Research Data Management using Software-as-a-Service

Vas Vasiliadis
vas@uchicago.edu

EDUCAUSE 2015
Thank you to our sponsors!
Agenda

• Research data management challenges
• Globus: a high-level flyover
• Accelerating and streamlining collaboration: transfer and sharing
• Enhancing reproducibility and discoverability: data publication
• Our sustainability challenge
• Campus deployment, security overview
• Leveraging the Globus platform
Who are you?
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 others can find/use/validate/reproduce it."
Research data management today

Index?
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. Peers, collaborators search and discover datasets; transfer and share using Globus.

7. Curator reviews and approves; data set published on campus or other system.

8. Publication Repository.

- Only a Web browser required
- Use storage system of your choice
- Access using your campus credentials
Globus delivers...
Big data transfer, sharing, publication, and discovery...
...directly from your own storage systems...
...via software-as-a-service
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
Our focus: User Experience

flickr  ...for your photos

Google  ...for your office docs

NETFLIX  ...for your entertainment

globus  ...for your research data
Accessing Globus and Moving Data
Example: Scaling up

Move datasets to supercomputer, national facility

Move results to campus (…)

Edison
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 vas#ebs endpoint to your laptop
Sharing Data
Lowering collaboration overhead

- Grant collaborators access to data on systems without requiring local accounts
- No need to replicate or move data to separate system/cloud just for sharing
- Researchers manage “virtual” ACLs...
- Respect local system access controls
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
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
Enhancing reproducibility and discoverability
# 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

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

No curation
- Automated dataset acceptance

Identify...
- Handle

High durability, low cost store

Glacier

Raw NGS output
- Automated dataset acceptance
- Source environment
  - Instrument, timestamp,…
  - Unique ID
- No curation
- 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...

Identify...
- URL

Moderate durability/cost

Moderate durability/cost

Automated curation
- Machine validated
- Exception review

Processing metadata...

Identify...
- URL

Moderate durability/cost
Downstream analysis

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

Team review
- Any collaborator may approve

Identify...
- Globus share

Widely accessible stores

Any collaborator may approve

Globus share
Peer reviewed paper

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

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

(Re)format...

Replicated, public repositories

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

Persistent identifier
- DOI

- PDF/A
- HDF
- ...

- Fully described...
- (Re)format...
- Replicated, public repositories
- Persistent identifier
- Review → Update → Resubmit cycle
- (Re)format...
Globus publication - Initial release

<table>
<thead>
<tr>
<th>Identifier</th>
<th>URL</th>
<th>Handle</th>
<th>DOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Standard</td>
<td>Domain-specific</td>
<td>Custom</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Description</th>
<th>None</th>
<th>Acceptance</th>
<th>Human-validated</th>
<th>Machine-validated</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Curation</th>
<th>None</th>
<th>Anonymous</th>
<th>Public</th>
<th>Embargoed</th>
<th>Collaborators</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Access</th>
<th>Anonymous</th>
<th>Public</th>
<th>Embargoed</th>
<th>Collaborators</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Preservation</th>
<th>Transient</th>
<th>Project Lifetime</th>
<th>Archive</th>
<th>“forever”</th>
</tr>
</thead>
</table>
Publish a dataset

1. Go to trial.publish.globus.org
2. Log in, click “Submit a New Dataset”
3. Select either of the Open Trial collections and continue
4. Accept the license terms
5. Enter required metadata to describe the dataset
6. Assemble data set from the vas#ebs endpoint (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
Reproducibility example

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 example

Figure 1

Start by loading some boiler plate: matplotlib, numpy, scipy, json, functools, 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. Slicet 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 slicet import CachedSlicet
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/"
```
Repository planning

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
Globus: today and tomorrow
<table>
<thead>
<tr>
<th>Major Services</th>
<th>Total PB Transferred</th>
<th>Files Processed</th>
<th>Registered Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>118 PB</td>
<td>20 billion</td>
<td>31,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>National Labs</th>
<th>Endpoints</th>
<th>Active Daily Users</th>
<th>Uptime</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>10,000</td>
<td>~350</td>
<td>99.95%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Institutional Subscribers</th>
<th>Largest Single Transfer</th>
<th>Longest Continuously Managed Transfer</th>
<th>Federated Campus Identities</th>
</tr>
</thead>
<tbody>
<tr>
<td>35+</td>
<td>1 PB</td>
<td>3 months</td>
<td>130</td>
</tr>
</tbody>
</table>
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
Demonstration:
Globus management console
Demonstration:
Bridging to Cloud Storage
- Amazon S3: supported
- Ceph: coming soon
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 with root privileges
   - Update package repos
   - Install packages
   - Setup Globus Connect Server

2. Log into Globus as end user/researcher

3. Access newly created endpoint

4. Transfer a file
Endpoint activation using MyProxy OAuth

1. Access endpoint
2. OAuth redirect
3. campus username password
4. username password
5. username password
6. certificate
7. certificate
8. certificate
9. Authorization (resolve local user)
10. Access files
11. Transfer request
12. Data transfer

Globus transfer and sharing hosted service

Control channel authorization

session certificate

OAuth Server

MyProxy Online CA

PAM

GridFTP Server

GridFTP Server

Remote cluster with Globus Connect Server or laptop/PC with Globus Connect Personal

Campus Cluster

Local Authentication System (LDAP, RADIUS, Kerberos, ...)

Local Storage
Common Configurations

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

Science DMZ + Globus

Details at: fasterdata.es.net
Use(r)-appropriate interfaces

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

laptop:~ ssh vas@cli.globusonline.org
$ Welcome to globusonline.org, vas. Type 'help' for help.
$ endpoint-modify vas#ebs --organization="University of Chicago"
$

Web

CLI

Globus service

Rest API
Quick look:
Globus Command Line Interface (CLI)
Globus Platform-as-a-Service

- Globus APIs
  - Data Publication & Discovery
  - File Sharing
  - File Transfer & Replication
  - Identity, Group, and Profile Management

- Globus Connect

- Globus Toolkit

Globus Toolkit

- XSEDE
- NERSC
- Michigan
- NCAR
- University of Exeter
- KBase
- Indiana University
Building bridges to global communities
What is the RDA?

• Free and open access to 600+ datasets for climate and weather research
• Worldwide usage
• Multiple data access pathways
  – HTTP (wget, cURL, etc.)
  – OPeNDAP, WCS, WMS
  – Web services (CLI, API)
  – Analysis on HPC systems (NCAR users)

Courtesy of Thomas Cram, NCAR
RDA Usage

- **2014**
  - 17+ PB virtual processing
  - Web downloads: 7300 users, 750 TB served
  - 45,000 custom orders, 4000 users, 380 TB served

Courtesy of Thomas Cram, NCAR
Globus @ RDA

• Single shared endpoint
• Data copied to subdirectories under endpoint source path
• Allow read permission to subdirectories under the shared endpoint
• ACLs managed programatically via Globus CLI
RDA Alternate Identity login

Courtesy of Thomas Cram, NCAR
Welcome to the NCAR RDA OAuth for MyProxy Client Authorization Page. The Client below is requesting access to your account. If you approve, please sign in with your RDA email/username and RDA password.

Client Information
Name: Globus Online
URL: https://www.globusonline.org

NCAR RDA Email/Username: tcram@ucar.edu
NCAR RDA Password: **********
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