Integration of archiving into RDM infrastructure

Matthew Addis
Arkivum
IDCC 2015
Overview

• Intro
• EPrints/Arkivum demo
• Archivematica/Arkivum demo
• Summary

All about what we’ve done, what we found in practice
Arkivum in less than 60 seconds

- SLA with 100% data integrity guaranteed
- World-wide professional indemnity insurance
- Long term contracts for enterprise data archiving
- Fully automated and managed solution
- Audited and certified to ISO27001
- Data escrow, exit plan, no lock-in
Example Customers

- FCA
- TATE
- OXFORD FERTILITY UNIT
- RGB Building Supplies
- NEWVOICEMEDIA
- Neilson Financial Services
- PHLEX The TMF Experts
- OXFORD Molecular Diagnostics Centre
- The University of York
- Queen Mary University of London
- Queen's University Belfast
- UNIVERSITY OF LEEDS
- UCL
- University of East London
- UNIVERSITY OF SOUTHAMPTON
- University of Hertfordshire
- Loughborough University
- Aston University
Data Archiving - Integrations

- Archimatica
- eprints
- figshare
- DSpace
- Symplectic Elements
Three common themes

• Ensuring the right stuff makes it into the archive
  – Complete, correct, quality, chain of custody

• Managed access to archived data
  – User experience, monitoring and control

• Simple to use for researchers
  – Easy to use, one place to go, automation
Archiving

Long-term safety and security
- Data reuse or compliance
- Data is immutable, replicated, managed
- Auditable integrity and authenticity

Easy access
- Quick to retrieve data when it's needed
- Doesn't have to be instant

Low cost
- Way to cope with data growth
- Free up expensive resources
- Eliminate the ‘cost of loss’
http://datablog.is.ed.ac.uk/2013/12/06/the-four-quadrants-of-research-data-curation-systems/
EPrints demo

Institution

Web Browser

http

EPrints Server

CIFS

REST API

A-Stor appliance

Astor Data Centre

Escrow (simulated)

Arkivum Service

Local files
The diagram illustrates the four quadrants of research data curation systems:

- **PRIVATE**
  - **CRIS**
  - **Data Vault**
  - **Deposit files**
  - **Link**
  - **Copy**

- **PUBLIC**
  - **Data Asset Register**
  - **Data Repository**
  - **Deposit files**
  - **Link**
  - **Deposit**

**SYNC METADATA**

**METADATA**

The diagram shows the flow of deposit files and links between CRIS, Data Asset Register, Data Vault, and Data Repository, indicating the processes involved in curation systems.
Role of the CRIS or IR

- Gateway active ↔ archive
- Single place to go
- Management of metadata
- Quality control
- Decision points and workflows
  - Effective use of resources
  - Review and approval
  - Budget and charging
  - Ethics and compliance
EPrints: deposit into the archive

Editor

Review

Approve

Researcher

Metadata

Files

Files

Files safe

Delete originals

Researcher files

EPrints

EPrints Storage

Arkivum Appliance

Arkivum Service

Clear cache

Clear cache

Files

Files

Files safe

Files safe
Putting data into the archive: chain of custody

- Send data to archive
- Generate checksum for data in archive
- Compare with checksum in IR or CRIS metadata
- Update metadata that transfer is successful
- Apply archive safety mechanisms
  - Replication, WORM, access control
- Update metadata to complete the audit trail
- Remove local copy (if desired)
EPrints: getting stuff out

1. Request data
2. Wait
3. Review request
4. Approve
5. Restore data
6. Ready
7. Get data
8. Ready
9. Get data
10. Cache/delete data
Getting the data out of the archive:

managed access

• Open access: no license, no barriers
• Unrestricted use, but request for access
• Unrestricted use, but charge for access
• Restricted use: request/approve access
• Embargoed: no access for set period
• Locked down: no access

“It may be reasonable for research organisations to require persons requesting access to specific research datasets to do any of the following before granting access: register; create an account; prove identity; accept terms and conditions of use of the data.”
Archivematica

• Preparing content for preservation and access
• Automatic deposit into the archive
• End-to-end chain of custody
Archivematica demo

Institution

Web Browser

http

Archivematica Server

CIFS

REST API

A-Stor appliance

Escrow (simulated)

Astor service Data Centre

Arkivum Service

Local files

sftp client
Archivematica

Admin
- Review
- Approve

Archivematica Pipeline
- Deposit files
- Temporary storage

AtoM
- DIP
- AIP

Arkivum Preservation Storage
- Public
- Retrieve file

Researcher
Summary

• Managed deposit
  – Check data and metadata is complete & correct
  – Chain of custody

• Managed access
  – Help user understand what’s happening
  – Monitoring and control

• Simple to use
  – Easy to use, one place to go
  – Automation plus checkpoints
Questions

matthew.addis@arkivum.com