Developing a Data Vault

Stuart Lewis
University of Edinburgh

Lorraine Beard
University of Manchester

Mary McDerby
University of Manchester

Robin Taylor
University of Edinburgh

Thomas Higgins
University of Manchester

Claire Knowles
University of Edinburgh

Abstract

Research Data is being generated at an ever-increasing rate. This brings challenges in how to store, analyse, and care for the data. A component of this problem is the stewardship of data and associated files that need a safe and secure home for the medium to long-term.

As part of typical suites of Research Data Management services, researchers are provided with large allocations of 'active data store'. This is often stored on expensive and fast disks to enable efficient transfer and working with large amounts of data. However, over time this active data store fills up, and researchers need a facility to move older but still valuable data to cheaper storage for long-term care. In addition, research funders are increasingly requiring data to be stored in forms that allow it to be described and retrieved in the future. For data that can't be shared publicly in an open repository, a closed solution is required which can make use of off-line or near-line storage for cost efficiency.

This paper describes a solution to these requirements, called the Data Vault. Further details about the system can be found at http://datavaultplatform.org/, and the software downloaded from https://github.com/DataVault/datavault.

Background

At the end of the research process when the data associated with an investigation is complete, there are currently four main options that can be taken:

- **Share the data:** Where possible, some of the data may be shared openly online through a data repository. However often some data may not be shared due to sensitivity reasons, and sometimes it is only the final curated dataset and supporting code plus documentation that is stored in a repository, not all of the other useful files that are associated with the data;

- **Do nothing:** Often the data is left on the storage device used during the investigation, and left there 'just in case'. This is a sub-standard solution, as the data is not described or recorded, and is using expensive online storage;
- **Take a manual backup:** It is quite common for researchers to take manual backups of their data, using devices such as USB hard drives, or DVDs. Whilst this can free-up the storage, the data isn't necessarily recorded or described (other than a handwritten note on the DVD), and isn't routinely checked for media degradation;
- **Delete the data:** If data storage space becomes a problem for future investigations, sometimes data is simply deleted. No record of this action is kept, and there is no ability to retrieve the data in the future.

Across the world, research funders are introducing requirements for better research data management practices, including the long-term storage of completed research. Typically these specify a number of years that the data should be retained for, or in some cases, such as the UK's Engineering and Physical Science Research Council (EPSRC) Research Data Expectations[1] the requirement is to store data for ten years from last date on which access to the data was requested by a third party, which introduces a requirement not just to store data, but also to track its usage.

Of the four options above, other than the first, these are not ideal practices. Data is either lost, at risk, and is usually not described or recorded anywhere. During conversations between the Universities of Edinburgh and Manchester in 2014, it became apparent that both institutions had identified the need for long-term, efficient, and economic long-term storage of data that could not be made open, either due to security or sensitivity issues, or because there was more that needed to be stored than just the final shareable dataset. The notion of the Data Vault was developed during these conversations.

The word 'vault' was used for a number of reasons. Firstly the term archive is already used, and can have many different meanings. Secondly, the project often draws an analogy between the Data Vault system to that of a bank vault. Analogous factors include:

- To put objects into a bank vault, you need to register with the bank, and provide some information about what you are depositing (Metadata collection).
- The bank keeps a record of what is stored in the bank vault (A Data Catalogue is created).
- The bank records deposits into the vault, and subsequent accesses to the vault (Deposit and Retrieval Audit Logs).
- What is put into the bank vault is not changed by the bank. For example an antique wedding ring will not be converted into a more modern design ring (No format migration - data is left as-deposited).

[1] EPSRC Research Data Expectations: https://www.epsrc.ac.uk/about/standards/researchdata/expectations/

- By placing something into a bank vault, you are handing over responsibility for its care to the bank. By handing over data to the institution, it can fulfil its responsibility (to funders) to preserve the data.
- To retrieve an item from the bank vault, an appointment must be made with the bank. Therefore access is not immediate (Data stored on tape does not provide instant access).

Project

At the end of 2014 the UK's Jisc initiated a new programme called 'Research Data Spring'[2], part of the wider Research Data at Risk stream of work[3] with the aim of 'Realising a robust and sustainable research data management infrastructure and services to enrich UK research'.

The first stage of this programme was to make use of the IdeaScale tool to allow proposals to be made that address issues in research data management[4]. The idea of developing a Data Vault was proposed. Following the proposals of ideas, the UK community was then able to vote on these, and those projects that received a large number of votes were invited to develop the ideas further and pitch them in a 'dragons den' style event in February 2015. The Data Vault project received funding, and the project was initiated.

The first round of funding under this programme was for three months. During this time staff from the Universities of Edinburgh and Manchester (Library, IT, RDM teams) scoped the project, made initial designs with input from RDM, IT infrastructure, and Library teams, and developed a working prototype of a Data Vault.

Projects were then able to pitch for further funding for four months in June, and then for six months in December. At each programme meeting, the projects had to present their updates and future plans to a panel of invited experts, who were then able to make future funding decisions. The project is now in Phase three, which lasts from February 2016 until July 2016.

Elements of the project have not only included the development of the Data Vault, but also a website (http://datavaultplatform.org/), a logo, and community engagement events to engage the potential user community of the Data Vault to ensure that it met not only the requirements of the founding institutions, but also of the whole community.

[2] Jisc Research Data Spring: https://www.jisc.ac.uk/rd/projects/research-data-spring
[3] Jisc Research at Risk: https://www.jisc.ac.uk/rd/projects/research-at-risk
[4] Research Data at Risk IdeaScale (archive): http://web.archive.org/web/20150123031008/http://researchatrisk.ideascale.com/

Community Engagement

It is recognised that whilst this product is being developed to address specific needs at the universities of Manchester and Edinburgh, it is also sensible and right to develop a product that can be reused, and possibly further developed, by other parties. To that end, community engagement has been an integral part of the development process.

As the development was funded as part of the Jisc Research Data Spring initiative, the project team were required to attend and participate in a number of related events that provided the first opportunity to expose the project plan and early product development to an audience of contemporaries. This included other universities, and some commercial companies who seek to integrate with the Data Vault, either as an archival storage provider, or using the API to directly store data in the vault.

In addition, two community engagement events were organised by the team during 2015, one at the University of Manchester in October, and one at the University of Edinburgh in November. The main aim of these events was to share progress and gather feedback to inform the ongoing development. The attendee list at both events was a mixture of researchers, technical staff, and other interested parties from both local institutions, as well as UK wide Universities, national organisations and commercial enterprise. At each event the product was demonstrated, and the attendees were able to test it themselves. Later the attendees split into breakout groups to discuss particular aspects of the product e.g. technical design, and metadata collection.

As the product reaches the later stages of development we intend to undertake further community events, but the focus will change to be about publicising the product and encouraging reuse, both by the user and development communities. It should be noted that good practice has been followed throughout with regards to Open Source development in order to facilitate the building of a community of developers.

Use cases

The initial use cases for the Data Vault were gathered at the very start of the Data Vault project:

- 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.

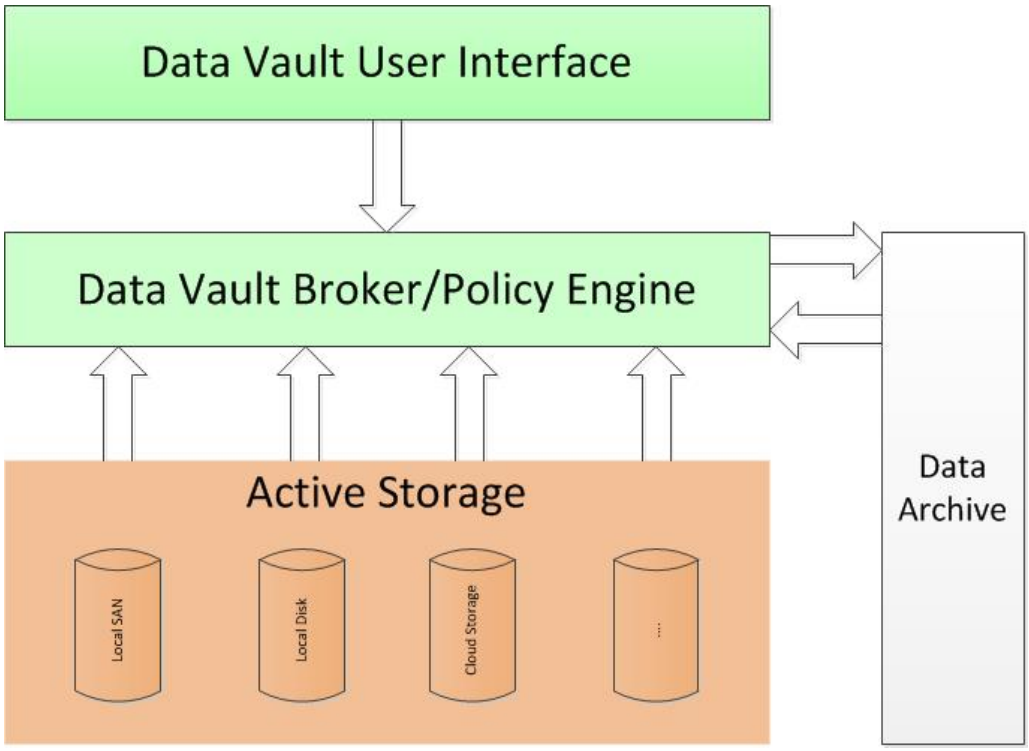
The community engagement events helped to build upon these initial use cases, but the main message from the community was that it was important to provide a service like Data Vault at this stage in the landscape. Researchers may be encouraged to just 'dump' all their data into archive (short term step). But in the long term this may build up some problems. The following outlines some of the main use cases outlined by the user community:

- Big data is being created at core facilities within institutions. The data is processed and it is the processed outputs which are shared with the researchers involved in the investigation. The core facilities need to archive the primary data so that they can repeat the analysis if need be, to satisfy funders' requirements but also the data can be used to undertake comparative analysis investigations as new algorithms come along.
- It should be possible to link the Data Vault with an institutional Data Catalogue or Repository. Such a link would be used to provide information around the re-use of the published data and reset the retention clock. Thereby enabling the services/institutions to understand the long term value of what is in the archive.
- If data that is currently held in an existing repository needs to be retained it should be possible to move it to the Data Vault for long term storage and management.
- To be able to curate and manage all other research outputs that are not shareable and publishable and in doing so providing a secure home which is linked to the published outputs and provides the long term care and retention required.
- The ownership of the research outputs should be handed over to the institution in archiving outputs so that support for the long term care, avoidance of orphaned data (for example staff leave, research centres shut down etc) and retention can be undertaken at the support service level (big issue with long term retention - nobody knows about the data!).
- The Data Vault should enable long term access to research outputs that are archived for current researchers (whom are data owners), data managers, funders and researchers who have left institutions.
- Some data may be highly sensitive whilst other data in the same collection may not be restricted at all. The Data Vault approach should be aware of such requirements and enable the silo-ing (also known as data pooling) of such data to enable restricted access, as well as appropriate metadata to support the need. Encryption would also be required at point of archive.

Technologies

Data Vault is a web-based software platform which allows researchers to create vaults and deposit data within them. Metadata is captured during this process to aid future

discovery, management, and retention decisions to be made. Researchers can deposit data into a vault from a variety of sources such as their institutional active file store or from cloud storage services.



The platform is organised around a 'broker' web service which handles requests to create and manage vaults and deposits. The broker coordinates the actions of 'workers' that undertake the tasks of adding and retrieving data from the long-term storage. The broker can coordinate multiple workers, depending upon the resources available to undertake the copying, processing, and packaging of data. The broker communicates with the workers using a message queue (RabbitMQ) so that actions can be queued for when a worker next becomes available. This mechanism allows the system to scale as required, by adding more workers across one or more machines. The archiving and retrieval of data is therefore undertaken asynchronously and does not require active participation from the researcher once initiated.

Workers are configured to connect to different active data storage and long-term storage systems using a plug-in system. This allows the platform to integrate with numerous data storage systems which may be either locally attached, or use different technologies (tape, disk, HSM), or be hosted in the cloud (for example using Amazon Glacier).

The broker receives requests to undertake actions via a RESTful API. A default customisable web user interface is available with the code, allowing for user management and administration of the system as well as archiving functions. However other systems can be built to interact with the system, and to send requests to archive or retrieve data. Examples could include electronic lab notebooks, or laboratory equipment that may wish to archive all raw data as it is generated.

Example API usage:

HTTP POST to /vaults	Creates a vault with associated metadata. Generates a new vault ID "123" and returns the ID to the caller.
HTTP POST to /vaults/123/deposits	Creates a deposit into the new vault. The broker will instruct a worker to ingest data from the specified path.

When processing a deposit request the relevant files are copied from user's active file store to a working area where some basic preservation activities are undertaken, such as file format identification (Apache Tika) and the creation of fixity checksums in md5 or sha1 format. The data is packaged along with associated metadata using the bagit format. The bagged data is then bundled into a TAR (Tape ARchive) file and copied to the long-term storage system. Once this activity is completed the user is notified, allowing them to delete the original copy of the data. If in the future the user wishes to access the data, the platform allows for a retrieval request to be made. This process will restore a copy of the data from the long-term storage system, check the validity of the package using the fixity values provided by bagit, and copy it back to the user's active file store.

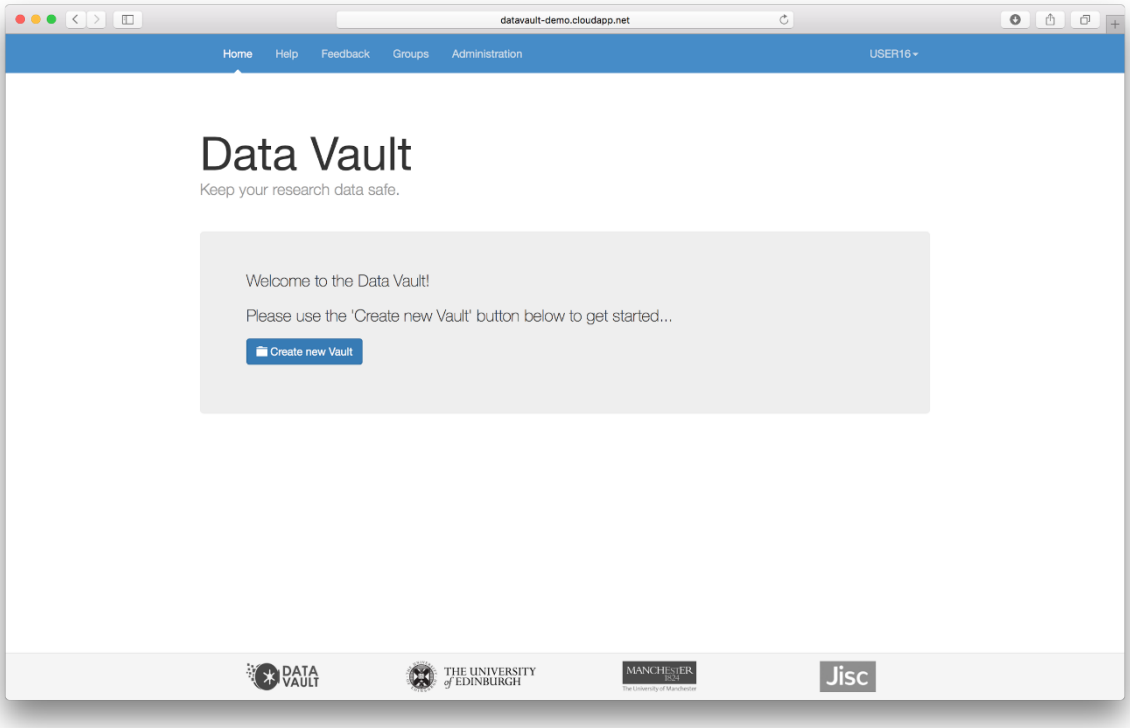
When collecting metadata the system can draw upon existing sources of information such as a current research information system (CRIS) in order to help populate some metadata fields automatically. For example, a default retention period of ten years could be suggested if the data was created as a result of funding from EPSRC. A retention policy causes data to be flagged for review, and possible deletion, once the time since last access passes the specified threshold. This last access date is calculated using the date of deposit, and the dates of any subsequent data retrieval requests, all of which are logged automatically.

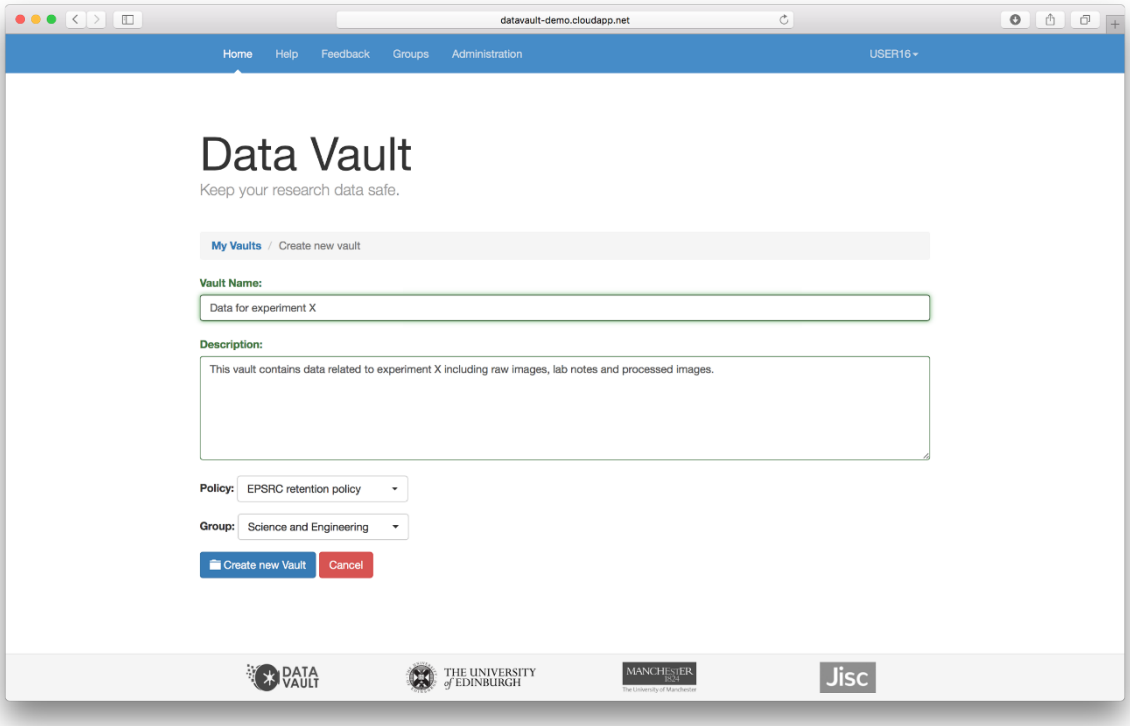
Next steps

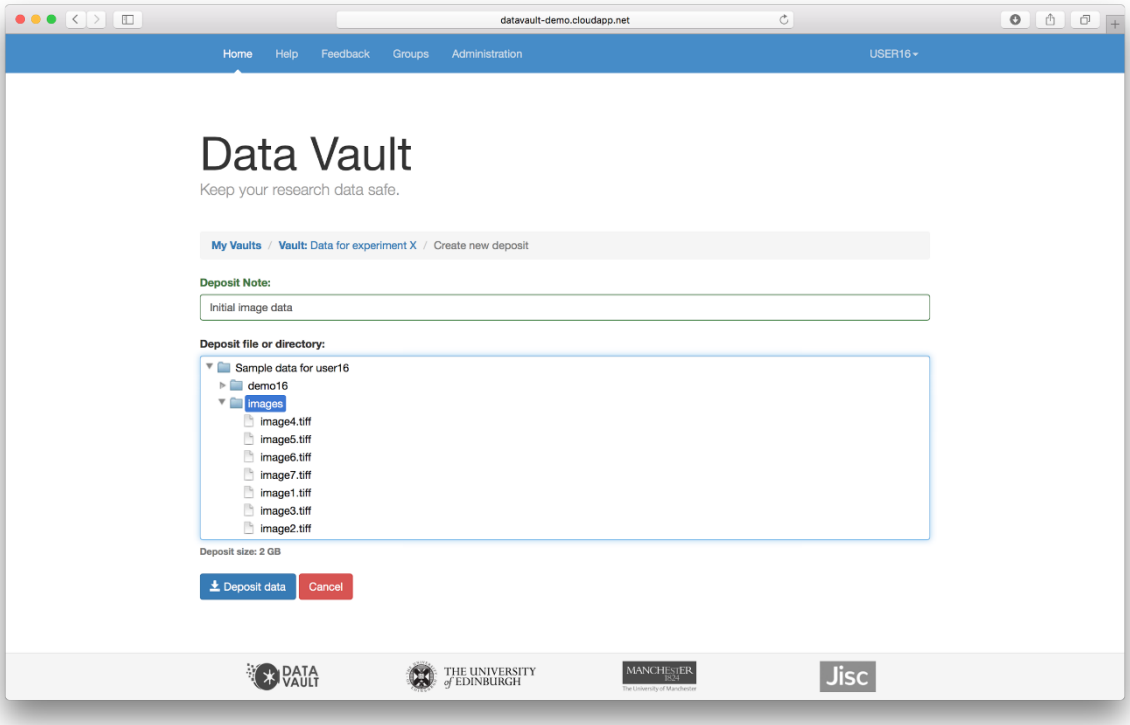
As mentioned, the Data Vault project has been funded for a third phase under the Jisc Research Data Spring programme. The aim is to develop and deliver the first full version of the system by the second quarter of 2016. Further community events will be held thereafter to demonstrate the system as well as train any interested parties in implementing the system and using it.

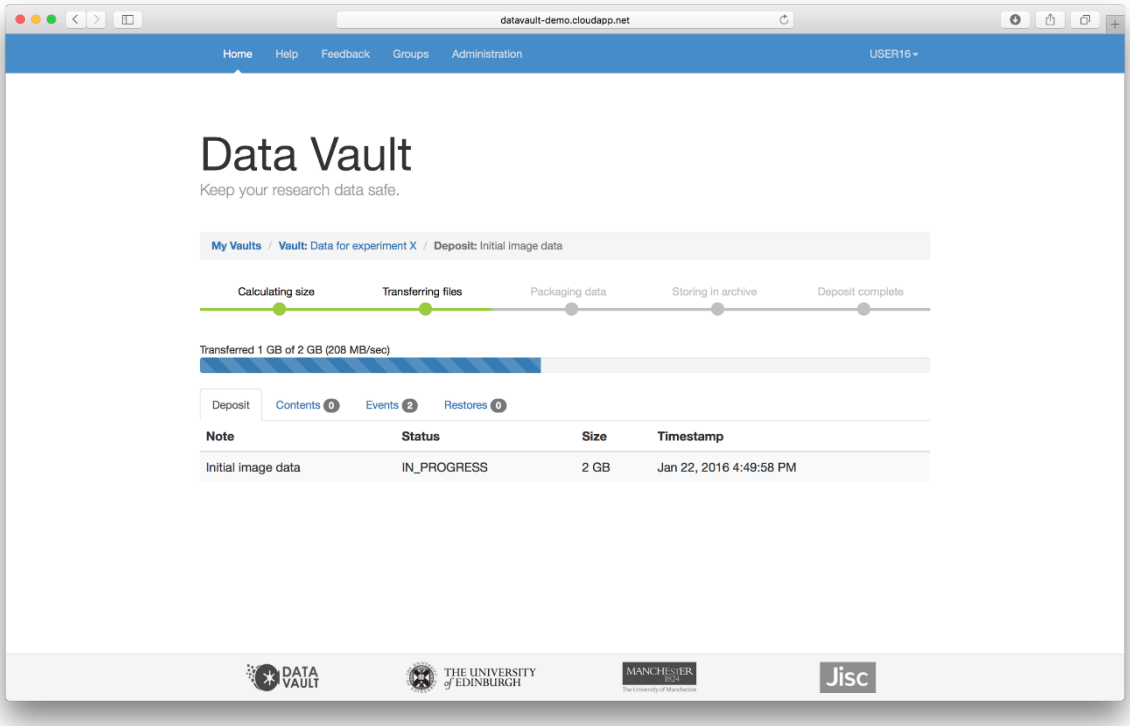
The Data Vault platform will help in playing an essential part in the long-term curation of Research Data, and will be particularly useful in cases where perhaps the data is not suitable for sharing. The management of this data, along with relevant metadata and retention policies, will help to ensure this data remains safely stored, and reviewed when appropriate.

Example deposit workflow






datavault-demo.cloudapp.net
Home
Help
Feedback
Groups
Administration
USER16
Data Vault
Keep your research data safe.
My Vaults / Vault: Data for experiment X / Create new deposit
Deposit Note:
Initial image data
Deposit file or directory:
Sample data for user16
demo16
images
image4.tiff
image5.tiff
image6.tiff
image7.tiff
image1.tiff
image3.tiff
image2.tiff
Deposit size: 2 GB
Deposit data
Cancel
DATA VAULT
THE UNIVERSITY of EDINBURGH
MANCHESTER
Jisc


datavault-demo.cloudapp.net
Home
Help
Feedback
Groups
Administration
USER16
Data Vault
Keep your research data safe.
My Vaults / Vault: Data for experiment X / Deposit: Initial image data
Calculating size
Transferring files
Packaging data
Storing in archive
Deposit complete
Transferred 1 GB of 2 GB (208 MB/sec)
Deposit
Contents 0
Events 2
Restores 0
Note
Status
Size
Timestamp
Initial image data
IN_PROGRESS
2 GB
Jan 22, 2016 4:49:58 PM
DATA VAULT
THE UNIVERSITY of EDINBURGH
MANCHESTER
Jisc