What is a Data Vault?

Mary McDerby, Tom Higgins
Research IT and Library, University of Manchester
Overview

• Research Data Spring Project
• Project Team
• The Problem
• What is Data Vault
• Phased approach and progress
• Demonstration
• Summary
Project Team

• Project members:
  • University of Edinburgh
  • University of Manchester
  • Other active contributors welcome

• Project leads on both sites (institutional contribution)

• Developers x 1 per partner, half funded by JISC, half locally funded

• Plus each sites RDM Teams, Domain Architects and metadata experts
Research Data Management Services

Data Management Support

- Data Management Planning
- Active Data Infrastructure
- Data Stewardship
Research Data Management Services

Data Management Support

Data Management Planning

Active Data Infrastructure

Data Stewardship
Problem Summary

• Active data storage being used for everything
• No interface to archiving (good RDM practice)
• Archiving undertaken on unsustainable/non-compliant storage solutions (DVDs / USB drives)
• No consideration for curation and archiving, data description, retention and review
• Items of long term value not actively identified and realised
Active Data Storage Problem

![Bar chart showing number of shares and static usage for different categories: FLS, EPS, MHS, HUM.](image-url)
Data Stewardship

• DataVault
  • Long term archival storage
  • Write once file-system
  • Secure storage
  • Policy around retention and review
  • Assurance
  • Auto-versioning
  • First envisaged a few years ago…
What is the DataVault - Analogies

https://www.flickr.com/photos/timshelyn/4125480767
What is the DataVault - Analogies

https://www.flickr.com/photos/brookward/8457736952
What is the DataVault?

• It’s a platform (a bit of web-based software), not a storage system
  • You still need an archival storage system

• It’s doesn’t undertake any active preservation activities
  • It does collection metadata and package the data for long-term storage
Where does it sit?

- Active Storage
- Lab Equipment
- Other Media
- Archival Storage
Solution
Information Architecture

Users → Vaults → Deposits → Bag → Archival Storage

Use cases

• A paper has been published, and according to the research funder’s rules, the data underlying the paper must be made available upon. It is therefore important to store a date-stamped golden-copy of the data associated with the paper. Even if the author’s own copy of the data is subsequently modified, the data at the point of publication is still available.

• Data containing personal information, perhaps medical records, needs to be stored securely, however the data is ‘complete’ and unlikely to change, yet hasn’t reached the point where it should be deleted.

• Data analysis of a data set has been completed, and the research finished. The data may need to be accessed again, but is unlikely to change, so needn’t be stored in the researcher’s active data store. An example might be a set of completed crystallography analyses, which whilst still useful, will not need to be re-analysed.

• Data is subject to retention rules and must be kept securely for a given period of time, for example EPSRC funded data that needs to be stored securely for ten years.
Progress (Phase 1)

• Month 1
  • Project initiation
  • http://Libraryblogs.is.ed.ac.uk/jiscdatavault/
  • RDM User Group workshop – and requirements gathering and use cases
  • Investigation into storage systems, metadata, packaging formats

• Month 2
  • Platform architecture design (with domain architects and infrastructure support)
  • Designed APIs
  • User interface wireframes
  • Test cases

• Month 3
  • Hackathon and ongoing development of a proof of concept
  • Community Event - Research Data Spring 'Solutions for your needs' Lancaster

• July 2015 - Proof of concept delivered at the second JISC Research Data Spring sandpit
Progress (Phase 2)

• Months 4-7:
  • Critical Success Factor: Deliver a first version of a complete Data Vault
  • Implement remaining requirements
    • Authentication and Authorisation
    • Integration with more storage options
    • Management/monitoring interface
    • Example interface to CRIS (PURE)
    • Development of retention and review policy
    • Scalability testing
  • Two community events (England + Scotland)
DEMONSTRATION
Summary

• Research data needs a long term home
• The data vault platform is the only archive interface designed specifically with research data in mind
• Archiving infrastructure is maturing (e.g., Amazon Glacier/Arkivum/Local solutions etc) but there is still a requirement for a researcher oriented user interface
• Platform is extendible (storage and integration) and open source, and will be in use at Edinburgh and Manchester
• Github.com/datavault
Summary (2)

• Impact:
  • Encourage archiving and proper long term storage of data
  • This leads to better data management practice

• Benefits
  • Researchers
  • University HE Community
  • IT Services
Where to find our outputs

• Source code: https://github.com/DataVault
  • Broker / API
  • Web application

• Blog: http://libraryblogs.is.ed.ac.uk/jiscdatavault/
Acknowledgements

• The Data Vault Project Team (Edinburgh) – Stuart Lewis, Robin Taylor and Claire Knowles
• Microsoft – Kenji Takeda
• Arkivum – Matthew Addis, Chris Pates
• Everybody who attended our community events in Manchester and Edinburgh in October and November 2015!
Mary.mcderby@manchester.ac.uk
Thomas.higgins@manchester.ac.uk

@JISCDataVault
Our demo system (feedback welcome)
http://datavault-demo.cloudapp.net:8080/