How to Setup and Use Globus with AWS S3 Buckets

Last updated 5/5/22

Contents

Using Globus to Transfer Files with S3 Object Store Buckets

Globus is a Software-as-a-Service (SaaS) that provides file transfer (copy) and sharing services, as well as identity, profile, and group management. It provides a high performing and secure method to transfer data between endpoints. A Globus transfer handles all the difficult aspects of data transfer by optimizing bandwidth usage, managing security configurations, providing automatic fault recovery, and notifying users of completion and problems. More information about Globus data transfer can be found at **https://www.globus.org/data-transfer**.

What can I do with Globus?

1. **Transfer files:** From kilobytes to petabytes, with Globus you can efficiently, reliably, and securely move (copy) data between systems within your site or across an ocean
2. **Share files with others:** All you need is an email address to share data with colleagues – Globus manages authentication and access
3. **Publish data:** Easily turn your data into a curated collection with metadata and identifiers, discoverable by others, under your full control and on storage of your choice
4. **Develop applications and gateways:** open REST APIs and Python SDK empower you to create an integrated ecosystem of research data services, applications, and workflows

Once you have setup an AWS S3 object store bucket, you can install and begin using Globus.

Setting Up a Globus Account and Setup Globus Connect Personal

1. Go to https://www.globus.org/globus-connect
2. Select **Get Globus Connect Personal**
3. On the right side, select **Install Globus Connect Personal** based on your workstation's operating system (Mac, Windows, Linux)
4. Follow the onscreen instructions.

Create a Duke Account

5. Register for an account at https://app.globus.org - select Duke University from the drop-down menu of institutions with organizational Globus accounts and Continue.

6. You will be redirected to Duke's login page. Use your credentials (NETID and Password) to login.

7. Once you've logged in with your organization, Globus will ask if you'd like to link to an existing account. If this is your first time logging in to Globus, click "Continue." If you've already used another account with Globus, you can choose "Link to an existing account."

8. You may be prompted to provide additional information such as your organization and whether or not Globus will be used for commercial purposes. Complete the form and click "Continue."

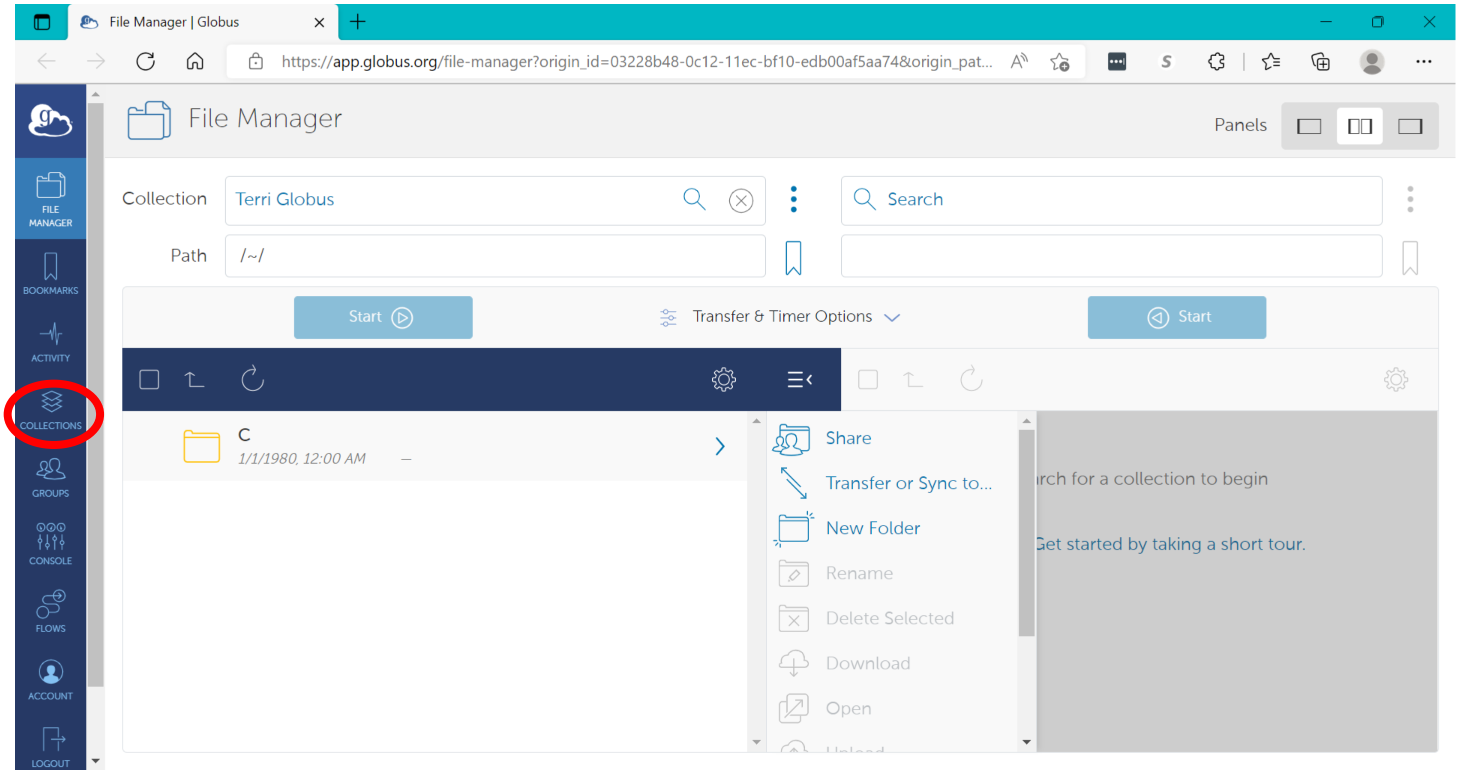
9. Finally, you need to give Globus permission to use your identity to access information and perform actions (like file transfers) on your behalf.

10. After you've signed up and logged in to Globus, you'll begin at the **File Manager**.

Setting Up a Collection of Data on the DHE AWS S3 Object Store Buckets

To copy and move data between S3 object store buckets or from a network mapped Isilon File Share to/from a Duke AWS S3 Object Store bucket, you need to setup a Collection on the **Duke Health Proof of Concept – Cloud S3** node.

Click "**Collections**" in the left-hand pane.

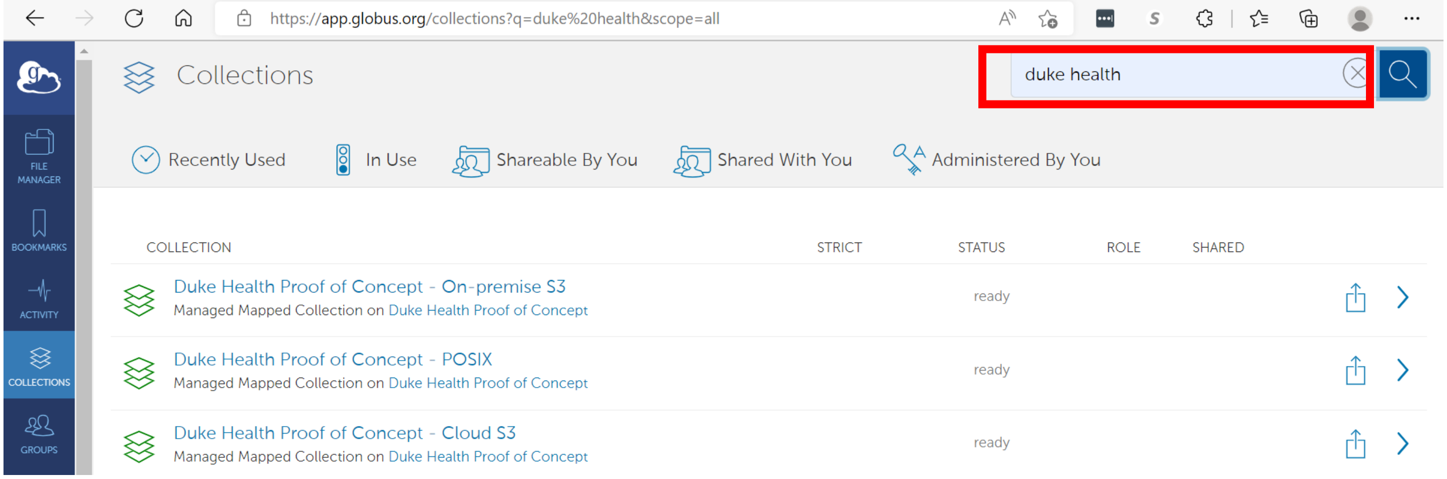


Collection: Terri Globus

Path: /~/

C	1/1/1980, 12:00 AM	–

On the Collections page, type in "Duke Health Proof of Concept – Cloud S3" in the search bar at the top of the page.



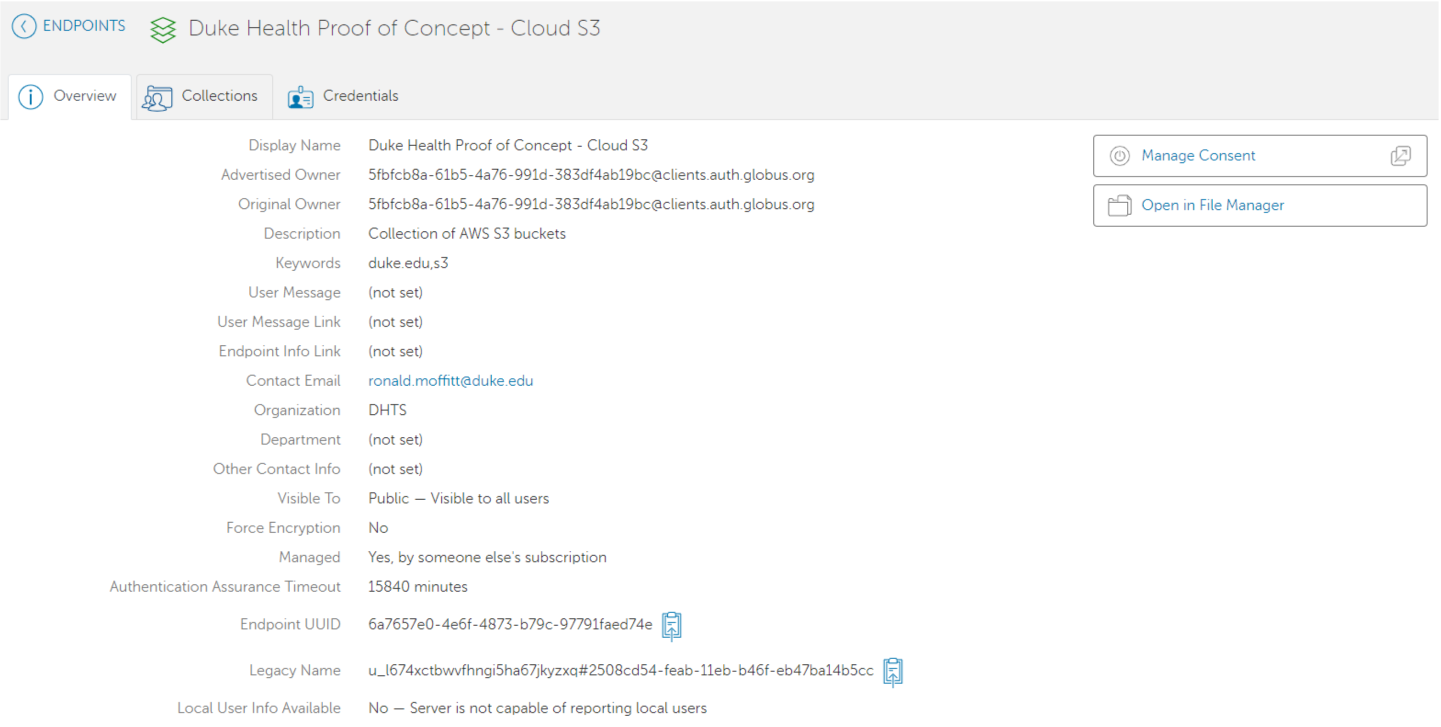
Search: duke health

COLLECTION	STRICT	STATUS	ROLE	SHARED
Duke Health Proof of Concept - On-premise S3 Managed Mapped Collection on Duke Health Proof of Concept		ready		
Duke Health Proof of Concept - POSIX Managed Mapped Collection on Duke Health Proof of Concept		ready		
Duke Health Proof of Concept - Cloud S3 Managed Mapped Collection on Duke Health Proof of Concept		ready		

Other popular Duke Collections include:

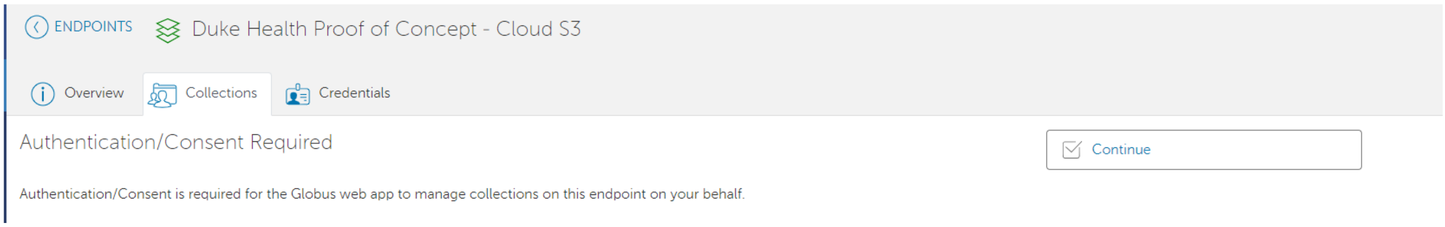
- **Duke Health Proof of Concept (for access to the S3 object store buckets)**
- Duke Compute Cluster (DCC) Data Transfer Node
- Duke Research Data Repository (RDR) Data Transfer Node
- Duke HARDAC Globus Transfer Node

Select the **Duke Health Proof of Concept – Cloud S3** node.



Display Name	Duke Health Proof of Concept - Cloud S3
Advertised Owner	5fbfcb8a-61b5-4a76-991d-383df4ab19bc@clients.auth.globus.org
Original Owner	5fbfcb8a-61b5-4a76-991d-383df4ab19bc@clients.auth.globus.org
Description	Collection of AWS S3 buckets
Keywords	duke.edu,s3
User Message	(not set)
User Message Link	(not set)
Endpoint Info Link	(not set)
Contact Email	ronald.moffitt@duke.edu
Organization	DHTS
Department	(not set)
Other Contact Info	(not set)
Visible To	Public — Visible to all users
Force Encryption	No
Managed	Yes, by someone else's subscription
Authentication Assurance Timeout	15840 minutes
Endpoint UUID	6a7657e0-4e6f-4873-b79c-97791faed74e
Legacy Name	u_l674xctbwvfhngi5ha67jkyzxq#2508cd54-feab-11eb-b46f-eb47ba14b5cc
Local User Info Available	No — Server is not capable of reporting local users

1. Select **Collections**.
2. On the Authentication/Consent Required page, select **Continue.**



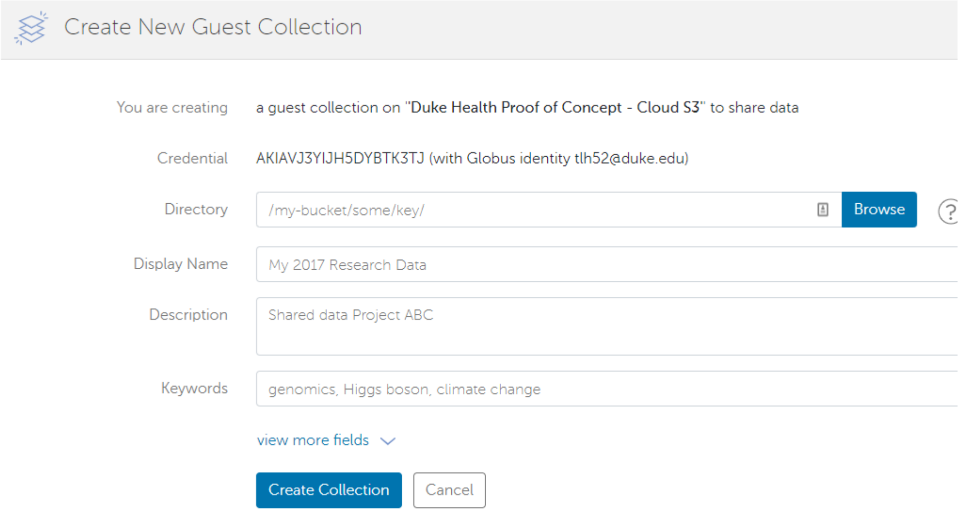
Authentication/Consent Required

Authentication/Consent is required for the Globus web app to manage collections on this endpoint on your behalf.

3. Enter your S3 Credentials (keys) for the AWS S3 Cloud bucket. (API id and secret key from Portunus https://portunus-eso.duhs.duke.edu/portunus/keys)
4. Select **Add a Guest Collection**.
5. Under Create New Guest Collection, enter the appropriate field values for:

   Directory: /bucket name/ *example: /dh-dusom-oasis-twest/*
   Display Name: *recommend to be same name of bucket*

6. Select **Create Collection**.



Create New Guest Collection

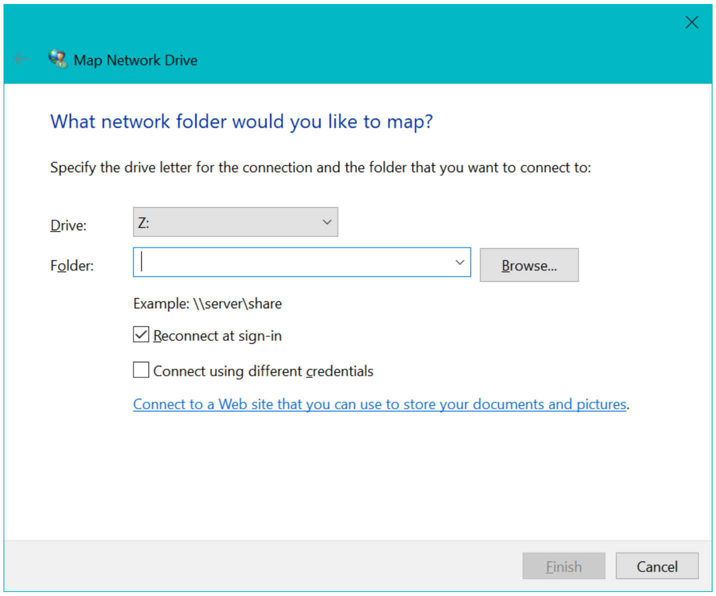
You are creating	a guest collection on "Duke Health Proof of Concept - Cloud S3" to share data
Credential	AKIAVJ3YIJH5DYBTK3TJ (with Globus identity tlh52@duke.edu)
Directory	/my-bucket/some/key/
Display Name	My 2017 Research Data
Description	Shared data Project ABC
Keywords	genomics, Higgs boson, climate change

To Setup Isilon File Share as a Collection

To setup an Isilon File Share as a collection to which you can move data to/from the AWS S3 bucket, you need to make sure that the Isilon File Share is mounted to the device on which you have the Globus client installed.

To Mount a Network Drive on Windows:

1. In Explorer, right click on "This PC" and select, "Map network drive…".
2. Select a Drive letter that is not being used; in the Example below: "Z".
3. In Folder: enter the path to the network file share; e.g., \\ ifs\Duke University Health Systems\DUSOM\West_globusarchive\All_Staff\
4. Select Finish.



Map Network Drive

What network folder would you like to map?

Specify the drive letter for the connection and the folder that you want to connect to:

Drive: Z:

Folder: | Browse...

Example: \\server\share

☑ Reconnect at sign-in

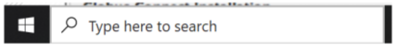
☐ Connect using different credentials

Connect to a Web site that you can use to store your documents and pictures.

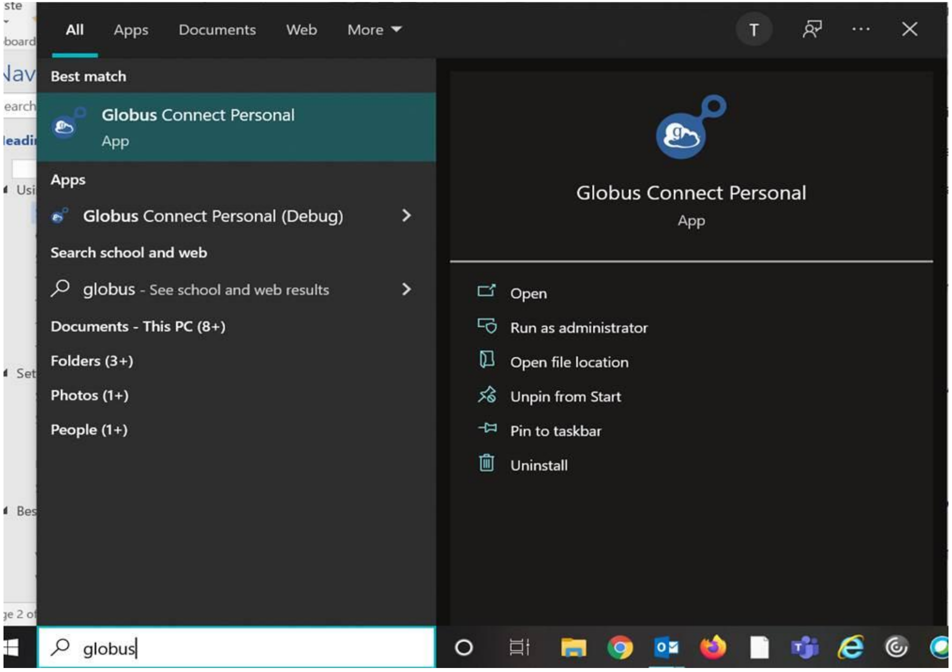
Finish Cancel

To Add an Isilon File Share to Globus:

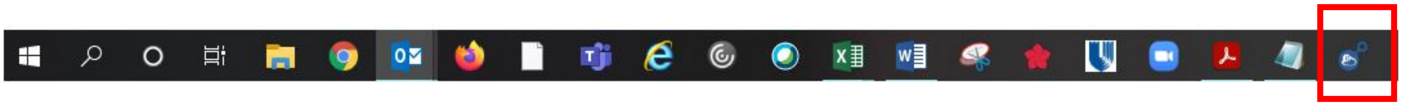
1. Go the **Start** menu (bottom left icon) in lower tray/menu and by the search, type Globus.



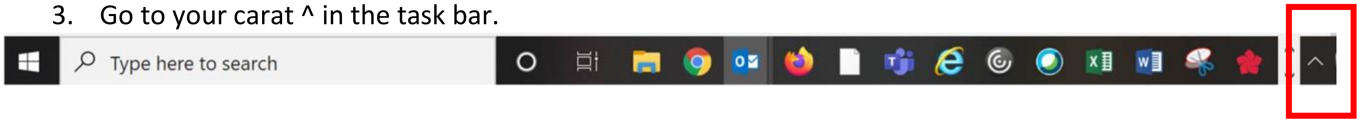
2. When Globus comes up, select Pin from taskbar and then close the black box.



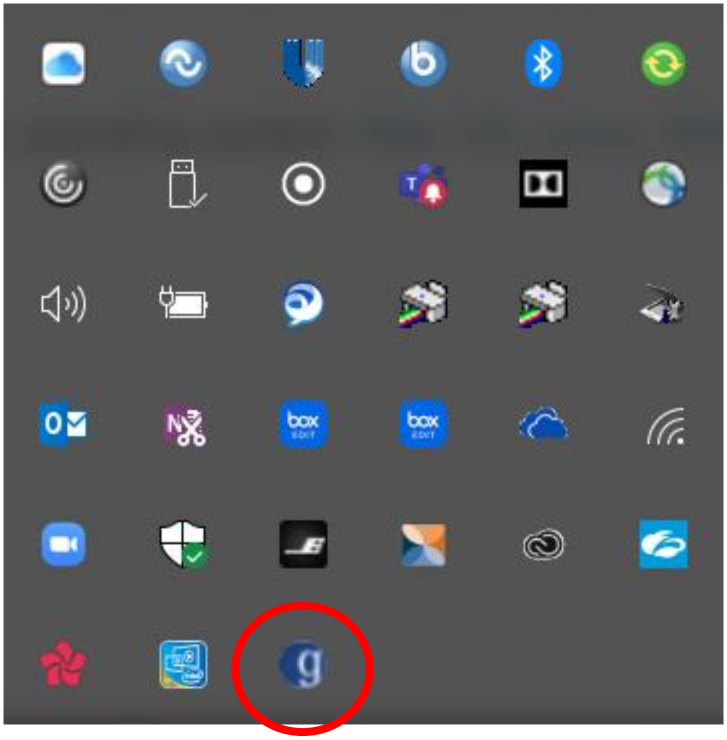
The Globus icon should appear in your task bar.



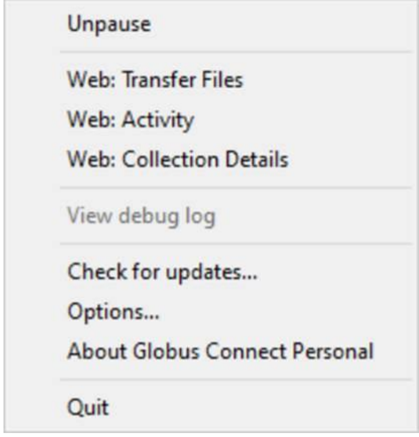
3. Go to your carat ^ in the task bar.



4. Select the carat and in the box of icons listed, LEFT click the Globus icon.



5. You should get a pop-up box that looks like this.



Unpause

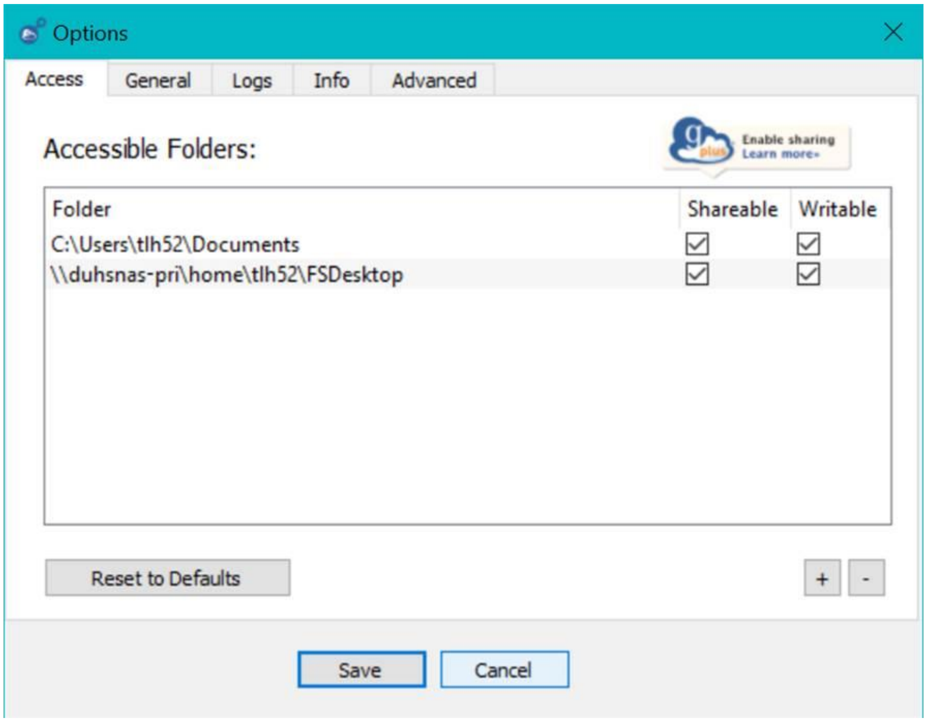
Web: Transfer Files
Web: Activity
Web: Collection Details

View debug log

Check for updates...
Options...
About Globus Connect Personal

Quit

6. Select Options and in the Access tab. click the + sign to add the path for the Isilon File Share to have access to it so that you can transfer files to/from it.



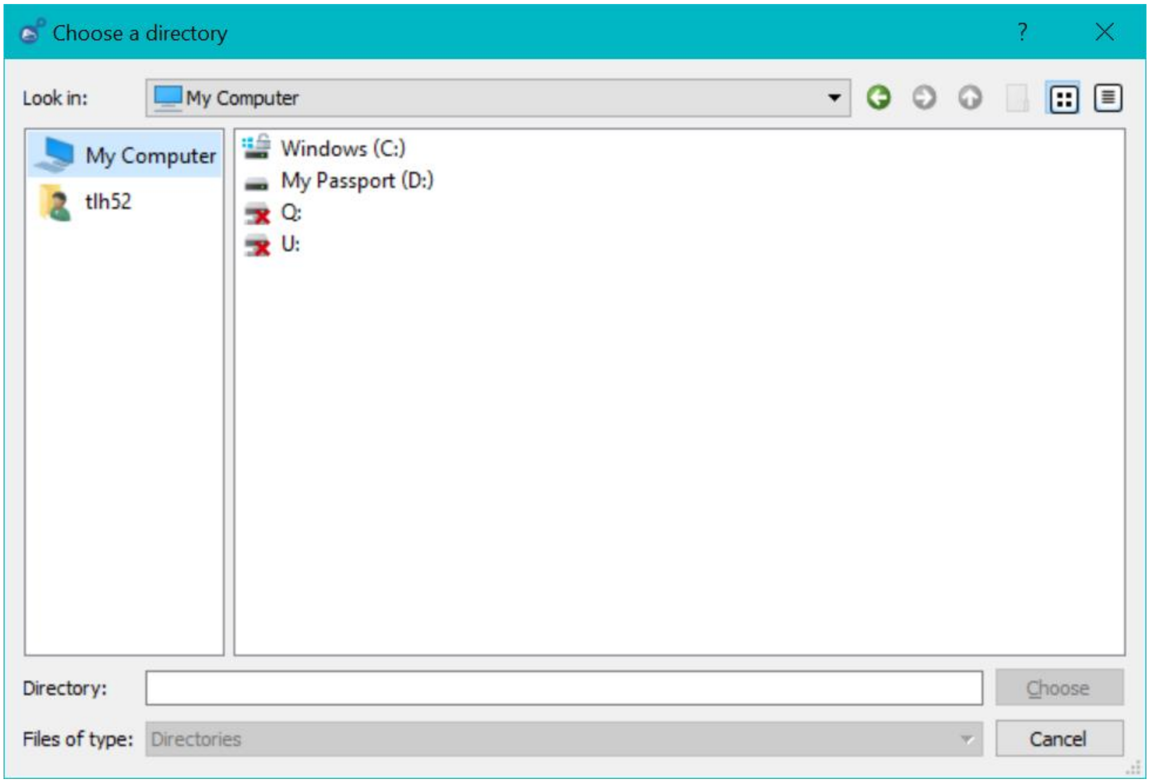
Options

Access | General | Logs | Info | Advanced

Accessible Folders:

Folder	Shareable	Writable
C:\Users\tlh52\Documents	☑	☑
\\duhsnas-pri\home\tlh52\FSDesktop	☑	☑

7. Select "My Computer" and then select the Directory for which you want to create a collection (from which to move data to/from).



Choose a directory

Look in: My Computer

- My Computer
- tlh52

- Windows (C:)
- My Passport (D:)
- Q:
- U:

Directory:

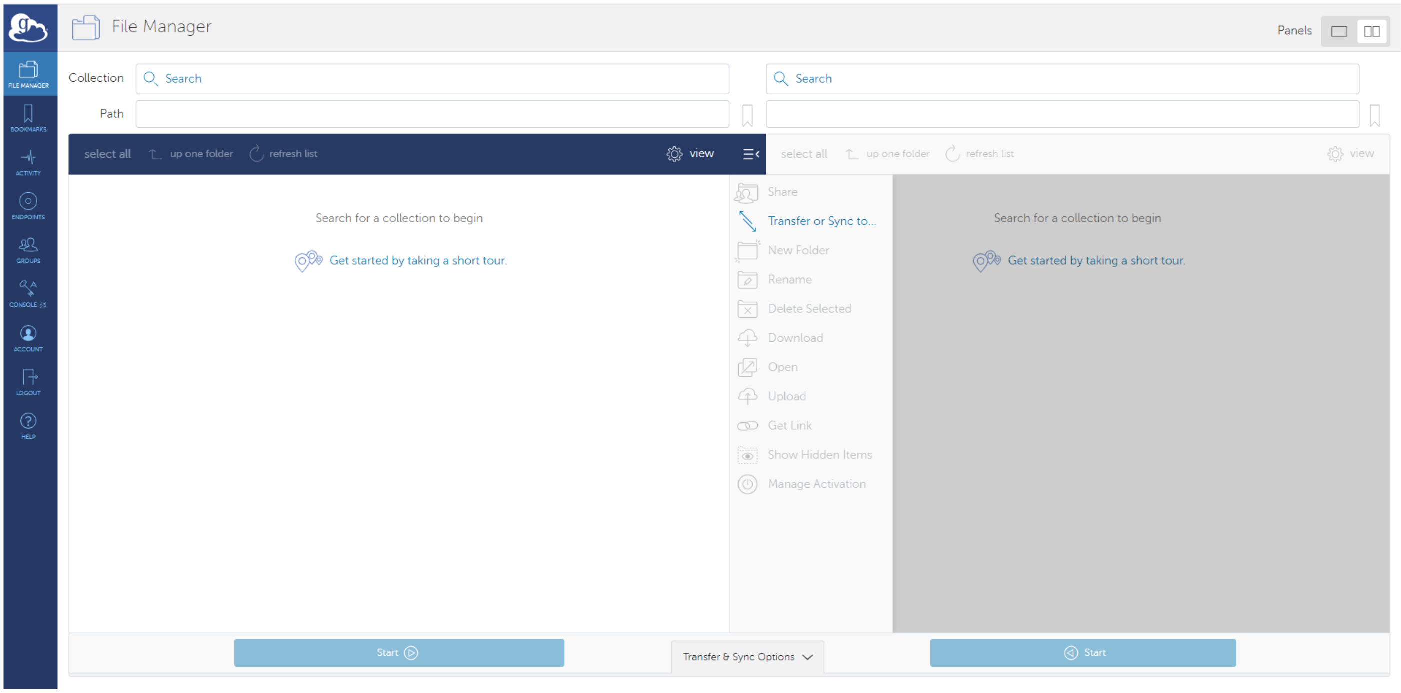
Files of type: Directories

8. Save your changes.
9. Go back to the carat ^, and right click it.
10. Then select Web: Transfer Files – and Globus will launch for you to follow the steps to setup the file transfers.

To Transfer (Copy) Data Between Collections Using Globus

1. Go to File Manager in upper left blue menu.

NOTE: You do NOT need to be on the Duke Health VPN to transfer (copy) data via Globus.



2. Select or enter the name of the source **Collection**; e.g., as in example below “Cell Biology Buckets”.

3. Select or enter the desired **Path** of the data to transfer; e.g., bucket name and path of the data; e.g., /dusom-cellbio-test/Telehealth.

4. On the right side, enter or select the name of the Destination **Collection** and the bucket name; e.g., Collection = Starfish Bucket; Path = /starfishmg/.

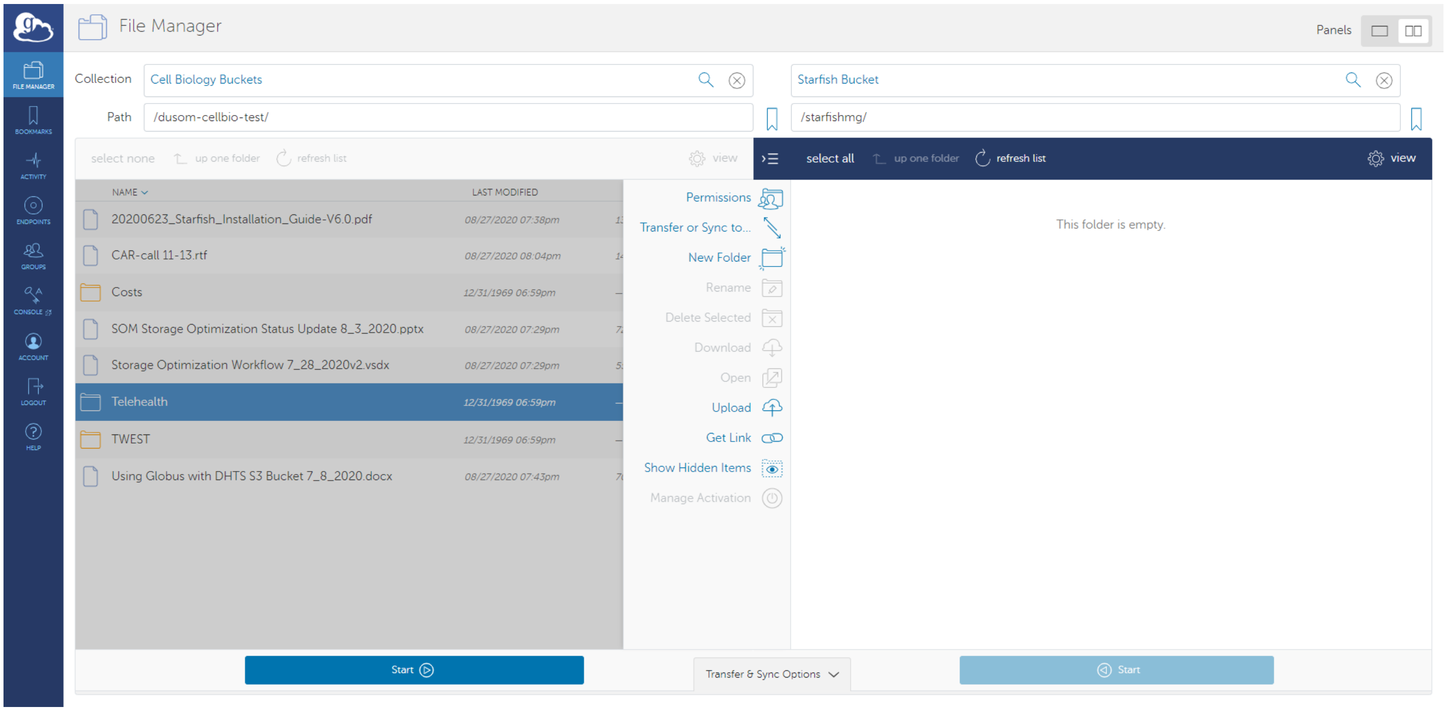
5. From the middle screen menu, select, “**Transfer or Sync to**”.

6. Select **Start** and the transfer will begin.

7. To track progress, you can select the **Activity** icon in the far left blue menu panel.

NOTE: you can log out of Globus and the transfer will continue. You do not need to monitor it.

8. You will receive an e-mail alerting you the status of the transfer, success or failure, and transfer details such as amount of data copied, transfer rates and the time taken to complete the transfer.



Collection: Cell Biology Buckets

Path: /dusom-cellbio-test/

NAME	LAST MODIFIED
20200623_Starfish_Installation_Guide-V6.0.pdf	08/27/2020 07:38pm
CAR-call 11-13.rtf	08/27/2020 08:04pm
Costs	12/31/1969 06:59pm
SOM Storage Optimization Status Update 8_3_2020.pptx	08/27/2020 07:29pm
Storage Optimization Workflow 7_28_2020v2.vsdx	08/27/2020 07:29pm
Telehealth	12/31/1969 06:59pm
TWEST	12/31/1969 06:59pm
Using Globus with DHTS S3 Bucket 7_8_2020.docx	08/27/2020 07:43pm

Collection: Starfish Bucket

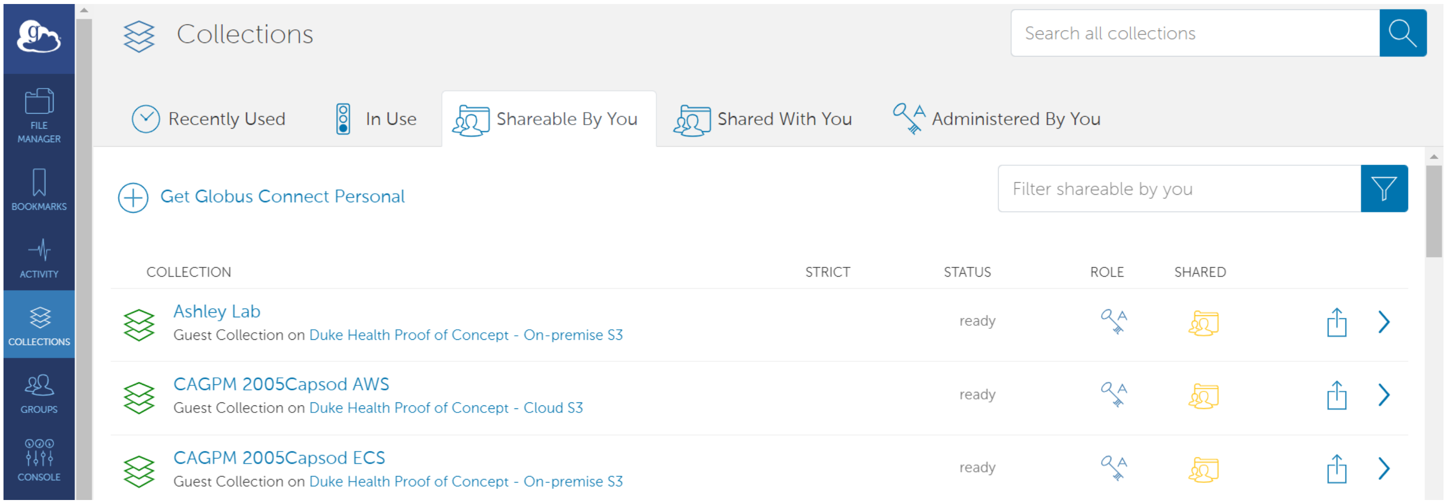
Path: /starfishmg/

This folder is empty.

To Share Files in S3 Buckets with Collaborators

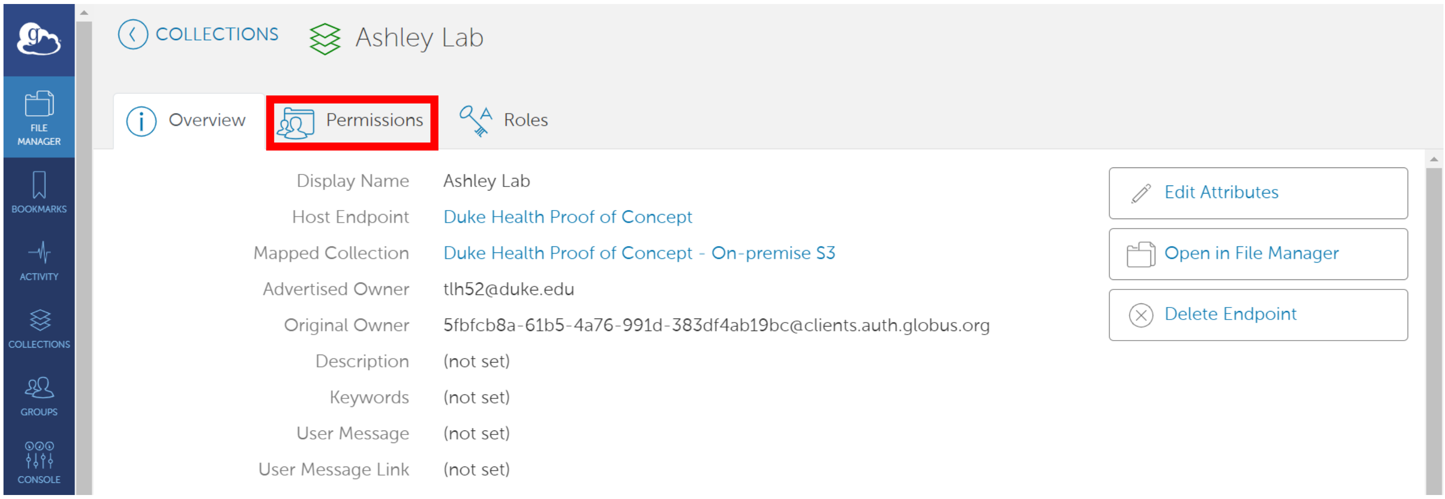
Once you have created your Collections, you can share data with collaborators.

5. Choose Collections on the left pane.
6. Select **Shareable by You** and the list of Collections that you can share will display.
7. Select the Collection from which you want to share data.



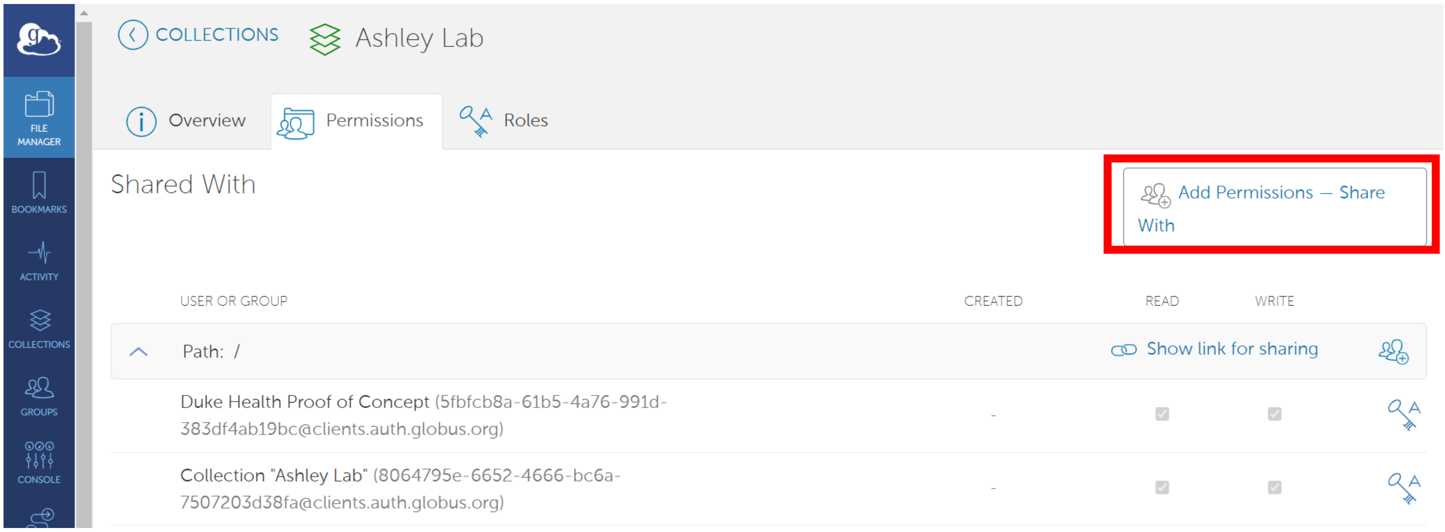
COLLECTION	STRICT	STATUS	ROLE	SHARED
Ashley Lab Guest Collection on Duke Health Proof of Concept - On-premise S3		ready		
CAGPM 2005Capsod AWS Guest Collection on Duke Health Proof of Concept - Cloud S3		ready		
CAGPM 2005Capsod ECS Guest Collection on Duke Health Proof of Concept - On-premise S3		ready		

8. On the next screen select **Permissions.**



Display Name	Ashley Lab
Host Endpoint	Duke Health Proof of Concept
Mapped Collection	Duke Health Proof of Concept - On-premise S3
Advertised Owner	tlh52@duke.edu
Original Owner	5fbfcb8a-61b5-4a76-991d-383df4ab19bc@clients.auth.globus.org
Description	(not set)
Keywords	(not set)
User Message	(not set)
User Message Link	(not set)

9. On the Shared With Screen, select "Add Permissions- Share With" on the right hand side.

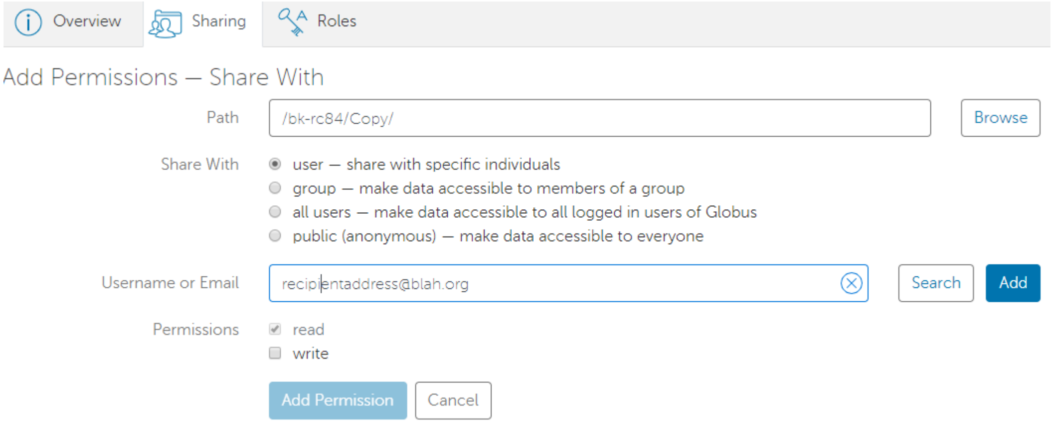


USER OR GROUP	CREATED	READ	WRITE
Path: /			
Duke Health Proof of Concept (5fbfcb8a-61b5-4a76-991d-383df4ab19bc@clients.auth.globus.org)	-	☑	☑
Collection "Ashley Lab" (8064795e-6652-4666-bc6a-7507203d38fa@clients.auth.globus.org)	-	☑	☑

On the **Add Permissions-Share With Screen**:

10. Choose the folder you want accessible via the share by clicking the **Browse** button. Drill down into the directory structure and locate the folder you wish to share (**highlight the folder**) click "**Select**" button.

11. Enter the **E-mail address of the recipient** you wish to share the data with, only the contents of this folder will be visible to the recipient and typically read only permissions is adequate for sharing. See below.



Add Permissions — Share With

Path: /bk-rc84/Copy/

Share With:
● user — share with specific individuals
○ group — make data accessible to members of a group
○ all users — make data accessible to all logged in users of Globus
○ public (anonymous) — make data accessible to everyone

Username or Email: recipientaddress@blah.org

Permissions:
☑ read
□ write

To Copy Files from the HARDAC Cluster

1. Select the "HARDAC Data Directories Gateway (POSIX)" link for directories found under /data on the HARDAC cluster.

2. Select the "HARDAC DHTS-hosted NFS volumes (POSIX)" link for directories found under /nfs on the HARDAC transfer node (hardac-xfer.genome.duke.edu).

3. For the Base Directory, enter a path to which you have read/write permissions, starting with the lab directory name (ie, shenlab, youlab, gersbachlab, etc).

   DO NOT enter the full path as it would be returned by 'pwd' in an ssh session on HARDAC. The /data directory is already set as the root for the Data Directories gateway, and the /nfs directory is already set as the root for the NFS volumes gateway. Entering a path starting with /data or with /nfs will result in a "path not found" error.

   The path entered will be the base directory of your collection. You will not be able to navigate above it from your collection, though you will be able to navigate within it.

   Use discretion in selecting the base directory - do not overshare!

   ONLY CREATE COLLECTIONS USING PATHS TO WHICH YOU HAVE READ AND WRITE ACCESS.

   Base Directory examples:
   * shenlab/<subdirectory>
   * gersbachlab/<subdirectory>/<sub-subdirectory>

4. Fill in the "Create a Guest Collection" form. Note that "Default Directory" is optional, as are keywords and description. When finished, click the "Create Collection" button at the bottom of the form.

5. With the collection created, you can add users/permissions for sharing, or transfer data into or out of the collection's directory path from other users' collections that are shared with you.

Setting Up a Globus Endpoint on a Linux (Ubuntu) VM

The following will walk through how to configure the Globus system to add an endpoint for file management, how to install and configure the globus CLI (if preferred), and how to map globus endpoint to a common folder using scripted setup (or manual by following this guide).

Setup of Globus Personal Endpoint

1. You must have an account set up on Globus already. If you need to set this up, follow the guide here, which will walk you through how to create an account. https://wiki.duke.edu/pages/viewpage.action?spaceKey=HAR&title=Using+Globus+for+data+transfer
2. Login to your target server via SSH or locally open a terminal on your linux system
3. Make a target shared directory: in our case, we've generated /mnt/Globus
    a. $ sudo mkdir /mnt/Globus
4. Install the globusconnectpersonal client:
    a. Install baseline apt sources needed: **sudo apt-get install tk tcllib**
    b. cd /opt
    c. wget https://downloads.globus.org/globus-connect-personal/linux/stable/globusconnectpersonal-latest.tgz
    d. tar xzf globusconnectpersonal-latest.tgz
    e. cd globusconnectpersonal-*
    f. ./globusconnectpersonal
        i. (run this without sudo, as it will launch the correct path installer and generate a login key for you to connect your globus endpoint with)
        ii. Paste the key that you get after clicking the link (and signing in, if you haven't already) into terminal to link your endpoint
        iii. Name your connection the same as the system hostname for ease of access/management unless a preferred shortcode is desired.
    g. By default, globus endpoint will share your home directory. Run the next line to link it to /mnt/Globus instead:
    h. ./globusconnectpersonal -restrict-paths rw/mnt/Globus -start &
5. You may need to re-authenticate your globus link periodically - you can re-run ./globusconnectpersonal from /opt/globusconnectpersonal-3.1.1 at any time to retrieve a new sign in key.

Now that your globusconnectpersonal endpoint is online, you can navigate to: https://app.globus.org/endpoints?scope=administered-by-me/ to see the endpoint and use the 'file browser' on the right side to navigate the published folders (presently just /mnt/Globus/…)

Setup of Globus CLI

1. Ssh to target server
2. Install Virtualenv if not installed:
    a. Virtualenv --version
    b. (if you don't get an output from the above, run the next line)
    c. pip install --user virtualenv
3. setup the virtual environment to host 'globus' command:
    a. virtualenv "$HOME/.globus-cli-virtualenv"
    b. source "$HOME/.globus-cli-virtualenv/bin/activate"
    c. pip install globus-cli
    d. deactivate
    e. export PATH="$PATH:$HOME/.globus-cli-virtualenv/bin"

    f. echo 'export PATH="$PATH:$HOME/.globus-cli-virtualenv/bin"' >> "$HOME/.bashrc"

4. if you want to share this CLI tool with other users:
    a. cd /usr/local/bin
    b. sudo ln -s ~/.globus-cli-virtualenv/bin/globus globus
    c. which globus (should now report that it found the symlink in /usr/local/bin)
5. you can call the newly installed command with 'globus --h' for usage and command options

User Permission Setup for Shared Folder Access

The folder /mnt/Globus can be modified to accept non-sudo access user modification via the application of group management and ACL adjustments:

1. navigate to /mnt
    a. cd /mnt
2. create a user group for shared permissions:
    a. sudo addgroup globus_users
3. grant ownership of the folder and sub-contained data to the globus_users group:
    a. sudo chown -R root:globus_users /mnt/Globus
    b. sudo setfacl -m g:globus_users:rwx /mnt/Globus/
4. add target users to the newly created group:
    a. sudo usermod -aG globus_users <netID>

Script for User Addition to Globus Endpoint

Copy and paste the below into a shell script called: globus_configuration.sh (or similar) and place it in the common dir where globusconnectpersonal-* is installed for common access. Run this script without sudo.

➔ To make executable: sudo chmod a+x ./globus_configuration.sh
➔ To run: cd /path/to/script && ./globus_configuration.sh

```
#!/bin/bash
#Scripted install and configuration wizard for Ubuntu Globus
#endpoint enrollment and setup
#Written by William Russell - DHTS (Oasis-Computing) 03/03/20
echo "Now running Globus endpoint setup"
read -p "is this the first time setting up your account? (y/n)" -n 1 -r
echo  ""
if [[ $REPLY =~ ^[Yy]$ ]]
then
        echo "now setting up account, please click on login link and paste API key below..."
        sleep 3
        cd /opt/globusconnectpersonal-3.1.1
        ./globusconnectpersonal
        sleep 2
        echo "Now launching globus with target dir: /mnt/Globus"
        ./globusconnectpersonal -restrict-paths rw/mnt/Globus -start &
else
        echo "Now launching globus with target dir: /mnt/Globus"
        ./globusconnectpersonal -restrict-paths rw/mnt/Globus -start &
fi
echo "globus should now be available at: https://app.globus.org/endpoints?scope=administered-by-me/"
echo "you may also utilize the CLI interface with the command 'globus -h'"
sleep 2
exit 0
```

Best Practices and Other Considerations for Using S3 Object Store

- Bucket names cannot be changed. If you need to change a bucket name, you essentially need to create a NEW bucket and move the data from the old bucket to the new bucket.
- As a "user" of the object store, you only have read/write access to a bucket. The reason for this is that the object store is not like a file share; e.g., permissions are not assigned at the folder level. Everyone who has access to the object store can see all the data in it – and delete permissions have to be very cautiously provided. (If you need to delete data from the S3 object store, contact the ServiceDesk at 919-684-2243 or via https://duke.service-now.com/sp?id=index)
- Buckets for a lab should be assigned to an administrator or individual who plans to remain with the lab/department for an extended period of time, versus a post doc who may move from lab to lab. Data is retrievable for an indefinite period of time. To transfer or change ownership of a bucket, contact the ServiceDesk at 919-684-2243 or via https://duke.service-now.com/sp?id=index.

Working with Small Objects

A small object is less than 100 KB. Object storage uses "box-carting" for data writes of small objects by aggregating multiple small data objects queued in memory and then writing them in a single disk operation, up to 2 MB of data. This improves performance by reducing the number of roundtrips to process individual writes to storage. If there is an option in the application to define a size, choose a larger size such as 1 MB rather than 64 KB or a value that aligns to the object storage internal buffer size of 2 MB for performance.

Working with Large Objects

- With large objects (> 100 MB), use the multipart upload feature to allow pause and resume uploads.
  - Object storage internal buffer size is 2 MB. For < 1 GB, use multiples of 2 MB (e.g., 8 MB).
  - Object storage chunk size is 128 MB. For > 1 GB, use 128 MB part size.
- Performance throughput can be improved by parallelizing uploads within the application.
- Use Byte Range reads for parallel downloads of large objects.
- Use APIs that allow for easy upload & download; e.g., In Java: TransferManager. In .NET: TransferUtility.

Compression

Object storage has a basic built-in compression mechanism at the platform level – not the file level. If it determines data is already compressed, it will not re-compress data. If a more sophisticated/specific compression is required for objects, consider using client-side compression. Some S3 Browser clients support compression.

Because compression is at the platform and not the file level, you will not likely see any difference in the sizes of the files when you move them to the object store. (One great thing to note: if you have images – like MRIs, etc. where compression could degrade them, this does not happen. All data is accessible in its native format when retrieved from the object store.)

Retention and Expiration

Retention means you cannot update or delete the object until the retention period ends. There are three ways to assign retention:

- At the bucket level (compatible with generic S3).
- At the policy defined at namespace and assigned to objects (e.g., email = 5 years, documents = 3 years).
- Explicit retention period at an object level. When retention is defined in multiple places, the longest time wins. Objects are automatically deleted when the expiration time is reached.