑÈ Á
ÁÁ
ääb'ää*↔↔^İÁU*æ'↔↑æ^Áää\áÁ*á'←á&æÁ}↔\áÁáÁb↑á→Á^|↑ääáÁ~ääâ↔→æbÈ ÁÁ
ÁÁ
'^\æ^R~ää→İÁää\ábæ\ Á
ÁÁ
{↔b↔â→æÔ↔→æbİÁá→

```

```

\↔→æİÁĖá\áÁ*á'←á&æÁ}↔\áÁá→Áâ↔→æbÁ{↔b↔â→æÁ↔^ÁÔ}ääáÁá'áÁ}↔\áÁääá↔↔↔^á→ Á
↑æ\áääá\áÁ
ÁÁ
'ääá\~äíÁR~|bæÉÁR↔'←æ]İÁR~|bæÉÁR↔^^↔æİÁĖ|'←ÉÁĖ~^á→ääÁÑÈ Á
ÁÁ
ääb'ää*↔↔^İÁU*æ'↔↑æ^Áää\áÁ*á'←á&æÁ}↔\áÁáÁ^|↑ääáÁ~ääâ↔→æbÈÁá→Á~Ááæ Á
ää}^→~ääää→æÁ\ää~|æÁÔ}ääáÈ ÁÁ
ÁÁ
{↔b↔â→æÔ↔→æbİÁá→ Á
Á
'↔\á\↔↔^İÁR~|bæÉÁR↔'←æ]İÁR~|bæÉÁR↔^^↔æİÁĖ|'←ÉÁĖ~^á→ääÁÑÈÁĖĖİJÁDÁ'æbæää'á Á
ää\áÁää~↑Á\ääÁĖ↔b^æ]Ášã~↓æ'\^ÁŮ↔↔{æáb↔↔\Á~ääÁÔ}→ Á
ÁÁ
b|á↓æ'\İÁOáá\~~^bİÁĖ↔b^æ]ÉÁŮá→\İÁN^↔↑á\↔↔^ Á

```

```

\↔→æİÁĖá\áÁ*á'←á&æÁ}↔\áÁbæ→æ'\æää↔→æbÁ{↔b↔â→æÁ↔^ÁÔ}ääá Á
ÁÁ
'ääá\~äíÁR~|bæÉÁR↔'←æ]İÁR~|bæÉÁR↔^^↔æİÁĖ|'←ÉÁĖ~^á→ääÁÑÈ Á
ÁÁ
ääb'ää*↔↔^İÁU*æ'↔↑æ^Áää\áÁ*á'←á&æÁ}↔\áÁáÁb↑á→Á^|↑ääáÁ~ääâ↔→æbÈÁU~↑æ Á

```

à↔æbÁà~ãÁ*ãæbæã{á\↔~^Á~^→]Á½Á^~\Á\~ÁâæÁ↑áääæÁá{á↔→áâ→æÁ↔^ÁÔ]ääáÊÁ\áääæ Á
 ~*]Áâ↔→æbÁá{á↔→áâ→æÁâ~ãÁã~}^→~áääÊ Á
 ÁÁ
 {↔b↔â→æÔ↔→æbÍÁâ↔→æFÊ↓*&ÍÁâ↔→æGÊ*ääÍÁâ↔→æÇÊ→~& Á

Á

\↔→æÍÁÊá\áÁ*á'←á&æÁ}↔\áÁ^~Áâ↔→æbÁ{↔b↔â→æÁ↔^ÁÔ]ääá Á
 ÁÁ
 'ääá\~ãÍÁR~|bæÊÁR↔←æ]ÍÁR~|bæÊÁR↔^↔æÍÁÊ|'←ÊÁÊ~^á→ääÁÑÊ Á
 ÁÁ
 äæb'ã↔*\↔~^ÍÁU*æ'↔↑æ^Áää\áÁ*á'←á&æÁ}↔\áÁáÁb↑á→Á^|↑âæää~àÁâ↔→æbÊÁÔ↔→æb Á
 à~ãÁ*ãæbæã{á\↔~^Á~^→]Á½Á'~*↔æbÁ^~\Á\~ÁâæÁ↑áääæÁá{á↔→áâ→æÁ↔^ÁÔ]ääáÊÁ↓|b\Áá Á
 ↑æ\áää\áÁää~ãääÊ Á
 ÁÁ
 '~^\'æ^\R~ää→ÍÁää\ábæ\ Á
 ÁÁ
 {↔b↔â→æÔ↔→æbÍ

Nxi	Xxi	jZ{{Zx'	unjbt/5x<xnVZX xZ'
Unni	ZfZI {'	jbtZI VZ'	uxZyZI {Mlml
Unni t aMli{Zx'	`ZI ZxM ni {ZI {'	k Mli'	xZ' jMlml
Vhl_ZxZI VZ<MliZx'	` bXMI VZ'	k ZZ{b`<MliZxy'	xZunx{'
Vhl_ZxZI VZ° Uy{xWl{'	aMIXUnni	k nfb`*k MIZ'	yn_{, MZ'
Vhl_ZxZI VZ<nyf{Zx'	b{ZxI Z{< UjbtMlml	k ybt@nxZ'	yn IX'
XMMVYZ{'	lm xl Nf' x{bVjZ'	l Z, yjZ{{Zx° x{bVjZ'	.
XbtMxM'	jZMl b`2 MZxbtjy'	uan{n`xMlia'	.

Table 1: Acceptable content model types

..
 ..
 ..
 ..

NWj{xNW'	VhfZxMVZ'	aNy' nXk M'	byGZxybnl 5_	xZw' bZy'
NWZyy?ba{y'	VZMZx'	aNy<M'	jM' MZ'	xba{y'
NWx NjZ Z{anX'	VZMInx'	aNyGZxybnl '	jbnZl yZ'	xba{y} njXZx'
NWx Nj-kZbnXBk'	XMZ'	bXZ {bZx'	k ZXbMInx'	yn xVZ'
NWx Nj-knbV'	XMZ' WZu{ZX'	b y{x V{bnl NjZ Z{anX'	k ZXb k '	yuMbnj'
Nj{Zx MjZ'	XMZl nutxba{ZX'	by' nXk M5_	k nXbZx'	y UfZV{'
MjXZl VZ'	XMZ@ Uk k{ZX'	by<M{5_	uxnfZl MIVZ'	{MjZ5_ l nl {Zl {y'
MjNjNjZ'	XZyVbnl{bnl '	by?Z_ZxZl VZXt'	u UjbyZx'	{Zk unxNj'
Ubnjbn' xNlaBb kMbn l'	ZX V{bnl lZfZj'	by?ZujMZXt'	xZ_ZxZl VZy'	{kZj'
Vnl _nXk yAn'	Z. {Zl {'	by?Zw' bZXt'	xZjMbnl '	{tuZ'
Vnl {xNj {nx'	_nXk M'	by ZX'	xZujMZY'	fNjX'

Table 2: Dublin Core metadata acceptable in description files

á^{a..*}_tn| aNy' V{Mbnl yn_ MuMZXinxuMIZxyxZjMZX{n'XMIM{aM'tn| MZ XZunybnl `S_nx'
 b y{MIVZ b` kbnZ{Z.._nXk MStn| k M' MXX {aby{aZyZ {n'tn| x_njXZxb MjZ VjJZX'
 -M{Mbnl y'Ubnl@'5{aZx'_nXk MjyMZ l n{ V| xxZl {jt'y| uunx{ZX^a'

á^{a..*}_Z{aZxZ by' b _nXk Mbnl 'tn| , bya {n'uMjy{n{aZ xZunybnlxt y{M_b {aZ BI bZxybnl' lbnVtS
 VZMZ' MIn{aZxytk ujZ {Z. { _bjZ VjJZX' -xZMk Z'q. { @b 'tn| xLn.._njXZx^a AaZ {Z. {b' {aby'
 _njXZx', bj' UZ' fbybnjZ {n{aZk , aZl {aZt Vnk Z {n'VaZM {aZ xZunybnlxt nUfZV{ {aM'by'
 VZMZx' Mx{aZt VMI {aZl Vnl {M' 'tn| b_l ZVZyyM^a'

á^{a..*}_tn| MZ yMby_ZX {aM'ZfZxt{ab` byMj'tn| , Ml {k {n'UZS| yZ {aZ k.n. : @aMZ@nu{bnl '_nx'
 tn| x_njXZx{n'yaMZ k , ká {aZ Zk Nj' MXXZy [MabZk MbMl a|jj'W|j](#) i^a'

á^{a..*}yyZl {bnj}tS{aM'y'kE*_tn| VaZM 'UWI _nXk {tk Z {n{tk Z'tn| , bj' yZZ {aM'tn| x_njXZxl M Z'
 aNy' {aZ y{M| yn_tn| xk MZxbjIMuZI XZx^a @nS_nxb y{MIVZSMnjXZxVjJZX' -2 tXMIM^o, bj' `n'
 {axn| `a'M | k UZx'n_y{MZYy{M{b` , ká -2 tXMIM^o uxnvZyyb` @MIXZl Xb` , ká -2 tXMIM^o
 ° Nj' uxnvZyyb` Vnk ujZ{Z@5l VZ tn| xZMva {abyjM' y{MZ'tn| VMISyan|jX'tn| , byaSXZjZ'
 {aZ'_njXZx' Xn'l n{XZjZ{Z k' | l {b' {aZl E

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

AabyZVnl X'nu{bnl 'Nj'n, y'tn| {n'VZMZ k | j{bnjZ'nUfZV{y'_nXk {aZ' bjZy{aM'tn| XZunybnl^a
 5u{bnl Nj}tSMl'tnxNj'n_{aZ' bjZy' b {aZyZ'nUfZV{y'VM UZ' VhuZx{n{aZ' xZunybnlxt'_nx'Xn, l jnM^a'

Ü³...t xZMZX MnjXZx b` tn| x`En.:yuWZ b` , ab`a {n`MyZk UjZ`tn| x`Vhl {ZI {a`

Ý³...t nut b` {n`{aby_njXZx{aZ`_bZy{aM|`tn|` , bya`{n`XZuny`a`

Þ³...Jn| k | y| l n,` MXX`{n`{aZ`En.:_njXZx`Mhk k MyZuMMZX`fny| Zy`yYf`_bZ`{aM|` , bjj`
XZ{Zxk b`Z`an,` `tn| x`Vhl {ZI {`by`uxnVZyyZX`{an|` `a`n| x`M|`{nk MZX`yty{Zk` a`°` `Yyf`_bZ`by`
k ny|`ZNY`b` VxZMZX`b` MyuxZNYaZZ`{`Mduj`b`M|`

† xZMZ`MyZ`{n`_aZ`NXb` `y`b` `xn,` `Ü`VhxZyuni`Xb` ` `n`{aZ`k`Z`{NXMM`lZjXy`{aM|`tn|` `b`{ZI`X`
uNy`y`b` `b`{n`{aZ`yty{Zk` `MIX`{aZ|` `Mn,` `n`_k`Z`{NXMM`l`nx`ZMwa`_bZ`{n`UZ`XZuny`aZ`X`a`

- Jn| k | y| `by|`uNy`y`{aZ`_bZ|`Nk`Z`{aM|`ZMwa`xn,` `n`_k`Z`{NXMM`l`Mduj`bZy`{n`
- Jn| k | y|`uNy`y`M|`bZjZ`{aM|`XZY`x`bZy`{aZ`k`MZX`b`y`XZuny`aZ`X`b`{aM|`n`UzV`{`
- Jn| k | y|`uNy`y`{aZ`l`Nk`Z`y`Än`_{aZ`VxZM`nx`y`Än`_{aZ`k`MZX`b`y`b`ZMwa`n`UzV`{`b`{aZ`
_nxk`-1Ny|`l`Nk`Z`S`by|`l`Nk`Z`®nx`-1Ny|`l`Nk`Z`S`by|`l`Nk`Z``{`l`Ny`y`Ä`2`|`{`bZjZ`M|`{anxy`
_nx`{aZ`y`Nk`Z`k`MZX`b`y`yan|`jX`UZ`y`ZuMM`ZX`U`y`Zk`b`hjn|`y`a`
- "Mwa`_bZ`xn,` `k`|`y|`Vhl`{Nk`M`f`bZjZ`_bZy`®ZI`{xt`b`{aZ`_bZ`by`{n`UZ`M`f`bZjZ`_nx`
Xn,` `l`jnNk`b`)`†XxM`AaZ`_f`bZjZ`_bZy`ZI`{xt`yan|`jX`UZ`y`Z`{`n`-Ny|`®`b`M`nut`n`_{aZ`_bZ`
by`{n`UZ`k`NXZ`M`f`bZjZ`_nx`Xn,` `l`jnNk`a`_`l`bZjZ`_UjM|`S|`n`Vhut`_bZ` ,` bjj`UZ`
M`f`bZjZ`a`
- Jn| k`M`uNy`y`M`Vhl`{ZI`{2`nXZj`®`_bZjX`{aM|`XZY`x`bZy`{aZ`{tuZ`n`_k`MZX`b`y`UZ`b`
XZuny`aZ`X`a`_`{aby`_bZjX`by`nk`l`{ZXS`{aZ`XZ`M|`j`by`-XMMVZ`{`®`_Z`Z`jby`b`ANUjZ`Ü`Ä`
- `k`M`Vhl`{Nk`NX`l`b`l`Ny|`z`|`Uj`b`†`nx`k`Z`{NXMM`l`ZjXy`y`Z`jby`b`ANUjZ`Ý`Ä`_nx`
b`y`{MIV`M`KZY`x`bZy`{n`_{aZ`k`MZX`b`y`XZuny`aZ`X`nx`MyZV`b`b`V`l`M|`l`{aM|`tn|`
 ,`n|`jX`j`b`Z`|`y`Z`b`x`Z`Z`l`V`y`a`3n`{`Ny|`z`†`_bZjXy`M`Z`V|`xxZ|`{†`tk`ujZk`ZI`{Z`X`a`

AaZ`_njn,` `b` `by`M|`Z`Nk`ujZ`n`-XZY`x`bZy`{n` `Yyf`®`Ny`j`Nk`n|`{`b`MyuxZNYaZZ`{`Mdu`a`3n`Z`
{aM|`{aZ`_bZjX`j`MdujZy`UZ` `b` ,` l`a`M`n,` Zx`V`y`Z`jZ`{Z`X`a`

ä↔æ^á↑æÄ	\↔→æÄ	´äää\~äÄ	äeb`ä↔*↔~^Ä	b á↓æ´\Ä	´~^æ^R~ää→Ä	{↔b↔ä→æÖ↔→æ bÄ
↔´\ ääÈ↓ &Ä	R Ä*↔´\ ääÄ	R~ bæÈÄ R↔´←æ Ä	Çää↔→&Ä R↔^↔→æÄ	R↔^↔→æÄ R~ bæÄ	*á~\~&ää*áÄ	ÄÄ
b\~ä]È**\Ä	R Ä *ääææ^á\↔~^Ä	Ç ´←ÈÄ Ç~^á→ää NÈÄ	R Äääææää´áÄ }~ä←Ä	Sá \↔´á→Ä ää↔b\~ä]Ä	*ääææ^á\↔~^Ä	á→Ä
ää\áÈ[→bÄ	R Ä b*ääääbääæ\Ä	R~ bæÈÄ R↔^↔→æÌÄ R~ bæÈÄ R↔´←æ Ä	R Äääá\áÄ	ÄÄ	ÄÄ	á→Ä

°`yy|k`b` `{aM|`{aZ`l`Nk`ZX`_bZy`MZX`uxZyZI`{`b`tn|`x`njXZx`SuxnVZyyb` `{aby`XZY`x`bZy`{n` `Yyf`
 ,`n|`jX`Zy|`j`{`b`{axZZ`n`UzV`y`b`)`†XxM,` l`a`VhxZyuni`Xb` `XMM`b`{aZ`uxZyZy`f`M|`l`y`{nxZ`a`
AaZ`_by|`n`UzV`{` ,`n|`jX`Vhl`{Nk`{aZ`k`Z`{NXMM`l`nx`MuxZyZy`f`ZX`tk`MIZ`_bZS|`{`{aZ`tk`MIZ`
 ,`n|`jX`l`n`{`UZ`M`f`bZjZ`_nx`Xn,` `l`jnNk`a`AaZ`y`Z`Vhl`X`n`UzV`{` ,`n|`jX`XZY`x`bZy`{aZ`uxZyZI`{`M|`l`
MIX`k`MIZ`M`nut`M`f`bZjZ`_nx`Xn,` `l`jnNk`a`AaZ`_b`Ny|`n`UzV`{` ,`n|`jX`XZY`x`bZy`M`yuxZNYaZZ`{`n`
XMM`MIX`k`MIZ`M`nut`M`f`bZjZ`_nx`Xn,` `l`jnNk`{`aby` ,`n|`jX`UZ`x`Z`Vh`l`bZjX`Ny`M`XMM`VZ`{` ,`l`a`b`
{aZ`xZuny`aZ`X`nxt`f`Zl`{`an|` `a`{aZ`Vhl`{ZI`{2`nXZj` ,`Ny|`n`{`yuZV`bZ`X`a`

β^{a...*}_{aZxZ by b _nxk M{nl 'tn| } , bya {n'uMy{y{n {aZ xZuny{nx t y{M_b {aZ BI bZxy{t 1bXVtS
WZMZ M{n{aZxYtk ujZ {Z. { _bz VjZX -xZk Z^q. { @b 'tn| xLn. . _njXZx^aAaZ {Z. { b {aby
_njXZx, bj UZ fbyUjZ {n {aZk , aZl {aZt Vnk Z {n VaZM {aZ xZuny{nx t nUzV {aM by
WZMZx Mx {aZt VMI {aZl Vhl {M' tn| b_l ZVzyMt^a

ā^{a...*} _tn| MZ yMby bZx {aM ZfZxt {ab ` byM' tn| } , M { b' {n UZS| yZ {aZ Ln. :-@Mz @nu{nl _nx'
tn| x_njXZx {n yaMz b' , b'a {aZ Zk Nj MXXzyy MabZk MbMl a|jj^aWp| i^a

á^{a..."} yyZl {bY}tS{aM y bE*_tn| VaZM UMI _nxk {tk Z {n {tk Z 'tn| } , bj yZZ {aM 'tn| x_njXZxl M Z'
aM y {aZ y{M| y'n _tn| xk MZxbyIMuZI XZx^a@S_nxb y{M VZSM_njXZx VjZX -2 tXMIM^o , bj ` n'
{axn| ` a M | k UZx_n y{M Zyy{M{b ` , b'a -2 tXMIM^o uxnVzyy b ` @MIXZI Xb ` , b'a -2 tXMIM^o
° Nj uxnVzyy b ` Vnk ujZ {Z^o 5l VZ 'tn| xZMw {abyjM y{M Z 'tn| VMI Syan| jX 'tn| } , byaSXZjZ {Z'
{aZ _njXZx Xn l n {XZjZ {Z b' | l {b {aZl E

A hierarchy of files to be represented in a single object

Aaby_b Njnu{nl Mj n, y'tn| {n XZuny{ M abZxVat _njXZx y{x| V| xZAn _bz y^a AaZ _bz y , bj UZ i Zu'
{n Z {aZx b {aZ b _njXZx y{x| V| xZ , ab'a by { | x Zx b {n M' b' _bz^a AaZxZ , bj UZ My b ` jZ nUzV {b {aZ
xZuny{nx t {aM XZyWbZy {aZk MxSnu{nl Mj tS {aZ %w _bz VMI UZ VhulZx {n {aZ xZuny{nx t _nx'
Xn , l jnM^a

Ü^a t xZMZ M_njXZx b 'tn| xLn. . yuWZ b , ab'a {n MyZk UjZ 'tn| xVhl {Zl {^a

Ý^{a...t} nut b {n {aby _njXZx {aZ _njXZx {aM Vhl {M y {aZ abZxVat yn {aM 'tn| } Z { M_njXZx
, b'ab {aZ Ln. . _njXZx

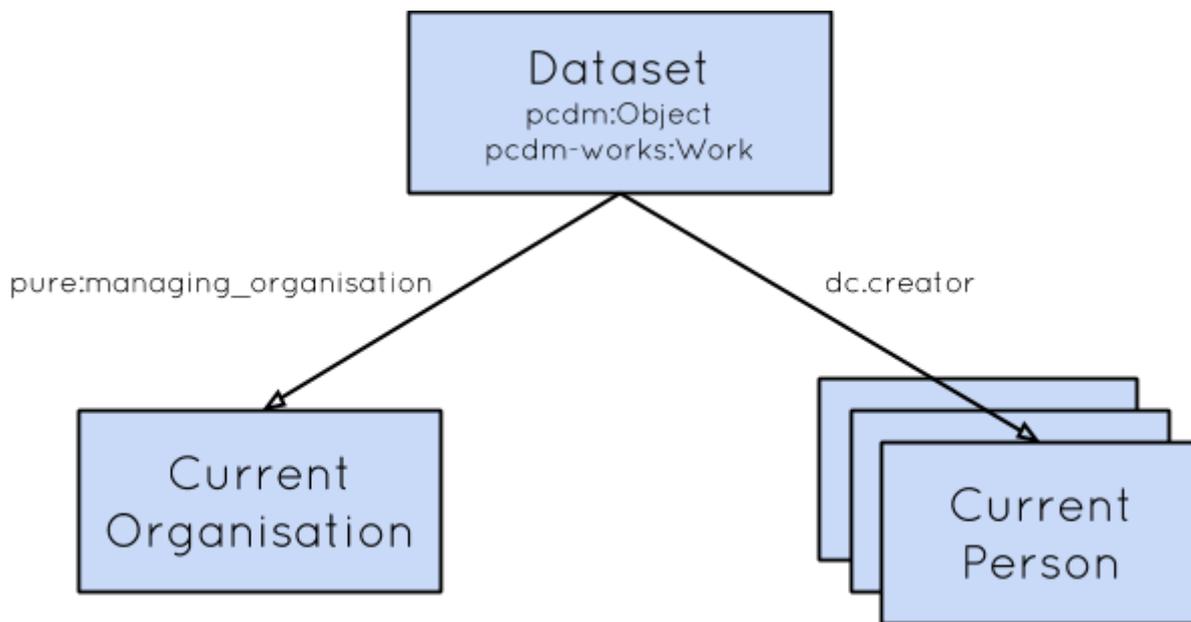
Þ^{a...Jn} | k | y{ l n , MXX {n {aZ Ln. . _njXZxnl Z'nxk nxZ ytk ujZ {Z. { _bz y {aM' , bj XZ {Zxk b Z'
an , 'tn| xVhl {Zl { by uxnVzyyZX {axn| ` a n| xM {nk MZX yty{Zk ^a l t ytk ujZ {Z. { ZX b nx
VMI UZ | yZx {n WZMZ {aZ _bz yS_nxb y{M VZ 3n {ZuM b 'H b Xn , ynxAZ. { ' Xb nl 'MI
° uujZ^a

t xZMZ M_bz VjZX -XZyWku{nl ^q. { @

- 'k | y{ Vhl {N b M' b' jZ {aM XZyWbZy {aZ k MZxby XZuny{ZX
- 'k | y{ Vhl {N b {aZl M Z yAn _{aZ WZMnx yAn _{aZ k MZxby b {aZ _nxk
-1M y{ M ZS by{ l M Z^o nx -1M y{ l M ZS' by{ l M Z ' * { b N y^a @2 | j { b jZ M {anx y an| jX
UZ yZuM MZX Ut yZk b h jnl y^a
- 'k | y{ Vhl {N b 'M- fbyUjZ' bZy^o Zl {x' {n yan , , aZ {aZx M hut' n _{aZ %w _bz yan| jX
UZ k MZ M' M jZ _nx Xn , l jnM' b) tXxM AaZ Zl {x' VMI UZ UjM i _bz l n'
M' M jZ Anxk M' Vhl {N b {aZ , nxX -Nj^o _bz {n UZ Xn , l jnM' M jZ^a
- 'k M' Vhl {N b 'M- Vhl {Zl {2 nXZj^o bZjX {aM XZyWbZy {aZ uxtk M t {tuZ n_k MZxby
UZ b ` XZuny{ZX^a* _{aby bZjX bynk b {ZX S {aZ XZ M j { by -X M V jZ { @Z j by { b ANUjZ Ü^a
- 'k M' Vhl {N b MXX b'nl Mj z | Ujb t nxZ k Z {M M M bZjX yZyZ j by { b ANUjZ Ý^a S_nx'
b y{M VZ MZyWku{nl n _{aZ k MZxby XZuny{ZX nx M yuZV b V b' M'nl {aM 'tn|
, n | jX j b Z | yZx b xZ_Zz l VZ y^a 3n { Mj z t _bz Xy MZ V | xZl j t tk ujZk Zl {ZX^a

AaZ _njn , b ` MZ fVbZ . M ujZ y n _XZyWku{nl ^q. { @ 3n {Z {aM {aZ bZjX j M UZ j y UZ b ` , b'a
Mj n , Zx V y Z j Z {Zx M X MZ _njn , ZX Ut M h jnl ^a

<ZnujZ'MIX'5x' MlbyMlml y'2 nXZj'



3Mk ZyuWZy''

Mk @ua{{u@pXjW^tnxi ^W^| i pml {njin` bZynMby^MabZk MlMlMl
 XjW@ua{{u@pXjW^tnxi ^W^| i pml {njin` bZyµ Zl ZxWlMl
 XV@ua{{u@pµ| xj^hx' µXWµ{Zxk yµ'
 bZl {bZxy@ua{{u@pµXjnV^ nµµµnMlW| jMµµXZl {bZxyµ'
 ynX' @ua{{u@pµWazk Mhx' µ'
 _nM@ua{{u@pµ k jl y^hk µ_nMµµµµ'

2 nXZjy''

ž MMYZf'

° ž MMYZ{ by^MkbyWZ{Z'yZ{n_xZjMZX^XMMlbyZy^AaZ' bZy^MM UZ'n_Ml†{tuZ^a'1 <B?" Mjyb` jZ'
 ž MMYZ{k Z{NXXMMZVhx^aMj^Mhl Z^i{n^nl Z^xZjMlml yabµ{n^Mē MMYZ{Mj^XZYWtbZX^aZx^a° ž MMYZ{
 XnZyl n{San, ZfZxSrequire Ml^Mvhk uMl tµ` <B?" xZVhx^a^1nVj^XZYWtbµ{bZ k Z{NXXMMMM UZ'
 NXXZX, k^a{aZ.Z.VZu{tµl n_u| xZB| bX, ab^a^must UZ^XZx^fZX_xnk <B?"^a'

† nXZ©

[a{{uy@pµ b^a| U^vk µXb bMj^tnxi µXjWb^tXxVµUjnUµk Mj^ZxjWµXjWb^tXxVµk nXZjyµ nxi yµXMMYZ{^xU'](#)

† jMyj	<xnuZx{t'	".uZV{ZX'5UfZV{AtuZ'	ByMjZ'
ž MMYZf'	xX_@tuZ'	B?*	XMM^XMMYZ{'
.	XV@bZjZ'	1bZxVjµ_@xµ` Ä	Ü;XMMWkZ k MlXMMhx^tÄ ' xnk <B?"'

·	XV ^u xZMInx	† xxZI { ' < Zxynl ' }	Ü ^{pa} Ü; XMIMWkZ k MIMMInxtÄ ' xnk ' < B? " ' .
·	bXZI { bZxy ^u nb	B?*	Ü ^{pa} Ü; XMIMWkZ k MIMMInxtÄ ' xnk ' < B? " ' .
·	XV ^u UjbyZx	1kZxNjI; @xb ` Ä	Ü; XMIMWkZ k MIMMInxtÄ ' xnk ' < B? " ' .
·	XVMNjMjZ	1kZxNjI; @xb ` Ä	Ü; XMIMWkZ k MIMMInxtÄ ' xnk ' < B? " ' .
·	Xj ^u W xZB bX	1kZxNjI; @xb ` Ä	Ü ' xnk ' < B? " ' .
·	u xZ ^u MIMb ` BI k	† xxZI { ' 5x MbyMlml ' }	Ü ' xnk ' < B? " ' .
·	u xZ ^u xZ1b i	B?*	Ü ^{pa} ' xnk ' < B? " ' .
·	XV ^u WZy?bafy	1kZxNjI; @xb ` Ä	Ü ' xnk ' < B? " ' .
·	Xj ^u Wk UM ^u n?ZjZMjZ ž MZ	1kZxNjI; @xb ` Ä	Ü
·	Xj ^u W ^u WZyy	1kZxNjI; @xb ` Ä	Ü ^{pa} Ü
·	Xj ^u WZ{ZI {lml <njbt	B? *j5UHzV{	Ü ^{pa} Ü
·	Xj ^u W ^u nx ^u XZ. b `	1kZxNjI; @xb ` Ä	Ü ' xnk ' < B? " ' .
·	Xj ^u WZy{x ^u lml ¼ n{Z y	1kZxNjI; @xb ` Ä	Ü ^{pa} Ü

	uVXk @Mj2 Zk UZx'	<Ml MZ'	Üpa'
--	-------------------	---------	------

ž MlMjMjMjZ' _xnk' <B?" U| {l n{V| xxZl {jt' MXXZx©

- <| UjMlMl yMlX'uxnZV'y
- <ZnujZ', k'a'xnjZyn{aZx{aMl' d xZMlnx-i
- ž ZyWku{tMl'
- (Zn' xMlAbMyl'
- AZk unxyl'
- ž MZ'n'uxnX| V{tMl'
- H nxi _jn, 'pGyWljkf'

ž MlM n{MlMjMjZ' _xnk' <B?" H ZU'@ZyfbZy©

- "k UM' n'
- t nl {Ml' uZxynl'
- ?ZjMlMl y{n n{aZxXMlMjZ'y
- ?Zy{xW{tMl y'l MxMlGZ'

† jMy'	<xnuZx{t'	".uZV{ZX'5UfZV'	ByMZ'
† xxZl {<Zxyn L'	xX_@tuZ'	B?*	ynx' @Zxynl' u xZ@ xZ<Zxynl'
	yi ny@xZ_1MlZj'	1kZxyl'@xW`Ä	Ü'
	_nM@Nk k†3Mk Z'	1kZxyl'@xW`Ä	Ü'
	_nM@bZl 3Mk Z'	1kZxyl'@xW`Ä	Ü'
	u xZ@ xZB bX'	1kZxyl'@xW`Ä	Ü'
	u xZ@ xZ{tuZ'	1kZxyl'@xW`Ä	Ü'

† jMy'	<xnuZx{t'	".uZV{ZX'5UfZV'	ByMZ'
† xxZl {5x' Ml byMlMl'	xX_@tuZ'	B?*	ynx' @x' MlbyMlMl' u xZ@ xZ5x' MlbyMlMl'
	yi ny@xZ_1MlZj'	1kZxyl'@xW`Ä	Ü'

	_nM@Nk Z`	1kZxyl;@xb`Ä	Ü`
	u xZ@ xZB B`	1kZxyl;@xb`Ä	Ü`
	u xZ@ xZ{tuZ`	1kZxyl;@xb`Ä	Ü`

<Nk MZ`

° <Nk MZ` xZuxZyZi {yMkZunyq`n_XMMhZy{aM`_nxk` uMk`n`_M}Mk` Zxž`MnyZ{a° <Nk MZ`, k@a`{aZ` xX_@tuZ`Nk` © xVabfNyM`_nxk` {bnl <Nk MZ`_ž` *Ä,` h} Vhl` {Nk` y|_bVZi` {k` Z{NkMn{n`bXi` {b} {aZ` jnVnlnl` n`_{aZ`y{nxZX`° <`jNk`B| B`Ny`Mk` b` k` | k` Ä° <Nk MZ`, k@a`{aZ` xX_@tuZ` Nk` @`byZk` b` Mlnl` °_nxk` {bnl <Nk MZ`_ž` *Ä,` k@a` Vhl` {Nk` xZ`ZxZi` VZy{n`{aZ`_hZy`Vnk` ux`yb` ` {aZ`ž` <` MlX`{aZ`Xk`B| B`Ny`Mk` b` k` | k` Ä° `yb` `jZ` <Nk MZ` Vn` UZ` Un{a`° <`MlX`ž` *a`° njXZx`y{x| V| xZ`_nx`{aZ` nxb`b` NyIXZunyq`yaNy` UZ` xZ`{Nk` ZX`b` y| Va`M,` M`{aM`k` Vn` UZ` xZ`VwZ`MZX`{n`{aZ`Zi` X`| yZxS`_nx` Z`Nk` ujZ`b` M`° xVabfZk` Ml`M` "A@`_hZ`°5`{aZ`_hZy`k` M` UZ`b` V| XZX`b` {aZ` <Nk MZ`S`_nx`Z`Nk` ujZ` y| Uk` bybnl` XnV| k` Zi` {Mlnl` nx`MkZ`Nk` Z`_hZ`°AaZyZ`yan| jX`UZ`X`yb`b` | byaZ`_nxk` {aZ`XMM`hZy` {aZk` yZj`fZy`°

† nXZ°

[a{fuy@u` k@a| U°Vnk` p`Xb` k`Ny`†nxi` p`Xj`b`a`†X`Ny`U`jn`U`k` Ny`Z`x`j`b`p`Xj`b`a`†X`Ny`k` nXZ`jy`u` nxi` y`u`Nk` MZ`°XU`](#)

† jNy`	<nxuZx{†`	". .uZV{ZX`5UkZV` AtuZ`	ByMZ`
<Nk MZ`	xX_@tuZ`	B?*	Xj`b`@Nk MZ` Nk` © ° xVabfNyM`_nxk` {bnl <Nk MZ` Nk` © ž` byZk` b` Mlnl` °_nxk` {bnl <Nk MZ`
	yi ny@xZ`_1MkZj`	1kZxyl;@xb`Ä	Ü`
	Nk` @k`B B`	1kZxyl;@xb`Ä	Ü°Ü° _nxk` °° xVabfZk` Ml`M`
	Nk` @k`@M y`	1kZxyl;@xb`Ä	Ü° _nxk` °° xVabfZk` Ml`M`
	Nk` @k`@Z`	1kZxyl;@xb`Ä	Ü°Ü° _nxk` °° xVabfZk` Ml`M`

