The University of Edinburgh identified a gap for preservation of large research data sets that do not need active preservation at ingest as the majority of the content will not be retained beyond the initial retention period. The complexity of content types, size and quantity of data also means that active preservation is not currently achievable with the resources available. The initial retention periods are set by the research funders and are generally 5-10 years. The solution provides permanent dataset identifier DOI and data discoverability through linkage to PURE (University’s Current Research Information System (CRIS) as well as storage on the university’s secure storage platform. Three copies are kept with one being held in the Cloud.

Whilst developing the DataVault, changes to data protection legislation across the European Union impacted system development. All datasets are now encrypted at rest.

This demonstration of the DataVault application will show the deposit of research data into a vault and discuss the system design and decisions behind the application, including:

- Data encryption and its implications
- Deposit size and benchmarking
- Minimum metadata standards
- Review workflow when the data owner has left the institution
- Infrastructure decisions

Encryption – required for legal reasons

Speed of deposit – the current solution requires a long time to transfer the data to the 3 storage locations

Issues with BagIt – have been ongoing during project and now looking at not using it at file level

Chunking – how deposits are chunked impacts on speed and currently still testing best solution

The DataVault is unique in that it is a system for long term storage of large and smaller datasets. It has the potential to provide an interim solution prior to active preservation. It enables users to free up active data and securely store multiple copies of data.

It has the potential to be used not just for long-term retention of research data but also for other types of data which require:

- Appraisal
- Review period
- Secure storage with three copies
- Discoverability
Encryption

Initially, data within the DataVault was not going to be encrypted. However, whilst developing the DataVault changes to data protection legislation across the European Union impacted system development. All datasets are now encrypted at rest.

Encryption is currently provided through an interim solution using Galois/Counter Mode (GCM) with the Bouncy Castle API. This decision was made as a short-term solution as the institution has identified the need to have a key management system that can be used by a wide range of applications. Once the development of this has been completed this solution will be applied to the DataVault.

Why use DataVault?

DataVault is designed to provide a long-term, low-cost immutable storage solution for research data, which is no longer active and not intended for publication.

The service provides:
- Secure data storage on the University's archival storage platform
- Full encryption for sensitive data
- Restricted access (you or nominated individuals only)
- Permanent dataset identifier DOI
- Data discoverability through linkage to a Pure record

The service enables researchers to:
- Comply with funder and University requirements to preserve research data for the long-term
- Confidently store their data for retrieval at a future date
- Designed for large datasets up to 10TB whilst also fulfilling the requirements of researchers with smaller datasets

Design Decisions

- DataVault can be used for large and small deposits and has been tested currently up to 2TB
- Testing will continue up to 10TB
- BagIt used - but issues with implementation
- Chunking of data
Once your data files are ready, the following steps are required:

- Create a record describing your dataset (i.e., your project) in Pure.
- Login to DataVault with EASE and create a Vault, to group together all the data corresponding to the dataset.
- Enter the details DataVault needs to connect to where your files are located on DataStore.
- Deposit data into the Vault you have created.

Retrieving data:

- If you are the Owner of a vault or one of the nominated Data Managers (not yet implemented), then you should be able to find the vault and retrieve any or all deposits in it through the website using your EASE login.
- A location will need to be identified where sufficient storage capacity is available.

Costs:

- There are charges for depositing data in the DataVault. They are applied proportionally to the volume of data and the period of storage.
- Resource must be in place before a deposit is made. In most cases, the Principal Investigator will be required to provide a grant ID against which the team will invoice for usage after deposits are made.
Demonstration

Review of DataVaults

All vaults have a review date. When a vault's review date is approaching, the vault owner will be notified and given the option to:

• delete any of the deposits
• extend the review date further into the future (and provide any associated extra resource to fund the continued storage of the deposits not tagged for deletion)

Should the vault owner no longer be an active member of staff, the nominated Data Manager will be asked to make these decisions. In the absence of any nominated Data Manager, the School Data Manager will be asked to make these decisions.

No data will be deleted without the agreement of at least two actors in the review process. The College and the University archiving service are also part of the chain of custody. In cases where the owner is no longer at the University, and where the School Data Manager has not provided resource for data to be kept, then College officers and/or the archiving service will be given the opportunity to input to the decision.

To ensure a valid and appropriate outcome to this review process, and to any requests for access to the data, it is vital that owners and depositors ensure they provide detailed information at the time of deposit about:

• any reasons for deleting data sooner than the funder’s default minimum retention period e.g. privacy of subjects or commercial sensitivity of the data;
• any reasons why the data should be kept longer than the funder’s default minimum retention period e.g. particular significance of the work or long-term potential for re-use.

Appraisal

The review process within DataVault offers the following from an archival perspective:

• Secure place to hold data for defined period of time
• Gives archivist’s time to review/appraise content where the archival value cannot be determined immediately
• Where file format identifiers(puid) are not yet available it provides time to review and archive at a later date when identification can take place
• Deposit of data/material which the archivist has not had time to review or come to decisions around retention
• Potential to store copies of software and other associated material with the data to facilitate review and decision making at a later date