Simplified Research Data Management with the Globus Platform

CNI Membership Meeting, Fall 2018 - Project Update

Vas Vasiliadis
vas@uchicago.edu
Topics

• What is Globus?
• Globus from a researcher’s perspective
• Common use cases: research data automation
• Data publication with Globus
• Sustainability – it’s in our DNA
Research data management today (*circa* 2008)

How do we...

...move, share, describe, discover, reproduce?

Facilitate data stewardship
Globus: A Brief History of Time

- Oct. 1998 – Globus Toolkit v1.0.0
- Nov. 2010 – Globus Online initial release
- Nov. 2013 – Sustainability model launched
- Dec. 2016 – 50,000 registered users, 200PB+ moved
- Jan. 2018 – Globus Toolkit support EOL
- Jan. 2019 - 100th subscriber signed, >50% sustainable
- ??? – Globus becomes fully self-sustaining
Globus… bridges data and people within and beyond organizational boundaries
Unified access to data across storage tiers
Sharing with collaborators, community

Public / private cloud stores

External campus storage

Project repositories, replication stores

Public repositories

Sharing with collaborators, community

Public repositories

Public / private cloud stores

External campus storage

Project repositories, replication stores

Sharing with collaborators, community
Globus: Core functions

1. Researcher initiates transfer request; or requested automatically by script, science gateway.
2. Globus transfers files reliably, securely.
3. Researcher selects files to share; selects user or group, and sets access permissions.
4. Globus controls access to shared files on existing storage; no need to move files to cloud storage!
5. Collaborator logs in to Globus and accesses shared files; no local account required; download via Globus.
6. Researcher assembles data set; describes it using metadata (Dublin core and domain-specific).
7. Curator reviews and approves; data set published on campus or other system.
8. Peers, collaborators search and discover datasets; transfer and share using Globus.

- Use a Web browser
- Access any storage
- Use an existing identity
Demonstration
Protected Data Support

Working with PHI, PII, or CUI? Need to manage HIPAA-regulated data? Globus has you covered!

LEARN MORE  REQUEST PRICING

Research data management simplified.

TRANSFER  SHARE  PUBLISH  BUILD
Log in to use Globus Web App

Use your existing organizational login

e.g., university, national lab, facility, project

University of Chicago

Didn't find your organization? Then use Globus ID to sign in. (What's this?)

Continue

Globus uses CILogon to enable you to Log In from this organization. By clicking Continue, you agree to the CILogon privacy policy and you agree to share your username, email address, and affiliation with CILogon and Globus. You also agree for CILogon to issue a certificate that allows Globus to act on your behalf.

Or

Sign in with Google

Sign in with ORCID ID
Welcome!

Please log in to CiLogon.

CiLogon facilitates secure access to CyberInfrastructure (CI).

To securely logout, you must quit your browser.

CNetID / UCHADID: demo
Password: ***********
Forgot your password?

LOG IN

Need help with your CNetID or UCHADID? If you are a University of Chicago faculty member, student, or staff member you may use myaccount.uchicago.edu to change your password, recertify your account, set up two-factor authentication (2FA), and more. If you have not claimed your CNetID, you may claim it at cnet.uchicago.edu. Further support is available through the ITS Service Desk by calling 773.702.5800, and selecting option "1" from the menu. In-person support is also available at one of the convenient walk-up TechBar locations.

If you are a hospital employee and you do not already have a CNetID you may use your UCHADID to log in. If you do not have a UCHADID, contact the Chicago Biomedicine Information Services (CBIS) Service Desk at 773.702.3456.

If you are an alumni account holder and have questions or need support, contact alumnitechsupport@uchicago.edu or call 1.877.292.3945 between the hours of 9 a.m. and 3 p.m. CST. If you are an alumnus and do not have a CNetID, you may claim it at cnet.uchicago.edu.
Click in the **Collection** field above to begin

Watch a two minute tour of what’s new
<table>
<thead>
<tr>
<th>Collection</th>
<th>Path</th>
<th>Date/Time</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>UChicago RCC Midway</td>
<td>/~/</td>
<td></td>
<td></td>
</tr>
<tr>
<td>active</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mbarchive</td>
<td>9/20/2017 9:57pm</td>
<td>4.09 KB</td>
<td></td>
</tr>
<tr>
<td>projects</td>
<td>8/30/2018 12:30pm</td>
<td>32.76 KB</td>
<td></td>
</tr>
<tr>
<td>sharing</td>
<td>11/14/2018 2:22am</td>
<td>4.09 KB</td>
<td></td>
</tr>
</tbody>
</table>
Please authenticate to access this collection

When you click continue you will be redirected to the endpoint's login page (you will be returned here once you have authenticated).

Continue
Transfer request submitted successfully. Task id: 94d0885a-fcbe-11e8-9344-0e3d676669f4
### Task Details

**Task Label**: NCAR RDA Dataset Archive to petrel#testbed

**Source**: NCAR RDA Dataset Archive  
owner: rda@globus.org

**Destination**: petrel#testbed  
owner: petrel@globus.org

**Task ID**: 2e97e4f2-f8a2-11e8-8cda-0a1d4c5c824a

**Owner**: Vas Vasiliadis (vas@globus.org)

**Condition**: SUCCEEDED

**Requested**: 2018-12-05 09:26 am

**Completed**: 2018-12-05 09:31 am

**Transfer Settings**
- verify file integrity after transfer
- transfer is not encrypted
- overwriting all files on destination

### Transfer Statistics

- Files: 4751
- Directories: 2
- Bytes Transferred: 1.17 TB
- Effective Speed: 4.53 GB/s
- Pending: 0
- Succeeded: 4756
- Cancelled: 0
- Expired: 0
- Failed: 0
- Retrying: 0
- Skipped: 0

[View debug data](#)
Create New Shared Collection

You Are Sharing: UChicago RCC Midway

Path: /~/sharing/

Share Display Name: CNI Member Share

Description: Demonstrating how to make data accessible to external collaborators.

Keywords: file share, attached storage

Next step — add access permissions

Create Share  Cancel
Add Permissions — Share With

Path: /

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

Identity/Email: cni@wossamotta.edu

Username: cni@wossamotta.edu

Send E-mail: cni@wossamotta.edu

Here are the data:

Permissions:
- read
- write

Add Permission  Cancel
<table>
<thead>
<tr>
<th>USER OR GROUP</th>
<th>READ</th>
<th>WRITE</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:cni@wossamotta.edu">cni@wossamotta.edu</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vas Vasilidis (<a href="mailto:vas@globusid.org">vas@globusid.org</a>)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Identities

A list of identities linked to your Globus account

<table>
<thead>
<tr>
<th>PRIMARY</th>
<th>IDENTITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>🏰</td>
<td><a href="mailto:vas@globusid.org">vas@globusid.org</a></td>
</tr>
</tbody>
</table>
|         | This id is managed by Globus ID, where you can change your password, update your profile, and manage SSH and X.509 keys.
|         | vas@globus.org   |
|         | vas@nersc.gov    |
|         | vas@uchicago.edu |
|         | vas@uchicago.edu@trda.ucar.edu |
|         | vas@xsede.org    |
|         | vasv@anl.gov     |
|         | vasv@umich.edu   |
Globus for high assurance data management

• Restricted data handling: PHI, PII, CUI
• Security controls: NIST 800-53, 800-171 Low
• Business Associate Agreement (BAA) w/UChicago
  – University of Chicago has a BAA with Amazon
• **Additional authentication assurance**
  – Per storage gateway policy on frequency of authentication with specific identity for access to data
  – Ensure that user authenticates with the specific identity that gives them access within session

• **Application instance isolation**
  – Authentication context is per application, per session

• **Encryption of user data in transit and Globus data at rest**

• **Detailed audit log (on data transfer nodes)**
Additional logins needed to use Globus Web App

Globus Web App wants you to authenticate with the following account(s) because:

"User must login with vas@uchicago.edu in order to access GCSv5.2 Globus Demo HA Mapped Collection."

- vas@uchicago.edu
Accessing Globus from your own storage
…client software that makes a storage system accessible via Globus
Globus Connect Personal

- Installers do not require admin access
- Zero configuration; auto updating
- No firewall changes required; handles NATs
Globus Connect Server

- Installed and managed by sysadmin
- Default access for all local accounts
- POSIX + connectors
- Native packaging
  - Linux: DEB, RPM

docs.globus.org/globus-connect-server-installation-guide/
Use(r)-appropriate interfaces

Globus service

Web

CLI

Rest API

GET /endpoint/go%23ep1
PUT /endpoint/vas#my_endpt
200 OK
X-Transfer-API-Version: 0.10
Content-Type: application/json
...
Globus Command Line Interface

- Full-featured (web++)
- Uses Python SDK
- Open source

github.com/globus/globus-cli
docs.globus.org/cli
Globus is PaaS…
...for building science gateways, portals, and other web applications in support of research and education.
The Globus Platform

Globus APIs
- Transfer, Search, Identifiers, ...

Globus Auth
- (identity and access management)

Globus Connect
- Data Publication
- File Sharing
- File Transfer, Sync
- Data Automation

Partners:
- NCAR
- NERSC
- NEW YORK GENOME CENTER
- NIH
- NCSA
- KBase
- Welcome Trust Sanger Institute
- Extreme Science and Engineering Discovery Environment

...
Globus Auth

- Foundational Identity and Access Management (IAM) service
- Protects REST API communications
- Enables login for diverse app ecosystem, no new identity required
- Employs least privileges security model
Globus helps automate data flows
Automated data movement

Campus Lab

Recurring transfers with sync option

Scheduled backup

My Drive/backups
- /20170930
- /20170929
- /20170928
- ....
Streamlined data distribution

My Drive/projectX
__/source
__/pipe0001
__/pipe0002
__/....

Discover and access via data portal

Secure sharing with research community

HPC resource, Campus storage, ...

Modern Research Data Portal

Globus Transfer API
API reference for transfer and sharing functions.

Globus Auth API
API reference for authentication and authorization.

Frequently Asked Questions
When all else fails...
Data distribution example: NCAR RDA

### NCEP Climate Forecast System Version 2 (CFSv2) Monthly Products

**ds094.2**

For assistance, contact Bob Dattore (303-497-1825).

#### Data Description

<table>
<thead>
<tr>
<th>Data Description</th>
<th>Data File Downloads</th>
<th>Customizable Data Requests</th>
<th>Other Access Methods</th>
<th>NCAR-Only Access</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Web Server Holdings</td>
<td>Subsetting</td>
<td>THREDDS Data Server</td>
<td>GLADE File Listing</td>
</tr>
<tr>
<td><strong>Globus Transfer Service (GridFTP)</strong></td>
<td></td>
<td></td>
<td></td>
<td>GLADE File Listing</td>
</tr>
<tr>
<td>Union of Available Products</td>
<td>Web File Listing</td>
<td>Get a Subset</td>
<td>TDS Access</td>
<td>HPSS File Listing</td>
</tr>
<tr>
<td>Diurnal monthly means</td>
<td>Request globus Invitation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular monthly means</td>
<td>Get a Subset</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRON</td>
<td>Web File Listing</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Description

- **Home**: Link to the homepage of NCAR's Research Data Archive.
- **Find Data**: Search for specific datasets.
- **Ancillary Services**: Access additional services related to data.
- **About/Contact**: Contact information and other details.
- **Data Citation**: Instructions for citing the data.
- **Web Services**: Various services for accessing the data.

### Access Methods

- **Web Server Holdings**
- **GridFTP** (Globus Transfer Service)
- **SUBSETTING**
- **THREDDS Data Server**
- **Central File System (GLADE) Holdings**
- **Tape Archive (HPSS) Holdings**

### Available Products

- **Diurnal monthly means**
- **Regular monthly means**

**For Technical Assistance:**

- **Send an email** to tuecke@uchicago.edu

---

**Disclaimer:** Information is subject to change. Always verify data availability and access methods directly from the NCAR RDA website.
Reliable instrument data egress

NGS and high-res Imaging
(APS, ALS, CryoEM, fMRI, …)

Stage data for downstream analysis

My Drive/FASTQ
/cohort_0_0
/cohort_0_1
/cohort_0_2
….
Instrument data egress example

- Kasthuri Lab at Uchicago: brain aging and disease
- Construct connectome (map neuron connections)
Kasthuri Lab neuroanatomy reconstruction pipeline

1 Imaging
2 Acquisition
3 Pre-processing
4 Preview/Center
5 User validation
6 Reconstruction
7 Publication
8 Visualization
9 Science!

UChicago

APS

Lab Server 1
Lab Server 2

ALCF

PETREL

JLSE
Data Management Plan enablement

Diverse storage systems

Dataset assembly, description, curation

Access via persistent identifier

http://hdl.handle.net/11466/OMN5BFB
Globus Data Publication V1

- Cloud-based web app
- BYO storage
- User-managed collections
- Select pre-defined schema
- Handle, DOI persistent identifiers
- >2000 users, >600 datasets

publish.globus.org
Many variations of data publication...

**Citable Data**
- Standard metadata
- Persistent identifiers
- Durable storage

**Institutional Data**
- Many domains
- Custom metadata
- Locally managed storage

**Community Data**
- Agreed schema
- Larger datasets
- Fine grained metadata
Active Research Data

- Less standard and evolving schema
- Data organized independent of storage
- Support active collaboration
- Location agnostic identifiers
Publication v2 Platform

- Decompose Globus turnkey solution into microservices
- Enable flexible re-composition and adaptation of services
- Support extension and enhancement of publication flows

**Automate**
- Get credentials

**Transfer**
- Create folder
- Set ACL
- Transfer data

**Describe**
- Get metadata

**Identify**
- Mint persistent identifier

**Search**
- Catalog
Globus Search service

- Hosted, scalable service for research data discovery
- Schema agnostic
- Fine grained access control
- Plain text search
- Faceted search
- Rich query language
Globus Identifiers service (limited beta release)

- Issue persistent identifiers...
- ...within *your* namespace, with access control
- Identifiers have...
  - ...link to data
  - ...landing page
  - ...visibility
  - ...checksum
  - ...extensible metadata
  - ...versioning
Globus Automate (coming soon)

• Composition and execution service for automating research data management

• Higher level flow description language and authoring tools

• Pluggable API to integrate any actions
  – e.g. automated validation, metadata extraction

• Flexible invocation of actions: user or event driven
Globus platform applications
Jupyter + Globus for interactive data science at scale

3.2M materials data

2PB, 80Gbps store

Cooley: 290 TFLOPS
Genotype imputation: Wellcome Sanger
National Resource Access

Jetstream Web App would like to:

- Access all Jetstream resources

By clicking "Allow", you allow Jetstream Web App to access your information and services. You can rescind this permission at any time.

Is there a reason why you would not want to grant access to all Jetstream resources?

Compute Canada has partnered with Globus to offer this high performance file transfer service. Calcul Canada s’est associé à Globus pour vous offrir ce service de transfert de fichier à haute performance.

Log in to use Compute Canada Globus Web App

Use your existing organizational login

- e.g. university, national lab, facility, project, Google or
- Globus ID

(Your Globus username and password used prior to February 13, 2016 is now Globus ID)

- WestGrid

Continue

Didn't find your organization? Then use Globus ID to sign up.
Identity Management

The new Systems Biology Knowledgebase (KBase) is a collaborative effort designed to accelerate our understanding of microbes, microbial communities, and plants. It will be a community-driven, extensible and scalable open source software framework and application system. KBase will offer free and open access to data, models and simulations, enabling scientists and researchers to build new knowledge and share their findings.

Accelerate research on campus by providing connective services for local, cloud and national cyberinfrastructure

Connecting Science to Cycles and Data
Ci Connect offers campus IT organizations connective services for cyberinfrastructure. End-users quickly gain access to shared campus clusters while gaining access to distributed resources of national HPC or HTC facilities (such as the Open Science Grid) or cloud-based resources. Ci Connect offers several capabilities based on Globus, HTCondor and other HTC and data sharing technologies that augment your vision of providing scientific computational resources to research communities you support.

Reliable, High Performance CI Services
Ci Connect provides services to CI providers at US universities and laboratories following a simple set of principles:
Globus PaaS developer resources

Python SDK

Sample Application

Jupyter Notebook

Requirements
- You need to be in the tutorial users group for sharing: https://www.globus.org/app/groups/50b6a29c-63ac-11e4-8062-22000ab68755
- Installed Globus Python SDK

```
In [15]: from __future__ import print_
tutorial_endpoint_1 = "ddb59a40-6d04-11e5-ba46-22000b92c6ec"  # endpoint "Globus CLI"
tutorial_endpoint_2 = "ddb59a40-6d04-11e5-ba46-22000b92c6ec"  # endpoint "Globus CLI"
tutorial_users_group = "50b6a29c-63ac-11e4-8062-22000ab68755"  # group "Tutorial"
```

Configuration
First you will need to configure the client with an OAuth2 access token. For the purpose of this tutorial, you can use the website. Click the “Jupyter Notebook” option and copy the resulting text below, or click on “Globus CLI” and

```
In [16]: transfer_token = None  # if None, tries to get token from ~/.globus.cfg file
```
...on sustainability
Globus by the numbers

- 1,923 most shared endpoints at a single institution
- 500 PB moved
- 70+ petabyte movers
- 80 billion files processed
- 120,000 users
- 1,800 active server endpoints
- 94 subscribers
- 20,400 active personal endpoints
- 559 identity providers
- 1 PB largest single transfer to date
- 8,300 active shared endpoints
- 99.9% availability
Thank you to our sponsors...
...and THANK YOU, subscribers!
Globus sustainability model

- **Standard Subscription**
  - Sharing, data publication
  - HTTPS access
  - Console, usage reporting
  - Priority support
  - App integration support

- **High Assurance subscription**
  - App instance isolation
  - Additional authentication assurance
  - Audit logging
  - NIST 800-53, NIST 800-171 (+ BAA)

- **Branded Web Site**
- **Premium Storage Connectors**
- **Alternate Identity Provider** (InCommon is standard)
Support resources

• Globus documentation: docs.globus.org
• Community email list: developer-discuss@globus.org
• Helpdesk and issue escalation: support@globus.org
• Customer engagement team
• Globus professional services team
  – Assist with portal/gateway/app architecture and design
  – Develop custom applications that leverage the Globus platform
  – Advise on customized deployment and integration scenarios