Digital Preservation Requirements for Research Data Management

Jenny Mitcham, University of York
Chris Awre, University of Hull
DP and RDM

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

• A lot! DP is an essential component of RDM

• Why?
  – The stick - Funder requirements/mandates (e.g., EPSRC, Horizon 2020)
  – But also...
Impact of no DP in RDM

...but we probably all have our own examples of data loss through lack of preservation planning or actions
It started with a Christmas wish...

Digital Archiving at the University of York

WEDNESDAY, 11 DECEMBER 2013

My digital preservation Christmas wish list

All I want for Christmas is a digital archive.

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

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

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box we can put data in, a place where we put data when we have finished using it, a place where data will be stored but no longer maintained.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

My requirements are listed below. Feedback is most welcome

<table>
<thead>
<tr>
<th>#</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INGEST</strong></td>
<td></td>
</tr>
<tr>
<td>I1</td>
<td>The digital archive will enable us to record/store administrative information relating to the Submission Information Package (information and correspondence relating to receipt of the SIP)</td>
</tr>
<tr>
<td>I2</td>
<td>The digital archive will include a means for recording decisions regarding selection/retention/disposal of material from the Submission Information Package</td>
</tr>
<tr>
<td>I3</td>
<td>The digital archive will be able to identify and characterise data objects (where appropriate tools exist)</td>
</tr>
<tr>
<td>I4</td>
<td>The digital archive will be able to validate files (where appropriate tools exist)</td>
</tr>
<tr>
<td>I5</td>
<td>The digital archive will support automated extraction of metadata from files</td>
</tr>
<tr>
<td>I6</td>
<td>The digital archive will incorporate virus checking as part of the ingest process</td>
</tr>
<tr>
<td>I7</td>
<td>The digital archive will be able to record the presence and location of related physical material</td>
</tr>
</tbody>
</table>

| **DATA MANAGEMENT** | |
| DM1 | The digital archive will generate persistent, unique internal Identifiers |
| DM2 | The digital archive will ensure that preservation description information (PDI) is persistently associated with the relevant content information. The relationship between a file and its metadata/documentation should be permanent |
| DM3 | The digital archive will support the PREMIS metadata schema and use it to store preservation metadata |
| DM4 | The digital archive will enable us to describe data at different levels of granularity – for example metadata could be attached to a collection, a group of files or an individual file |
| DM5 | The digital archive will accurately record and maintain relationships |

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

- The Open Archival Information System Reference Model (OAIS)
- ISO16363 (Audit and Certification of Trusted Digital Repositories)
- Work of others
  - Chris Fryer’s business case work for the SPRUCE project
  - University of Leeds RoaDMaP project
- Discussion with colleagues and the wider community
...then “Filling the Digital Preservation Gap” came along

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

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

Archivematica as part of a Hydra preservation workflow in Hull

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

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

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

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

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

It looks something like this...

Workflow deposit tool
Metadata record QA-queue Repository
Proto-queue

DIP processor

Ingest folders Arrangement tool Archivematica

DIP

Legacy content for AIPs

Fedora PID

Reprocess for DIP

AIP store

HTML embedded content

Your thoughts and questions are welcome!
c.awre@hull.ac.uk
r.green@hull.ac.uk
hydra.hull.ac.uk
Making DP work for RDM

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

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