DATA VAULT

Stuart Lewis

Deputy Director, Library & University Collections
The University of Edinburgh

Robin Taylor

Library Software Developer
The University of Edinburgh

@JiscDataVault

DATA VAULT

THE UNIVERSITY of EDINBURGH

MANCHESTER 1824
The University of Manchester

Jisc

Introducing the DataVault

- Practice paper:

1. What is the DataVault
2. Use cases
3. How does it work
4. Demo
5. What do you think?

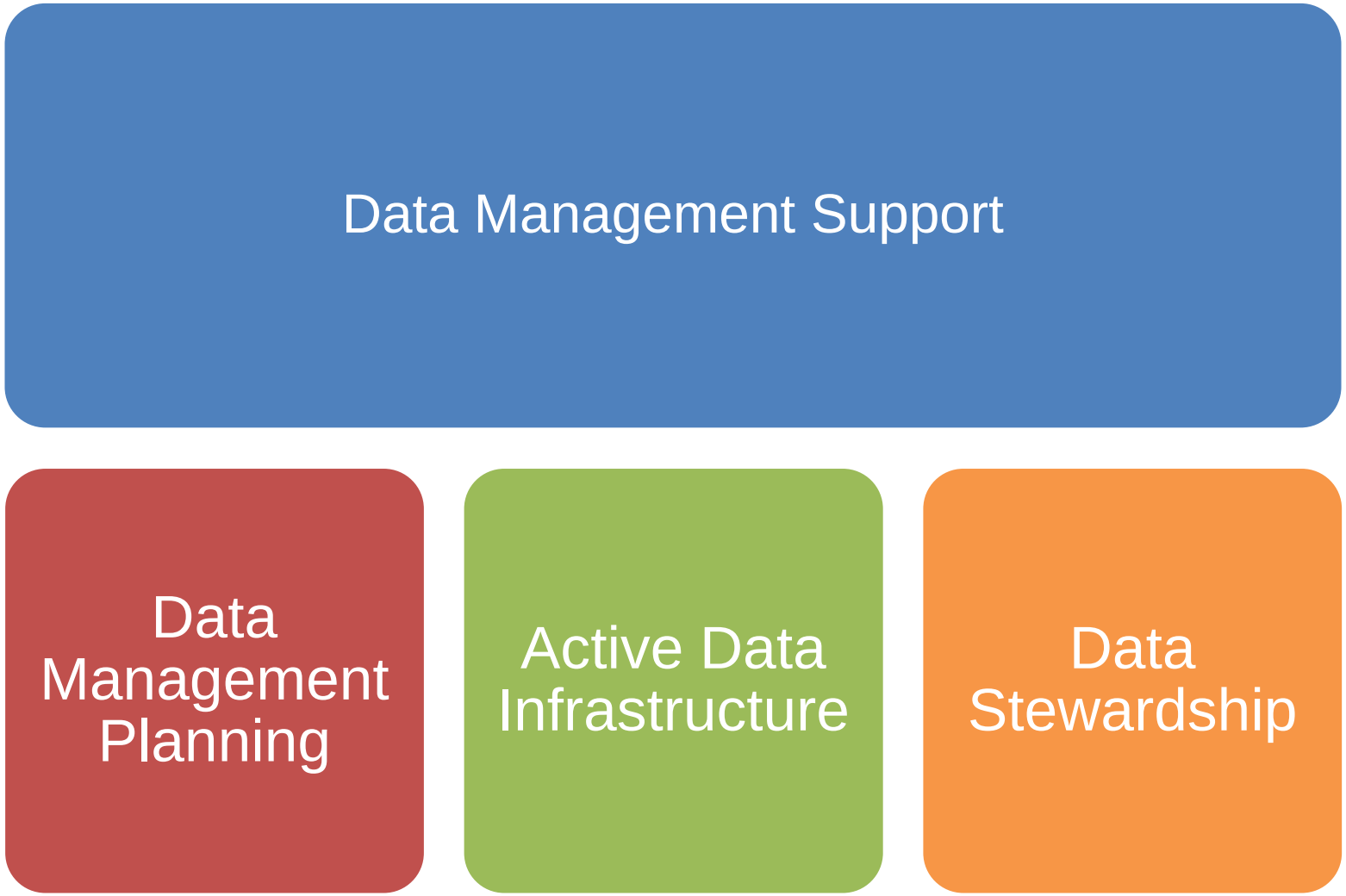
DATA VAULT

THE UNIVERSITY of EDINBURGH

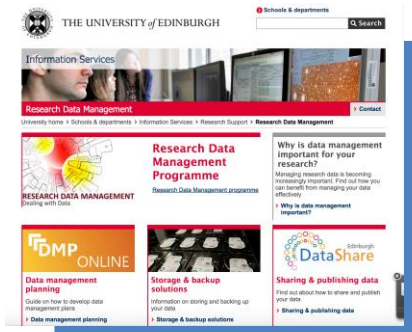
MANCHESTER 1824
The University of Manchester

Jisc

Research Data Management Services



Research Data Management Services



Data Management Support

DMP ONLINE

Data Management Planning

DataSync Edinburgh

Active Data Infrastructure

GitLab

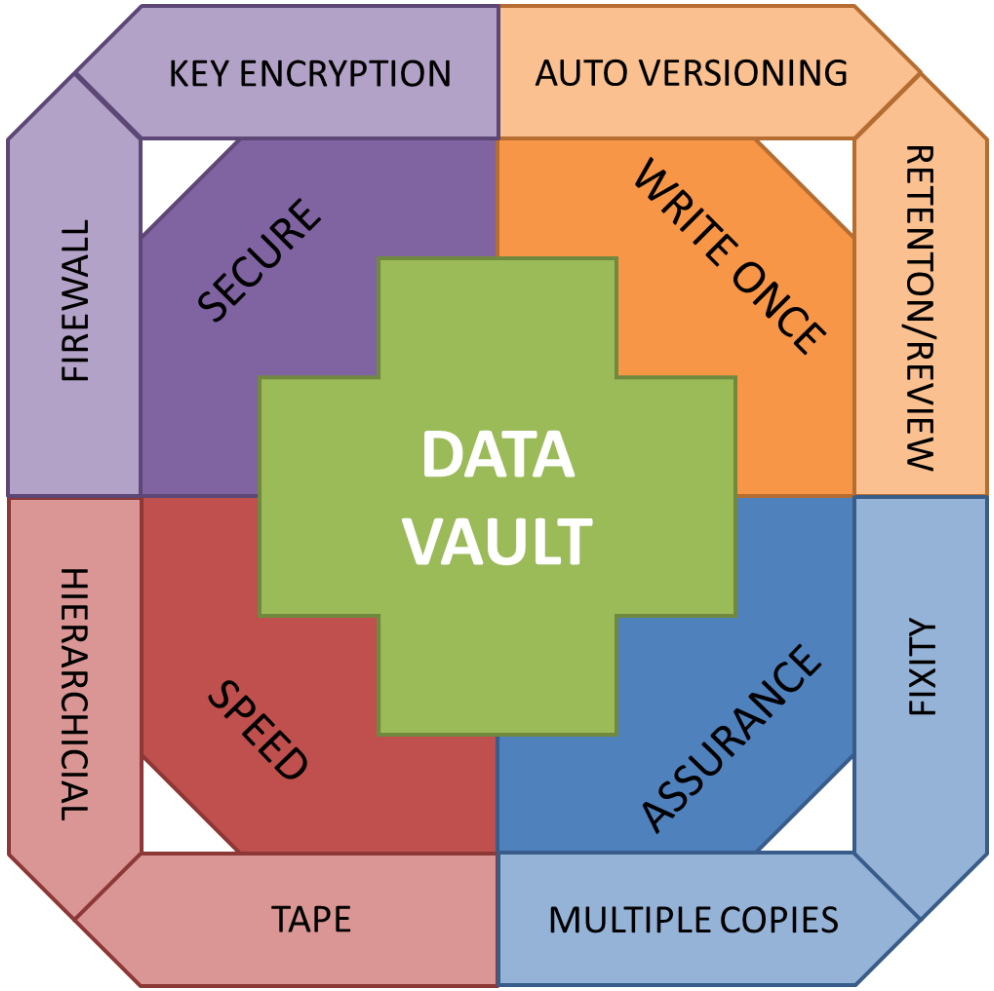
Pure

Data Stewardship

DataShare Edinburgh

Data Stewardship

- DataVault
  - Long term archival storage
  - First envisaged a few years ago…



DATA VAULT

THE UNIVERSITY of EDINBURGH

MANCHESTER 1824
The University of Manchester

Jisc

What is the DataVault - Analogies

DATA VAULT

THE UNIVERSITY of EDINBURGH

MANCHESTER 1824
The University of Manchester

Jisc

What is the DataVault - Analogies

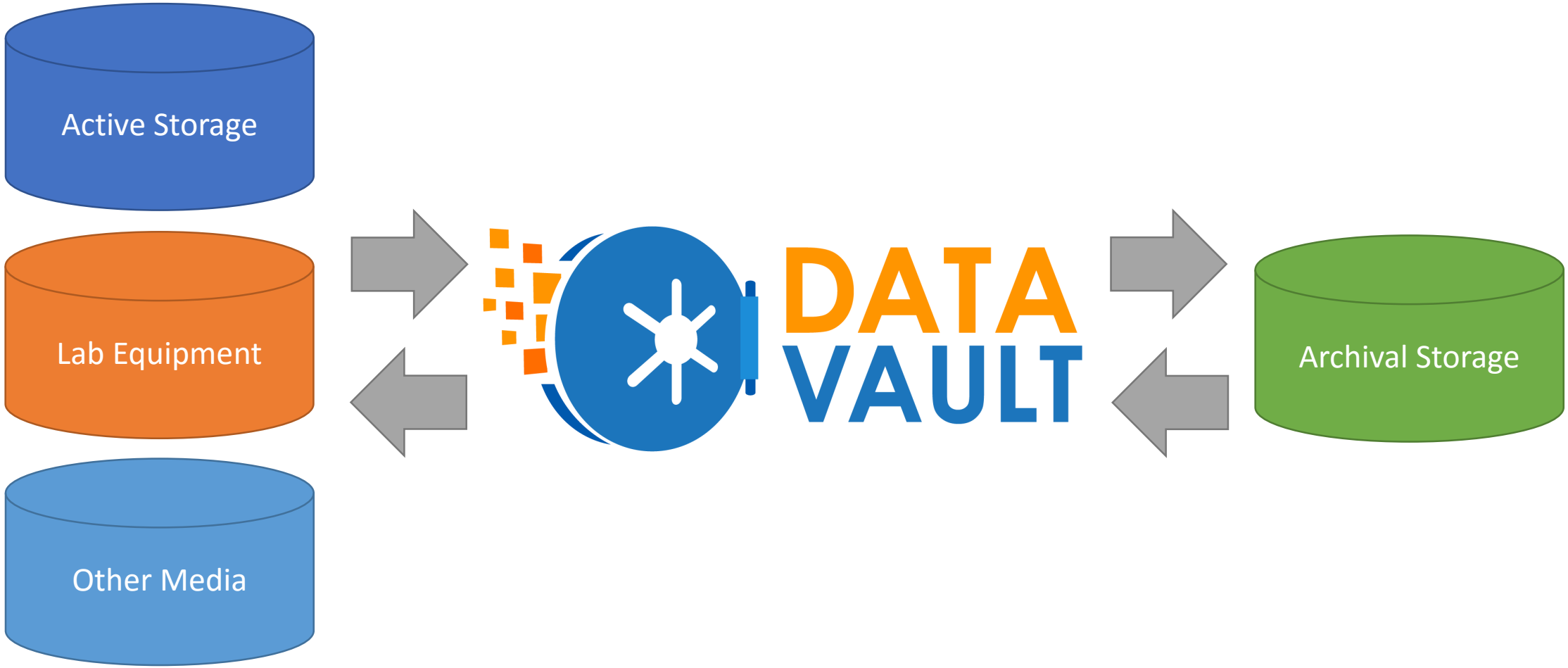
DATA VAULT

THE UNIVERSITY of EDINBURGH

MANCHESTER 1824
The University of Manchester

Jisc

Where does it sit?



DATA VAULT

THE UNIVERSITY of EDINBURGH

MANCHESTER 1824
The University of Manchester

Jisc

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 is provided via a RESTful API
  - But also comes with a default web user interface application
- It’s doesn’t undertake any active preservation activities
  - It does collect metadata and package the data for long-term storage

DATA VAULT

THE UNIVERSITY of EDINBURGH

MANCHESTER 1824 The University of Manchester

Jisc

DataVault use cases

- Store golden-copy data so that it can’t be changed
- Storing finalised data that can’t be openly shared (or ‘everything else’)
- Retaining old data in case it can be re-analysed later on
- Retention to meet funder requirements

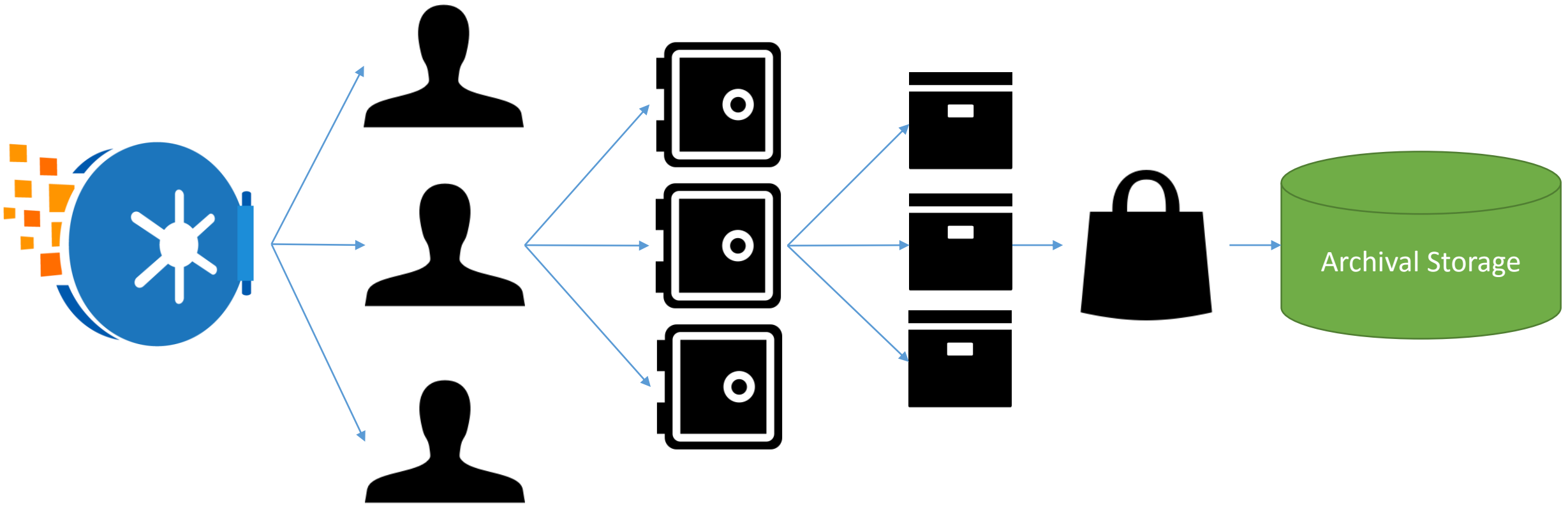
DATA VAULT

THE UNIVERSITY of EDINBURGH

MANCHESTER 1824 The University of Manchester

Jisc

Information Architecture



Icons: https://thenounproject.com/edward/, https://thenounproject.com/pixelbazaar/, https://thenounproject.com/useiconic.com/, https://thenounproject.com/bickov/

Technical Architecture

datavault-webapp

datavault-broker

RabbitMQ

datavault-worker(s)

DATA VAULT

THE UNIVERSITY of EDINBURGH

MANCHESTER 1824
The University of Manchester

Jisc

Project funding

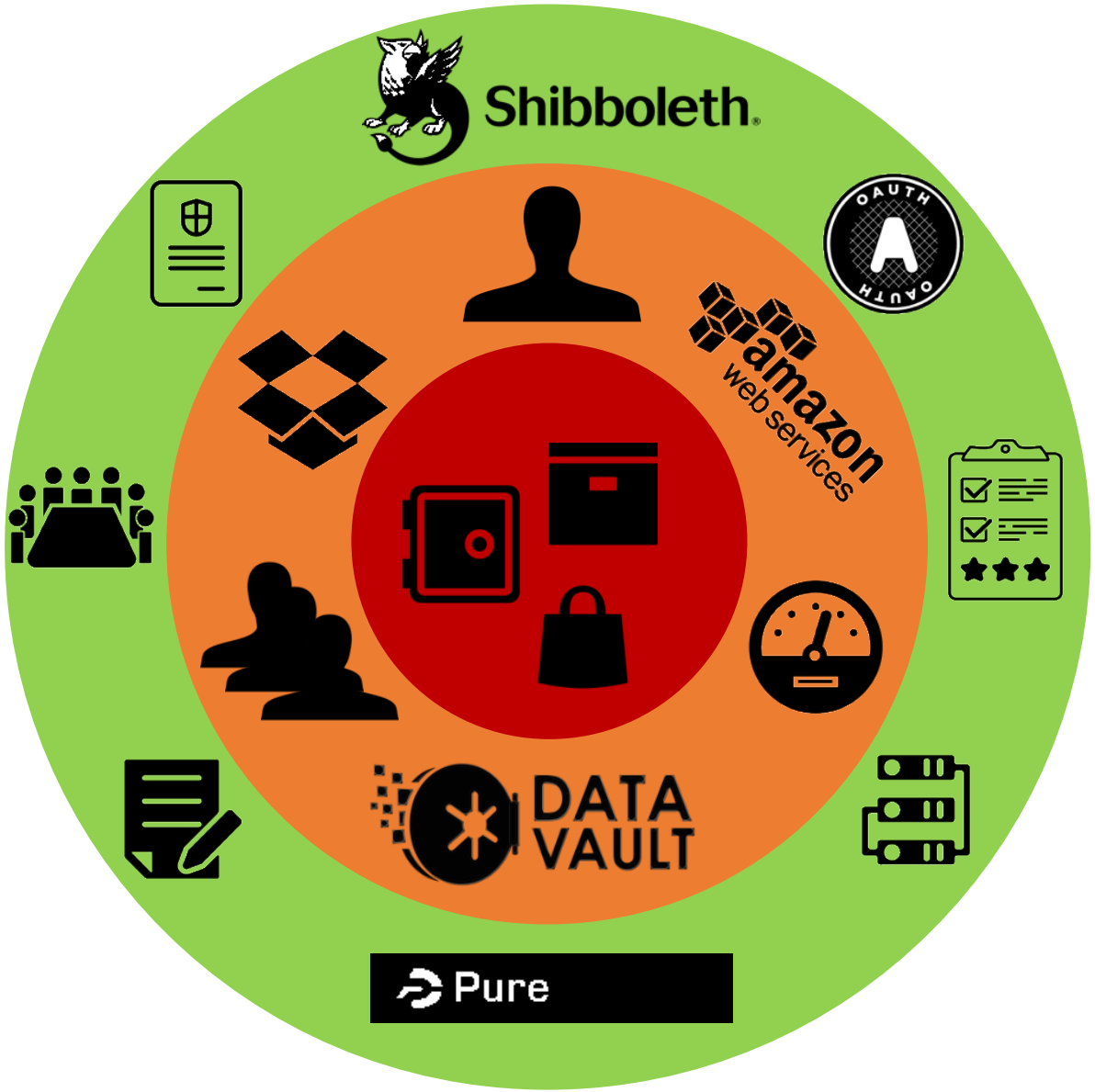
- Funded by Jisc
- DataSpring ‘Research at Risk’ programme
- Three phases
- Ideascale → pitches

DATA VAULT

THE UNIVERSITY of EDINBURGH

MANCHESTER 1824
The University of Manchester

Jisc



- Phase 1 (3 months)
- Phase 2 (4 months)
- Phase 3 (6 months)

Sprint 2 - April w2-4

Deposit/Vault size (bytes) incorrect - database update.
4 1/1 TH

Application logging infrastructure
3 1/1 TH

Use of the phrase 'retrieve' rather than restore
1 1/1 RT

Investigate UoE SSH key setup
1/1 RT

First version of help text / user guide
1/1

Arrange engagement event - 29th June (or 30th)

Add a card…

Sprint 3 - May w1-2

Onboarding: First time login UI - setup storage, connect to dropbox etc.
3 0/1 RT

Remove or fix "default" file store configuration (added for testing).

System Administrator / Data Manager (sub-admin)
4 0/1 RT

Tidy log file names
1 1/1 TH

Shibboleth (Spring security?)
1 0/1 RT

Clean-up /users endpoint

Add a card…

Sprint 4 - May w3-4

Email admin when job fails
1 0/1

Check boundaries and limits (e.g. max length of vault name in db, user agent length)
0/1

Technical scenario testing
0/3

Email user when deposit is complete.
1 0/1

Starting the worker process seems to require an internet connection to download/verify the spring XSD. I think it's possible for this to be checked against the local Jar instead.
1 0/1

Add a card…

Sprint 5 - May w5-Jw1

Metadata service for integration with the CRIS system
8 0/2 RT

Store raw CRIS metadata in bag metadata folder
1 0/1 RT

Metadata entry page as part of vault creation flow. Automatic fill of fields from PURE.
1 0/1

Out the box retention policies
2 0/3

User interface tooltips and inline text help.
0/1

Review file picker - confirm file

Add a card…

DataVault / datavault

Unwatch 5 Star 1 Fork 5

Code Issues 18 Pull requests 1 Wiki Pulse Graphs Settings

JISC DataVault project — Edit

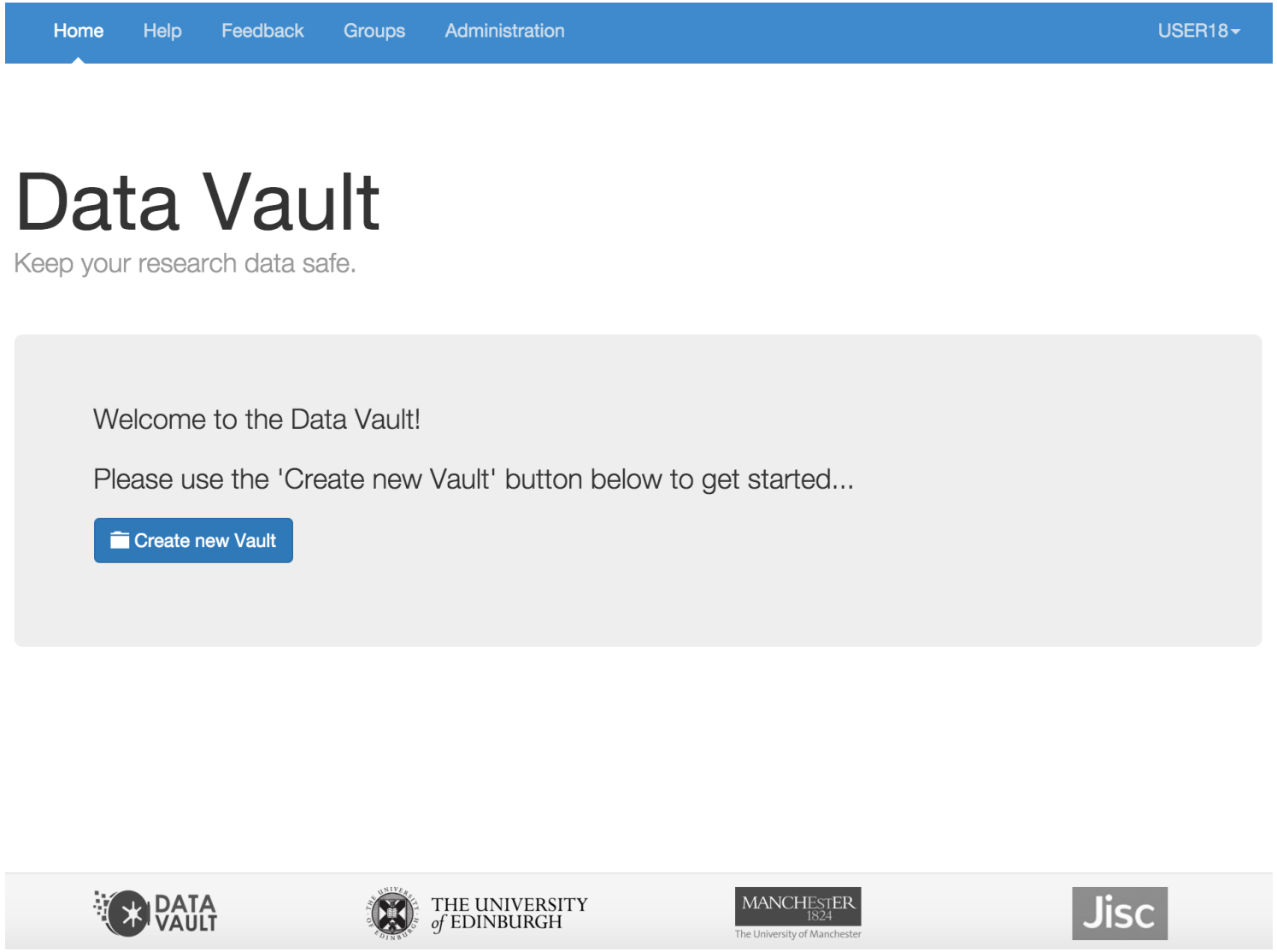
791 commits **1** branch **2** releases **4** contributors

Branch: **master** New pull request Create new file Upload files Find file Clone or download

tomhigginsuom Merge pull request #293 from tomhigginsuom/master ... Latest commit `ea625e2` an hour ago

datavault-assembly	Remove file count debug method	13 days ago
datavault-broker	Basic admin user search	10 days ago
datavault-common	Basic admin user search	10 days ago
datavault-webapp	Change info icon colour	2 hours ago
datavault-worker	Remove file count debug method	13 days ago
.gitignore	update .gitignore	2 months ago
.travis.yml	Travis on non-container hosts	3 months ago
LICENSE.md	Add MIT licence	11 months ago
README.md	Implement sustainability recommendations	11 days ago
logo-dvsmall.jpg	Trivial change to move the README image	4 months ago
pom.xml	Add organisation details to the master pom	18 days ago

Demonstration time…



Demonstration time…

- Try it yourself…

http://demo.datavaultplatform.org/

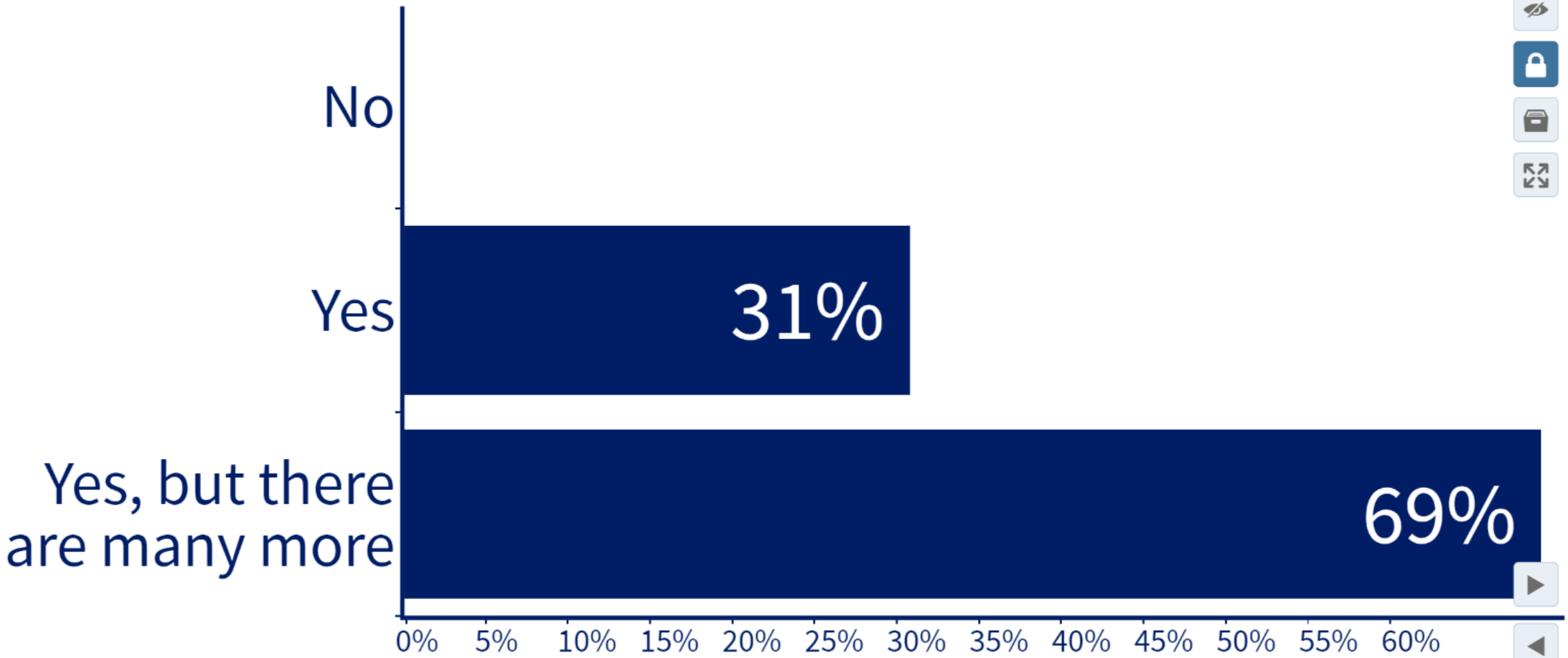
What do you think?

- Do you agree with the use cases?
- Do you see a need for such a tool?
- Do you think the tool approaches preservation appropriately?
- Do you think it is missing any obvious features?

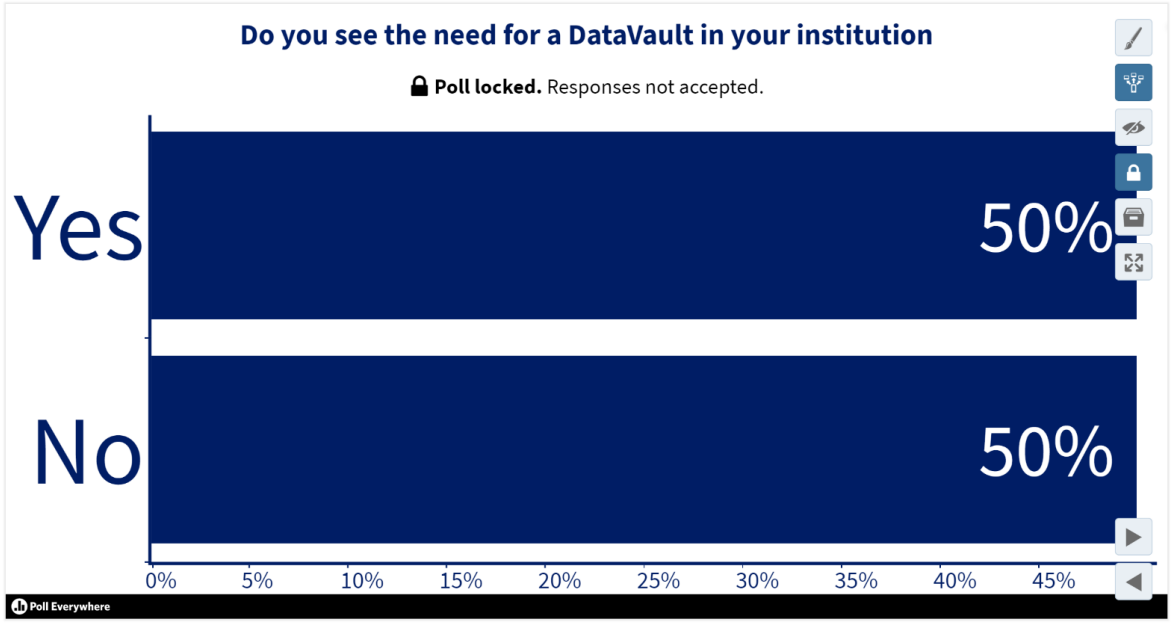
- Are you going to investigate it further!
  - Please talk to us!

Have we presented valid use cases?

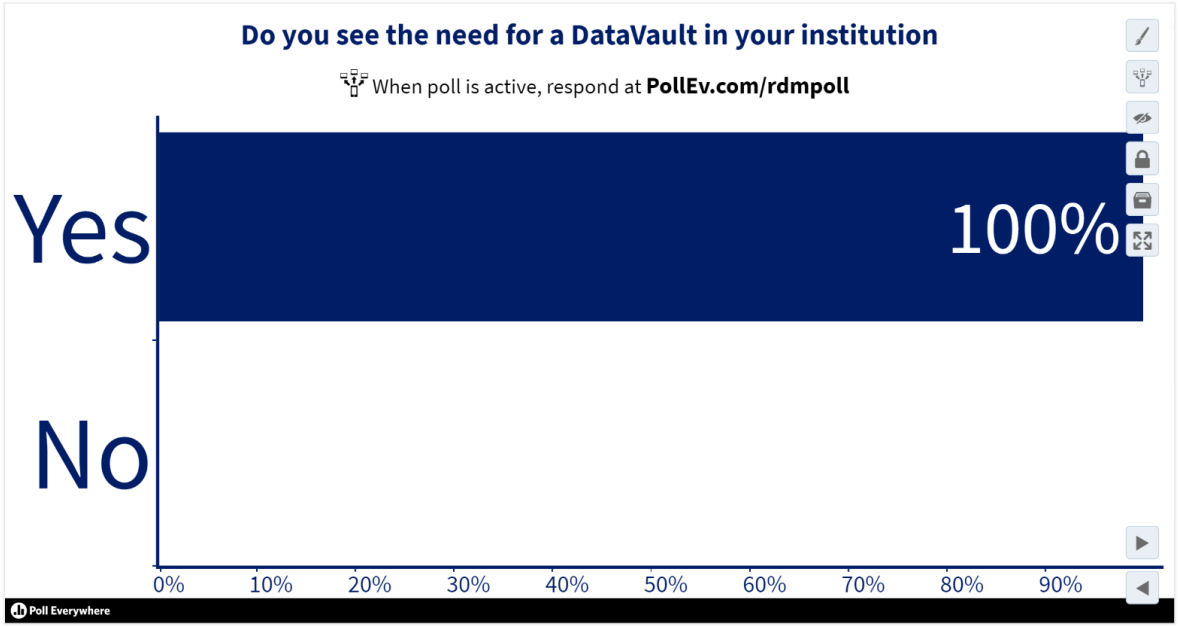
Poll locked. Responses not accepted.



IDCC (February 2016)



Jisc Research Data Network (May 2016)



Do you think the DataVault does enough 'preservation'

Poll locked. Responses not accepted.

"I don't feel like I have enough information to answer the question."
3 months ago

"No, not as yet"
3 months ago

"For its limited purpose, yes."
3 months ago

"Its doing bit preservation which is necessary, but not sufficient."
3 months ago

"Unsure - need more information"
3 months ago

"depends on the use cases"
3 months ago

"Not sure..."
3 months ago

"Surely, more complex formats will become unusable over time ??"
3 months ago

"No"
3 months ago

"No, but is better than nothing. And couldn't it be expanded?"
3 months ago

Poll Everywhere

Upgrade to enable moderation

Do you think it is missing any obvious features?

Poll locked. Responses not accepted.

"Warning to depositor that certain formats might not work in the long term ?"
3 months ago

"Yes"
3 months ago

"So far, so good!"
3 months ago

"File format metadata?"
3 months ago

"Large datasets can be problematic to upload via https"
3 months ago

"researchers might think everything is fine now"
3 months ago

"How do you insure sustainability?"
3 months ago

"How many petabytes it can manage?"
3 months ago

"Warnings upfront to researchers"
3 months ago

"Does it record any metadata associated with events (e.g. PREMIS)?"
3 months ago

"Incremental updates."
3 months ago

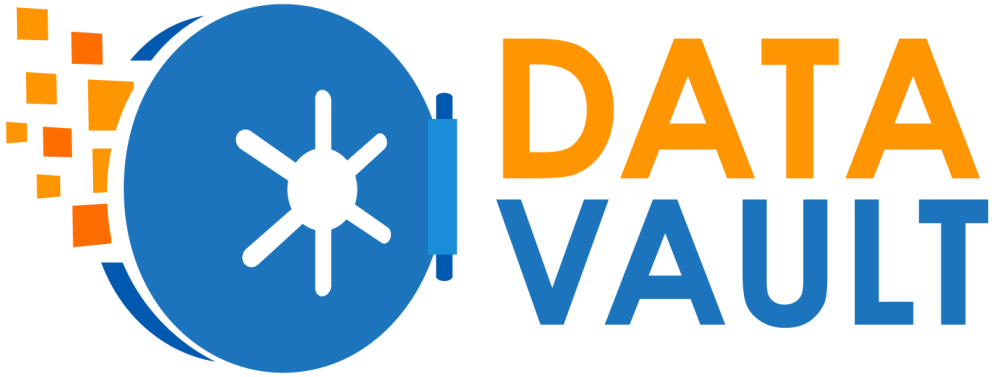
"costs?"
3 months ago

"It does what is says on the tin"
3 months ago

Poll Everywhere

Upgrade to enable moderation

Hot off the press…



Data Vault Implementation Workshop

Wednesday 29th June 2016, Central London (Islington)

15 spaces (1x RDM + 1x Tech)

Keep an eye out for the booking form / speak to me

DATA VAULT

THE UNIVERSITY of EDINBURGH

MANCHESTER 1824
The University of Manchester

Jisc

DATA VAULT

Stuart Lewis

stuart.lewis@ed.ac.uk

@stuartlewis

Robin Taylor

robin.taylor@ed.ac.uk

@JiscDataVault

DATA VAULT

THE UNIVERSITY of EDINBURGH

MANCHESTER 1824
The University of Manchester

Jisc