

Dataverse for the Canadian Research Community

Developing reusable and scalable tools for data deposit, curation, and sharing

Meghan Goodchild & Kaitlin Newson

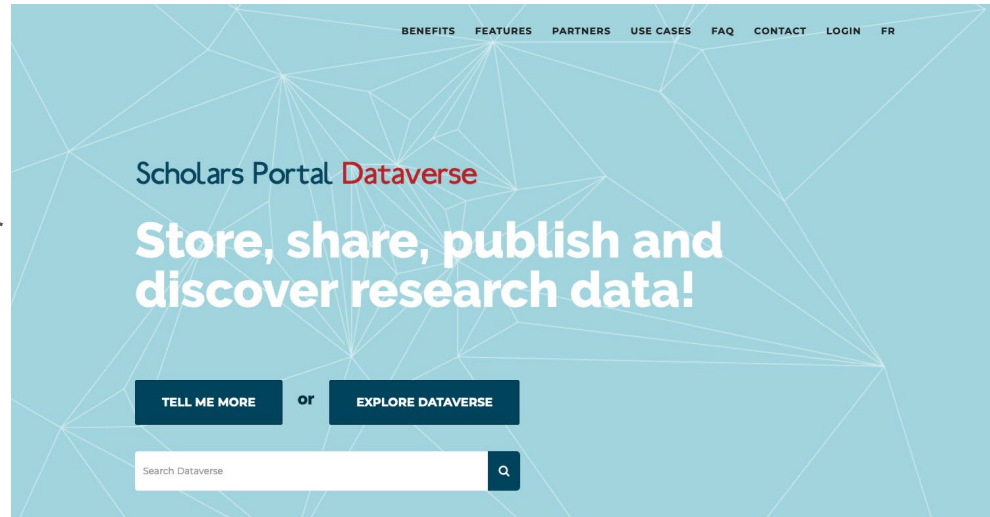
Canarie RDM Workshop
February 2020





What is Dataverse?

- Open-source research data repository software developed by IQSS at Harvard
- Store, share, publish and discover research data
- Hosted by Scholars Portal on behalf of 51 universities across Canada





Our team

Kate Davis - PI
Amaz Taufique- Technical Lead
Amber Leahey- co-PI (*on leave*)
Meghan Goodchild - Project Manager
Kaitlin Newson

Developers

Jayanthi Chengan
Sunil Manikonda
Victoria Lubitch

Systems

Bikram Singh
Sohaib Anwar
Carlos McGregor

Lee Wilson (Portage)



Our project

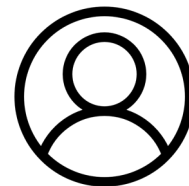
- Aim to make enhancements to the Dataverse platform to address the needs of a broad range of users across Canada
- 3 key focus areas:
 - Authentication
 - Data Curation
 - Scalability



canarie

1) Authentication

- “As a user from an affiliated university, I want to log into Dataverse using the same login as other services I access.”
- Integrate Dataverse with CAF Shibboleth Login for single-sign on
- Streamline login workflows



Shibboleth®

Scholars Portal Dataverse



Shibboleth process

- Phase 1: connected Dataverse sandbox to University of Toronto identity provider
- Phase 2: connected Dataverse sandbox to Canarie identity provider (R&S category)
- Phase 3: connected production Dataverse instance to Canarie identity provider (R&S category)

➔ Log In

Log in or sign up with your institutional account — [learn more](#). Leaving your institution? Please contact [Dataverse Support](#) for assistance.

Your Institution

- ✓ Please select...
- Carleton University
- McGill University
- University of British Columbia
- University of Guelph
- University of Toronto
- University of Victoria
- University of Western Ontario

[Continue](#)[Institution](#)

Other options

[Username/Email](#)[Sign In](#)



Launch & Next Steps

- Launched as part of Dataverse upgrade on October 31, 2019
- Planning webinar to improve uptake and clarify details
- Ongoing challenge of onboarding institutions using Canarie R&S category due to internal privacy assessment and approval processes

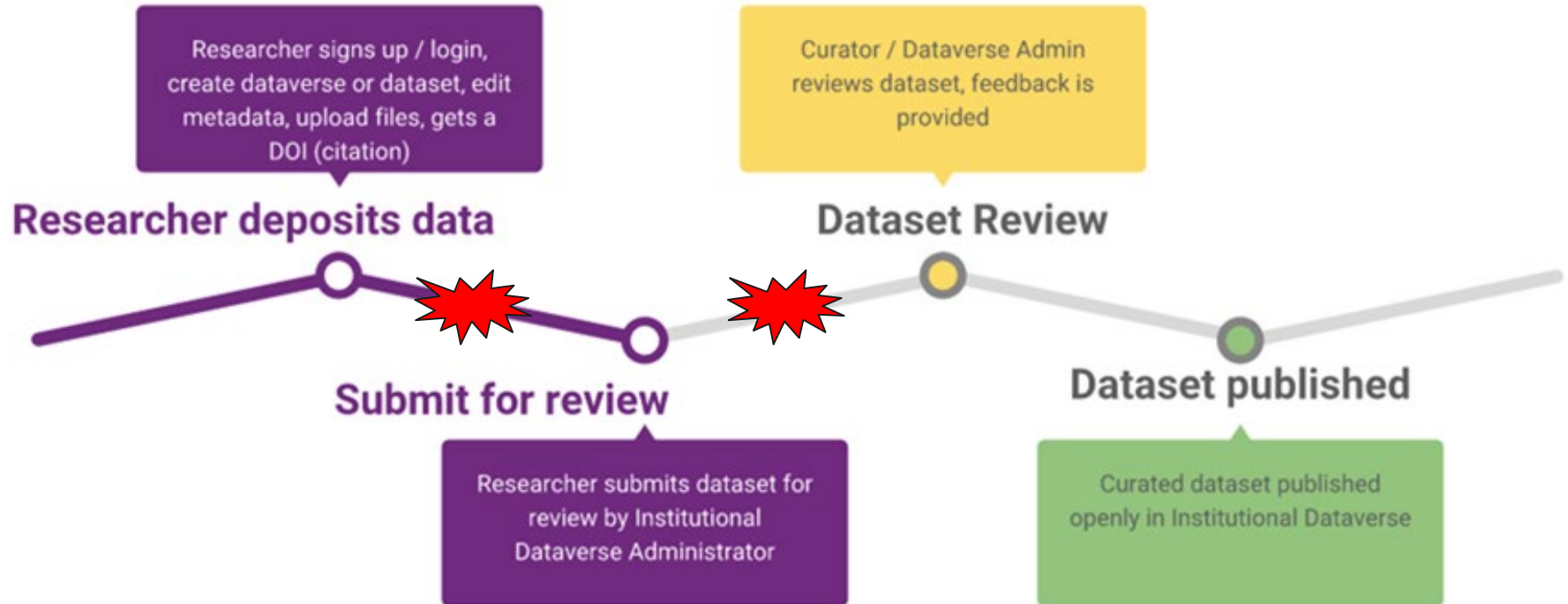


2) Data Curation

- “As a user, I want to add details about the variables in my dataset after I’ve uploaded it to Dataverse.”
- Developed the Data Curation Tool (DCT)
 - External web application that connects to Dataverse to create and edit metadata at the variable level using DDI standard
- Aim to improve data curation workflows and promote adoption of standards and best practices
- Works with tabular data files (CSV, SPSS, SAV, etc.)

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16
1	1.040e+14	4	26	104012	99	1	1973	2	1642	1040	1040	1	6	1	5	5
2	1.040e+14	1	19	104008	99	2	1943	1	99	99	99	1	1	6	4	5
3	1.040e+14	4	98	104003	99	1	1990	1	99	99	99	1	6	3	99	11
4	1.040e+14	2	98	104010	99	1	1983	2	1756	1040	1040	1	6	3	99	11
5	1.040e+14	1	18	104007	99	2	1927	1	99	99	99	1	1	6	4	1
6	1.040e+14	2	19	104007	99	1	1983	1	99	99	99	1	6	2	5	3
7	1.040e+14	4	15	104005	99	2	1970	1	99	99	99	1	1	2	4	6
8	1.040e+14	4	19	104006	99	1	1942	1	99	99	99	2	1	6	5	88
9	1.040e+14	4	15	104005	99	1	1965	2	1040	1276	1040	7	4	1	5	5
10	1.040e+14	7	15	104004	99	2	1955	1	99	99	99	1	1	2	5	5
11	1.040e+14	7	15	104003	99	2	7777	1	99	99	99	1	1	6	77	77
12	1.040e+14	4	18	104005	99	2	1938	1	99	99	99	1	3	6	5	5
13	1.040e+14	7	17	104005	99	1	1945	1	99	99	99	1	1	6	5	4
14	1.040e+14	4	18	104005	99	2	1949	2	1040	1380	1040	1	4	6	5	5
15	1.040e+14	2	15	104003	99	2	7777	1	99	99	99	1	1	2	4	5
16	1.040e+14	4	32	104007	99	1	1974	1	99	99	99	1	6	2	5	5
17	1.040e+14	4	98	104006	99	1	7777	1	99	99	99	1	6	3	99	11
18	1.040e+14	4	25	104010	99	2	1968	1	99	99	99	1	1	2	5	1
19	1.040e+14	4	40	104007	99	1	1967	1	99	99	99	1	1	2	6	2
20	1.040e+14	1	23	104004	99	1	1932	1	99	99	99	1	1	6	4	3
21	1.040e+14	2	18	104003	99	2	1965	1	99	99	99	1	6	2	5	5
22	1.040e+14	1	27	104011	99	1	1956	1	99	99	99	2	1	1	5	2
23	1.040e+14	4	18	104004	99	2	1923	1	99	99	99	1	3	6	3	3
24	1.040e+14	1	19	104006	99	2	1952	1	99	99	99	1	4	2	5	3
25	1.040e+14	3	16	104004	99	2	1947	2	1276	1040	1040	1	1	6	3	3

Curation Workflows



Data Curation Tool

Files

Metadata

Terms

Versions

+ Upload Files

1 File

LFS2016-01_PUMF_EN.tab

Tabular Data - 11.9 MB - Sep 19, 2019 - 6 Downloads

75 Variables, 101887 Observations - UNF:6:EerkQFr2ySzwCu4oV5jExA==

Configure

Explore

Download

Data Curation Tool

Labour Force Survey (Curated)

LFS2016-01_PUMF_EN.tab

Tester, Curation, 2019, "Labour Force Survey (Curated)", <https://doi.org/10.5072/FK2/YLJJAY>, Scholars Portal Dataverse, V1, UNF:6:EerkQFr2ySzwCu4oV5JExA== [fileUNF]

< Hide Groups

Download

Save

Add Group +

Search

Items per page: 25

1 - 25 of 75

< >

All Variables

Number of hours



ID

Name

Label

Weight

View



v14110

REC_NUM

Order of record in file



v14130

SURVYEAR

Survey year



v14106

SURVMNTH

Survey month



v14127

LFSSTAT

Labour force status



v14165

PROV

Province



v14168

CMA

3 largest CMAs



v14155

AGE_12

Five-year age group of respondent



v14136

AGE_6

Age in 2 and 3 year groups



Labour Force Survey (Curated)


LFS2016-01_PUMF_EN.tab

Tester, Curation, 2019, "Labour Force Survey (Curated)", <https://doi.org/10.5072/FK2/YLJJAY>, Scholars Portal Dataverse, V1, UNF:6:EerkQFr2ySzwCu4oV5jExA== [fileUNF]

< Hide Groups

Add Group +

All Variables

Number of hours 

Search

<input type="checkbox"/>	ID	Name
<input type="checkbox"/>	v14110	REC_N
<input type="checkbox"/>	v14130	SURVY
<input type="checkbox"/>	v14106	SURVM
<input type="checkbox"/>	v14127	LFSSTA
<input type="checkbox"/>	v14165	PROV
<input type="checkbox"/>	v14168	CMA
<input type="checkbox"/>	v14155	AGE_12
<input type="checkbox"/>	v14136	AGE_6
<input type="checkbox"/>	v14158	SEX



















Variable Information

ID	Name
v14158	SEX
<hr/>	
Label	
Sex of respondent	
<hr/>	
Literal Question	
<hr/>	
Interviewer Instructions	
<hr/>	
Post Question	
<hr/>	
Universe	
<hr/>	
Notes	
<hr/>	

 Download

 Save

Items per page: 25 1 - 25 of 75 < >

Weight	View	
		
		
		
		
		
		
		
		
		

Labour Force Survey (Curated)

LFS2016-01_PUMF_EN.tab

Tester, Curation, 2019, "Labour Force Survey (Curated)", <https://doi.org/10.5072/FK2/YLJJAY>, Scholars Portal Dataverse, V1, UNF:6:EerkQFr2ySzwCu4oV5jExA== [fileUNF]

< Hide Groups

Add Group +

All Variables

Number of hours

Search

☐ ID Name

☐ v14110 REC_N

☐ v14130 SURVY

☐ v14106 SURVM

☐ v14127 LFSSTA

☐ v14165 PROV

☐ v14168 CMA

☐ v14155 AGE_12

☐ v14136 AGE_6

☐ v14158 SEX

Age in 2 and 3 year groups

Sex of respondent

Download

Save

Items per page: 25 1 - 25 of 75 < >

Weight

View

AGE_6: Age in 2 and 3 year groups

Values	Categories	Count	Count Percentage(%)	Weighted Count
1	15 to 16	2,885	13.417	
2	17 to 19	4,224	19.644	
3	20 to 21	2,909	13.528	
4	22 to 24	4,219	19.621	
5	25 to 26	3,017	14.031	
6	27 to 29	4,249	19.76	

Data Curation Tool Codebook

Data Curation Tool Testing Dataset (ICPSR doi:10.5072/FK2/0TYIHL) (DCT Testing Dataset)

View: [Part 1: Document Description](#)
[Part 2: Study Description](#)
[Part 3: Data Files Description](#)
[Part 4: Variable Description](#)
[Part 5: Other Study-Related Materials](#)
[Entire Codebook](#)

Document Description	
Citation	
<i>Title:</i>	Data Curation Tool Testing Dataset
<i>Identification Number:</i>	doi:10.5072/FK2/0TYIHL
<i>Distributor:</i>	Scholars Portal Dataverse
<i>Date of Distribution:</i>	2019-06-14
<i>Version:</i>	2
<i>Bibliographic Citation:</i>	Lubitch, Victoria; Leahey, Amber, 2019, "Data Curation Tool Testing Dataset", https://
Study Description	
Citation	
<i>Title:</i>	Data Curation Tool Testing Dataset
<i>Subtitle:</i>	DDI Test
<i>Alternative Title:</i>	DCT Testing Dataset
<i>Identification Number:</i>	doi:10.5072/FK2/0TYIHL
<i>Authoring Entity:</i>	Lubitch, Victoria (University of Toronto) Leahey, Amber (Scholars Portal)
<i>Producer:</i>	Leahey, Amber
<i>Date of Production:</i>	2019-05-22
<i>Grant Number:</i>	4445555
<i>Distributor:</i>	Scholars Portal Dataverse
<i>Date of Distribution:</i>	2019-06-14
Study Scope	
<i>Keywords:</i>	Astronomy and Astrophysics, test, smoking
<i>Topic Classification:</i>	Metadata



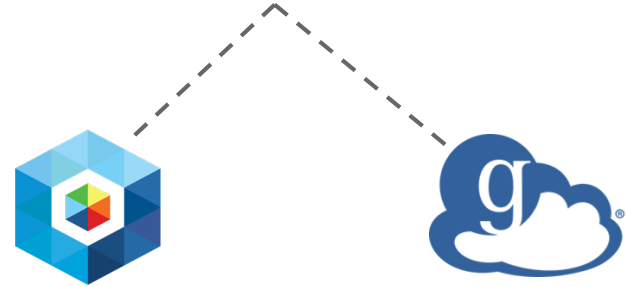
Launch & Next steps

- Launched as part of Dataverse upgrade on October 31, 2019
- Blog post: [Introducing the Data Curation Tool](#)
- Upcoming webinar on February 14 at 10AM EST (connection details TBA)

3) Scalability

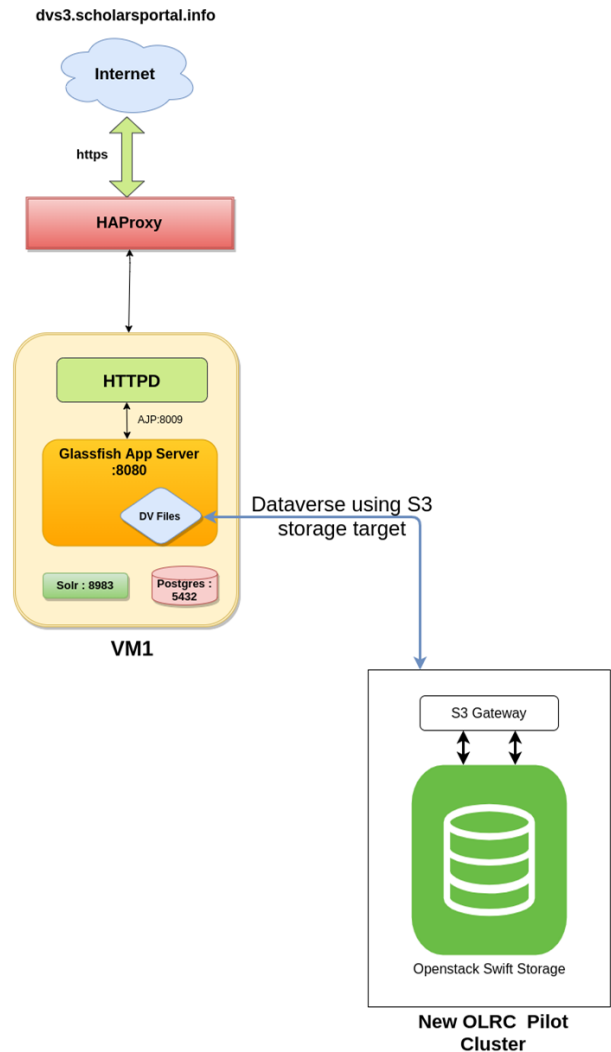
- “As a user, I want to upload files larger than 2.5 GB to Dataverse.”
- Optimize system architecture for scalable use
- Connect to existing Canadian data storage environments
- Support large files in upload/download contexts

Scholars Portal **Dataverse**



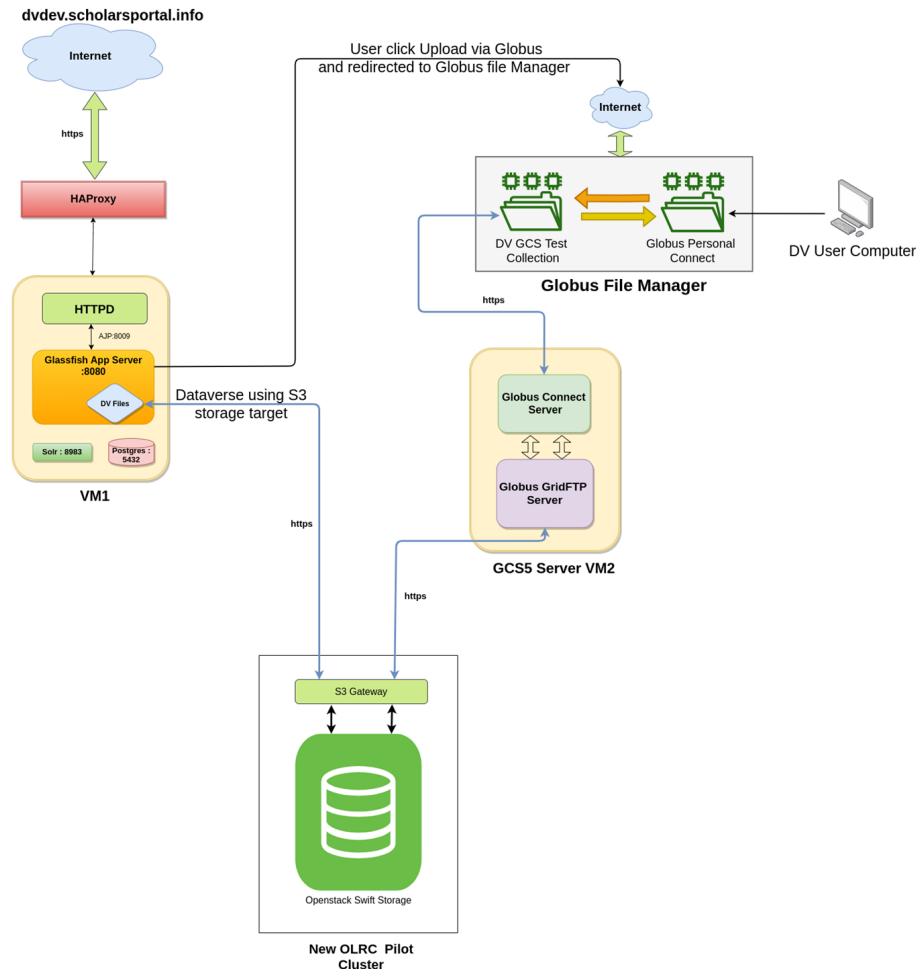
Scalability: Connect Dataverse with cloud storage

- Stood up test instance of Dataverse connected to the OLRC (Swift Object storage)
 - Amazon S3 emulation on top of OLRC using Dataverse S3 drivers



Scalability: Large -file support

- Modify Dataverse to work with Globus endpoints
- Modify UI to support downloading from Swift/Globus





Globus integration

- Phase 1: Proof of concept integration (Spring 2020)
 - Globus Connect Personal
- Phase 2: Tighter integration (Winter 2020)
 - Investigate designing specific UI for Dataverse using Globus APIs
 - Work with IQSS to include in core code
 - Aim to launch in production

Globus Demo

Dataverse - Globus option on “add data” page

Files

Upload with HTTP via your browser ^


Select files or drag and drop into the upload widget. File upload limit is 12.0 GB per file.

+ Select Files to Add

Drag and drop files here.

Upload with Globus ^

BEFORE YOU START: You will need to set up a free account with [Globus](#) and have [Globus Connect Personal](#) running on your computer to transfer files to and from the service.


 Upload with Globus

Once Globus transfer has finished, you will get an email notification. Please come back here and press the following button:

Globus Transfer has finished

Click here to view the dataset page: [Globus Test 2020-01-16](#) .

Globus Connect Personal



FILE MANAGER

BOOKMARKS

ACTIVITY

ENDPOINTS

GROUPS

CONSOLE

ACCOUNT

LOGOUT

HELP

File Manager

Panels

Collection Meghan Goodchild's Laptop

Path /~/OneDrive - Queen's University/Test Data/

Dataverse GCS test collection

/10.5072/FK2/KIDLKX/

select all up one folder refresh list columns view

dataverse_files (10).zip	12/11/2019 10:49am	74.85 MB
dataverse_files (300).zip	03/27/2019 10:38am	1.41 MB
eTheses	07/24/2018 02:45pm	—
Images	10/10/2019 09:55am	—
Images.zip	10/10/2019 09:58am	718.05 KB
LargeTest.zip	11/27/2017 02:55pm	5.26 GB
LargeTestFolder.tar.gz	11/28/2017 03:42pm	5.23 GB
MANIFEST.TXT	04/03/2019 08:08am	718 B
Queen's Journals	07/24/2018 02:44pm	—
Test 100 MB (2).zip	06/22/2018 07:46am	102.69 MB
Test AV 1GB.zip	05/25/2018 08:01am	1.05 GB
Test AV 2GB.zip	05/25/2018 08:01am	2.10 GB
Test AV 4 GB (2).zip	06/21/2018 09:13am	3.49 GB

Share

Transfer or Sync to...

New Folder

Rename

Delete Selected

Download

Open

Upload

Get Link

Show Hidden Items

Manage Activation

cached	01/20/2020 09:04pm	1.01 KB
json.cached	01/20/2020 09:04pm	4.46 KB
cached	01/20/2020 09:04pm	894 B
d	01/20/2020 09:04pm	2.81 KB
ed	01/20/2020 09:04pm	8.65 KB
ite.cached	01/20/2020 09:04pm	1.83 KB
ched	01/20/2020 09:04pm	627 B
ached	01/20/2020 09:04pm	2.78 KB
ached	01/20/2020 09:04pm	4.80 KB
rg.cached	01/20/2020 09:04pm	1.94 KB
p	01/20/2020 06:42pm	102.69 MB
	01/20/2020 06:41pm	1.05 GB
	01/20/2020 06:38pm	2.10 GB

Start

Transfer & Sync Options

Start

Dataset page post upload



Globus Test 2020-01-16

Version 2.0

Goodchild, Meghan, 2020, "Globus Test 2020-01-16", <https://doi.org/10.5072/FK2/KIDLKX>, Scholars Portal Dataverse, V2

Cite Dataset

Learn about [Data Citation Standards](#).

Dataset Metrics

0 Downloads

Description

test

Subject

Arts and Humanities

Files

Metadata

Terms

Versions

Search this dataset...

Find

Upload Files

Filter by

File Type: All

Access: All

Sort



1 to 8 of 8 Files

Edit Files

Download



AcademicKnowledgeProcess.PNG

application/vnd.dataverse.file-globus - 1.0 MB - Jan 23, 2020 - 0 Downloads
SHA-1: 7c4e62ed6bd22ace8aa7996679a983b9

Download through Globus

Download



CurationVsSelfDepositDCN.PNG

application/vnd.dataverse.file-globus - 262.9 KB - Jan 23, 2020 - 0 Downloads
SHA-1: e8f0ee2bf6767bcb3548031088b30caf

Download through Globus

Download



CurationVsSelfDepositDCNGraph.PNG

application/vnd.dataverse.file-globus - 147.7 KB - Jan 23, 2020 - 0 Downloads
SHA-1: 0726acd43dba00739e2761a5942065eb

Download through Globus

Download



Test 100 MB (2).zip

application/vnd.dataverse.file-globus - 97.9 MB - Jan 20, 2020 - 0 Downloads
SHA-1: 1fc35b98e03c69c90826acd875df3cf2

Download through Globus

Download

Globus considerations



Features

- Robust tool (100 GB file transfer)
- Currently being by Canadian researchers and FRDR

Considerations

- Internationalization of the Globus Connect UI
- Dataverse permissions cannot be applied to Globus on file level
- Syncing and checking for files uploaded/changed to Globus directory
- Challenge getting users comfortable with Globus
- Globus license



3) Scalability - Plans for production

- Upgrade to OLRC storage backend planned for Fall 2020
- Phase 2 of Globus integration planned for end of 2020



Challenges and lessons learned

- Challenge working with open-source community for developments to be incorporated into core code
- Need for usability testing during development workflow
- Internationalization considerations
- Promotion of developments to ensure uptake by researchers and institutions
- Plans for production:
 - Upgrade to OLRC storage backend planned for Fall 2020
 - Phase 2 of Globus integration planned for end of 2020



Thank you!

meghan.goodchild@queensu.ca

kaitlin.newson@utoronto.ca

Extra slides

- Architecture diagram of current stack

