Enhanced Research Data Management and Publication with Globus

Vas Vasiliadis
Jim Pruynne

Presented at OR2015
June 8, 2015
Presentations and other useful information available at

globus.org/events/or2015/tutorial

bit.ly/or2015-globus
Thank you to our sponsors!

U.S. DEPARTMENT OF
ENERGY

NSF

ALFRED P. SLOAN FOUNDATION

THE UNIVERSITY OF
CHICAGO

NATIONAL INSTITUTES
OF HEALTH

Argonne
NATIONAL LABORATORY

powered by
amazon
web services
Agenda

- Research data management scenarios and introduction to Globus
- Demonstrations and Exercises
  - Accessing Globus and moving data
  - Sharing data and group management
  - Data publication and discovery
  - Creating collections
- Campus deployment overview
- Globus today and tomorrow
Research data management scenarios and challenges
“I need to easily, quickly, & reliably move portions of my data to other locations.”
“I need to get data from a scientific instrument to my analysis system.”
“I need to easily and securely share my data with my colleagues at other institutions.”
“I need to publish my data so that others can find it and use it.”
Globus is...

Research data management...

delivered via SaaS
Globus delivers...
Big data transfer, sharing, publication, and discovery...
...directly from your own storage systems
Our focus: User Experience

flickr ...for your photos

Google ...for your office docs

Netflix ...for your entertainment

globus ...for your research data
Globus is SaaS

• Easy to access via Web browser
  – Command line, REST interfaces for flexible automation and integration

• New features automatically available

• Reduced IT operational costs
  – Small local footprint (Globus Connect)
  – Consolidated support and troubleshooting
Globus and the research data lifecycle

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.

• SaaS → Only a web browser required
• Use storage system of your choice
• Access using your campus credentials
Demonstration: Accessing Globus and Moving Data
Exercise 1: Sign up & transfer files

1. Go to: www.globus.org/signup
2. Create your Globus account
3. Validate e-mail address
4. Optional: Login with your campus/InCommon identity
5. Install Globus Connect Personal
6. Move files from esnet#... endpoint to your laptop
Demonstration: Sharing Data
Exercise 2: Share files

1. Join the “Tutorial Users” groups
   – Go to “Groups”, search for “tutorial”
   – Select group from list, click “Join Group”

2. Create a shared endpoint on your laptop

3. Grant your neighbor permissions on your shared endpoint

4. Access your neighbor’s shared endpoint
Demonstration: Group Management
Exercise 3: Create/configure group

1. **Create a group**
   - Go to globus.org/groups
   - Click “Create New Group”
   - Enter the group name and a short description
   - Set visibility to “all Globus members”

2. **Configure your group policies**
   - Select your group and click the “Settings” tab
   - Set requests to “a logged in Globus user”
   - Set approvals to “automatically if all policies are met”

3. **Ask your neighbor to join your group**

4. **Grant permissions to the group on your shared endpoint**

5. **Confirm your neighbor can access your shared endpoint**
“Appropriate” preservation
Globus data publication framework

### Identifier
- **URL**
- **Handle**
- **DOI**

### Description
- **None**
- **Standard**
- **Domain-specific**
- **Custom**

### Curation
- **None**
- **Acceptance**
- **Human-validated**
- **Machine-validated**

### Access
- **Anonymous**
- **Public**
- **Embargoed**
- **Collaborators**

### Preservation
- **Transient**
- **Project Lifetime**
- **Archive**
- **“forever”**
Raw NGS output

No curation
- Automated dataset acceptance

Minimal metadata...
- Source environment
  - Instrument, timestamp, ...
  - Unique ID

Identify...
- Handle

High durability, low cost store

Glacier
Upstream analysis

- Pipeline description
- Tool parameters
- Exec environment

Campus HPC

Automated curation
- Machine validated
- Exception review

Processing metadata...

Moderate durability/cost

Identify...
- URL
Downstream analysis

Optional metadata...
- “Implicit” metadata
- Description through organization

Widely accessible stores
- Any collaborator may approve
- Globus share
Peer reviewed paper

Formal, multi-step review
- Review → Update → Resubmit cycle

(Re)format...
- PDF/A
- HDF
- ...

Fully described...
- Dublin core metadata
- Domain metadata
- Provenance info

Persistent identifier
- DOI

Replicated, public repositories
# Globus publication - Initial release

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Supported in GA release</th>
<th>Consulting support</th>
<th>Planned</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handle</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Standard</td>
<td>Domain-specific</td>
<td>Custom</td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Curation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>Acceptance</td>
<td>Human-validated</td>
<td>Machine-validated</td>
</tr>
<tr>
<td>Curation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anonymous</td>
<td>Public</td>
<td>Embargoed</td>
<td>Collaborators</td>
</tr>
<tr>
<td>Access</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preservation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transient</td>
<td>Project Lifetime</td>
<td>Archive</td>
<td>“forever”</td>
</tr>
</tbody>
</table>
Demonstration: Data Publication and Discovery
Exercise 4: Publish a dataset

1. Go to trial.publish.globus.org
2. Log in, click “Start a New Submission”
3. Select the “Globus Demo Collection”
4. Agree to the license
5. Enter the required metadata
6. Assemble data set from the esnet#... Endpoints (or your own laptop if you installed Globus Connect Personal)
7. Complete the workflow and submit
8. Curators (a.k.a. presenters) will “review” your submission and publish
9. Search for your published dataset and browse the data
Example use cases
Mix-Model: Technology meets Pedagogy

• **Central IT provides infrastructure**
  – Storage, Computing, Cluster, Servers...

• **Library responsible for data stewardship**
  – Collection, Acquisition, Search & Discovery, Metadata, Preservation...

• **Staff: technologists + librarians and subject specialists**

Source: H. Mistry, New York University
Repository planning

Source: H. Mistry, New York University
Reproducibility

Post-processing pipeline; Investigator

Transform
Reduce
Transform
Reduce
Transform
Reduce
Transform
Reduce

Third-party validation

Visualize

Source: M. Hutchinson, R. Rosner, University of Chicago; Argonne; Image: UNM
Reproducibility

Figure 1

Start by loading some boiler plate: matplotlib, numpy, scipy, json, funtools, and a convenience class.

```python
In [1]: %matplotlib inline
import matplotlib
matplotlib.rcParams['figure.figsize'] = (10.0, 8.0)
import matplotlib.pyplot as plt
import numpy as np
from scipy.interpolate import interp1d, InterpolatedUnivariateSpline
import json
from functools import partial
class Foo: pass
```

And some more specialized dependencies:

1. Slict provides a convenient slice-able dictionary interface
2. Chest is an out-of-core dictionary that we'll hook directly to a globus remote using...
3. glopen is an open-like context manager for remote globus files

```python
In [2]: from chest import Chest
from slict import CachedSlict
from glopen import glopen, glopen_many
```

Configuration for this figure.

```python
In [3]: config = Foo()
config.name = "HighAspect/HA_visc/HA_visc"
config.arch_end = "maxhutch#alpha-admin/~/pub/"
```
Demonstration: Collection Configuration
Exercise 5: Create a collection

1. Create a new collection

2. Enter metadata
   1. Name and submission license (required)
   2. Description (optional)

3. Enter the endpoint for collection storage: globuspublish#or2015-tutorial

4. Enter a prefix of your choice

5. Identifier: Select OR2015: bit.ly

6. Leave all other fields at default values

7. Select Curation Group (Tutorial Users)

8. Submit a dataset for publication into your new collection

9. Review and approve your neighbor’s submission
Campus Deployment Overview
Globus Connect Server

- Create endpoint in minutes; no complex software install
- Enable all users with local accounts to transfer files
- Native packages: RPMs and DEBs
Standard package installation

1. Install Globus Connect Server
   - Access server as user "clusteradmin"
   - Update repo
   - Install package
   - Setup Globus Connect Server

2. Log into Globus (using your Globus username)

3. Access the newly created endpoint (as user 'researcher')

4. Transfer a file
Common Configurations

- Change endpoint name
- Customize filesystem access
- Enable sharing; set path restrictions
- Integrate with campus Id system
- Scale your campus deployment
  - Data node(s)
  - Science DMZ
Typical deployment

Science DMZ + Globus

Details at: fasterdata.es.net
Demonstration: Globus Command Line Interface (CLI)
Globus: today and tomorrow
Globus today…

~ 100PB moved

>10,000 endpoints

>300 active users/day
We are a non-profit, delivering a production-grade service to the non-profit research community.
We are a non-profit, delivering a production-grade service to the non-profit research community.

Our challenge:

Sustainability
Globus Provider Subscriptions

- **Globus Provider Plan**
  - Shared endpoints
  - Data publication
  - Amazon S3 endpoints
  - Management console
  - Usage reporting
  - Priority support
  - Application integration

- **Branded Web Site**

- **Alternate Identity Provider** (InCommon is standard)

- **Mass Storage System optimization**

[globus.org/provider-plans](globus.org/provider-plans)
Demonstration:
Globus management console
Globus Platform-as-a-Service

- Data Publication & Discovery
- File Sharing
- File Transfer & Replication
- Identity, Group, and Profile Management
- Globus Toolkit
Some early Globus supporters
Enable your campus

- Signup: globus.org/signup
- Enable your resource: globus.org/globus-connect-server
- Need help? support.globus.org
- Subscribe to help make Globus self-sustaining globus.org/provider-plans
- Follow us: @globusonline
Thank you