Publishing Large Data on ESS-DIVE's Tier 2 Storage Service

Fianna O'Brien

Computer Systems Engineer

ESS-DIVE BERKELEY LAB NERSC NCEAS DataONE U.S. DEPARTMENT OF ENERGY Office of Science

September 2023 Webinar

Objectives

- What is considered “large data”?
- **Overview** data submission tools
- What are Tier 2 and Globus?
- Is my data suitable for large data support?

*This session is designed for **data managers and data contributors***

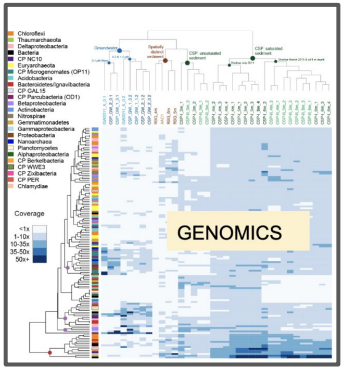
What is “Large Data” at ESS-DIVE?

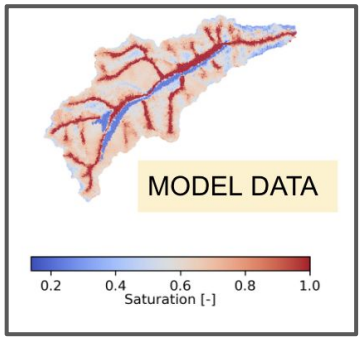
Large data defined

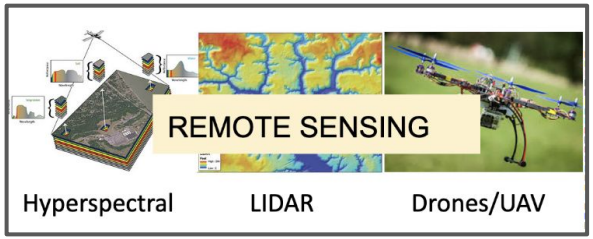
Any file or set of files whose total size is greater than **500GB** is considered **large data** by ESS-DIVE.

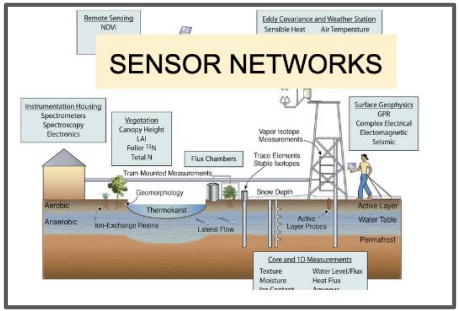
- Cannot be uploaded via the UI (limit 10 GB)
- Cannot be uploaded via the Package Service API (limit 500GB)
- Requires special handling

Common Types of Large Data







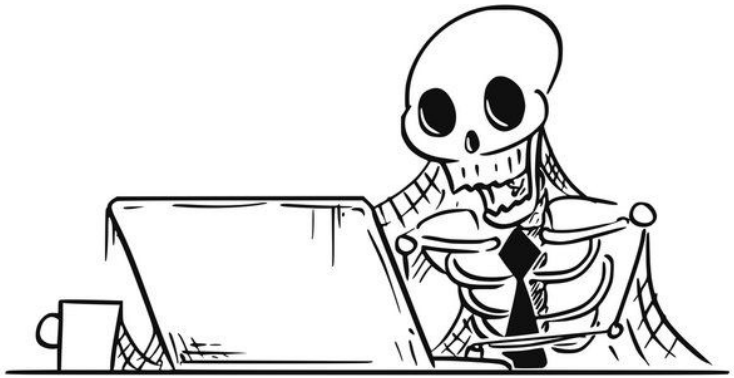


Under 500GB and still bottlenecked?

Users having difficulties with uploads can benefit from the large data tools, no matter the data size.

Reasons why data may be difficult to upload

- Available local memory is smaller than data volume.
- You have many files that are too large to be uploaded all at once but it’s too tedious to upload them in tens of batches
- Unstable Internet Connection creating timeout issues or errors



Overview of Large Data Support Tools

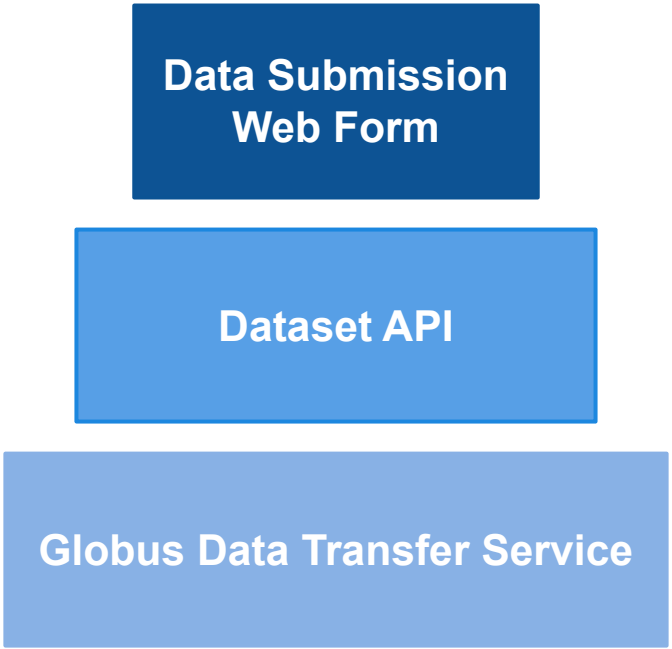
Today's webinar will focus on two new tools to the ESS-DIVE ecosystem:

- **Globus**: a data submission tool for moving or transferring large volumes of data
- **Tier 2**: ESS-DIVE's secondary storage location for publication and data access

Both are offline services that can be setup for a dataset by ESS-DIVE Support. Expect that publishing time using these tools will take longer than data uploaded directly.

ESS-DIVE's Data Submission Tools

Submission Tools



Data Submission Tools

Upload Method	File volume
Data Submission Web Form	< 10 GB
Dataset API	< 500 GB
Globus Data Transfer Service	> 500 GB

File Volume

Represents the amount of data each tool can upload at once.

For example, with the web form if you select one file at a time then each file can be up to 10GB. If you select multiple files at once, then the total volume can be up to 10GB

Data Submission Tools

Upload Method	File volume	No. of files
Data Submission Web Form	< 10 GB	< 100
Dataset API	10 - 500 GB	< 100
Globus Data Transfer Service	> 500 GB	> 100

No. of Files

Represents the number of files you may publish in a dataset.

Larger volumes become unwieldy in non-hierarchical UI.

Data Submission Tools

Upload Method	File volume	No. of files	Considerations
Data Submission Web Form	< 10 GB	< 100	Immediate upload. Easy to follow guidance throughout form.
Dataset API	10 - 500 GB	< 100	Takes initial investment of time to set up, but immediate upload. Technical skills helpful but not required (see tutorials).
Globus Data Transfer Service	> 500 GB	> 100	Longest process, and requires collaboration with Support Team and use of a separate desktop app. Can use if unfamiliar with scripting.

Globus: Data Transfer service/submission tool for large data

Data Submission Tools

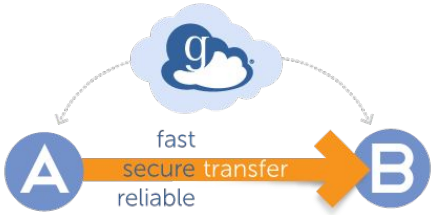
Upload Method	File volume	No. of files	Considerations
Data Submission Web Form	< 10 GB	< 100	Immediate upload. Easy to follow guidance throughout form.
Dataset API	< 500 GB	< 100	Takes initial investment of time to set up, but immediate upload. Technical skills helpful but not required (see tutorials).
Globus Data Transfer Service	> 500 GB	> 100	Longest process, and requires collaboration with Support Team and use of a separate desktop app. Can use if unfamiliar with scripting.

Globus Data **Transfer** Service

Globus is a free, non-profit cloud-based data transfer service designed to move significant amounts of data. ESS-DIVE is using this service to move data from your local desktop (or existing endpoint) to ESS-DIVE's extended resources.

- **Why?** Recommended by NERSC (LBL IT?): https://docs.nersc.gov/services/globus/
- Learn more: https://www.globus.org/

For anyone experiencing trouble uploading data files of any size, the ESS-DIVE team will work with you to upload those data files using Globus,



Large data from a **contributor** perspective

- Data is too large to download from storage location onto data manager's local system for upload
- Uploading large volumes of data on unstable internet connection is prone to failure
- Cloud storage service provider's API & tools aren't made accessible to help you move data

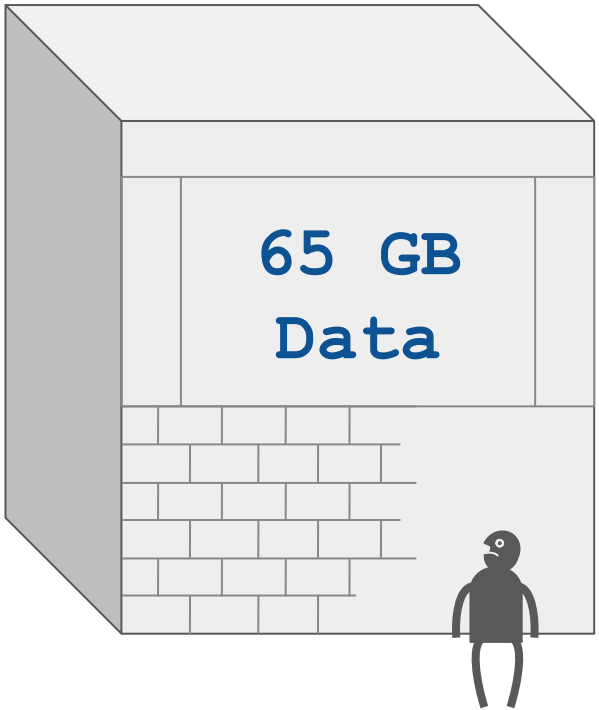
User Story

Project data was in online storage platform. It was too big to download by project data manager & too cumbersome to upload to ESS-DIVE.

Using large data tools, **ESS-DIVE team copied the data directly from online storage into ESS-DIVE**, cutting out the middleman and clearing the bottleneck.

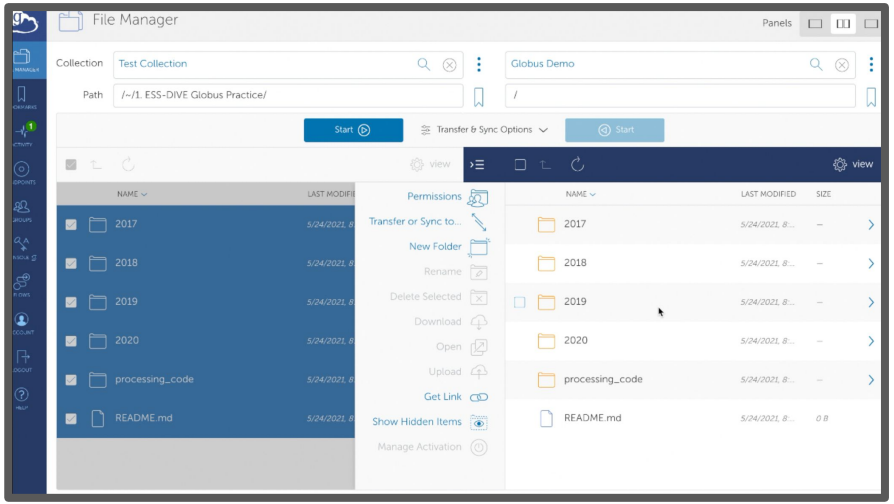
Large data from a **user** perspective

- Downloading large data files or volumes on unstable internet connection is prone to failure
- There's a mountains of files and the "Download All" button is blocked (limit 3GB) so they have to download the files individually



Why ESS-DIVE uses Globus

- Uses GridFTP, a high-performance data transfer protocol
- Maximizes bandwidth by tuning transfer parameters
- Automatic fault recovery
- Fire & Forget: Users are notified via email when data transfer is complete
- Allows for quicker and more reliable file transfers
- Allows for bulk transfer for data users

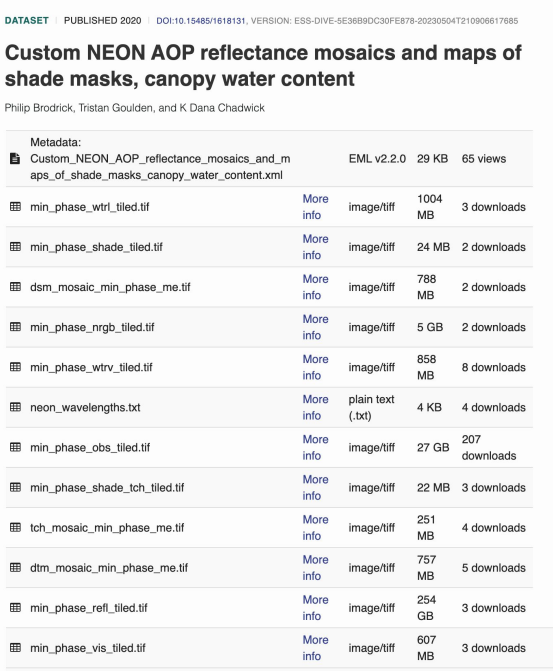


Globus File Manager Interface

How we use Globus

- Transfer data to ESS-DIVE for upload on Tier 1 (Web UI)
  - This ESS-DIVE dataset with a 254GB file was transferred via Globus https://data.ess-dive.lbl.gov/view/doi:10.15485/1618131

- Transfer data to Tier 2, ESS-DIVE's large data storage.



DATASET | PUBLISHED 2020 | DOI:10.15485/1618131, VERSION: ESS-DIVE-5E36B9DC30FE878-20230504T210906617685

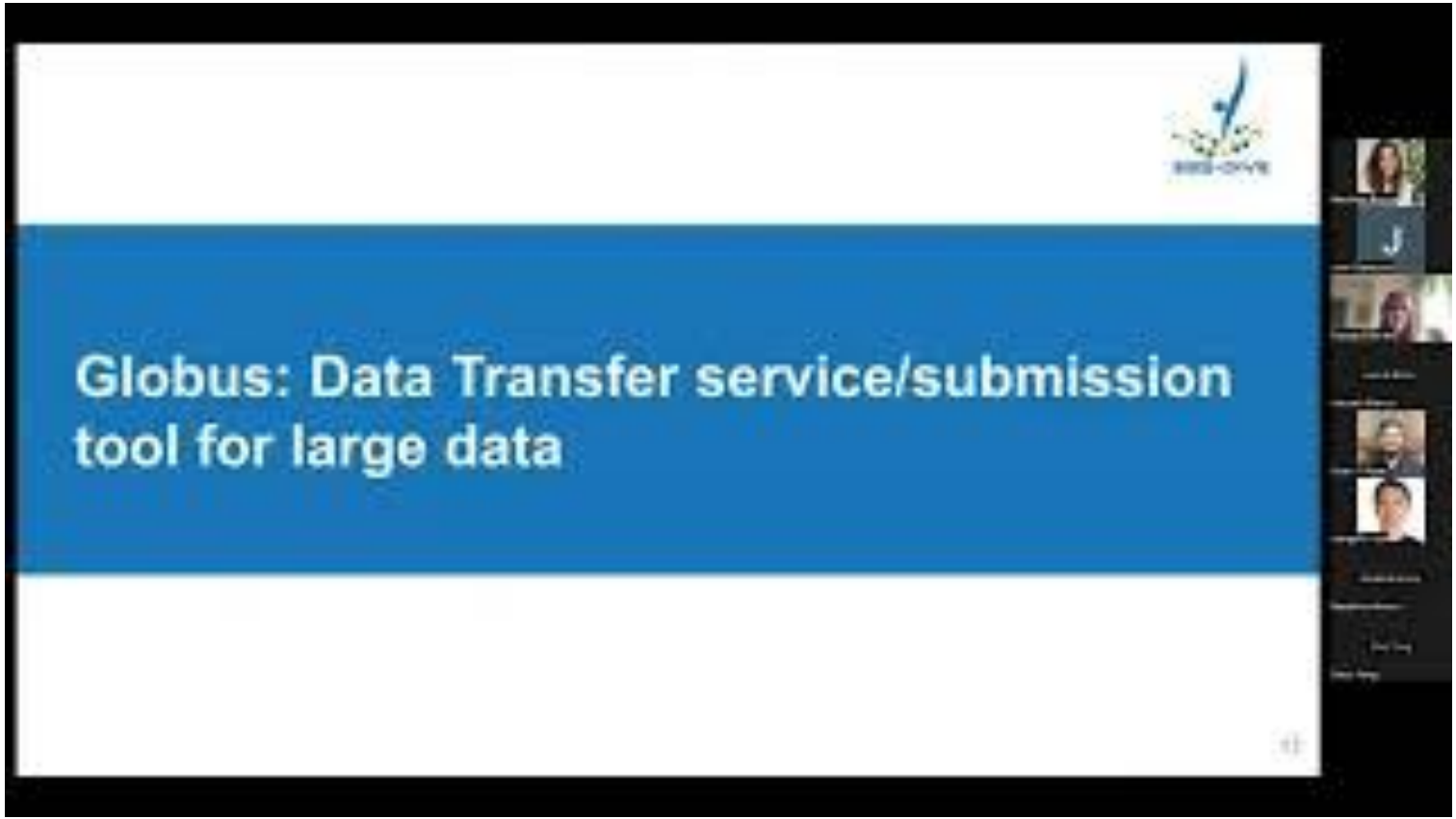
Custom NEON AOP reflectance mosaics and maps of shade masks, canopy water content

Philip Brodrick, Tristan Goulden, and K Dana Chadwick

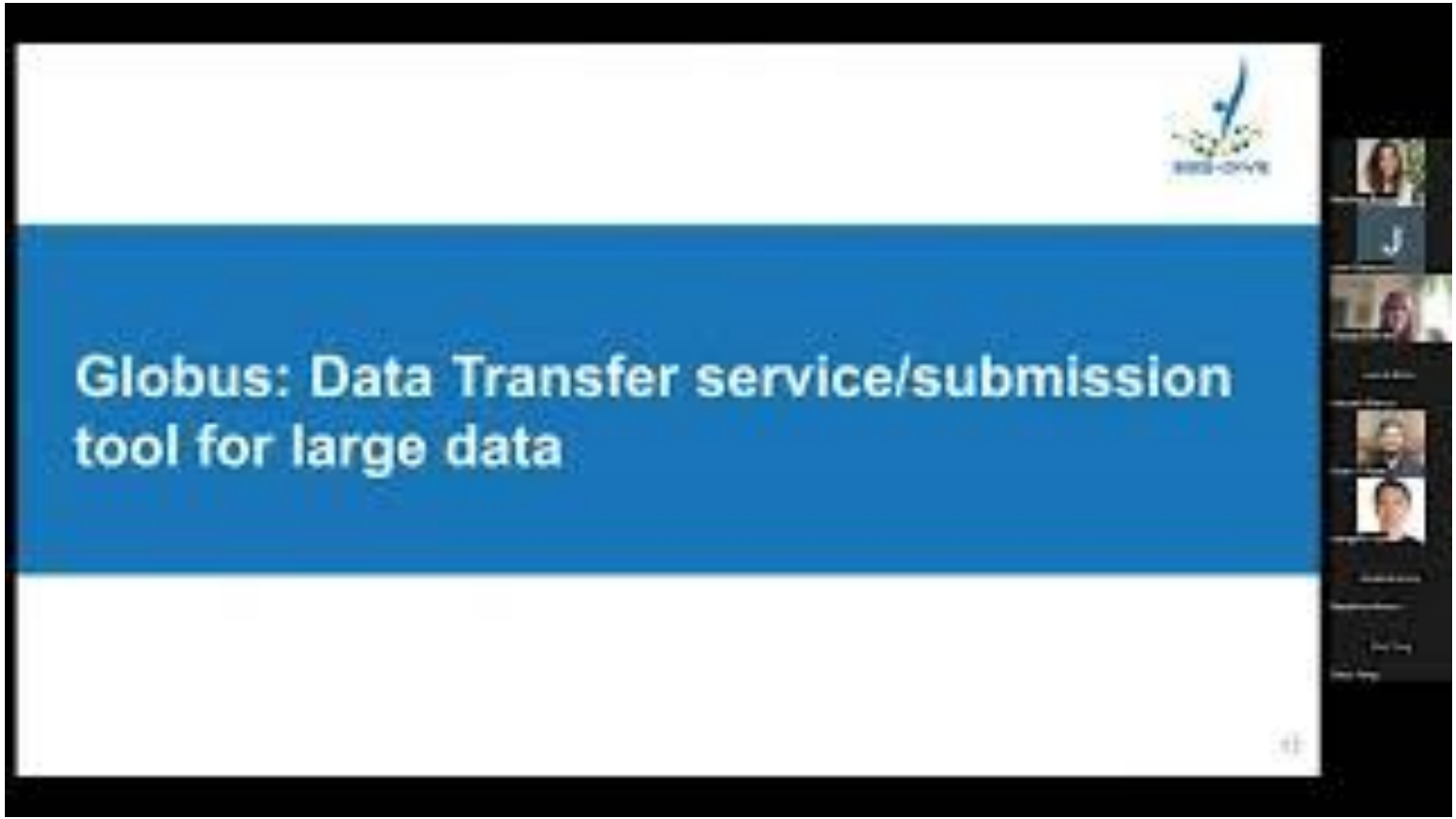
Metadata: Custom_NEON_AOP_reflectance_mosaics_and_maps_of_shade_masks_canopy_water_content.xml		EML v2.2.0	29 KB	65 views
min_phase_wtrl_tiled.tif	More info	image/tiff	1004 MB	3 downloads
min_phase_shade_tiled.tif	More info	image/tiff	24 MB	2 downloads
dsm_mosaic_min_phase_me.tif	More info	image/tiff	788 MB	2 downloads
min_phase_nrgb_tiled.tif	More info	image/tiff	5 GB	2 downloads
min_phase_wtrv_tiled.tif	More info	image/tiff	858 MB	8 downloads
neon_wavelengths.txt	More info	plain text (.txt)	4 KB	4 downloads
min_phase_obs_tiled.tif	More info	image/tiff	27 GB	207 downloads
min_phase_shade_tch_tiled.tif	More info	image/tiff	22 MB	3 downloads
tch_mosaic_min_phase_me.tif	More info	image/tiff	251 MB	4 downloads
dtm_mosaic_min_phase_me.tif	More info	image/tiff	757 MB	5 downloads
min_phase_refl_tiled.tif	More info	image/tiff	254 GB	3 downloads
min_phase_vis_tiled.tif	More info	image/tiff	607 MB	3 downloads

Data transfer to ESS-DIVE via Globus

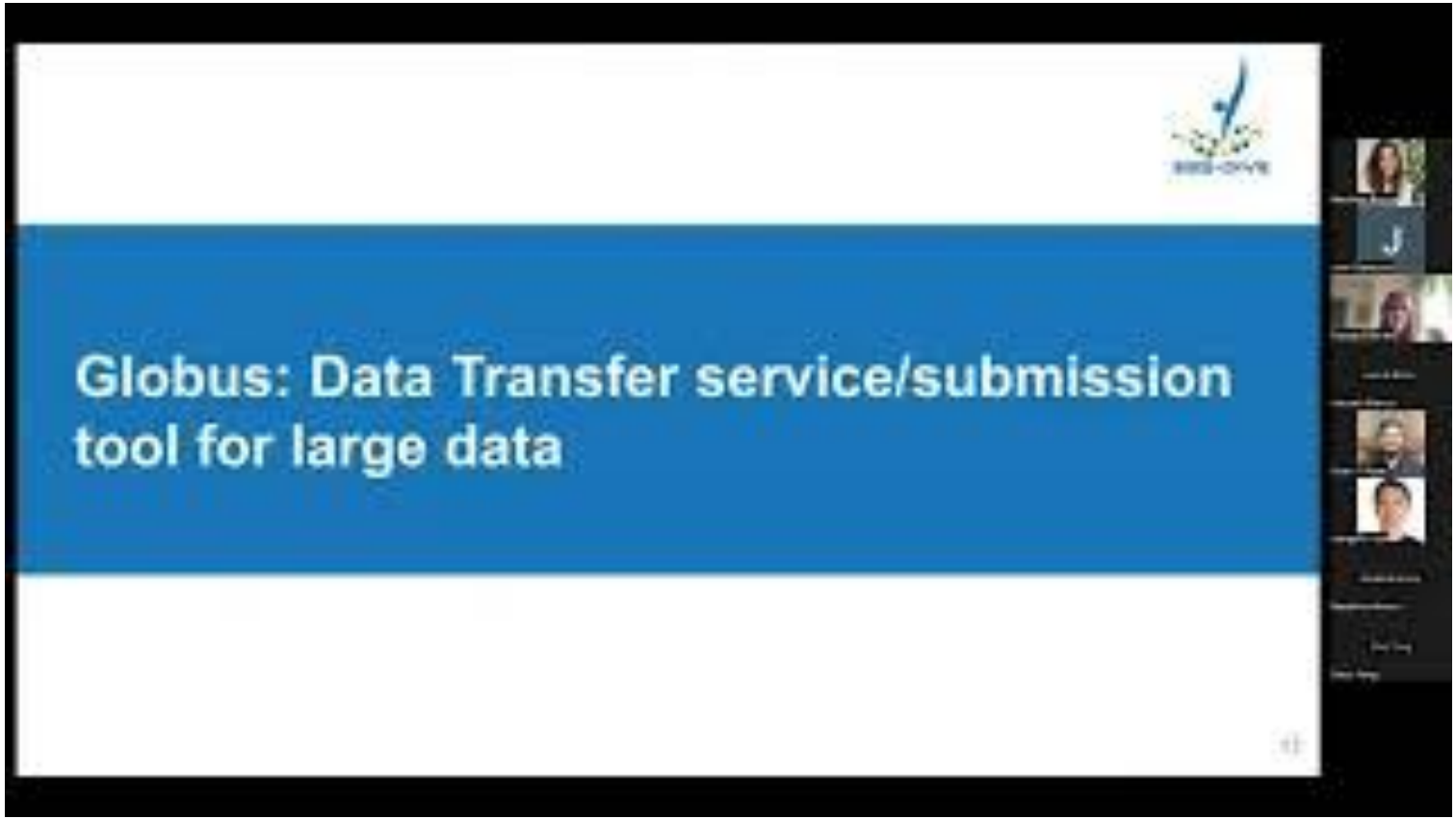
Using Globus: Creating a Globus Account



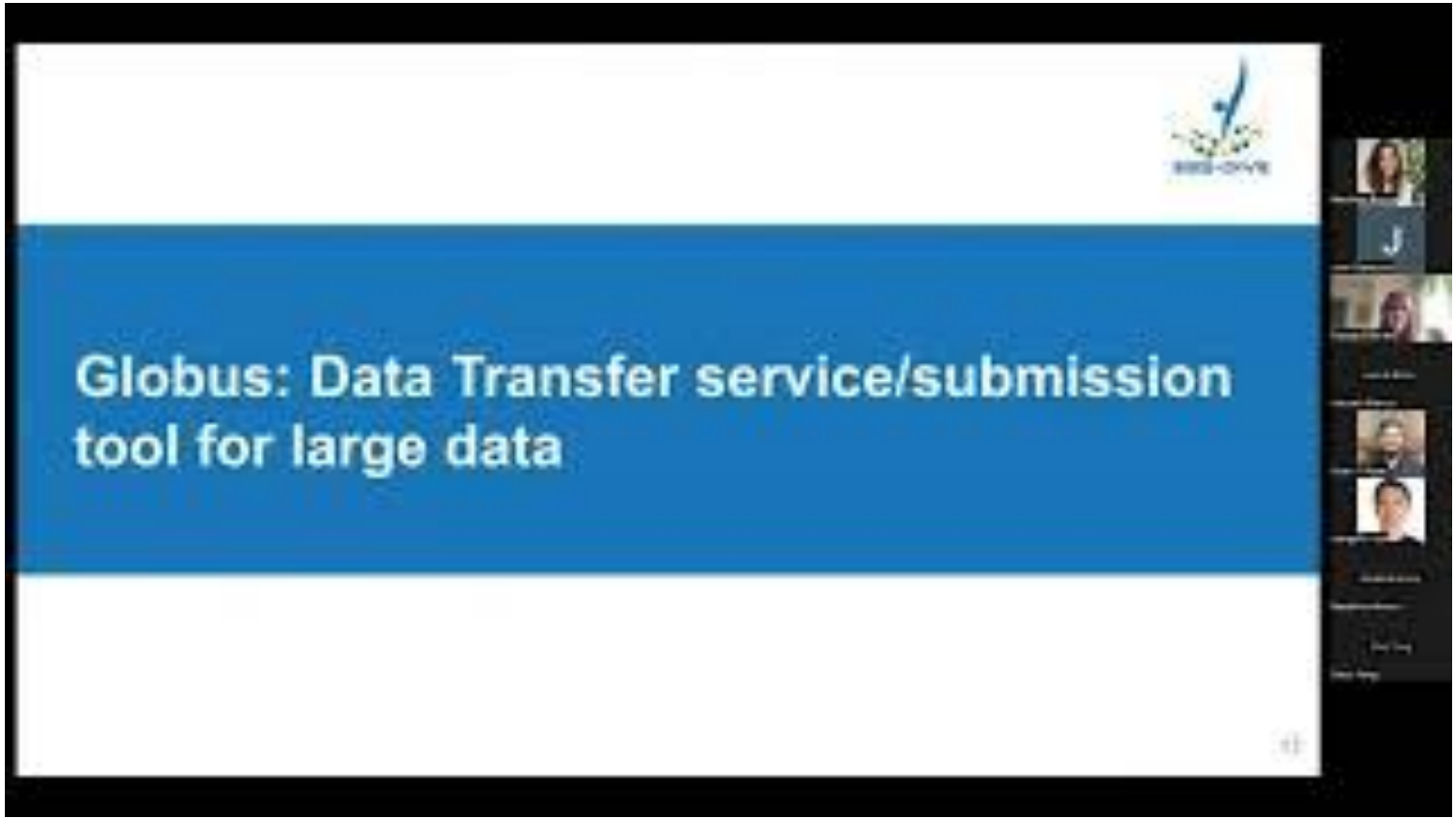
Using Globus: Setting up Globus Desktop



Using Globus: Creating a Collection



Using Globus: Sharing Data to ESS-DIVE

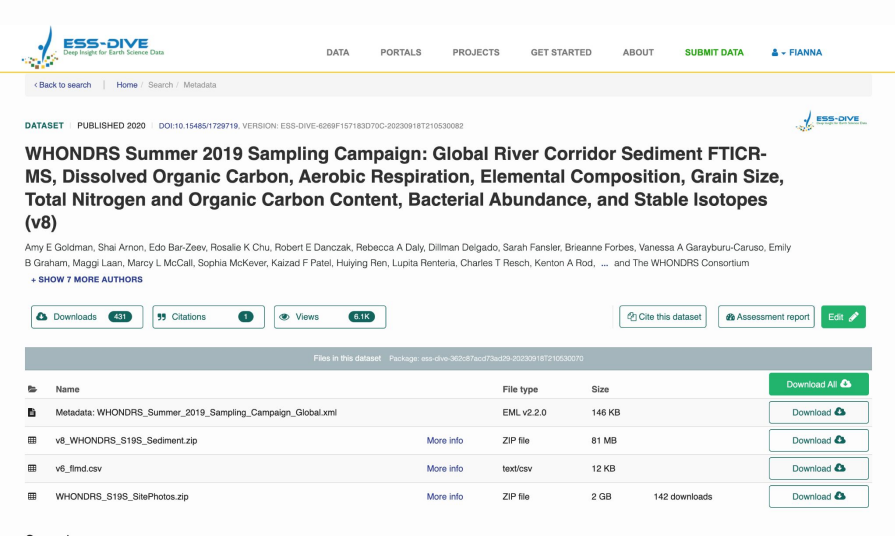


ESS-DIVE Data Storage

Tier 1 Data Storage

This is the storage underlying the ESS-DIVE website, as seen by most users.

- Data & metadata accessible together on metadata landing page.
- Access to tools and features on metadata landing page, including file level download metrics
- Data is not browsable & folders must be zipped.

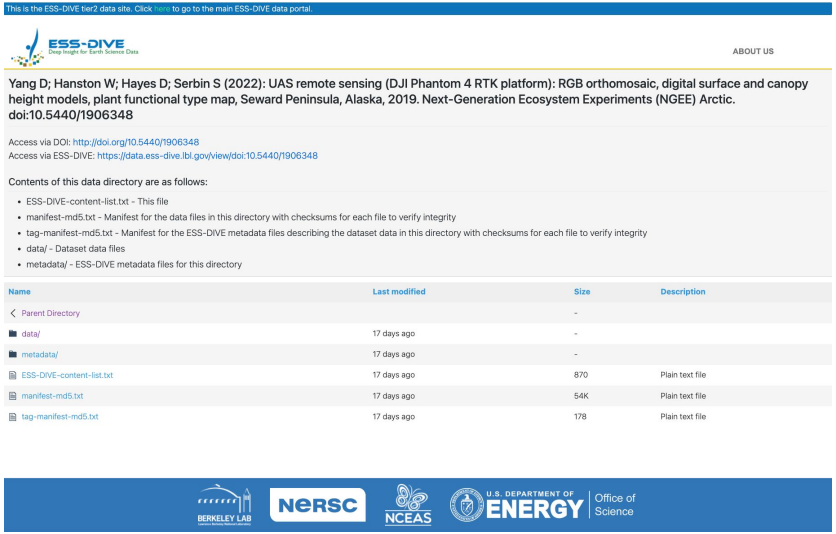


Tier 1 Data Storage

Tier 2 Data Storage

Instead of storing data directly on ESS-DIVE's dataset landing pages, very large, hierarchical datasets are stored on **Tier 2, ESS-DIVE's extended file storage.**

- Used for archiving data files greater than 500GB
- Tier 2 website supports browsing folder hierarchies
- Data downloaded from Tier 2 landing page or via Globus
- Tier 2 data is accessible via external data link on metadata landing page



Tier 2 Data Storage

Tier 1 vs. Tier 2 Data Storage

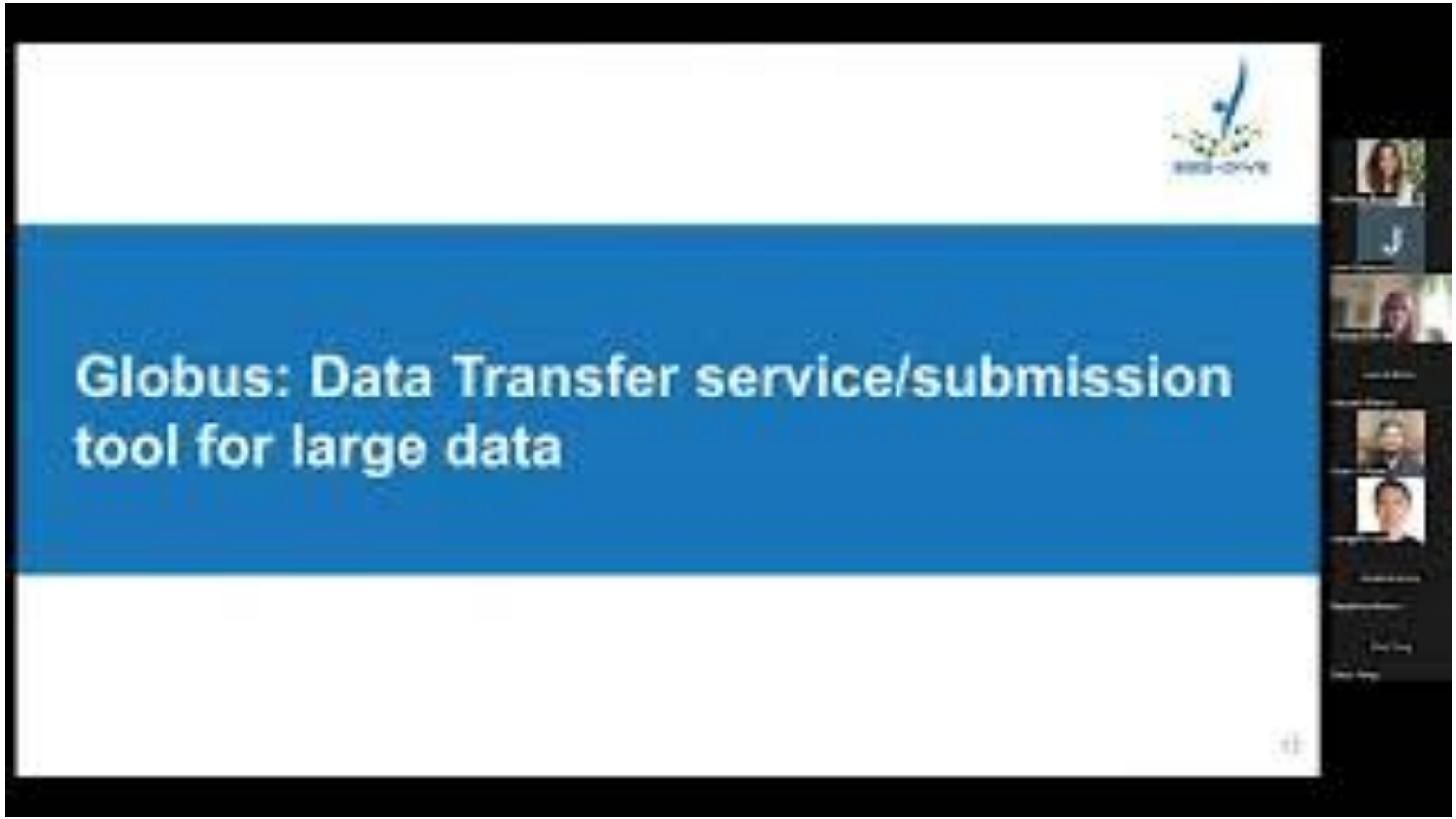
Feature	Tier 1	Tier 2
File size maximum	< 500 GB	> 500 GB
File Volume	< 100	> 100
Data Structure	Flat	Hierarchical

Tier 1 vs. Tier 2 Data Storage

Feature	Tier 1	Tier 2	Considerations
Metadata stored with data	**Yes**	**No** **Stored on Tier 1**	Full dataset metadata provides needed context for data users
Download Metrics	**Yes**	**No**	Download metrics help to highlight how your data is being used
New Features & Functionality	**High Priority**	**Lower Priority**	ESS-DIVE is constantly expanding and improving with new features on Tier 1 that may not be supported on Tier 2

Tier 2: Large Data Storage Service

Tier 2 Demo


Globus: Data Transfer service/submission tool for large data

Tier 2: When to use it

- If your data is large enough to require Tier 2, we are excited to support you.
- **We are constantly growing out Tier 1 services** and these are just our current constraints.

If you feel like your data is large enough to require special handling, reach out to the ESS-DIVE team to discuss if Tier 1 or Tier 2 is the best fit for your data.

Summary OR
Is my data suitable for large data support?

Summary

Globus is a means for supporting data **contributors** trying to upload large data

Tier 2 is a means for supporting data **users** accessing large data

Common types of sources of large data include model data, remote sensing, genomics, sensor network data, and more.

But what about my data?

Key is to talk to ESS-DIVE support and help us learn more about your data – we can help make the determination for you

How to ask ESS-DIVE Support

Reach out to ESS-DIVE support at **essdive-support@lbl.gov**

Provide the answers to the following questions to help us determine if your data requires large data support:

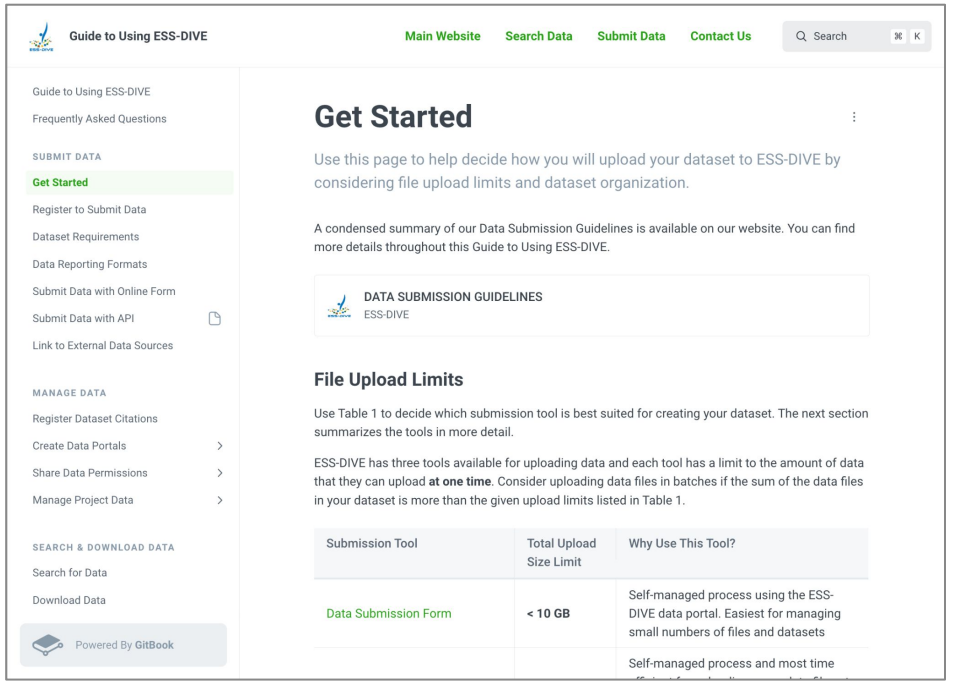
- What's the total file volume of your dataset?
- How many files are in your dataset? What's the range of file sizes?
- Is the data structure hierarchical? Can the data structure be easily flattened?
- Where is your data stored currently (e.g. local desktop, cloud, Google Drive)?
- What’s your available local storage and processing?
- Is your local internet connection strong enough to prevent timeouts?

Large Data Support Documentation

Large Data Support Documentation coming soon

- Contacting ESS-DIVE about large data support
- Using Globus for data transfer
- Navigating Tier 1 & Tier 2
- Management & Preservation of Tier 2
- Downloading Data from Tier 2 & Globus

Find our documentation at docs.ess-dive.lbl.gov



ESS-DIVE Documentation Website

Any Questions?

Photo by Towfiqu barbhuiya on Unsplash

Connect With Our Team!

To get help:

ess-dive.lbl.gov

ess-dive-support@lbl.gov

docs.ess-dive.lbl.gov

To stay updated:

ess-dive-community@lbl.gov

@essdive

https://bit.ly/essdiveMailingList

Acknowledgements

Advisory Groups: ESS-DIVE Archive Partnership Board, ESS Cyberinfrastructure Working Groups

Funding: EESSD Data Management