Data Transfer Workshop

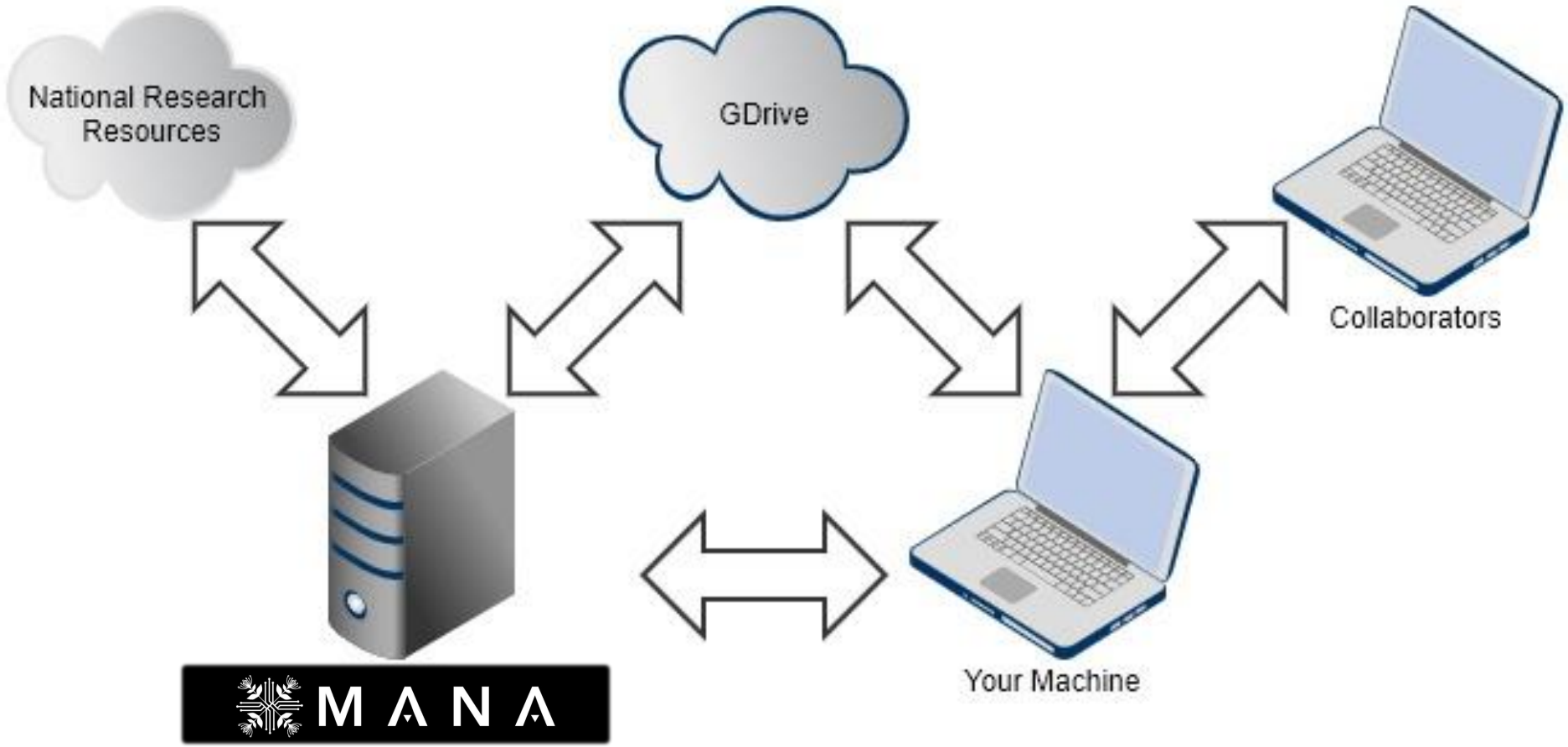
Sean Cleveland

Adriana Comerford

UNIVERSITY *of* HAWAI'I®

HAWAI'I DATA SCIENCE

Introduction



Globus is a service that makes it easy to move, sync, and share large amounts of data.

- Globus will manage file transfers, monitor performance, retry failures, recover from faults automatically when possible, and report the status of your data transfer.
- Globus uses GridFTP for more reliable and high-performance file transfer, and will queue file transfers to be performed asynchronously in the background.

Globus was developed and is maintained at the University of Chicago and is used extensively at supercomputer centers and major research facilities. https://globus.org

When To use Globus

To transfer or share data between two Globus managed endpoints (e.g. two multi-user systems at different universities, each running a Globus server)

To transfer data between a managed endpoint (e.g. UH-HPC) to a Globus Connect Personal endpoint (e.g. your desktop)

Globus Plus

For some kinds of data transfer or sharing, you need Globus Plus. The UH Globus subscription includes Globus Plus for all users, but you need to to request a Globus Plus invite.

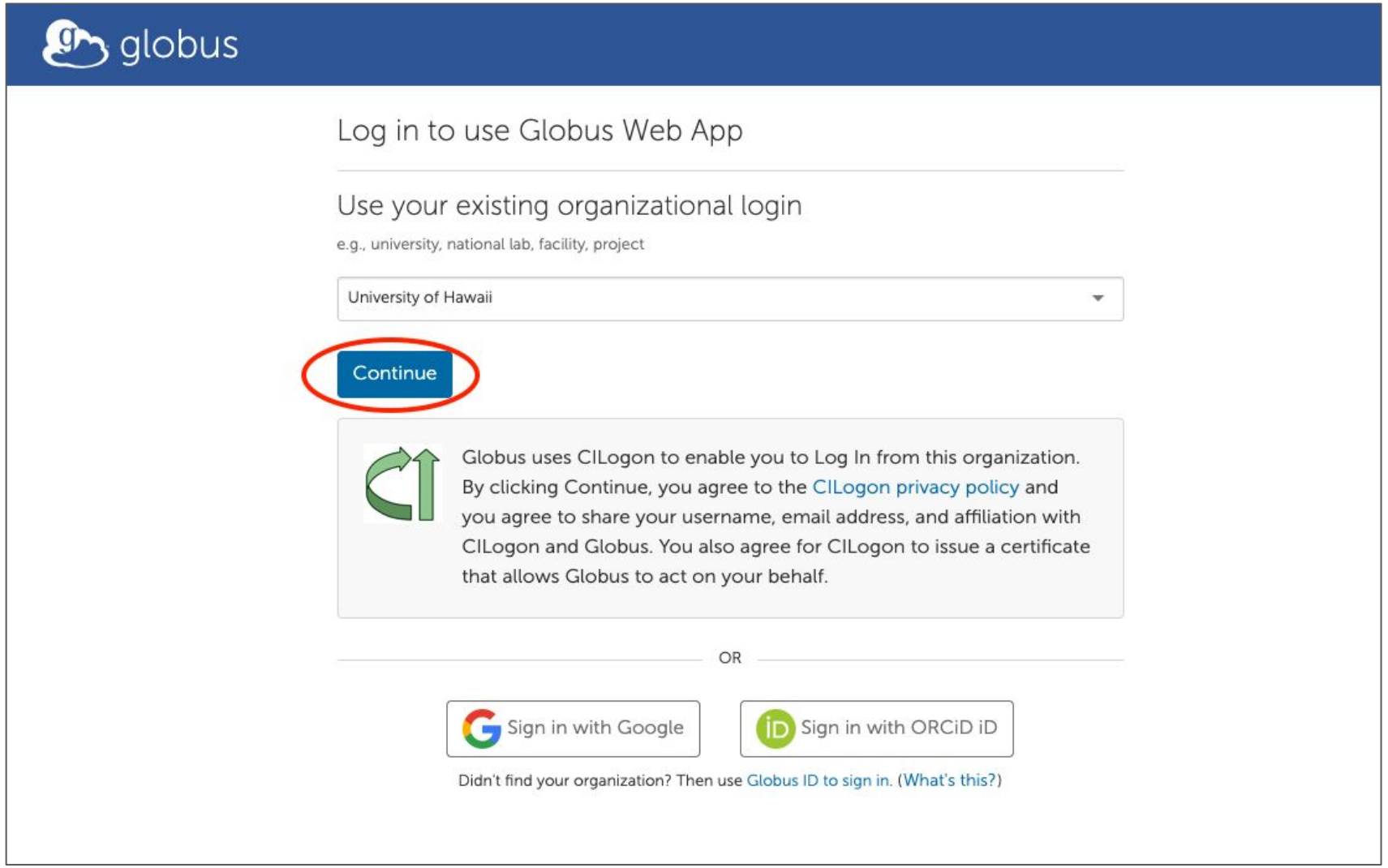
- To transfer data between two Globus Connect Personal endpoints (e.g. your desktop system and a laptop).
- To share data from a Globus Connect Personal endpoint (e.g. your desktop system)
- To transfer data from/to a shared endpoint that is hosted on a Globus Connect Personal endpoint.
- Note 1: if your collaborator needs Globus Plus to download data, and is not at UH, we cannot provide Globus Plus to that person.
- Note 2: By default, files on a Globus Connect Personal endpoint (e.g. your laptop or desktop) may not be shareable. You will need to configure that via the instructions at these links: Linux, Mac, Windows.

Transfering Files with Globus

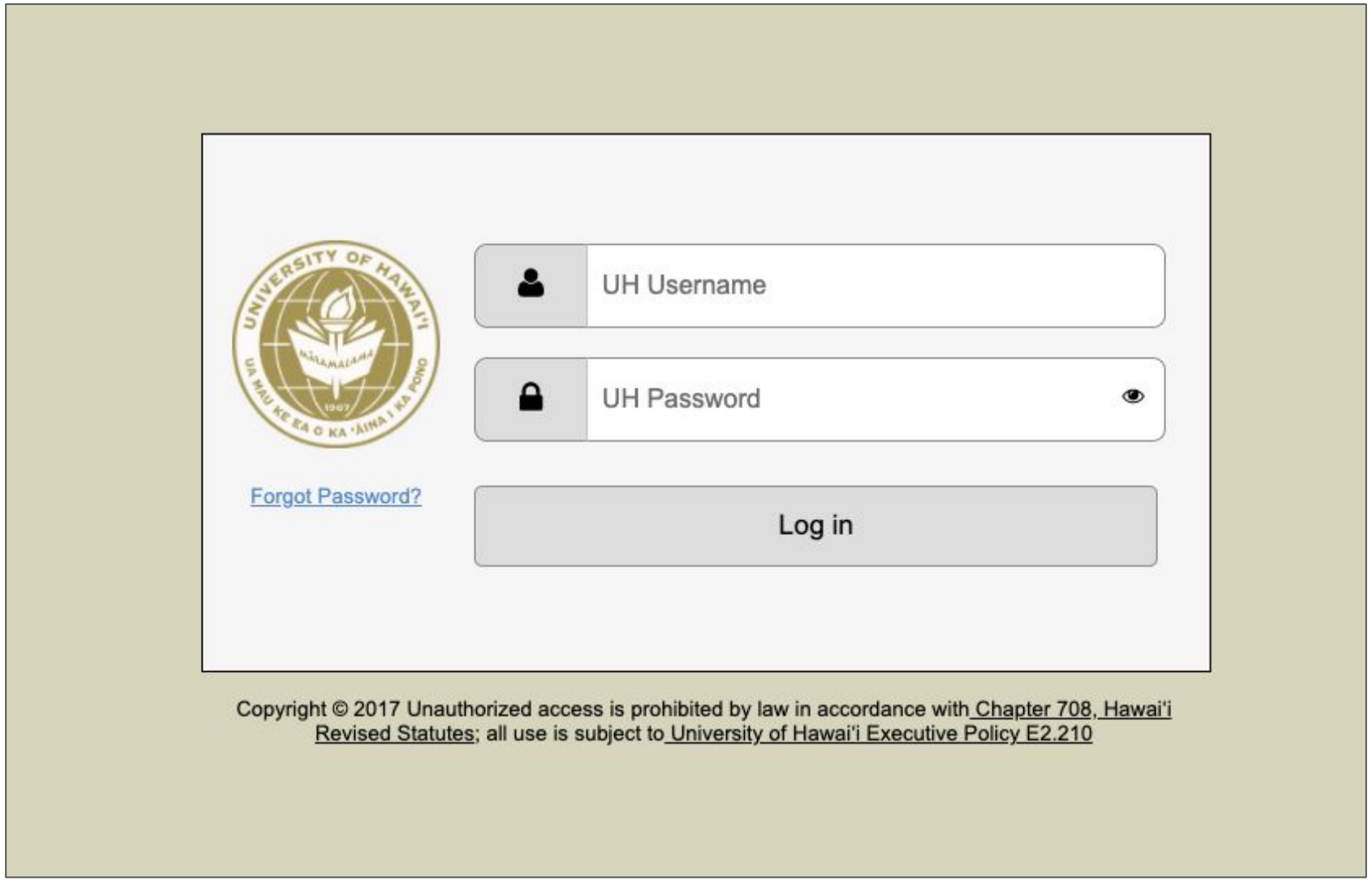
- Creating a Globus Account Using Your UH Credential
- Installing Globus Connect Personal
- Transferring Data from and to Mana

Creating a Globus Account Using UH Credential

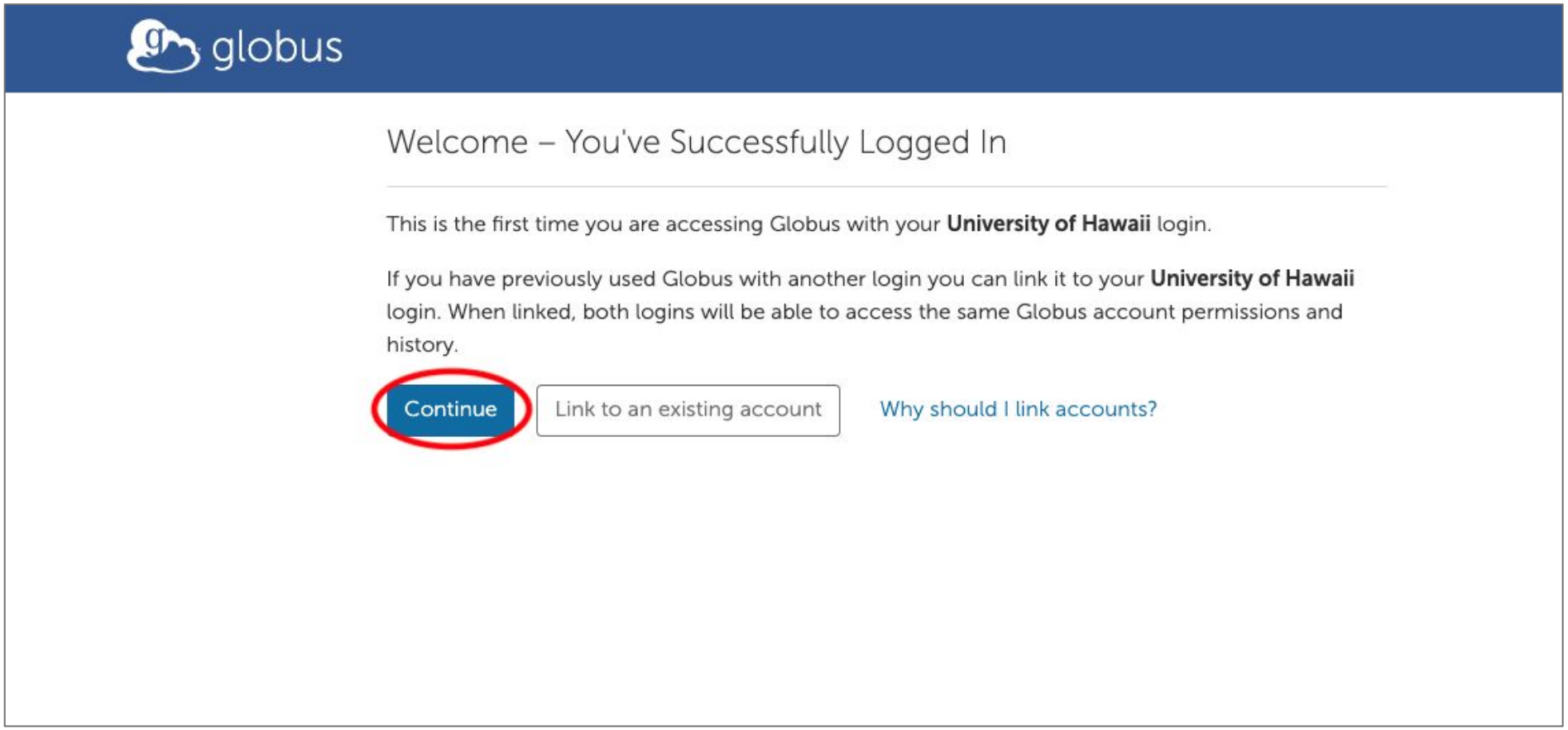
Visit www.globus.org and click "Login" at the top of the page. On the Globus login page, type in University of Hawaii. When you find it, click Continue.



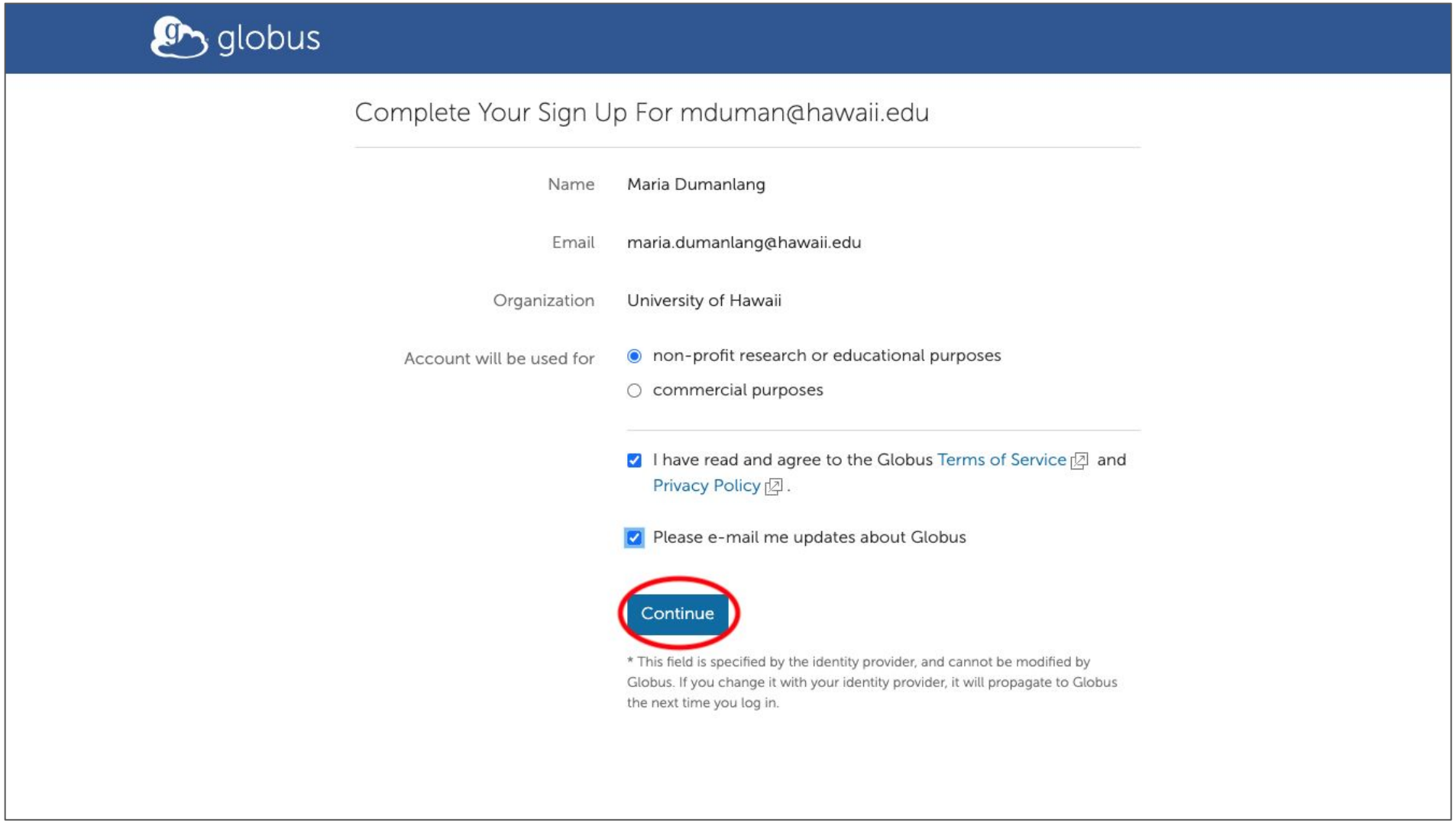
You'll be redirected to your UH login page. Use your UH credentials to login.



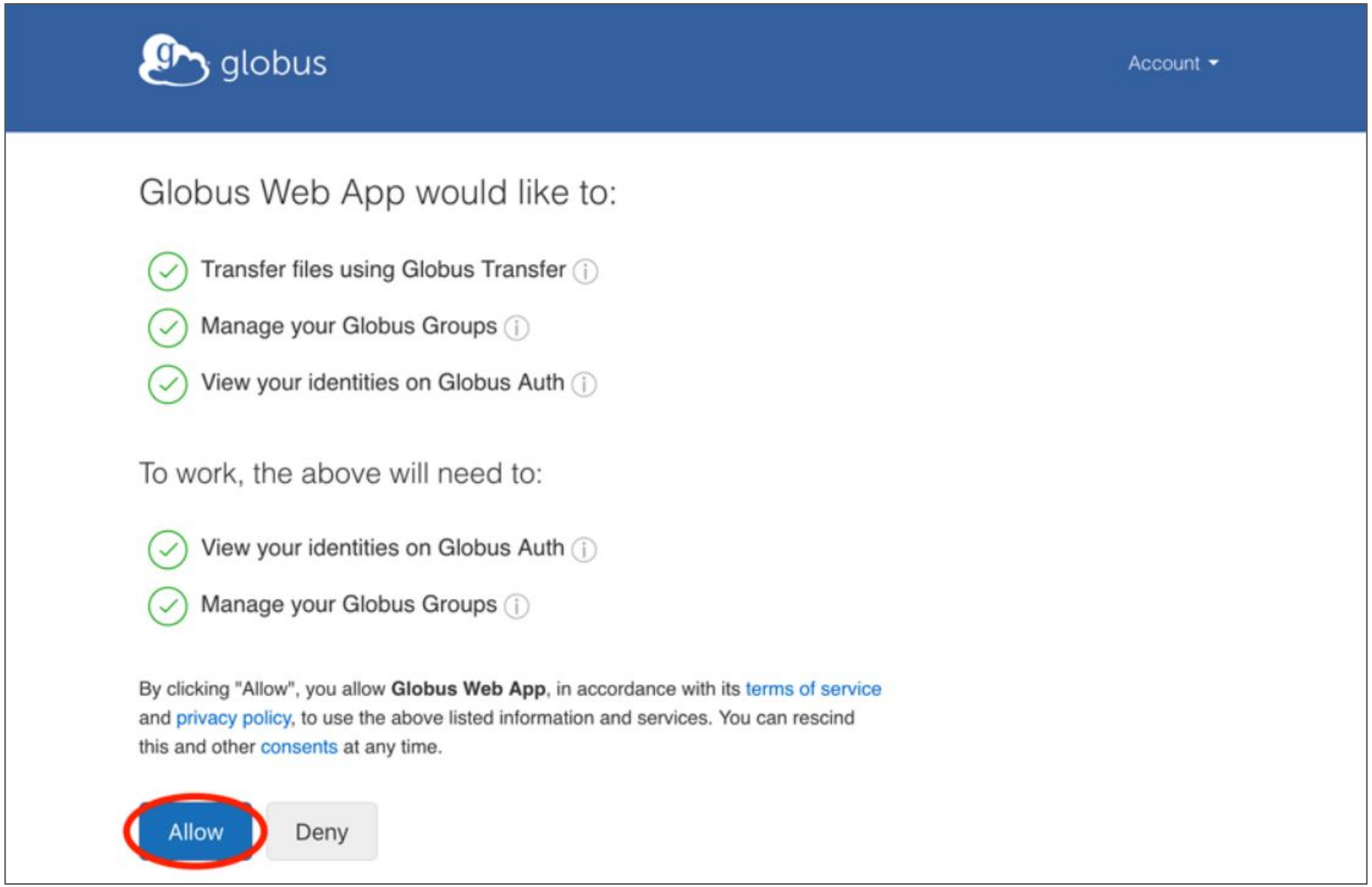
Some organizations will ask for your permission to release your account information to Globus. Once you've logged in with your UH credentials, Globus will ask if you'd like to link to an existing account. If this is your first time logging in to Globus, click "Continue." If you've already used another account with Globus, you can choose "Link to an existing account."



You may be prompted to provide additional information such as your organization and whether or not Globus will be used for commercial purposes. Click on non-profit research or educational purposes. Complete the form and click "Continue."



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



Globus Connect Personal

Globus Connect Personal turns your laptop or other personal computer into a Globus endpoint with just a few clicks. With Globus Connect Personal you can share and transfer files to/from a local machine—campus server, desktop computer or laptop—even if it's behind a firewall and you don't have administrator privileges.

Globus Connect Personal puts the power of Globus on your computer.

- Dramatically increases data transfer speeds over scp and other transfer tools.
- Automatically suspends transfers when computer sleeps and resumes when turned on.
- Installs in seconds using native operating system install packages.
- Works with firewalls that block incoming connections, and behind most NATs.
- Uses proven Globus infrastructure for security and authentication.

Installation Instructions

Install Globus Connect Personal

Globus Connect Personal for Mac for Mac OS X 10.7 or higher (Intel only)

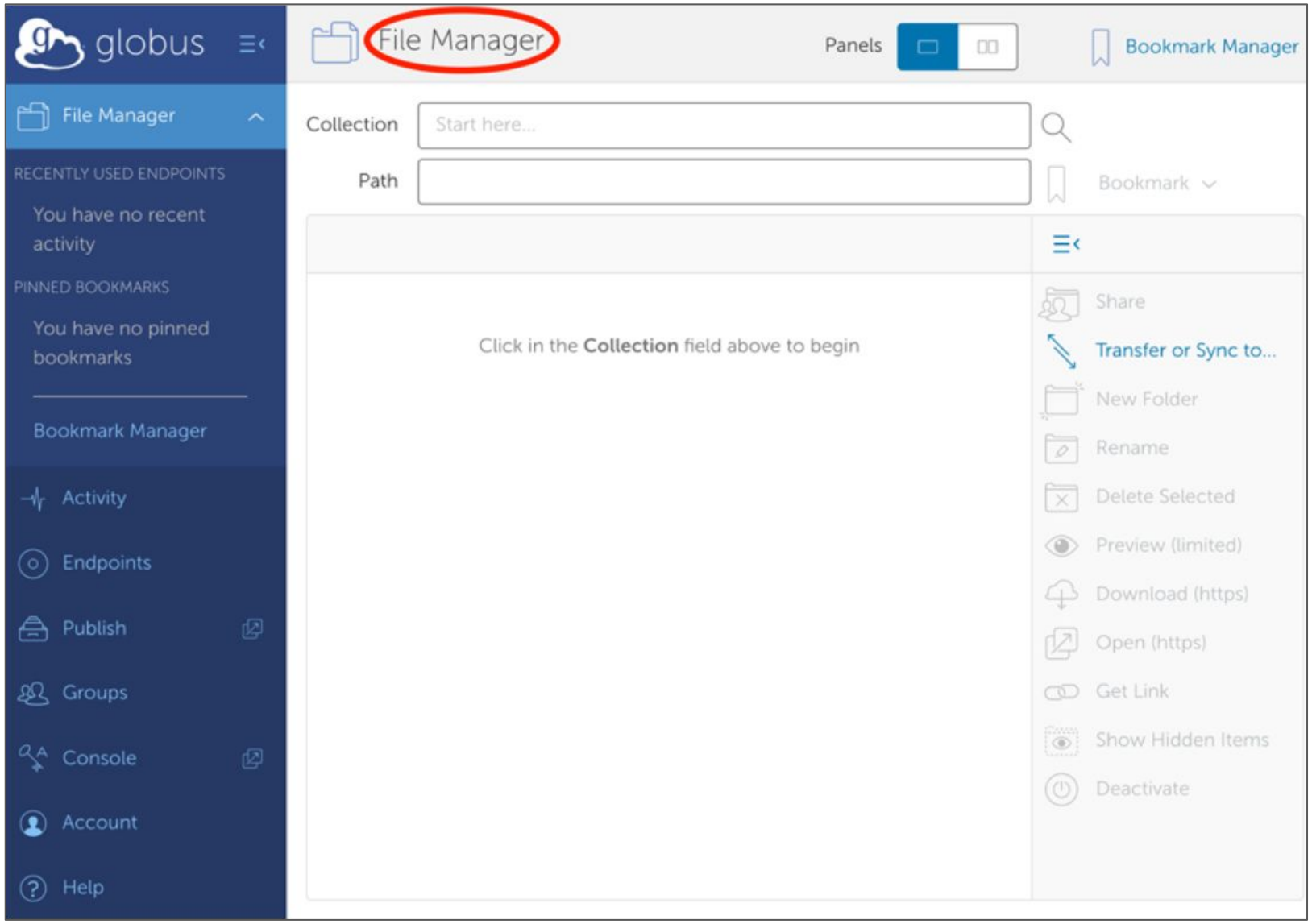
Globus Connect Personal for Linux for common x86-based distributions

Globus Connect Personal for Windows for recent Windows versions

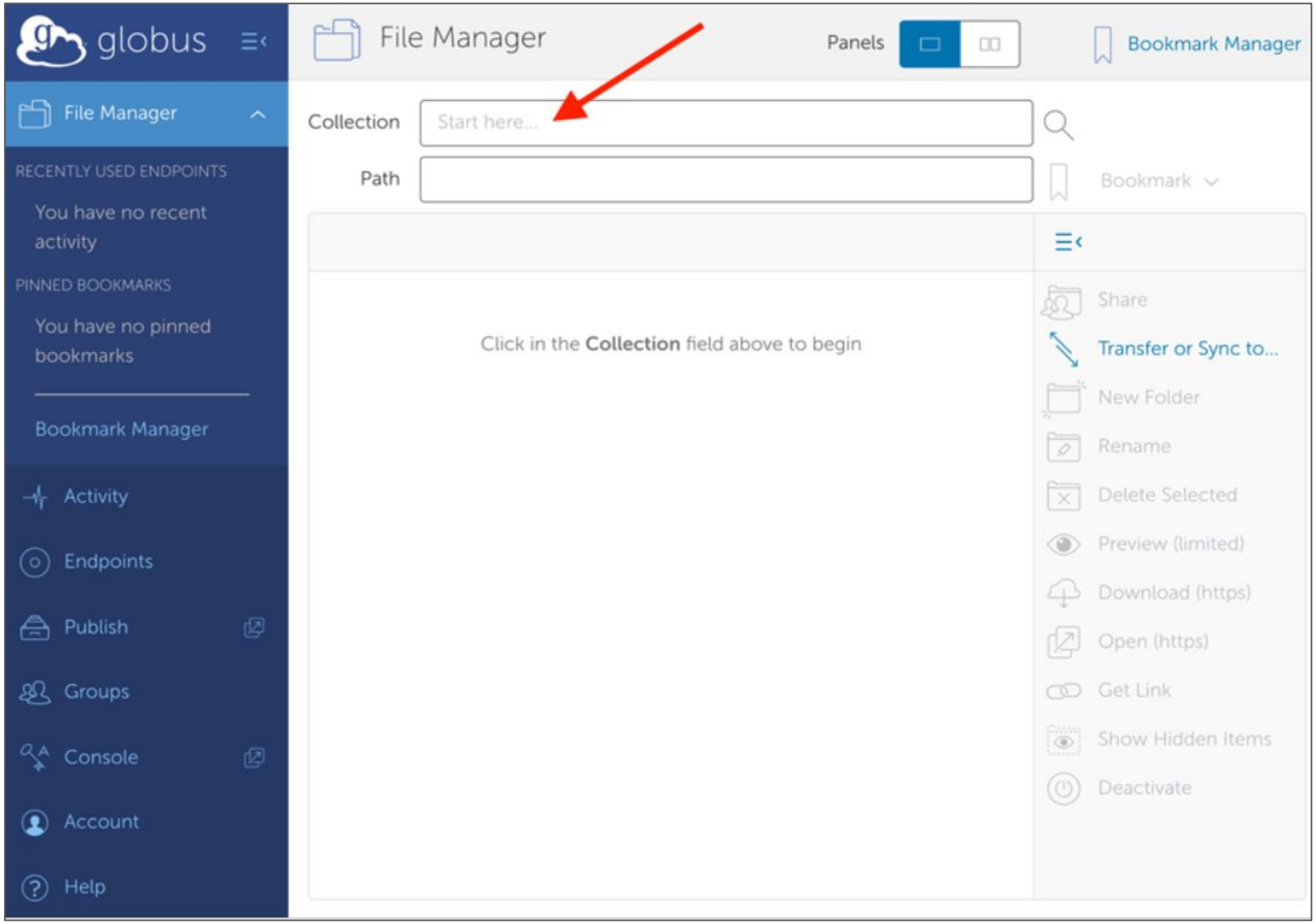
Transferring Data from and to Mana Using Globus Connect Personal

The File Manager

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



The first time you use the File Manager, all fields will be blank.



Tip

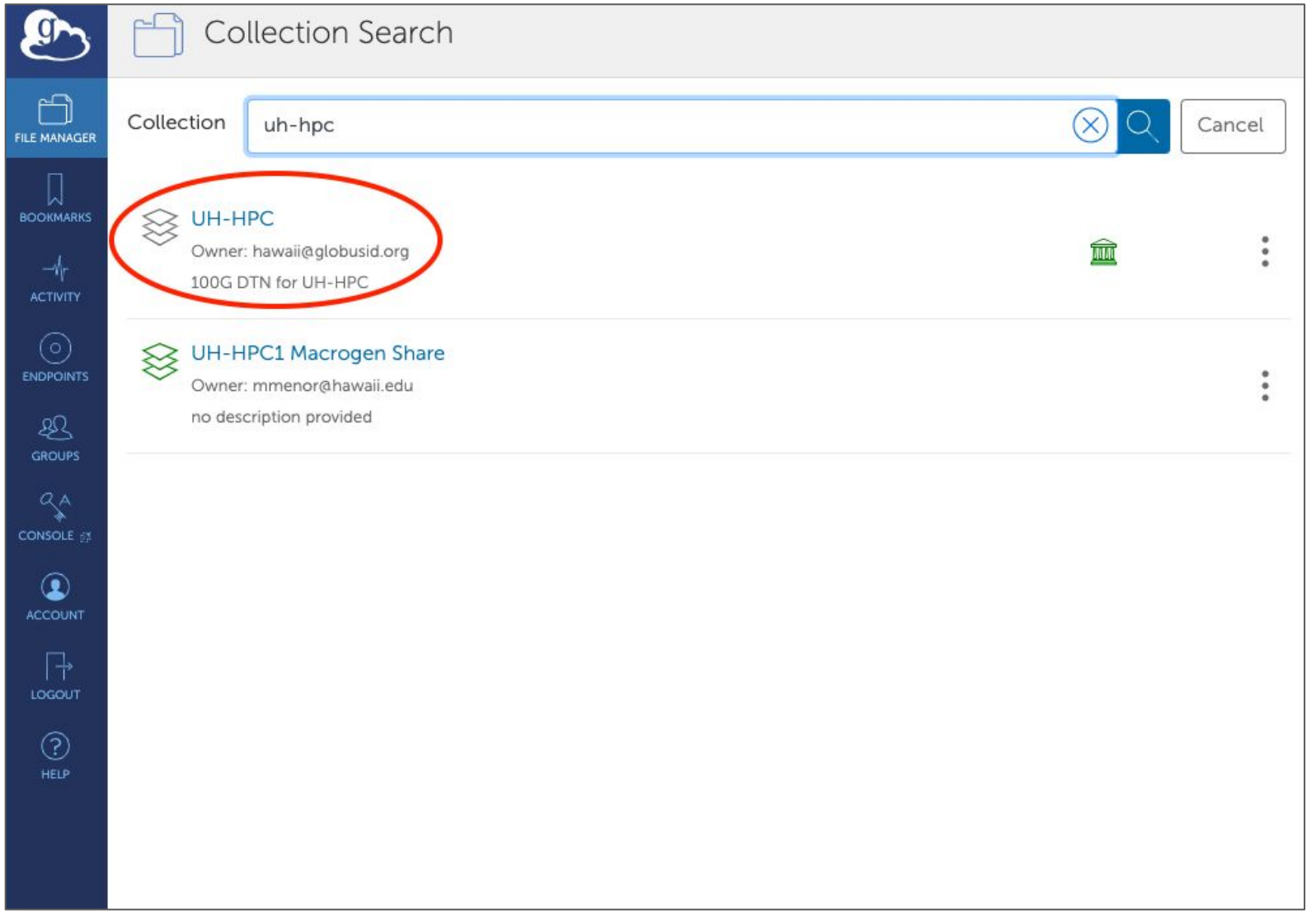
Key Concept: *Collection*

A collection is a named location containing data you can access with Globus. Collections can be hosted on many different kinds of systems, including campus storage, HPC clusters, laptops, ***Amazon S3 buckets, Google Drive*** *(these are “premium” connectors so seperate subscription is required)*, and scientific instruments.

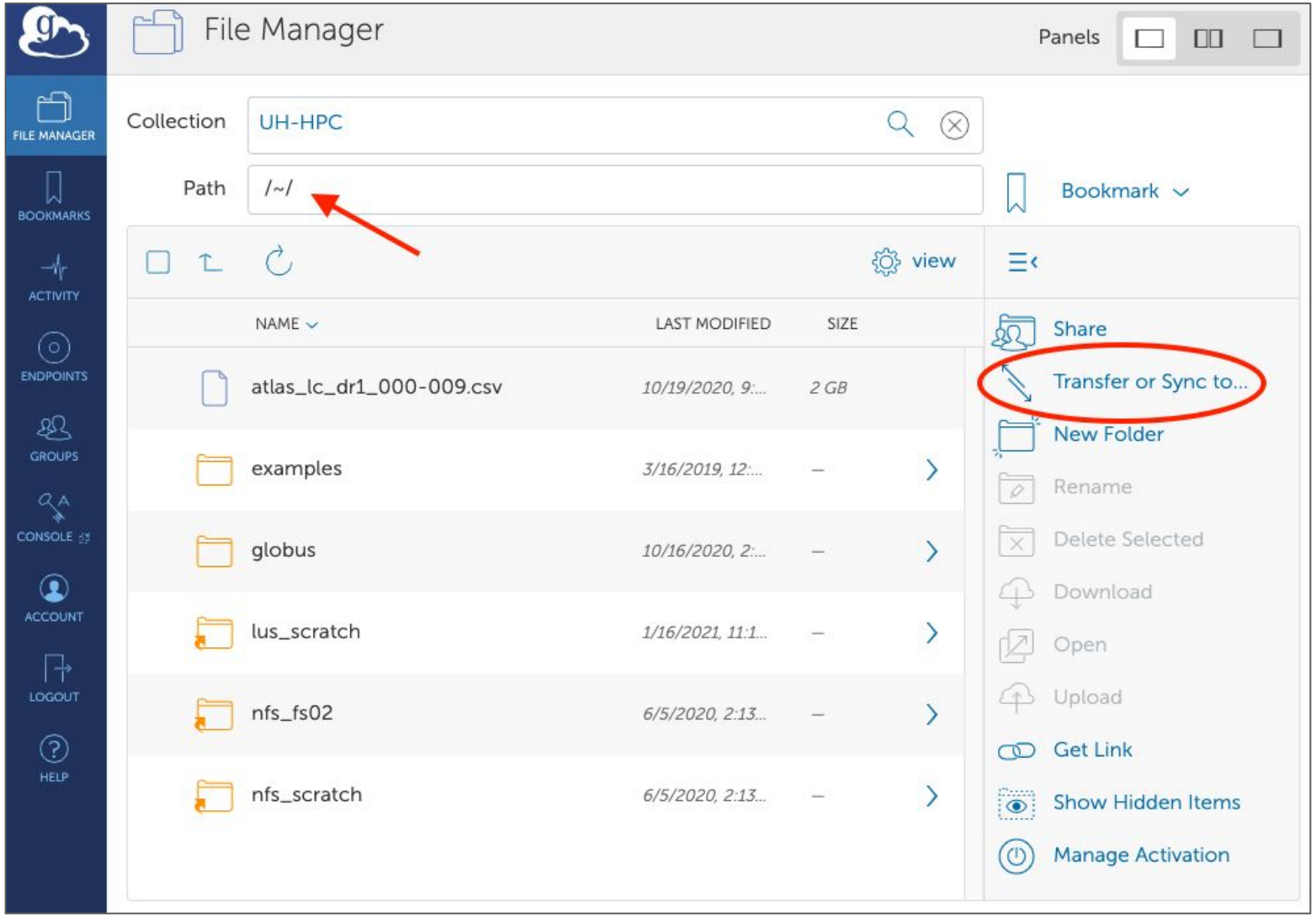
When you use Globus, you don’t need to know a physical location or details about storage. You only need a collection name. A collection allows authorized Globus users to browse and transfer files. Collections can also be used for sharing data with others and for enabling discovery by other Globus users. Globus Connect is used to host collections.

Access a collection

Click in the Collection field at the top of the File Manager page and type ”UH-HPC". Globus will list collections with matching names. Choose UH-HPC.

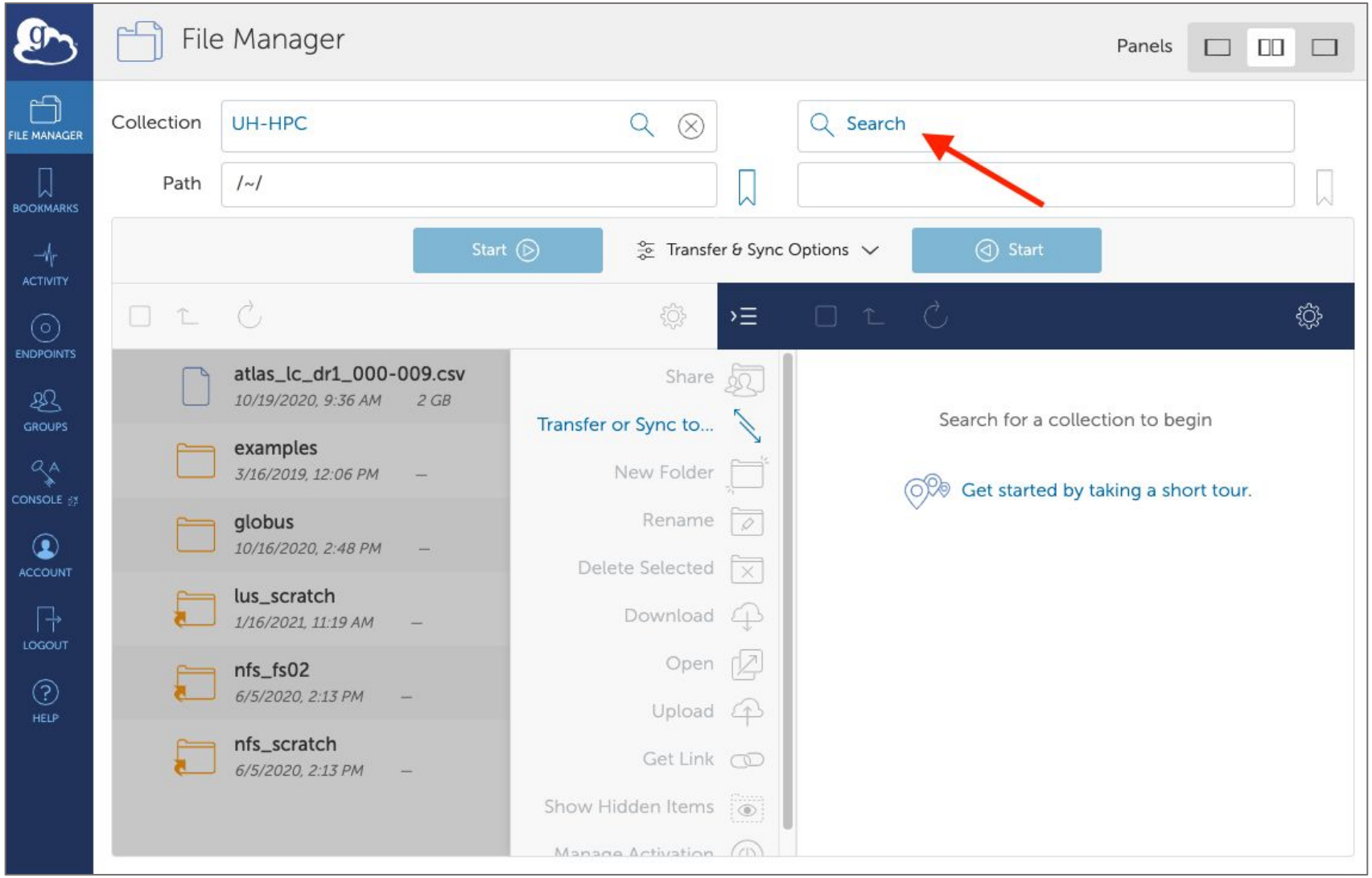


Globus will connect to the UH-HPC collection and display the default directory, /~/. This is your home directory in the Mana Globus endpoint. Click on **“Transfer or Sync to”.**

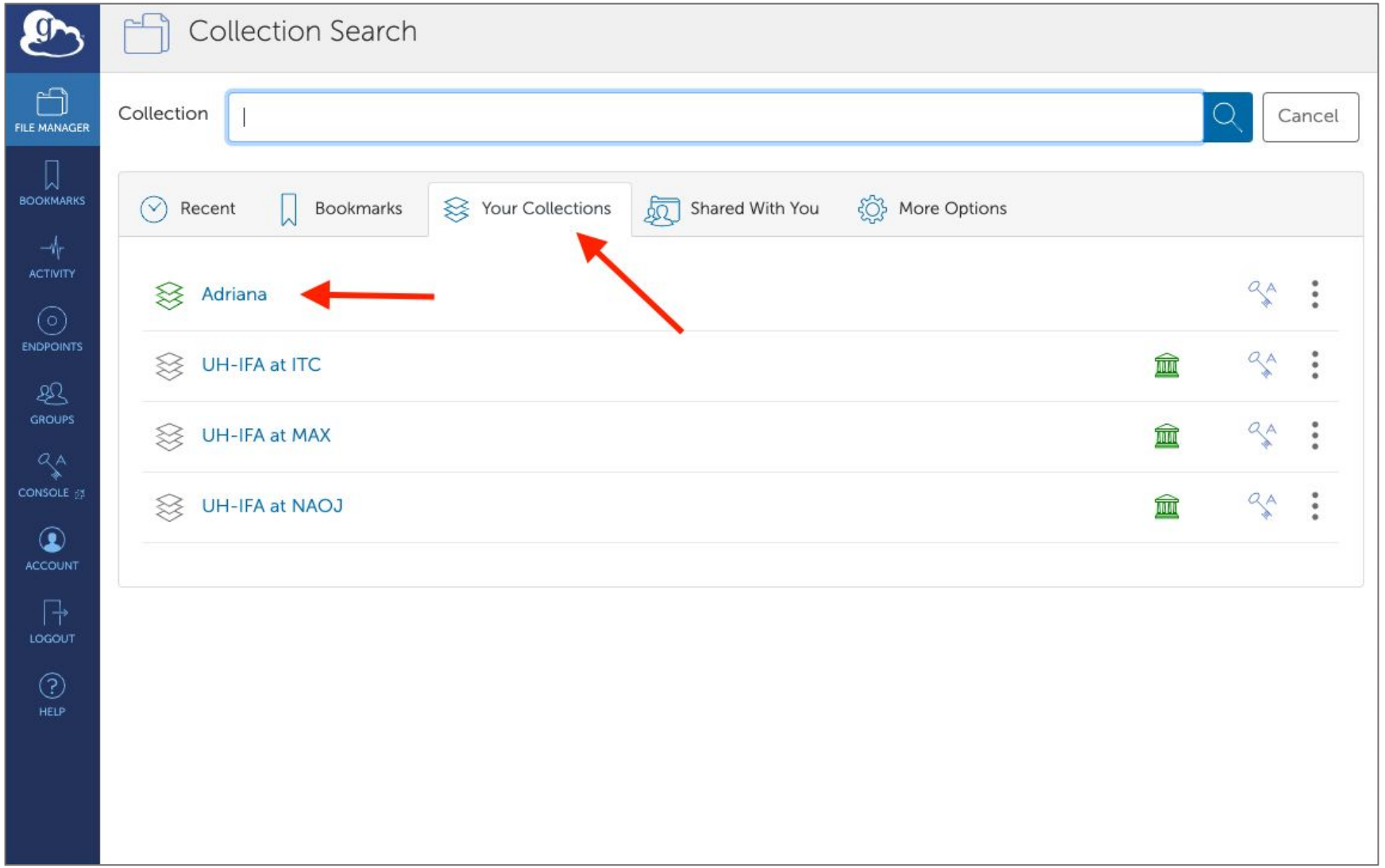


Request a file transfer

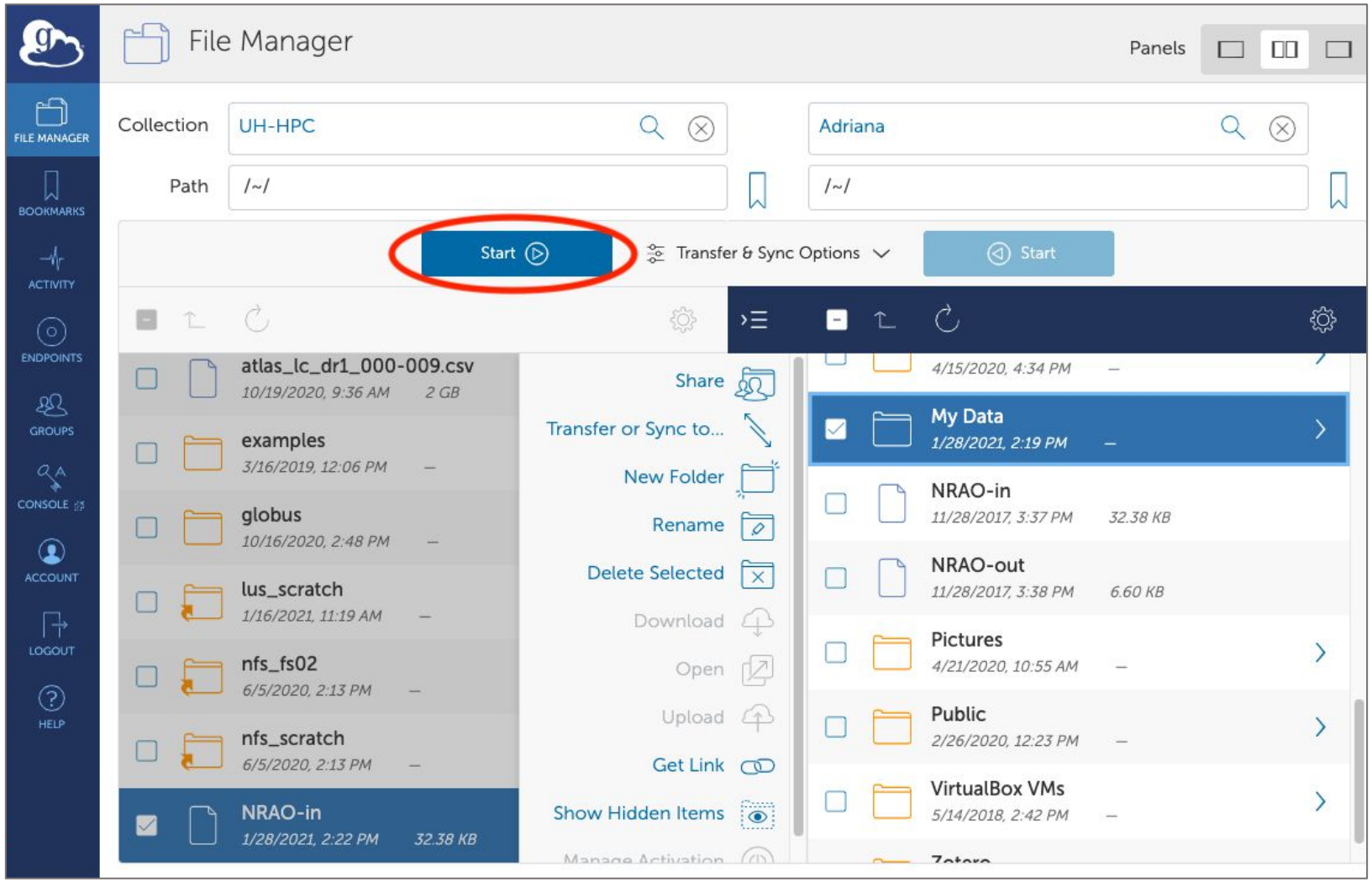
A new collection panel will open, with a ”Search" field at the top of the panel. Click on it.



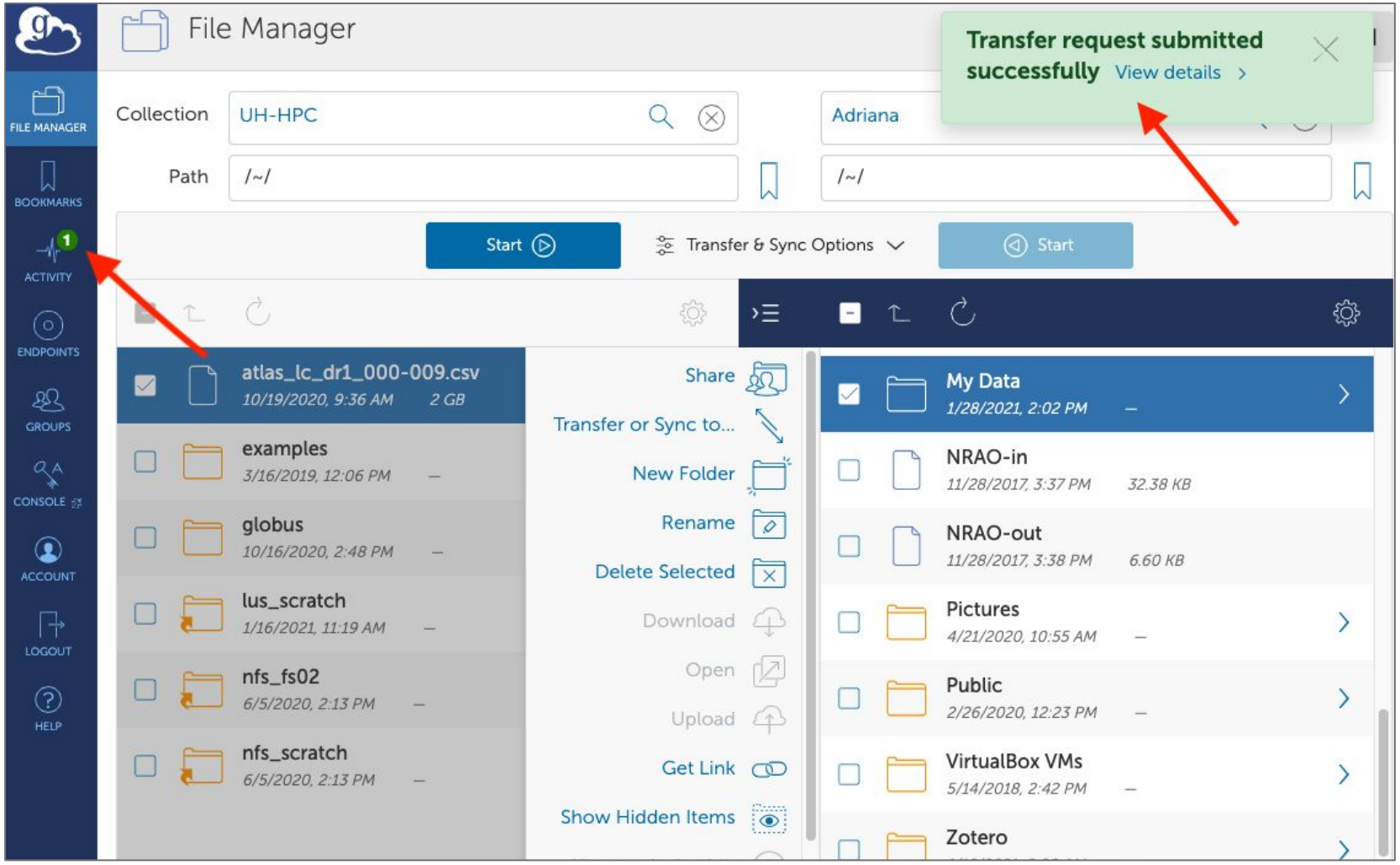
Click on “Your Collection” tab. Find the Globus Connect Personal endpoint you created earlier and click on it.



On the left collection, UH-HPC, select the file you would like to transfer. The Start> button at the top of the panel will activate. Click the Start> button to transfer the selected files to the collection in the right panel.



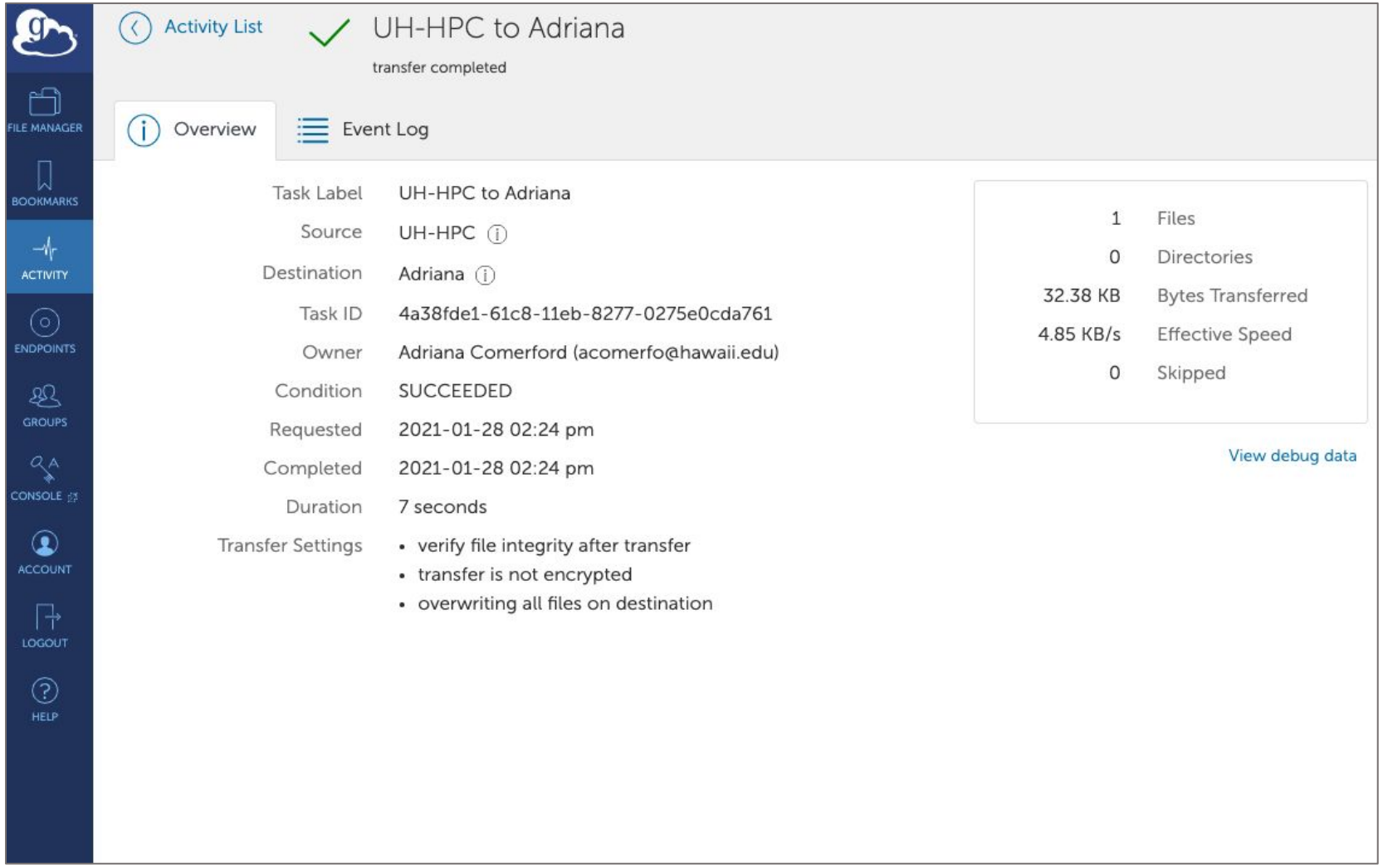
Globus will display a green notification panel—confirming that the transfer request was submitted—and add a badge to the “Activity” item in the command menu on the left of the page. Click Activity in the command menu on the left of the page to go to the Activity page.



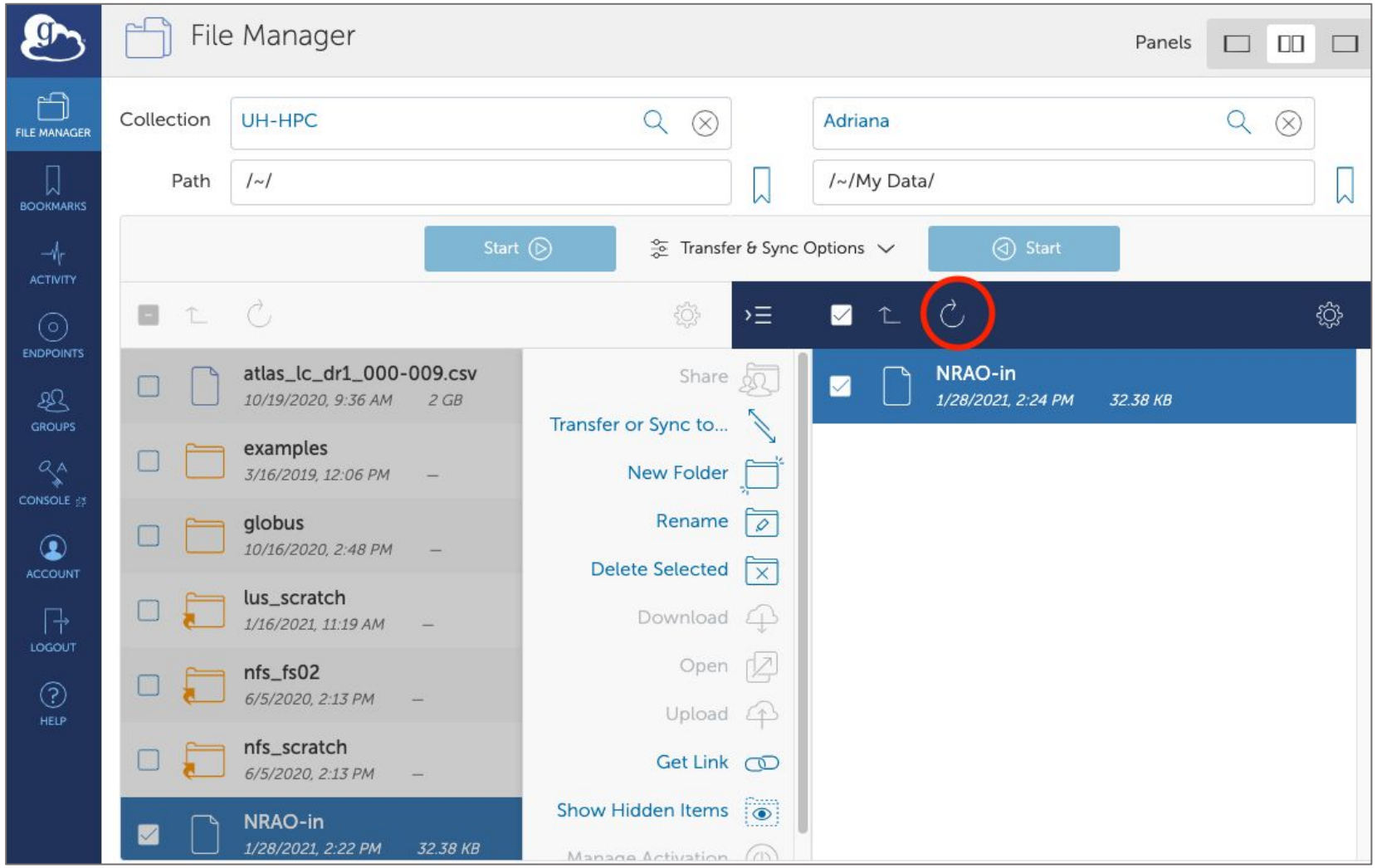
Confirm Transfer Completion

On the Activity page, click the arrow icon on the right to view details about the transfer. You will also receive an email with the transfer details.

Activity Page



If you notice that the transferred files are not listed in the right panel with your Globus Connect Personal collection. Click the refresh icon (circular arrows) at the top of the collection panel to see the updated contents.



Questions about Globus

Rclone is a free utility for syncing directories between object storage systems (such as Amazon S3, Dropbox, Google Drive etc) and file based storage (e.g. /home or /scratch).

https://rclone.org/

Note about this tutorial

Wherever this tutorials uses ‘>’ that means there is a command to execute on the terminal/shell

Mana and Rclone

Rclone is installed on the Mana Data Transfer Nodes and can be used in the command line via

```
> rclone
```

Configuring Rclone

Before you can use Rclone, you must configure it. This configuration step will set up access for the remote object storage system that you want to transfer data to and from.

In this tutorial we will configure Google Drive since UH has Google for Education there is “unlimited” storage available there and everyone at UH has it.

Get a shell session on Mana - your own terminal or you can use Open OnDemand via https://mana.its.hawaii.edu and select from the Menu

Clusters -> >_Mana_Shell_Access

SSH to DTN

From your terminal/shell ssh to one of the Mana DTNs

>ssh username@hpc-dtn1.its.hawaii.edu

Your may be prompted for your password depending on where you are SSHing from and you WILL be prompted for DUO two-factor verification.

```
[seanbc@login002 ~]$ ssh hpc-dtn1.its.hawaii.edu
Duo two-factor login for seanbc
Enter a passcode or select one of the following options:
 1. Duo Push to XXX-XXX-5555
 2. Phone call to XXX-XXX-5555
 3. SMS passcodes to XXX-XXX-5555 (next code starts with: 4)
Passcode or option (1-3): 1
Pushed a login request to your device...
Success. Logging you in...
[seanbc@hpc-dtn1 ~]$
```

Step 1 - Configure RClone to use GDrive

> rclone config

2021/01/29 16:16:25 Failed to load config file "/home/xxxx/.rclone.conf" - using defaults: open /home/xxxx/.rclone.conf: no such file or directory

No remotes found - make a new one

n) New remote

s) Set configuration password

q) Quit config

Type "n" to set up a new object storage system with which to transfer data.

Choose a name for the remote object storage system

You'll be prompted for the name of the remote object storage system, we use "rclone-gdrive" in this tutorial

name> rclone-gdrive

Select Storage System - Google Drive 13

10 / Encrypt/Decrypt a remote
  \ "crypt"
11 / FTP Connection
  \ "ftp"
12 / Google Cloud Storage (this is not Google Drive)
  \ "google cloud storage"
13 / Google Drive
  \ "drive"
Storage> 13

There will be a long list of options to choose from - Google Drive is 13

Select Google Client - we will use Rclone’s for now

Storage> 13

** See help for drive backend at: https://rclone.org/drive/ **

Google Application Client Id

Setting your own is recommended.

See https://rclone.org/drive/#making-your-own-client-id for how to create your own.

If you leave this blank, it will use an internal key which is low performance.

Enter a string value. Press Enter for the default ("").

client_id>

Client Secret - leave it blank

OAuth Client Secret

Leave blank normally.

Enter a string value. Press Enter for the default ("").

client_secret>

Pick Rclone’s scope - choose 1 - so we can read & write

```
Scope that rclone should use when requesting access from drive.
Enter a string value. Press Enter for the default ("").
Choose a number from below, or type in your own value
 1 / Full access all files, excluding Application Data Folder.
   \ "drive"
 2 / Read-only access to file metadata and file contents.
   \ "drive.readonly"
   / Access to files created by rclone only.
 3 | These are visible in the drive website.
   | File authorization is revoked when the user deauthorizes the app.
   \ "drive.file"
   / Allows read and write access to the Application Data folder.
 4 | This is not visible in the drive website.
   \ "drive.appfolder"
   / Allows read-only access to file metadata but
 5 | does not allow any access to read or download file content.
   \ "drive.metadata.readonly"
scope> 1
```

Set the “root” folder - Rclone can only read files in this folder and it’s sub-folders

ID of the root folder

Leave blank normally.

Fill in to access "Computers" folders (see docs), or for rclone to use

a non root folder as its starting point.

Enter a string value. Press Enter for the default ("").

root_folder_id>

Service Account - leave blank

Service Account Credentials JSON file path

Leave blank normally.

Needed only if you want use SA instead of interactive login.

Leading `~` will be expanded in the file name as will environment variables such as `${RCLONE_CONFIG_DIR}`.

Enter a string value. Press Enter for the default ("").

service_account_file>

Edit config? - No

edit advanced config? (y/n)

y) Yes

a non root folder as its starting point.

n) No (default)

y/n> n

Use Auto config - No

Remote config

Use auto config?

 * Say Y if not sure

 * Say N if you are working on a remote or headless machine

y) Yes (default)

n) No

y/n> n

Verification Code - go to “your” link in a browser & authorize then copy the code and paste it

Please go to the following link:
https://accounts.google.com/o/oauth2/auth?access_type=offline&client_id=202264815644.apps.googleusercontent.com&redirect_uri=urn%3Aietf%3Awg%3Aoauth%3A2.0%3Aoob&response_type=code&scope=https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdrive&state=5ewrjlkjsoeR349302323

Log in and authorize rclone for access

Enter verification code> 2334sdffouiewk32EESKDLF324234

Configure as team drive - No

Configure this as a team drive?

y) Yes

n) No (default)

y/n> n

Approve setup - Yes

[rclone-gdrive]

type = drive

token = {"access_token":"ya29.a0sdjlkfjwoerjkfsldalsfKSDSD_$49fKFJDIs02934dsFSDKL","token_type":"Bearer","refresh_token":"1//06sejlksfdjlsjfoERE034sREPKSLKCEPROE","expiry":"2021-02-02T17:39:51.393440573Z"}

---

y) Yes this is OK (default)

e) Edit this remote

d) Delete this remote

y/e/d> y

Finished Setup - Quit config

```
Current remotes:
Name                 Type
====                 ====
rclone-gdrive        drive
e) Edit existing remote
n) New remote
d) Delete remote
r) Rename remote
c) Copy remote
s) Set configuration password
q) Quit config
e/n/d/r/c/s/q> q
```

Lets list files from GDrive

‘lsf’ is how we list files using Rclone

```
> rclone lsf rclone-gdrive:/
```

Lets create a directory to transfer files to/from

Make a directory called “rclonefiles” using the “mkdir” command

> mkdir rclonefiles

Move into the directory we just created

> cd rclonefiles

‘cd’ is the change directory command

Create a test document for transfer

In google drive create a folder name it “rclonetest”. Within that folder create a new doc and call it “testfile”.

Now copy the directory contents from GDrive to Mana. ‘rclone copy’ has a source and destination required. GDrive being the source in the example below and the current directory (represented by the ‘.’) the destination

```
> rclone copy rclone-gdrive:/rclonetest .
```

This will copy the folder contents to the current directory - Note the ‘.’ at the end this is represents the current directory as the destination folder - we could also have used ~/rclonefiles or /home/username/rclonefiles as that same folder path.

Lets transfer a file from Mana to GDrive

Create a testfile2.docx on the Mana DTN by copying testfile.docx

> cp testfile.docx testfile2.docx

‘cp’ is the copy command in the terminal/shell

> ls

testfile2.docx testfile.docx

‘ls’ is how we list files in the terminal/shell

Now copy testfile2.docx to GDrive (the source is the Mana testfile2.docx and the destination is gdrive)

>rclone copy testfile2.docx rclone-gdrive:/rclonetest

You can check GDrive and the file should appear

RClone Copy

The ‘rclone copy’ command is the way to move files to or from GDrive.

The copy command on a folder will overwrite files that have the same name but if a files exists on the destination that isn’t in the folder being copied it will be retained on the destination (when we get to sync you will see a difference in this behavior)

RClone Sync

The sync command is useful to keep a folder on GDrive and somewhere else with identical contents - meaning that if the destination folder has files that do not exist on the source they will be removed (so be careful)

rclone sync source destination

Let remove testfile.docx and sync our rclonefiles folder to our GDrive rclonetest folder

> rm testfile.docx

> rclone sync ~/rclonefiles rclone-gdrive:/rclonetest

We should see in GDrive that now only testfile2.docx is there because the folders are in sync - Mana’s rclonefiles folder was the source so the GDrive rclonetest folder is now identical to rclonefiles

RClone large transfer - use nohup

For transfers that make take a long time that you do not wish to observe or that your connection might disconnect you should use ‘nohup’ so they run in the background until complete. Example of nohup ‘rclone copy’ below:

> nohup rclone copy source destination > nohup.out &

The ‘>’ after the destination will direct any standard output to be written to the nohup.out file and the ‘&’ on the end tells the shell to disconnect the command issued and run it in the background so you can still use your terminal/shell for other commands or exiting the session - the command issued with ‘nohup’ will continue to run.

Rclone documentation

More information about Rclone and Google Drive can be found here:

https://rclone.org/drive/#limitations

Questions