Northwestern

Northwestern
INFORMATION TECHNOLOGY

Advanced Topics in Globus: Automation with the Globus Timer, Globus CLI, and Globus SDK

Brian Roland
Data Management Specialist
Northwestern IT
Research Computing and Data Services

Agenda

- Challenges of Manual Data Management
- Globus Automation Opportunities
- Live Demo
  - Globus Scheduler & Timer
  - Globus CLI
  - Globus CLI in Bash script
  - Globus SDK
    - App Registration
    - Native App
    - Service Account App

The Challenge of Manual Data Management

In today's research landscape, scientific workflows often involve the movement of large datasets between various storage locations. Manually managing these data transfers can present a variety of challenges:

- Time-consuming file transfers
- Error-prone processes
- Difficulty in tracking transfers and task progress
- Limitations when integrating data movement into complex workflows

Globus Automation to the rescue!

Globus offers automation options that can be leveraged to help address the challenges of data management. You can improve the efficiency of your research workflows by leveraging:

- Timer functionality native to the browser interface for scheduling tasks
- Globus Command Line Interface (CLI) for scriptable transfers
- Globus Software Development Kit (SDK) for integrating Globus capabilities into your own applications.

Automating data transfers minimizes manual intervention, reduces errors, and ensures data movement happens seamlessly within your scientific pipelines, allowing you to focus on the core aspects of your research.

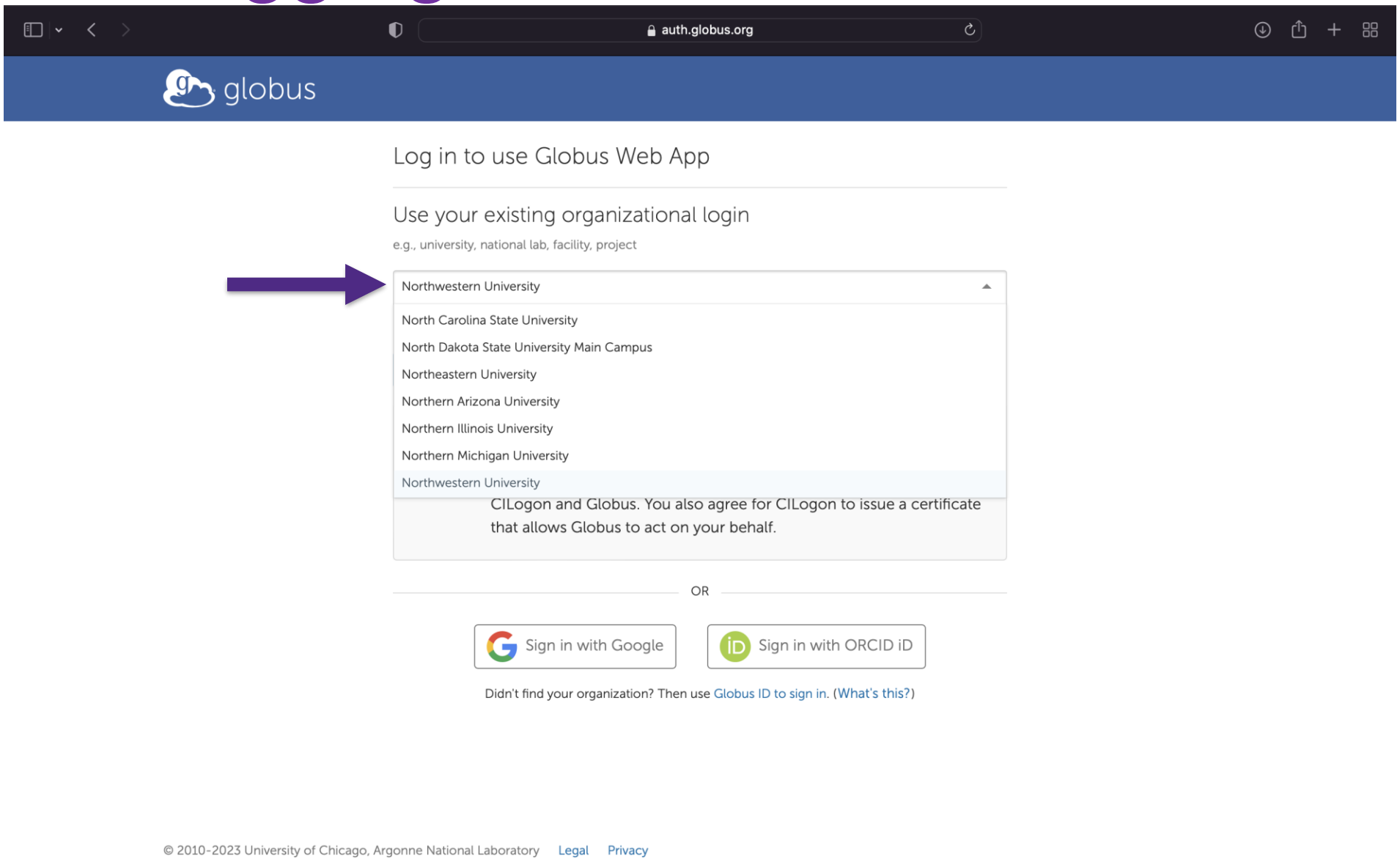
Automating Transfers with Globus Timers

Embedded within the web interface, Globus offers the timer feature for automating transfers

- Ideal for datasets requiring frequent synchronization or scheduled backups
- Flexible scheduling options including hourly, daily, weekly, monthly, or other custom time intervals

https://www.globus.org/

Logging in via NU Credentials

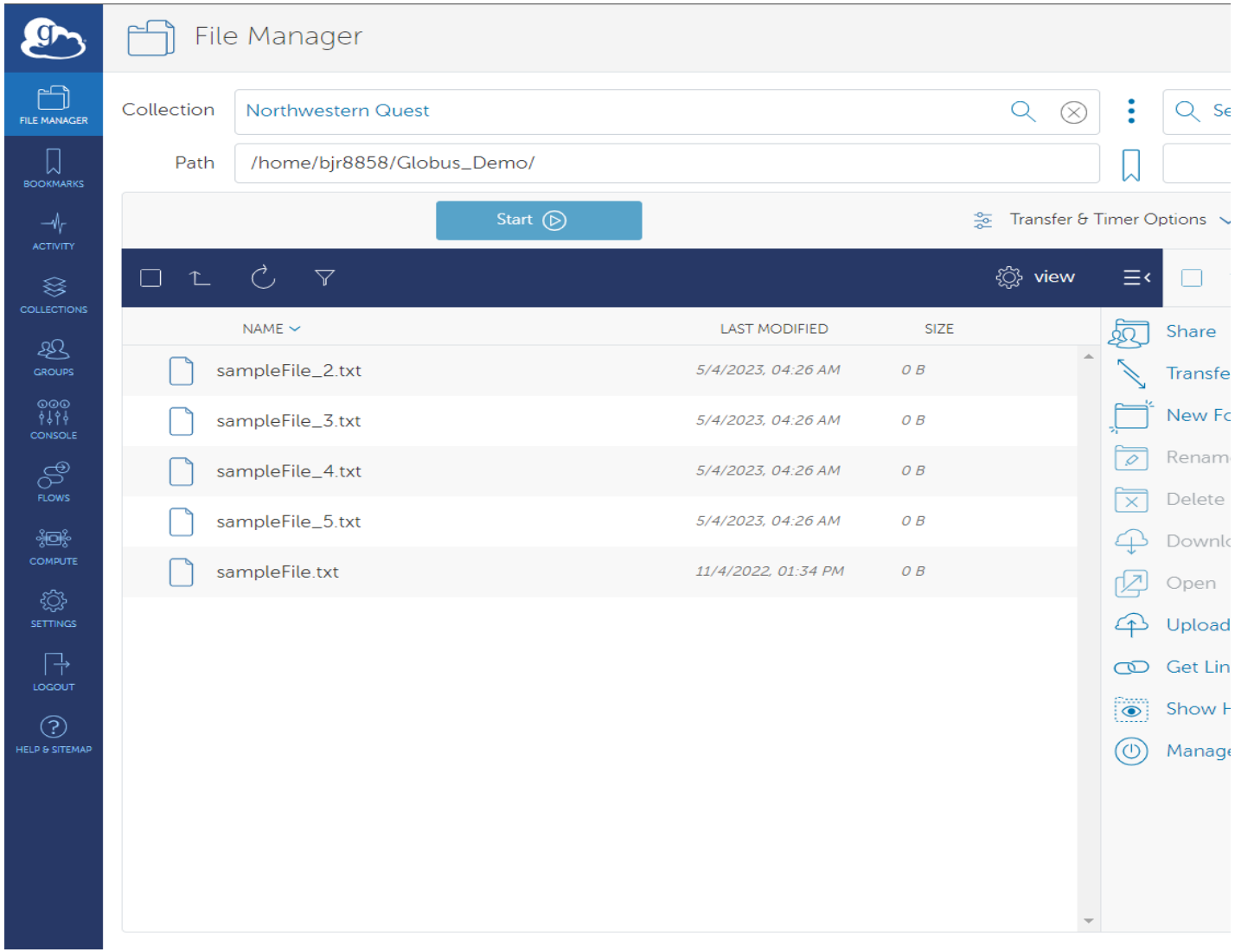


File Manager: Selecting Data Source

- To select a Collection to serve as the *data source*, click in the **Collection** text field and type in the name or the UUID of the collection

- Once the you select a Collection, you can navigate to the path they want to access manually or type it into the **Path** field.

- Globus will display the files and subdirectories in the path provided. You can select which files and directories you want to transfer from the data source.

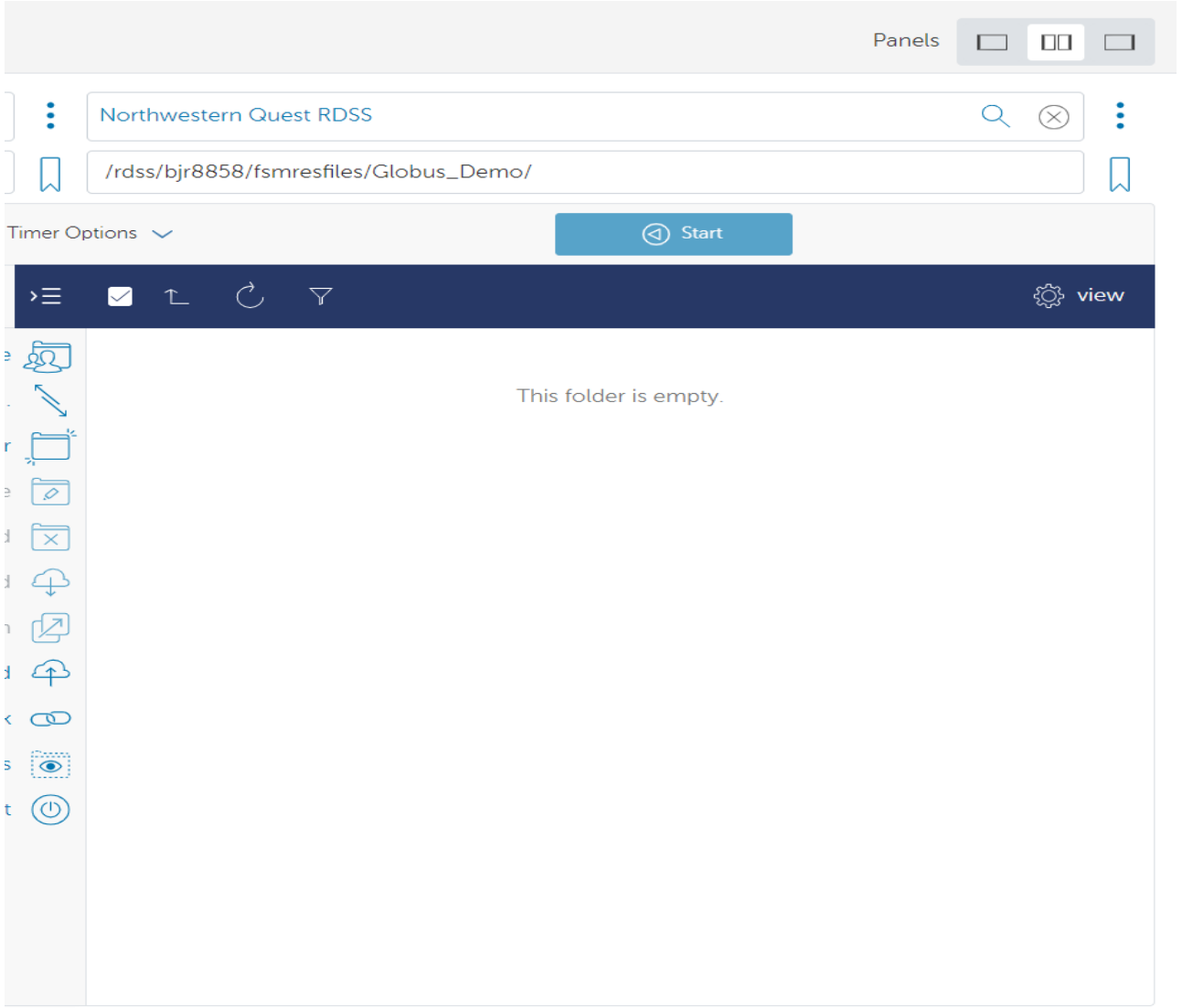


File Manager: Selecting Data Destination

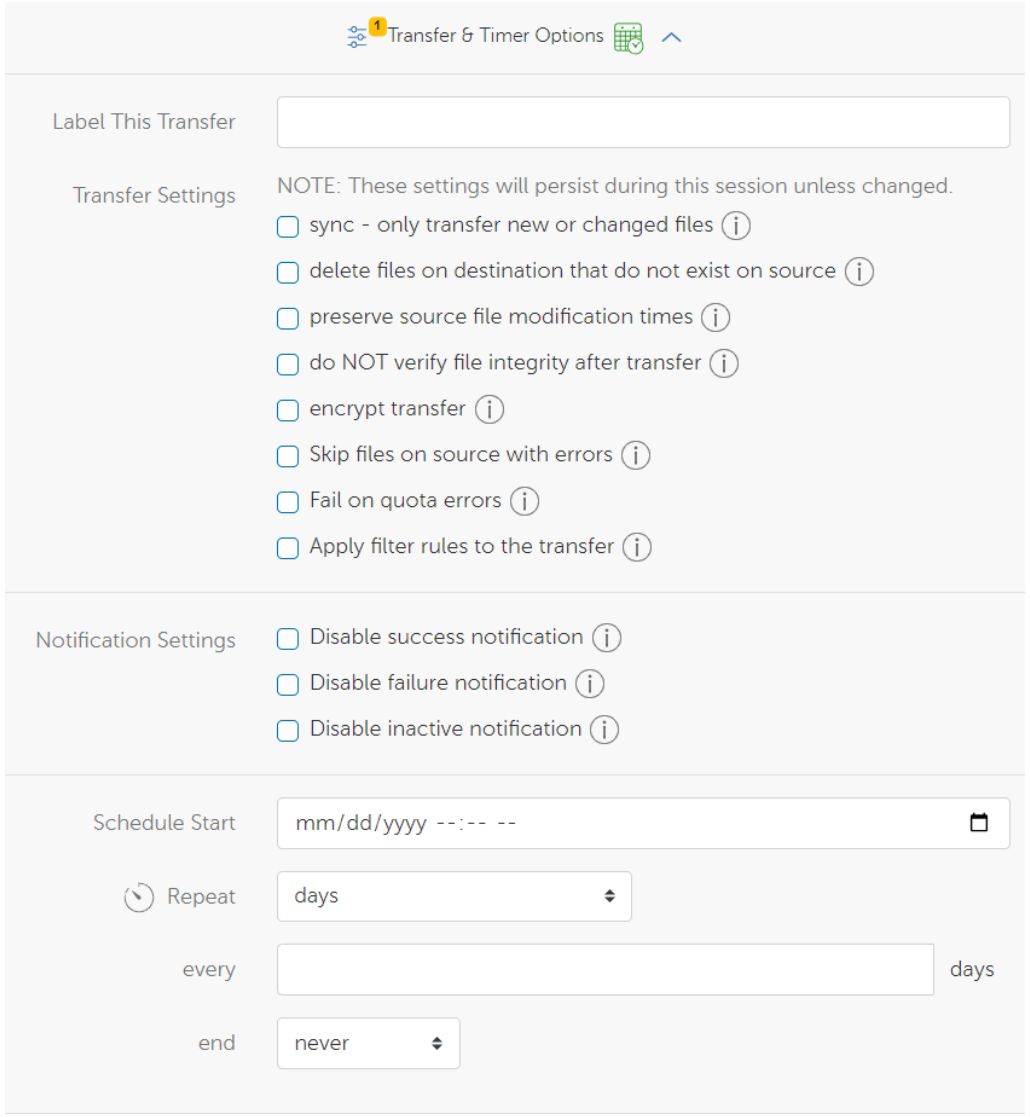
- When selecting the *data destination* you will need to select/enter a Collection name or UUID in the blank **Collection** field in the 2nd panel.

- Once the you select a Collection, you can navigate to the path they want to access manually or type it into the **Path** field.

- Globus will display any files and subdirectories that already exist in the path provided.



File Manager: Transfer & Timer Options



Recommended Transfer Options

- **Sync**: Data in your files will only be transferred if it meets a selected criteria when comparing the source to the destination of the transfer. Criteria include:
  - Modification time is newer
  - File size is different
  - File does not exist on destination
  - File checksum is different

- **Preserve Source file modification times**

- **Encrypt Transfer**: Data in your files will be encrypted while in transit. Both endpoints must support encryption to ensure a successful transfer.

- **Skip files on source with errors**: Files on source with "file not found" and "permission denied" errors will be skipped rather than causing the entire transfer procedure to fail.

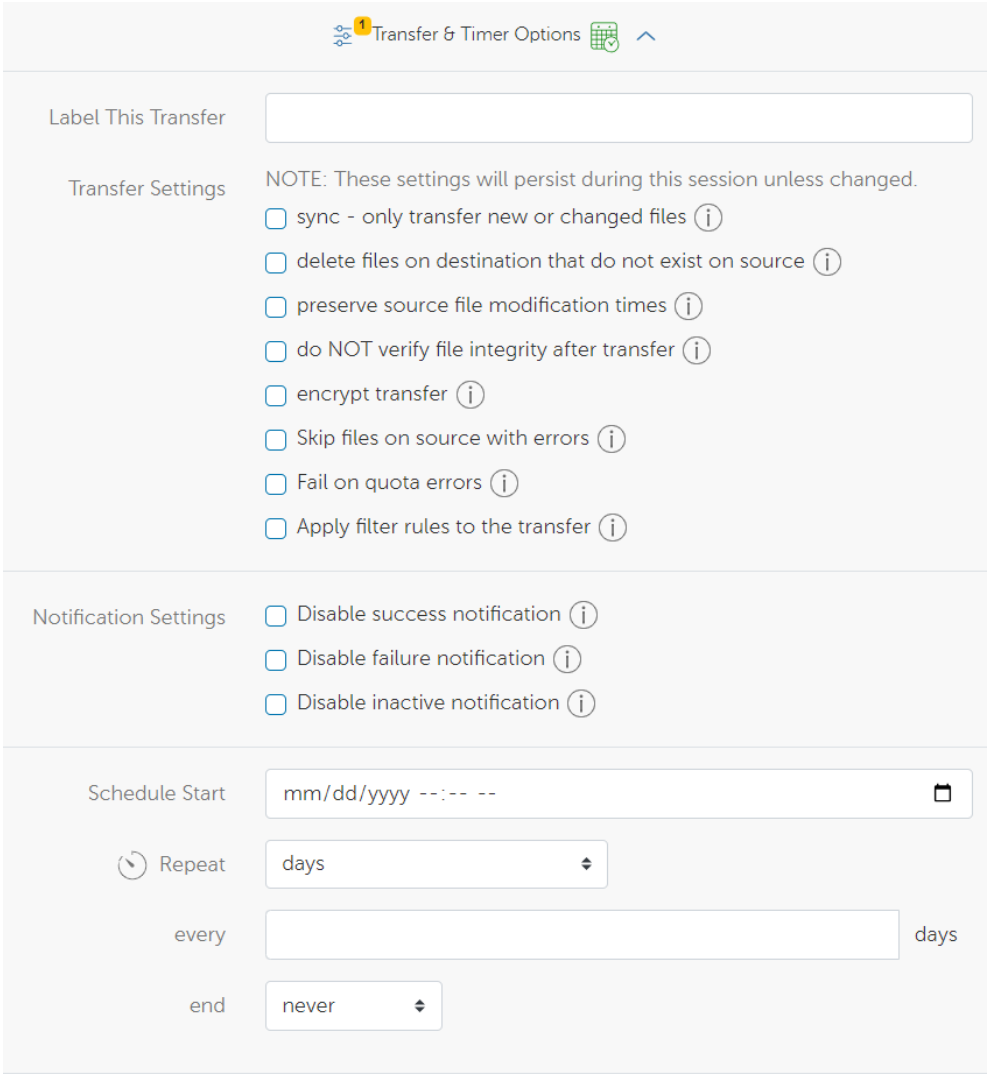
Automating and Scheduling Globus Transfers

To setup a scheduled or recurring transfer, you would start by following the same steps as a manual transfer by selecting a data source and destination. Next select the transfer settings.

To automate/schedule you would need to complete the additional Timer Options:

- **Schedule Start**: The date and time you want to schedule the automated transfer
- **Repeat**: Define the unit of measurement for how often you want the scheduled transfer to repeat. Options are days, hours, and minutes
- **Every**: Provide a numerical value to define the interval between recurring transfers based on the chosen unit of measurement
- **End**: Define the end date that the recurring transfers should stop

Scheduled and repeating transfer tasks can be viewed in the Timers tab of the Activity page.



Key Steps in Using Globus CLI and Globus SDK

- Authentication
- Token handling
- Define Transfer Task
  - Source
  - Destination
  - Options
- Submit Transfer

Authentication

Authentication is the process of verifying the identity of a user or system. It ensures that the entity accessing a system is who they claim to be. This process protects sensitive data and resources from unauthorized access.

- **Password-Based Authentication**: Users provide a username and password to log in.
- **Token-Based Authentication**: Uses tokens for identity verification.

What is a token?

- **Token**:
  - A token is a string of characters that serves as a means of authentication and authorization.
  - Tokens are used to grant access to resources and perform actions within a system.
  - They eliminate the need to repeatedly provide credentials.

What is an Auth Token?

- **Auth Token**:
  - An Auth Token, or Authentication Token, is used to verify the identity of a user.
  - It is issued after successful authentication (e.g., logging in with a username and password).
  - The Auth Token grants the holder permission to access Globus services and APIs.
  - It contains user identity information and is used to obtain additional tokens for specific services

What is a Refresh Token?

- **Refresh Token**:
  - A Refresh Token is a special token used to obtain new access tokens without requiring the user to re-authenticate.
  - It is a long-lived token that can be used to refresh an expired access token.
  - Issued alongside the access token during initial authentication.
  - When an access token expires, the refresh token is sent to the authorization server to obtain a new access token.

Define Source/Destination Collections

Name	UUID
Northwestern Quest	d5990400-6d04-11e5-ba46-22000b92c6ec
Northwestern Quest RDSS	8f796c9e-f5c8-11e5-9842-22000b9da45e
Northwestern AWS us-east-1 N. Virginia	bf0474ea-df2f-45f6-82e3-41d062d84ce7
Northwestern AWS us-east-2 Ohio	a14148f0-23dc-488f-affd-bd981f919b10
Northwestern AWS us-west-2 Oregon	880cd765-5d14-43f6-97a6-2dbb355f3560
Northwestern OneDrive Pilot	930c2fcb-416e-4540-a757-496f86acbe70

*It is recommended that you leverage Guest Collections for automation scripts

What are Optional Elements?

Additional parameters that can be specified to customize and control the behavior of Globus transfer tasks. They provide flexibility and additional functionality for managing file transfers.

Command Options

- -h, --help:
  - Displays help information for the Globus transfer command.
- -r, --recursive:
  - Transfers directories recursively, including all subdirectories and files.
- -F, --force:
  - Overwrites existing files at the destination with no confirmation prompt. (Use with caution!)
- -L, --no-overwrite:
  - Skips transferring files that already exist at the destination.

Task Configuration

- --label <label>:
  - Assigns a descriptive label to the transfer task for easier identification and tracking.
- --sync-level <level>:
  - Defines the synchronization level, controlling how the transfer handles existing files (e.g., update, delete, etc.).
- --verify-checksum:
  - Verifies that the contents of the file were not corrupted in transit by doing a checksum comparison.
- --notification <email>:
  - Sends email notifications about the transfer status (success, failure, etc.) to the specified address.

Task Configuration

- --encrypt:
  - Encrypts the transfer in transit.
- --preserve-mtime:
  - Preserves source file modification time.
- --skip-source-errors:
  - Skips files on the source that generate errors.

Globus CLI

The Globus CLI is a powerful tool that allows users to interact with Globus services directly from the command line, facilitating scriptable transfers and task management.

- Cross-Platform Compatibility. Available for Windows, macOS, or Linux
- Allows increased control and flexibility of all aspects of your transfers

Globus CLI Live Demo Prerequisites

- Prerequisites:
  - Install Globus CLI to Quest
  - Define source and destination collection variables based on UUIDs
  - Provide endpoint consent to use the CLI

Install Globus CLI to Quest

```
module load python-anaconda3/2019.10
conda create --no-default-packages --name globus pip
source activate globus
pip install globus-cli
```

Authentication and Consent with Globus CLI

Authentication is the first step in using the Globus CLI.

- Execute **globus login** to authenticate via your default web browser.
- This process will grant you an Auth Token and a Transfer Token, which are necessary for accessing Globus services.
- These tokens are stored locally and used by the CLI to authenticate your commands.

Define Source/Destination Collections

Name	UUID
Northwestern Quest	d5990400-6d04-11e5-ba46-22000b92c6ec
Northwestern Quest RDSS	8f796c9e-f5c8-11e5-9842-22000b9da45e
Northwestern AWS us-east-1 N. Virginia	bf0474ea-df2f-45f6-82e3-41d062d84ce7
Northwestern AWS us-east-2 Ohio	a14148f0-23dc-488f-affd-bd981f919b10
Northwestern AWS us-west-2 Oregon	880cd765-5d14-43f6-97a6-2dbb355f3560
Northwestern OneDrive Pilot	930c2fcb-416e-4540-a757-496f86acbe70

*It is recommended that you leverage Guest Collections for automation scripts

Provide endpoint consent to use the CLI

After you authenticate with globus, you must provide consent to use the endpoints that you will access via the Globus CLI

- globus session consent 'urn:globus:auth:scope:transfer.api.globus.org:all[*https://auth.globus.org/scopes/**d5990400-6d04-11e5-ba46-22000b92c6ec**/data_access]'

Build a Transfer Task (CLI)

The **globus transfer** command is the foundation for initiating a data transfer.

globus transfer [OPTIONS] SOURCE_ENDPOINT_ID[:SOURCE_PATH] DEST_ENDPOINT_ID[:DEST_PATH]

- Required Elements:
  - **SOURCE** and **DESTINATION** endpoints identified by their UUID
  - File and directory **PATHS**. These paths are relative to the selected endpoints
- Optional Elements:
  - Globus transfer **OPTIONS** as CLI parameters
- Reference:
  - https://docs.globus.org/cli/reference/transfer/

Checking Task Status/Results (CLI) Live Demo

- High level task status
  - globus task show <taskID>
- Successful transfers
  - globus task show <taskID> -t
- Skipped files
  - globus task show <taskID> --skipped-errors
- Event list
  - globus task event-list <taskID>
  - globus task event-list --filter-errors <taskID>
- Cancel Task
  - globus task cancel <TASK-ID>
- List
  - globus ls <UUID>:/Path/to/your/directory
- Reference:
  - https://docs.globus.org/cli/reference/task_show/
  - https://docs.globus.org/cli/reference/task_event-list/

Bash Script

```
#!/bin/bash
#SBATCH -A a9009              # Allocation
#SBATCH -p short              # Queue
#SBATCH -t 00:40:00           # Walltime/duration of the job
#SBATCH -N 1                  # Number of Nodes
#SBATCH --mem=1G              # Memory per node in GB needed for a job. Also see --mem-per-cpu
#SBATCH --ntasks-per-node=1   # Number of Cores (Processors)
#SBATCH --mail-user=brian.roland@northwestern.edu

module load python-anaconda3/2019.10
source activate globus
source_ep=0d2961d1-cb8a-4c10-afe9-1d1edb1ff9c2 ##Quest Bootcamp Demo
dest_ep=b306501e-284a-4013-8b8e-0769877d9823 #RCDS Bootcamp Demo
globus transfer -r $source_ep:~/myFiles/ $dest_ep:~/
```

What is the Globus SDK?

SDK stands for Software Development Kit. The Globus SDK provides a set of Python tools that allow you to:

- Automate Transfers: Instead of clicking through websites, write scripts to transfer files exactly when you need them, scheduled or on-demand.
- Integrate with your Applications: Embed Globus transfer capabilities directly within your own research tools and workflows.

Why use the Globus SDK?

- Custom Control: Go beyond the standard Globus web interface for tailored processes.
- Efficiency: Automate tasks that would be tedious to do manually, especially for many files.
- Power: Tap into the full potential of Globus for features like resuming failed transfers, monitoring progress, and more.

Globus SDK Live Demo

Prerequisites

- Register your applications via Globus Developer Console found in Globus Settings
  - Thick Client (Native App)
  - Script Client (Service Account App)
- Add application as user on Globus Guest Collections as needed

Registering application with Globus

Any applications or scripts that will leverage the Globus SDK need to first be registered with Globus via the Globus Developer Portal.

The process of registering your application with Globus is a two-phase process:

1. Define a Globus Project
2. Register the application under a Globus Project

Globus Project

A Globus Project is a way to Organize and manage related applications, endpoints, and other resources.

You can group multiple applications under a single umbrella for easier management

Create Project

- Access the Developer Portal:
  - Log in to the Globus Developer Portal
- Create a New Project:
  - Navigate to "Projects".
  - Click on "Create a Project".
  - Provide the required details such as Project Name and Contact Email.
  - Submit the form to create the project.

Register an App

After creating the Globus Project to house your application(s), you will next need to start the process of registering your application.

Globus will provide a list of application types that you can chose to register with along with a description of the application type.

Most research use cases will leverage:

- Service account or application credential for automation (Service Account App)
- Thick client or script that will be installed and run by users on their devices (Native App)

Native App

- An application that operates with user interaction
- Uses OAuth2 for user authentication and authorization.
- An application installed on a user's device (e.g., desktop, mobile).
- Acts on behalf of the authenticated user
- Client ID used as part of authentication process
- Use **NativeAppAuthClient** to start the OAuth2 flow.
- Redirects the user to the Globus login page for authentication and authorization code.

Service Account App

- An application that operates without user interaction.
- Typically used for backend services or automated tasks.
- Client ID and Client Secret used for authentication
- Does NOT act on behalf of other users.
  - Globus Guest Collections required to provide app access to Collection data
- Use **ConfidentialAppAuthClient** to authenticate with the Globus Auth service.
- Request an access token using the client credentials flow.

Register App

- Select the Project:
  - In the Globus Developer Portal, select the project under which you want to register the application.
- Navigate to Applications:
  - Click on the "**Apps**" tab within the project.
- Register a New Application:
  - Click on “**Add an App**".
  - Select App type and fill in the required details, including:
    - Application Name
    - Redirect URIs (https://auth.globus.org/v2/web/auth-code)
- Submit the form to register the application.
- Configure Application Settings:
  - After registration, configure additional settings such as OAuth2 settings, generate Client Secret, and more.

Globus SDK Live Demo

- Authentication
- Token Management
- Build Transfer Task
- Submit Transfer
- Check Results

References

- Globus Doc
- Globus CLI
- Globus SDK
- Globus github

Questions?

globus-help@northwestern.edu