

ARCHIVES

# “Filling the digital preservation gap” for Research Data

Jenny Mitcham, Julie Allinson - University of York  
Chris Awre, Richard Green, Simon Wilson - University of Hull

IDCC conference - 24 February 2016



Jisc

# Why do we need digital preservation?



# Why do we need digital preservation for research data?

- We can't ignore digital preservation – moving targets for data retention mean we need to take this seriously
- Funder requirements around retention:
  - **NERC** - data should be retained for a minimum of 10 years but for projects of major importance this may need to be 20 years or longer
  - **STFC** - expect data to be retained for a minimum of 10 years and data that cannot be re-measured should be retained indefinitely
  - **Wellcome Trust** – expect data to be kept for a minimum of 10 years but suggest longer periods for certain types of data

# Why do we need digital preservation for research data?

University of York RDM questionnaire 2013

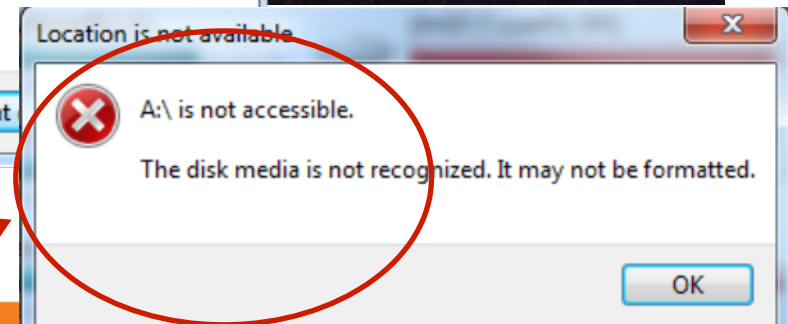
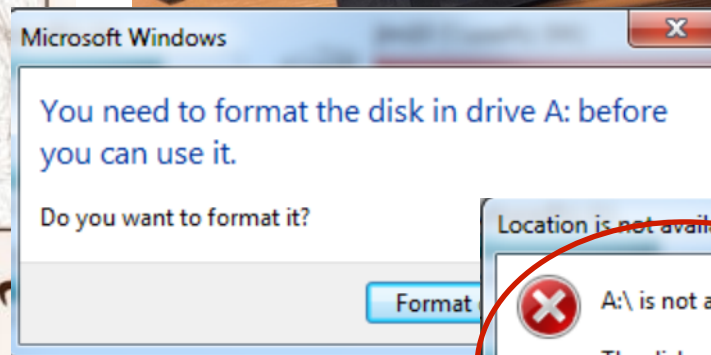
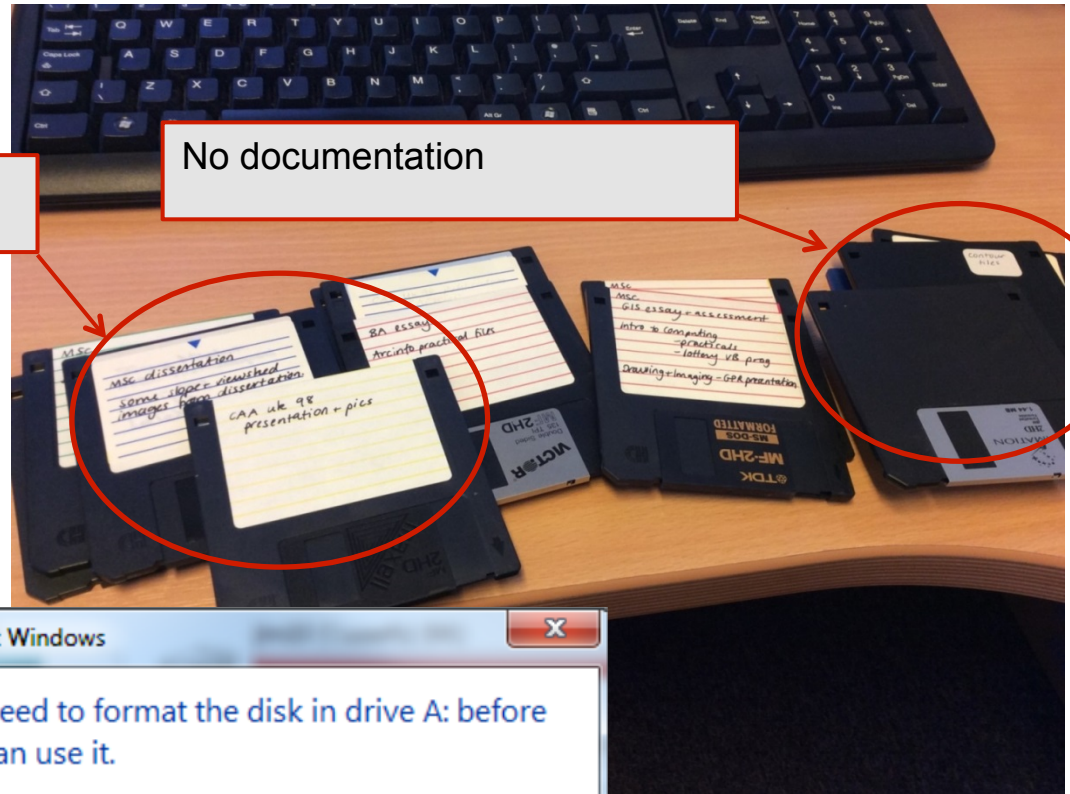
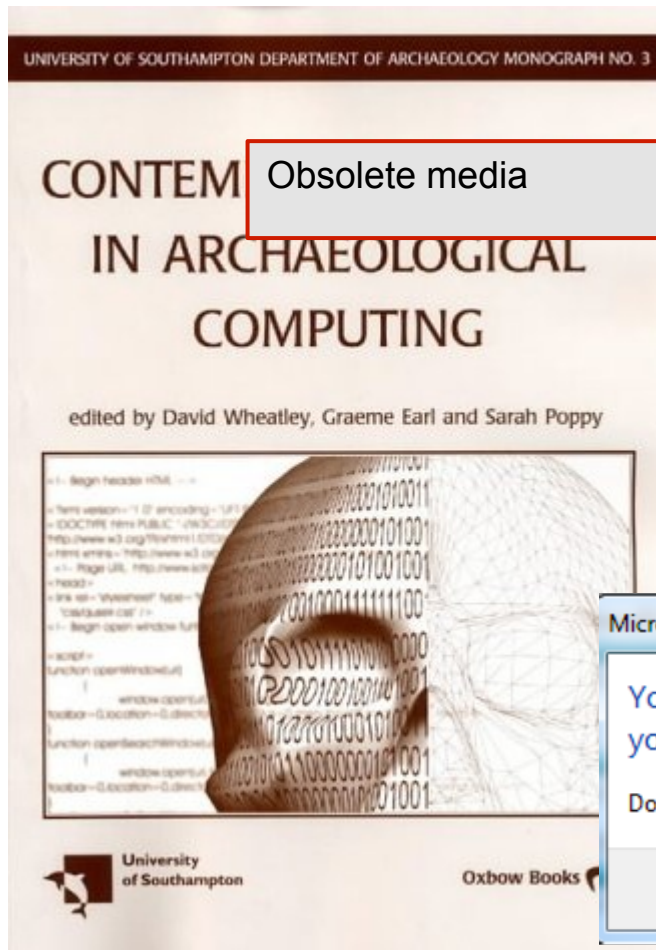
- Which data management issues have you come across in your research over the last five years?
  - **“Inability to read files in old software formats on old media or because of expired software licences”**
  - 24% of 181 researchers who answered this question admitted this had been a problem for them



What services would you like us to provide to help you manage and re-use your research data?



# This is my research data!



Data not accessible

Cryptic names

Missing information about software

DISTEX  
TILES  
ANTIQUIT.DOC  
BIB.DOC  
BIBANT.DOC  
CAAPAPER.DOC  
CAAPUB.DOC  
CVNOTT.DOC  
CVSMR.DOC  
CVSMR.TXT  
CVYORK.DOC

12/10/2015 16:04	File folder		
20/04/1998 17:15	Microsoft Word 97 - 2003 Document	57 KB	
20/04/1			
20/04/1	SU60.GZ	24/07/1997 11:10	GZ File 269 KB
24/02/1	SU62.GZ	24/07/1997 11:11	GZ File 410 KB
21/04/1			
08/01/1			
26/02/1			
21/04/1			
24/02/1			

Windows



Windows can't open this file:

File: SU60.GZ

To open this file, Windows needs to know what program you want to use to open it. Windows can go online to look it up automatically, or you can manually select from a list of programs that are installed on your computer.

What do you want to do?

- ☒ Use the Web service to find the correct program
- ☐ Select a program from a list of installed programs

OK

Cancel

Poorly organised

CAANEW.DOC (Protected View) - Microsoft Word

Protected View Editing this file type is not allowed due to your policy settings. Click for more details.

Backwards compatibility not assured

In Search of a Defensible Site :  
A GIS Analysis of Hillfort Placement  
CAA CONFERENCE PAPER 1998

The aim of the research I will describe in this paper was to use Geographic Information

# The Open Archival Information System

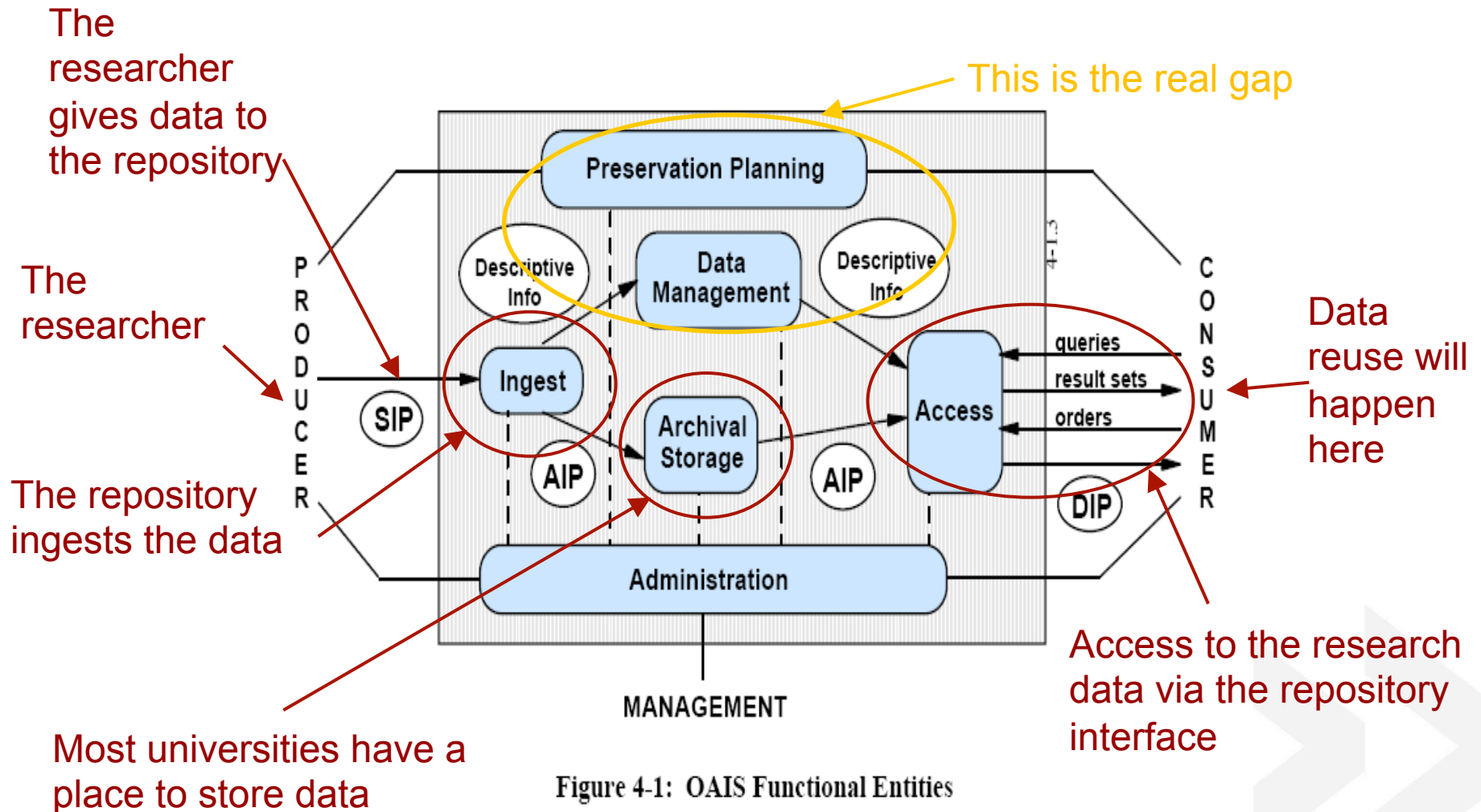


Figure 4-1: OAIS Functional Entities







# Visible v. invisible

Invisible

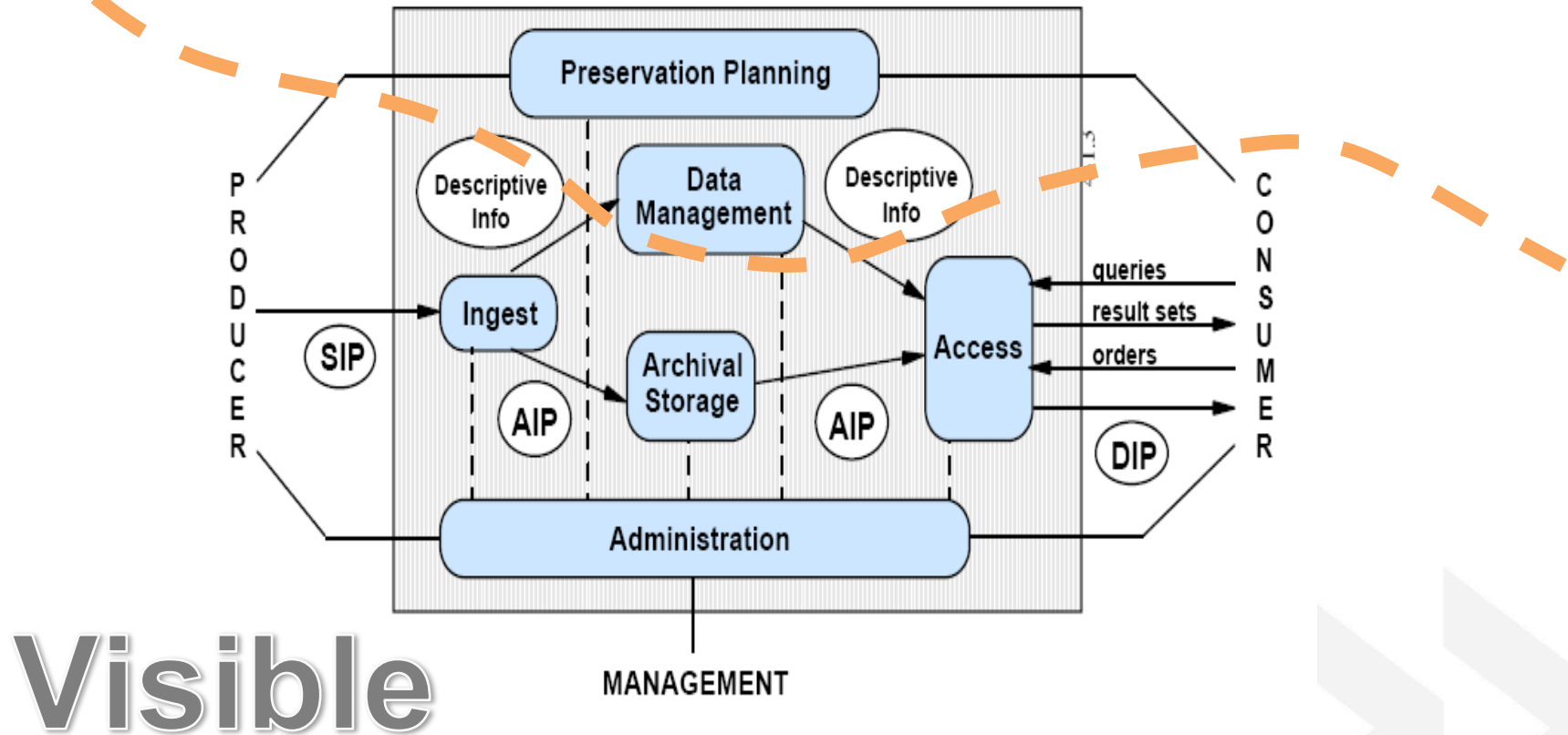
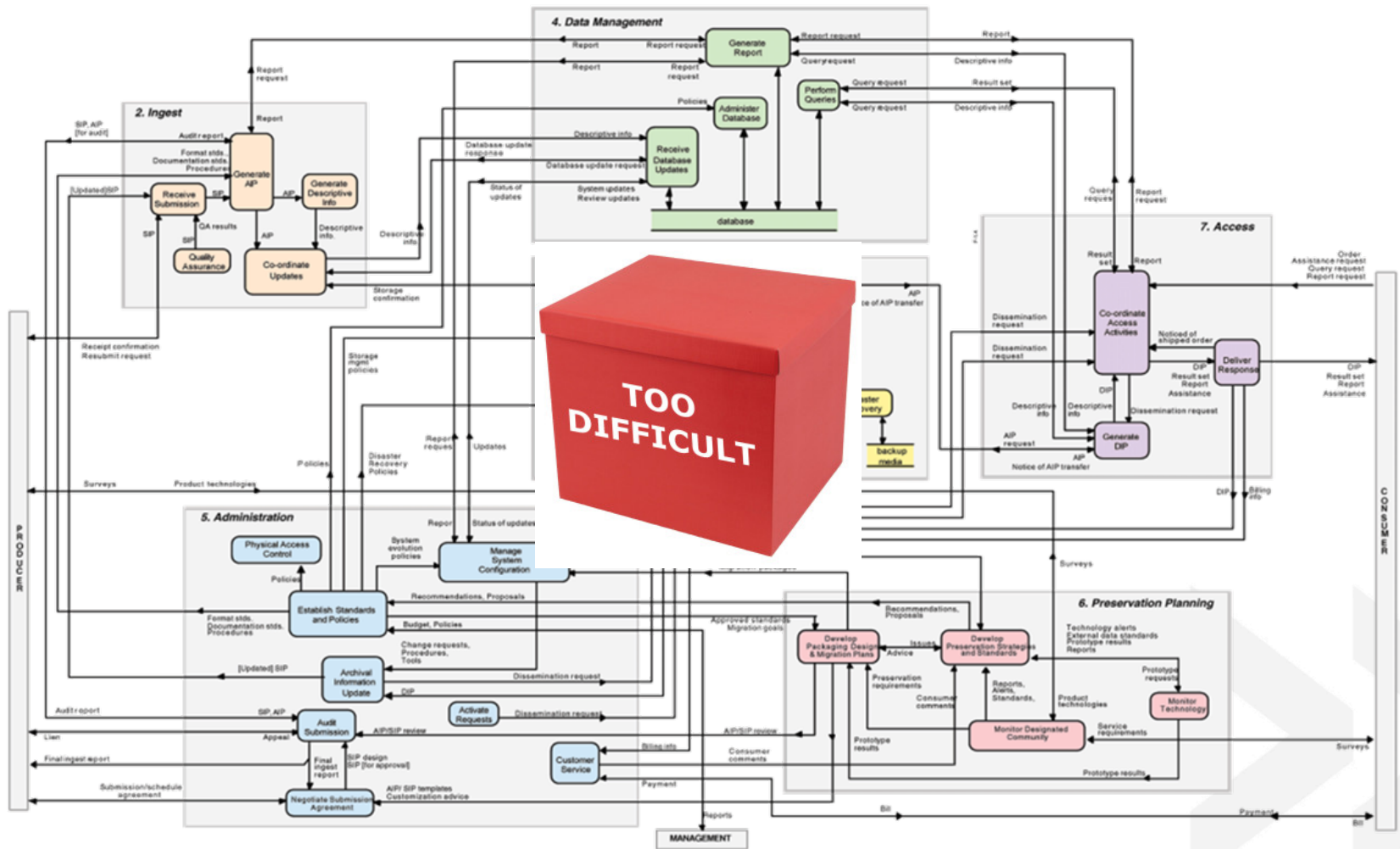


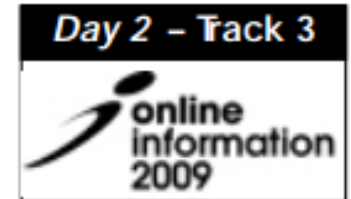
Figure 4-1: OAIS Functional Entities



# ...but we do need a pragmatic approach

Parsimonious preservation:  
preventing pointless processes!

(The small simple steps that take digital  
preservation a long way forward)



**Tim Gollins**

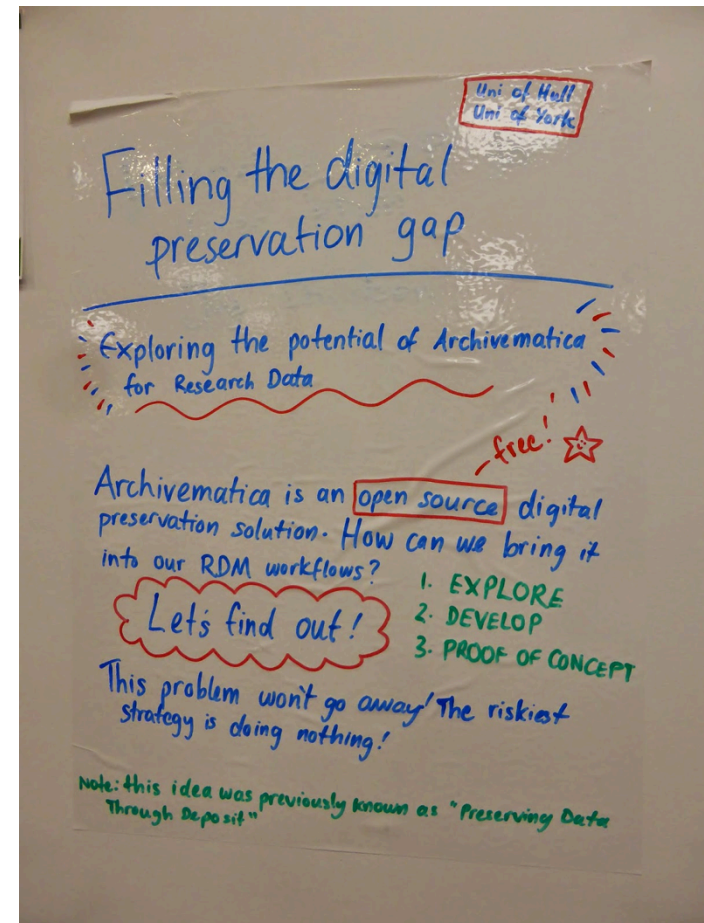
*Head of Digital Preservation, The National Archives, UK*

## Abstract

While there are many and varied threats to the successful curation of digital material, the impression given by the current generation of digital preservation systems and by much of the "received wisdom" in the digital preservation community is that imminent technological (software/data format) obsolescence is the primary threat. This gives rise to the belief that the only way to successfully start doing digital preservation is to invest in a technically complex, expensive, and difficult to operate integrated digital preservation system. This paper argues that, while the threat of technological obsolescence is real in some particular cases, a much more imminent threat is poor capture and inability to achieve safe and secure storage of the original material. By applying the principle of parsimony to digital preservation, institutions can find ways forward that are incremental, manageable and affordable, and which achieve the goal of securing our digital heritage for the next generation.

# Filling the digital preservation gap: Project aim

“...to investigate **Archivematica** and explore how it might be used to provide digital preservation functionality within a wider infrastructure for Research Data Management.”



# Project structure



- Phase 1 – **explore**: testing, research, thinking -produce a report (3 months)



- Phase 2 – **develop**: make Archivemata better for RDM, plan implementation - report (4 months)



- Phase 3 – **implement**: set up proof of concepts at York and Hull and further investigation of file format problem (6 months)



# The team

## University of Hull:

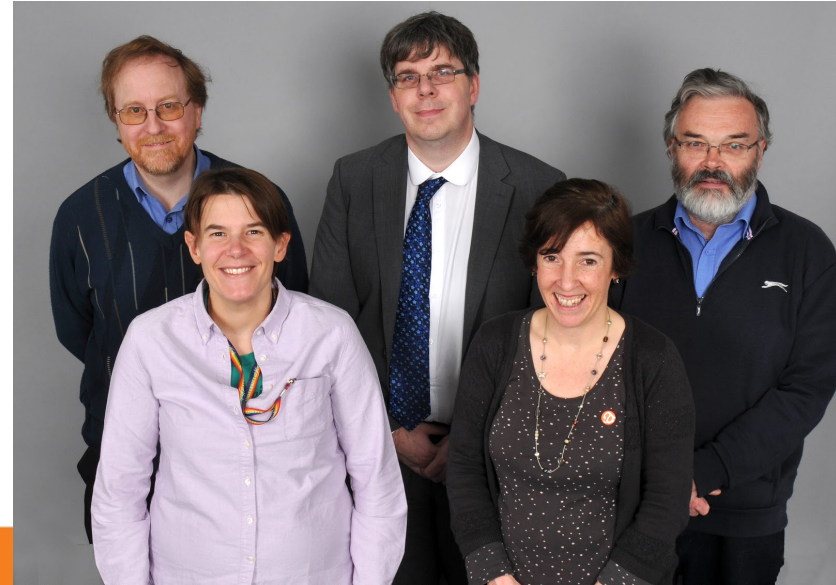
- Chris Awre – Head of Information Services, Library and Learning Innovation
- Richard Green – Independent Consultant
- Simon Wilson – University Archivist

## University of York:

- Julie Allinson – Manager, Digital York
- Jen Mitcham – Digital Archivist

## Artefactual Systems

**Funded by Jisc  
(Research Data Spring)**



# Why would we recommend Archivematica for RDM?

- It is flexible and can be configured in different ways for different institutional needs and workflows
- It allows many of the tasks around digital preservation to be carried out in an automated fashion
- It can be used alongside other existing systems as part of a wider workflow for research data
- It is a good digital preservation solution for those with limited resources
- It is an evolving solution that is continually driven and enhanced by and for the digital preservation community
- It gives institutions greater confidence that they will be able to continue to provide access to usable copies of research data over time

# The other side of the coin..

- It isn't a magic bullet
- There is no guarantee your data will be readable in the future
- It can only be as good as current digital preservation practice
- It can be fiddly to install correctly
- The GUI isn't that intuitive
- You need staff who understand it

UNIVERSITY of York  
  
UNIVERSITY OF Hull


**Filling the Digital Preservation Gap**

*A Jisc Research Data Spring project*

*Phase One report - July 2015*

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

# How have we improved Archivematica?

1. Enabled better workflows for RDM (producing a DIP on request)
  2. Allowing the DIP (access copy of data) to be more usable by different repository systems
  3. Helping reduce bottlenecks for big data (through choice of checksum algorithm)
  4. Workflows for unidentified files
  5. Enabling easier querying of data within Archivematica by third party applications
  6. Better documentation
- 

# Impact

Not all of the work we have sponsored is 'visual' but much of it is fundamental to the future development of Archivematica.

Our work has been enabling.

“The Jisc work has helped to modernise some of the internal infrastructure of Archivematica”

*Sarah Romkey, Artefactual Systems, 8th December 2015*



# 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.

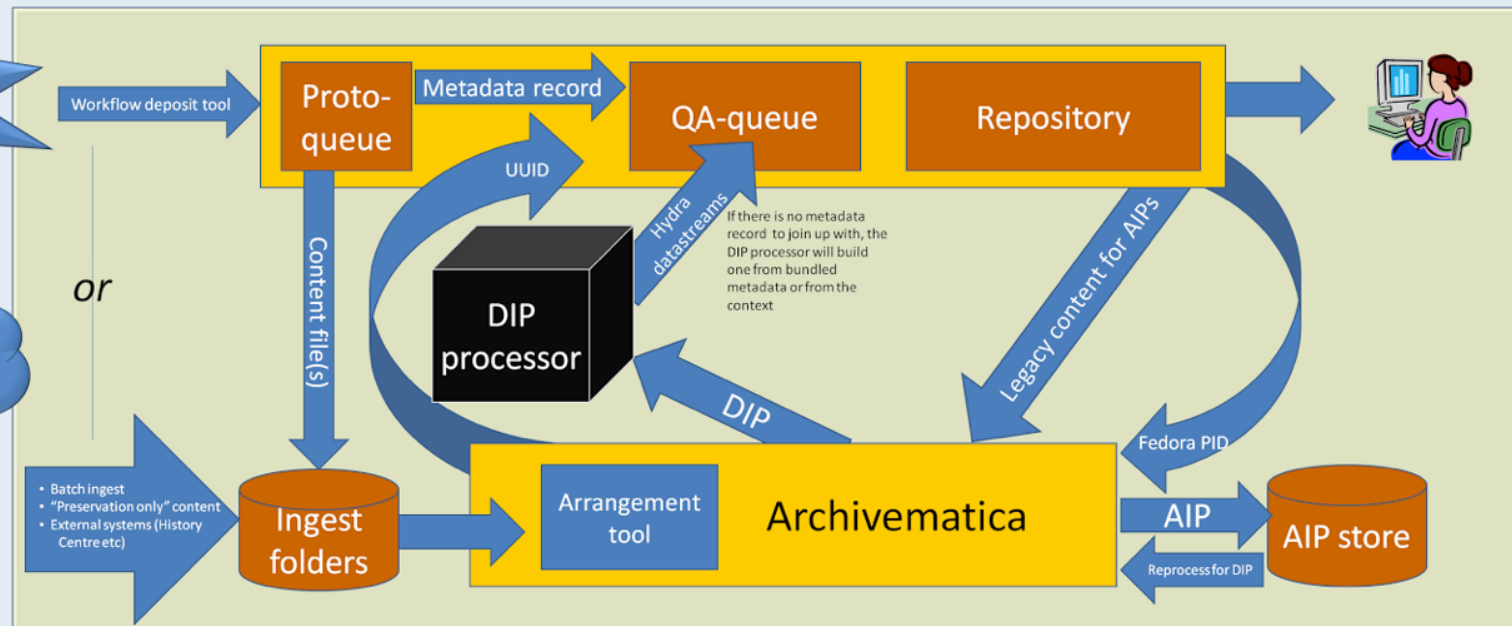
## What are we doing now?

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...



# Future plans

- In phase 3 of our project we intend to look more closely at the issue of unidentified files
- ...as well as creating our own proof of concepts of Archivematica at York and Hull
- York will also be working with Jisc as a pilot institution in their Shared Service initiative



# Where to find out more

The screenshot displays the Borthwick Institute for Archives website. The header includes the University of York logo and the text 'Borthwick Institute for Archives'. A navigation bar shows 'University | A to Z | Departments' and a breadcrumb trail '» Borthwick Institute for Archives'. A sidebar on the left lists various links like 'Borthwick Institute home', 'Our holdings', and 'Visiting us'. The main content area features a large orange banner for 'Filling the Digital Preservation Gap' with the subtitle 'Report on Archivemata for research data now available.' Below this, a 'Borthwick blog' section lists several articles, including 'Shedding new 'Lite' on Atkinson Brierley', 'The Sextoness of Goodramgate', 'Sledmere House - Rising from the Ashes', and 'Who came to see the Retreat? A look through the Retreat Visitors' Books'. A central white box highlights the 'Filling the Digital Preservation Gap' report, noting it is a 'Jisc Research Data Spring project' and a 'Phase One report - July 2015', authored by Jenny Mitcham, Chris Awre, Julie Allinson, Richard Green, and Simon Wilson. The right sidebar contains sections for 'ABOUT THIS BLOG', 'CONTRIBUTORS' (listing Nathan Williams, Julie Allinson, and Jenny Mitcham), 'BORTHWICK LOGO', 'SUBSCRIBE' (with a RSS feed icon), and 'TWITTER' (with a follow button for @Jenny\_Mitcham).

<http://www.york.ac.uk/borthwick/>

Do talk to me if you are interested in finding out more about this project

Useful links:

Project website: <http://www.york.ac.uk/borthwick/archivematica>

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

Archivematica: <https://www.archivematica.org/en/>

Phase 1 report <http://dx.doi.org/10.6084/m9.figshare.1481170>

Phase 2 report <https://dx.doi.org/10.6084/m9.figshare.2073220>

