Globus for Repositories

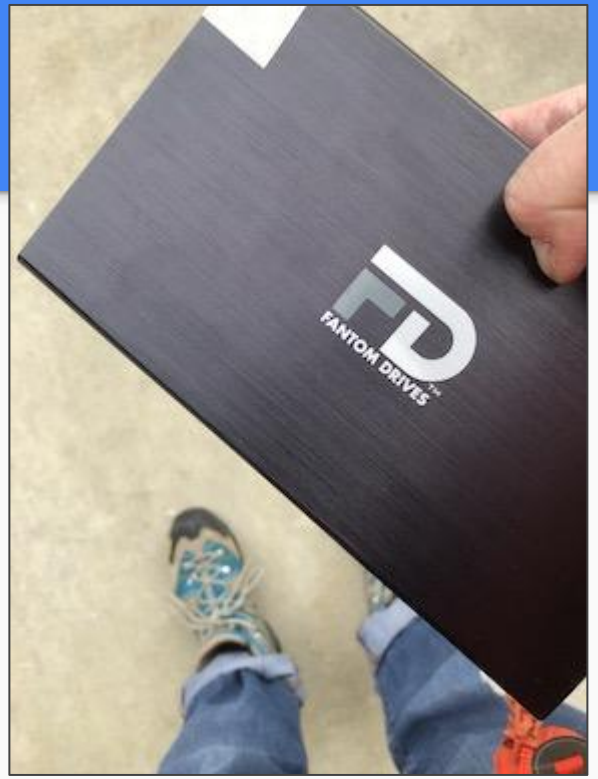
Susan Borda - Data Workflows Specialist
University of Michigan - Library

So you want to put your LARGE data set in the repository...

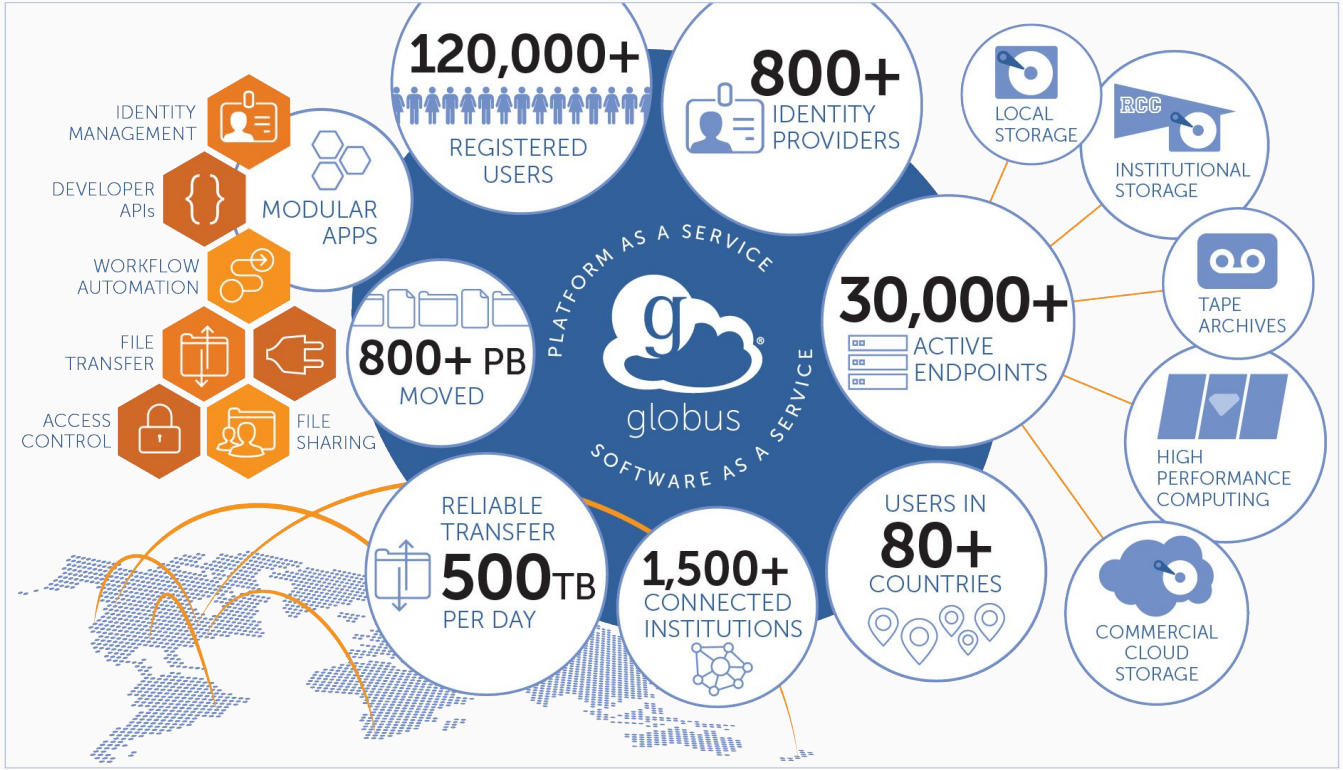
- Where is the data?
- How large is the data?
- ~~What kind of data is it?~~
- Where does it need to go?
- How to retrieve the data

Where is the data?

- External hard drive
- Personal server in their office
- Lab server
- On a centrally managed server
- Attached to an HPC environment



Still the fastest way to get 1TB across campus!

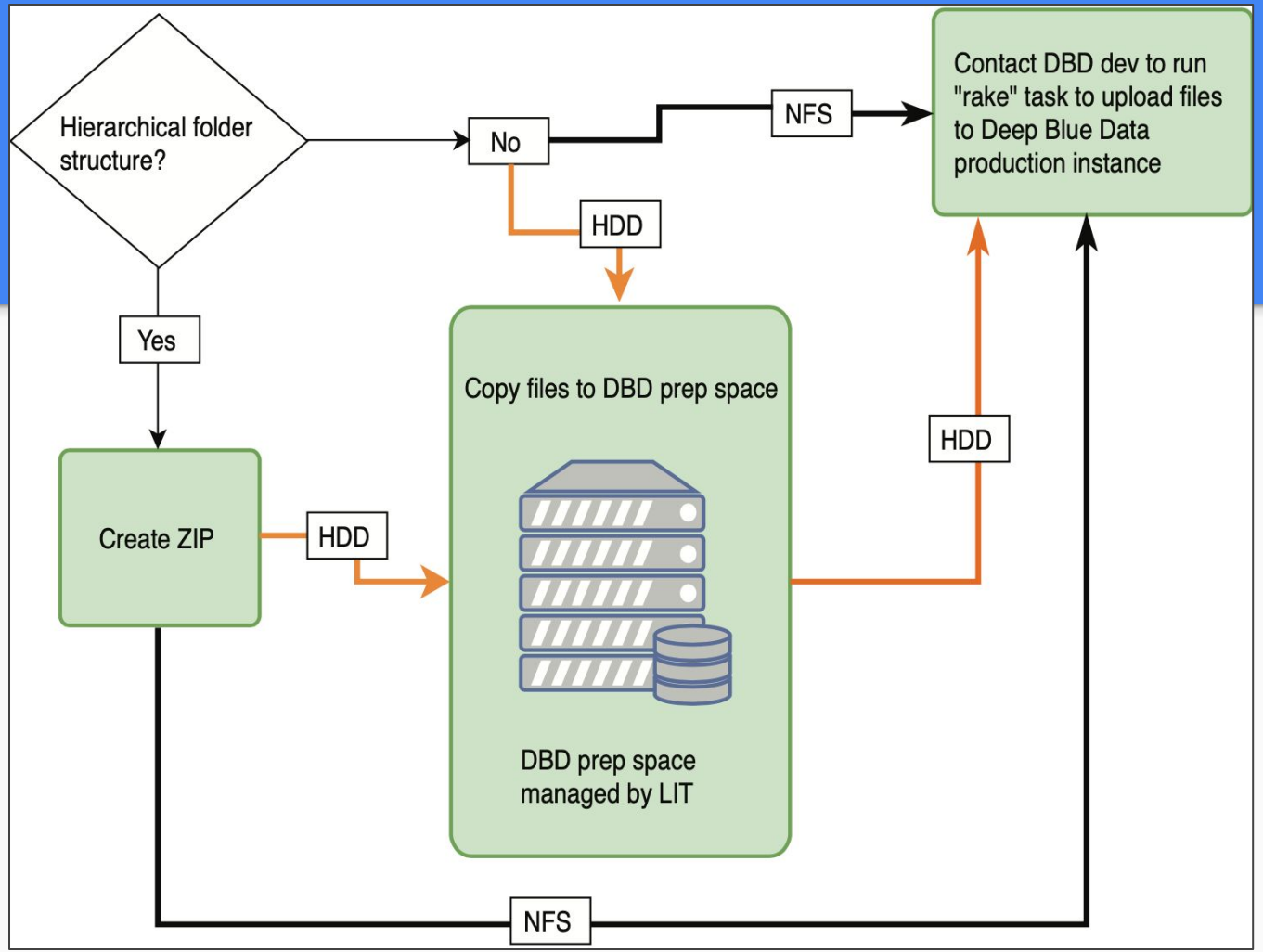


https://www.globus.org/

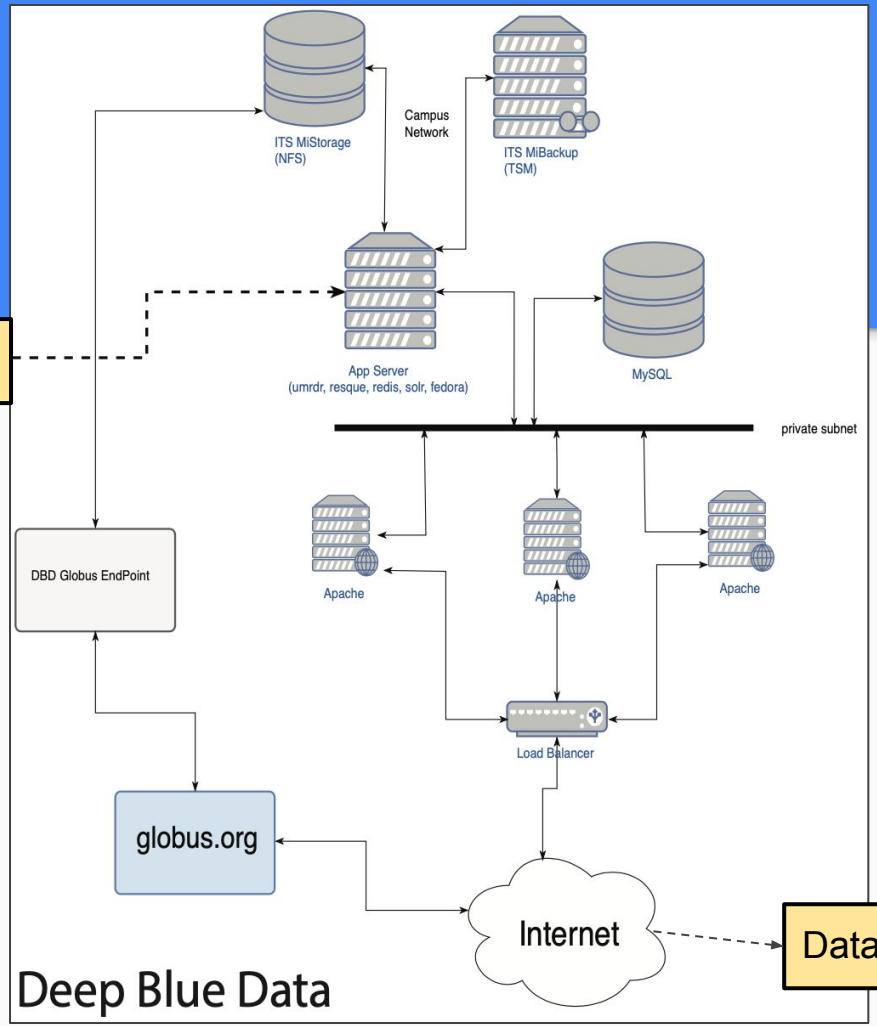
Helloooo Globus!!

There is NOT a direct connection between Globus and Deep Blue Data for data ingest/upload.

Folder structures need to be zipped or tarred first.



Data ingest



Data download

Globus is integrated for data-set download only!

Globus prevents the “Roach Motel” scenario.

Data checks in but doesn’t check out.

https://www.homedepot.com/p/Black-Flag-Roach-Motel-Glue-Traps-2-Count-HG-11020-1/204237338

In Deep Blue Data

Files (Count: 14; Size: 3.56 TB)

Title	Original Upload	Last Modified	File Size	Access	Actions
2010_gitm_input_files.tgz	2017-11-22	2017-11-22	102 MB	Open Access	Select an action
gitm_170809.tgz	2017-11-22	2017-11-22	26.9 MB	Open Access	Select an action
gitm_201001_all_output.tgz	2018-01-25	2018-12-13	312 GB	Open Access	Select an action
gitm_201002_all_output.tgz	2018-01-25	2018-12-14	273 GB	Open Access	Select an action
gitm_201003_all_output.tgz	2018-01-25	2018-12-14	315 GB	Open Access	Select an action
gitm_201004_all_output.tgz	2018-01-25	2018-12-15	302	Open Access	

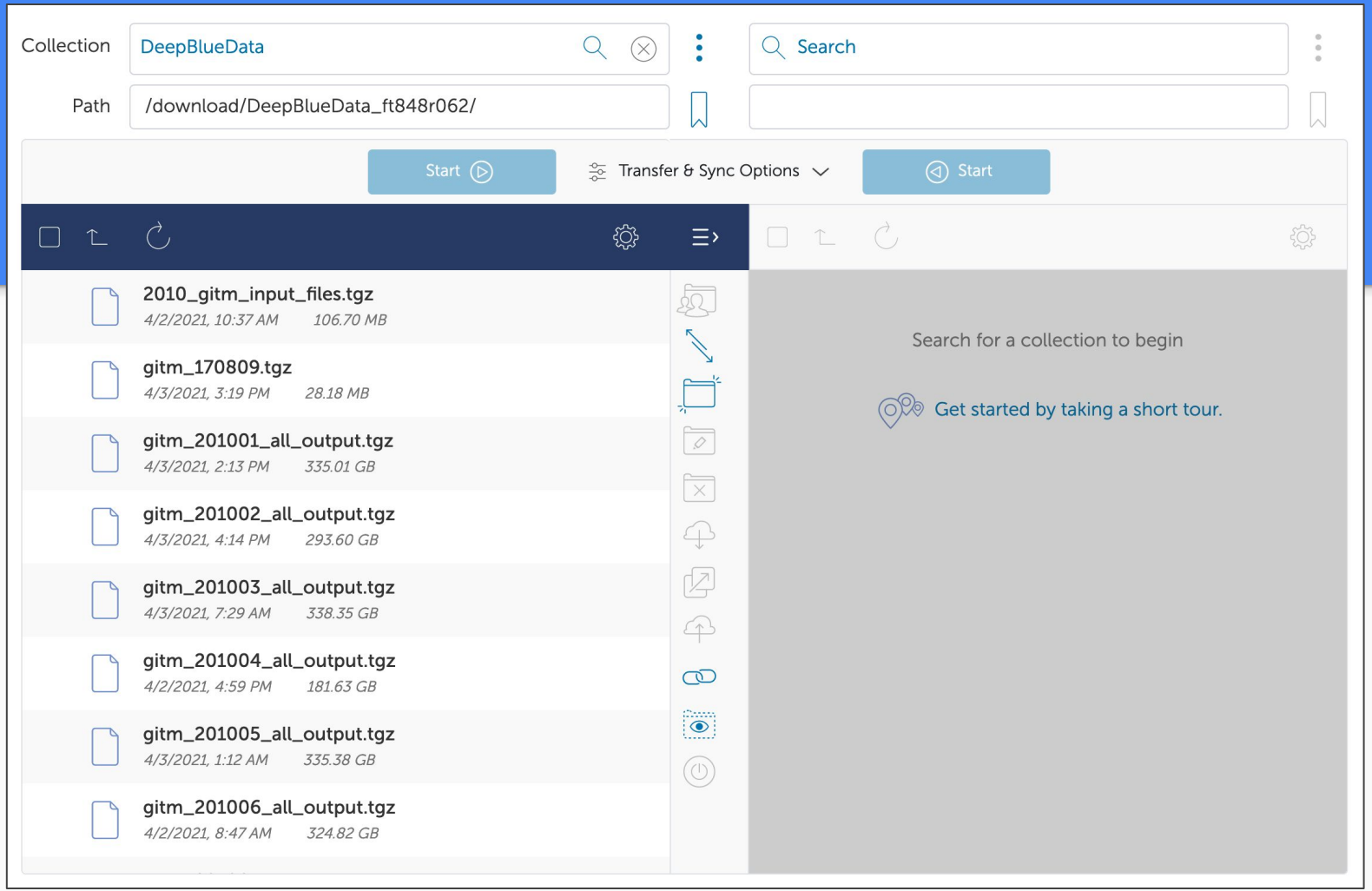
Download All Files (To download individual files, select them in the “Files” panel above)

Total work file size of 3.56 TB is too large to download directly. Consider using Globus (see below).

Files are ready Download Data from Globus

https://deepblue.lib.umich.edu/data/concern/data_sets/ft848r062?locale=en

In Globus

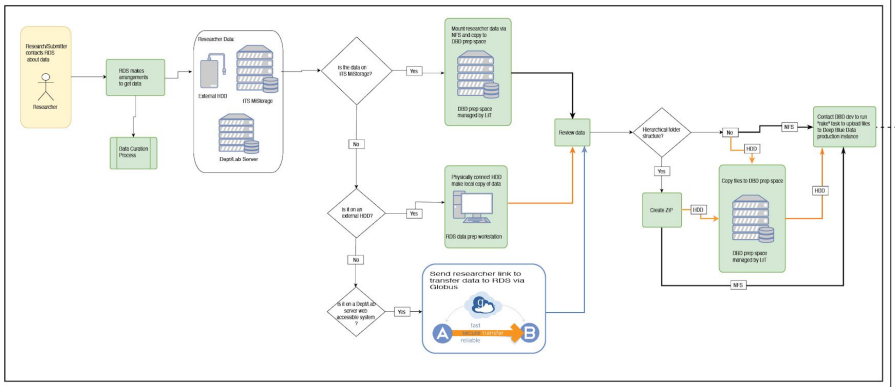

Collection
DeepBlueData
Search
Path
/download/DeepBlueData_ft848r062/
Start
Transfer & Sync Options
Start
2010_gitm_input_files.tgz
4/2/2021, 10:37 AM 106.70 MB
gitm_170809.tgz
4/3/2021, 3:19 PM 28.18 MB
gitm_201001_all_output.tgz
4/3/2021, 2:13 PM 335.01 GB
gitm_201002_all_output.tgz
4/3/2021, 4:14 PM 293.60 GB
gitm_201003_all_output.tgz
4/3/2021, 7:29 AM 338.35 GB
gitm_201004_all_output.tgz
4/2/2021, 4:59 PM 181.63 GB
gitm_201005_all_output.tgz
4/3/2021, 1:12 AM 335.38 GB
gitm_201006_all_output.tgz
4/2/2021, 8:47 AM 324.82 GB
Search for a collection to begin
Get started by taking a short tour.

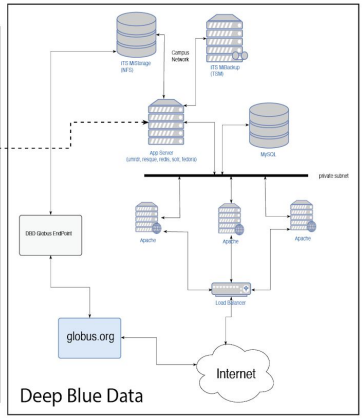
Data Sharing Workflow for Large Datasets (>1TB) with Globus

Susan Borda - Data Workflows Specialist University of Michigan - sborda@umich.edu

Research Data Services

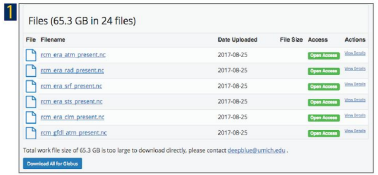
Workflow for Large Datasets





Globus File Download Process

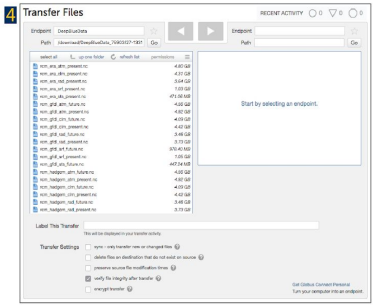
1



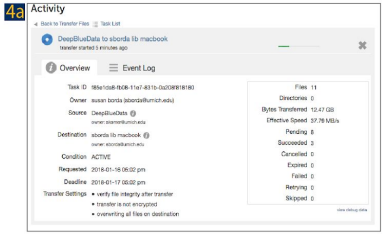
2

3

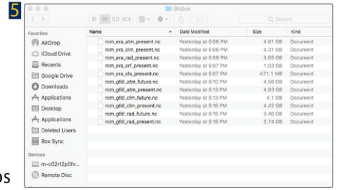
4



4a



5



Hyrax

Fedora

samvera

globus

Globus Code on Github:
https://github.com/mlibrary/deepblue/tree/master/app/jobs

Complete workflow

https://www.globus.org/

LIBRARY

Questions?
sborda@umich.edu