

Introduction to Research Data Management with Globus

Vas Vasiliadis
University of Chicago,
Argonne National Laboratory
vas@uchicago.edu

GlobusWorld
April 20, 2016





Presentation material available at
[**globusworld.org/tutorial2016**](https://globusworld.org/tutorial2016)



Thank you to our sponsors!



U.S. DEPARTMENT OF
ENERGY



THE UNIVERSITY OF
CHICAGO

Argonne

NATIONAL LABORATORY



powered by
amazon
web services



Agenda

- **Research data management challenges**
- **Globus: a high-level flyover**
- **File Transfer and Sharing: Accelerating and streamlining collaboration**
- **Data Publication: Enhancing reproducibility and discoverability**
- **Our sustainability challenge**
- **Globus campus deployment & intergation**
- **Deployment best practices: the Science DMZ**
- **Leveraging the Globus platform**



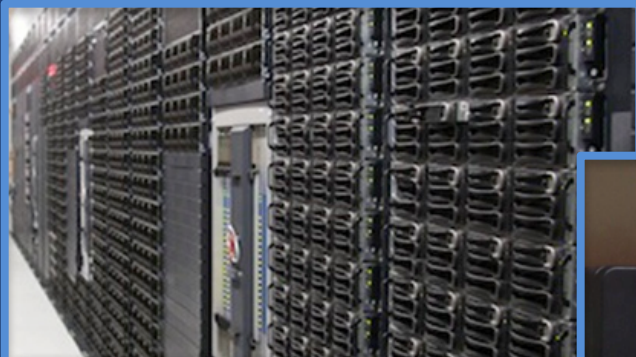
Who are you?



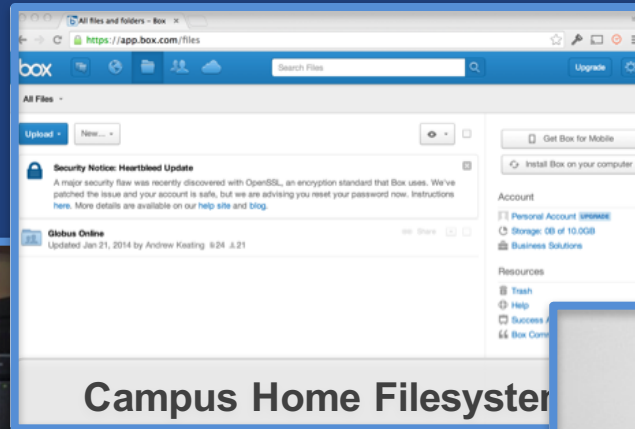
Research data management scenarios and challenges



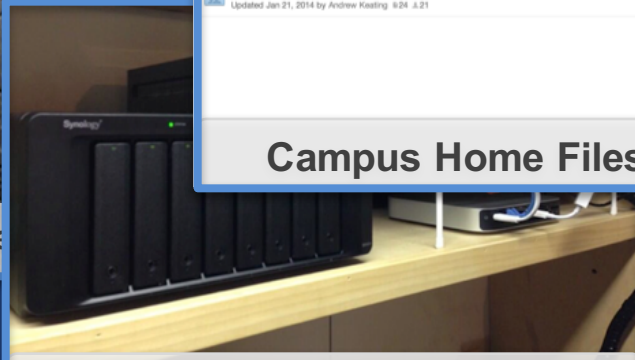
“I need to easily, quickly, & reliably move portions of my data to other locations.”



Research Computing HPC Cluster



Campus Home Filesystem



Lab Server



Personal Laptop



Desktop Workstation



XSEDE Resource



Public Cloud

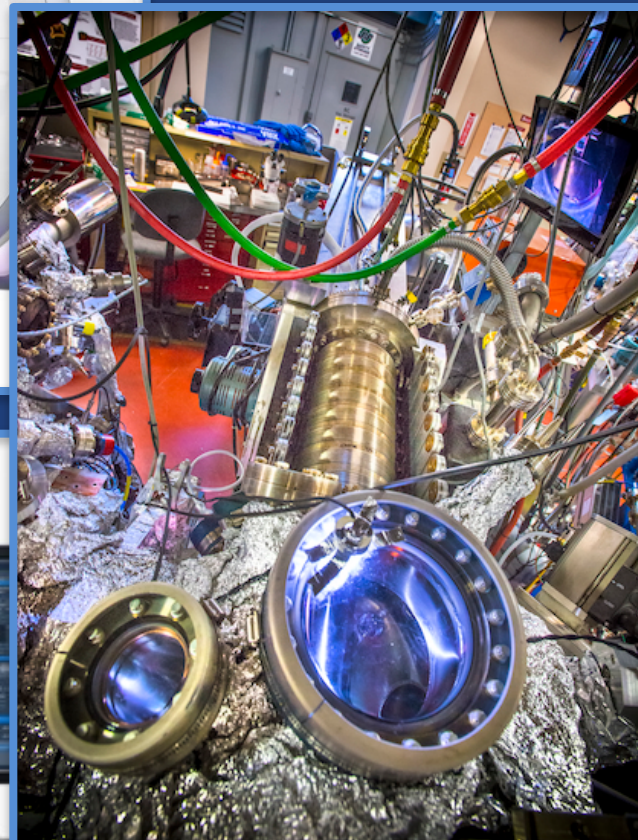


“I need to get data from a scientific instrument to my analysis system.”

MRI



Advanced
Light Source



Next Gen
Sequencer



Light Sheet Microscope



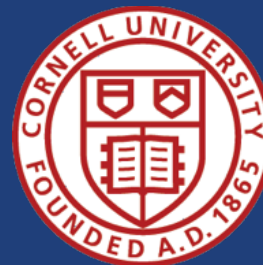
“I need to easily and securely share my data with my colleagues at other institutions.”



NCAR



Computation
Institute





“I need to publish my data so others can find/use/validate/reproduce it.”

Reference
Dataset



Scholarly
Publication



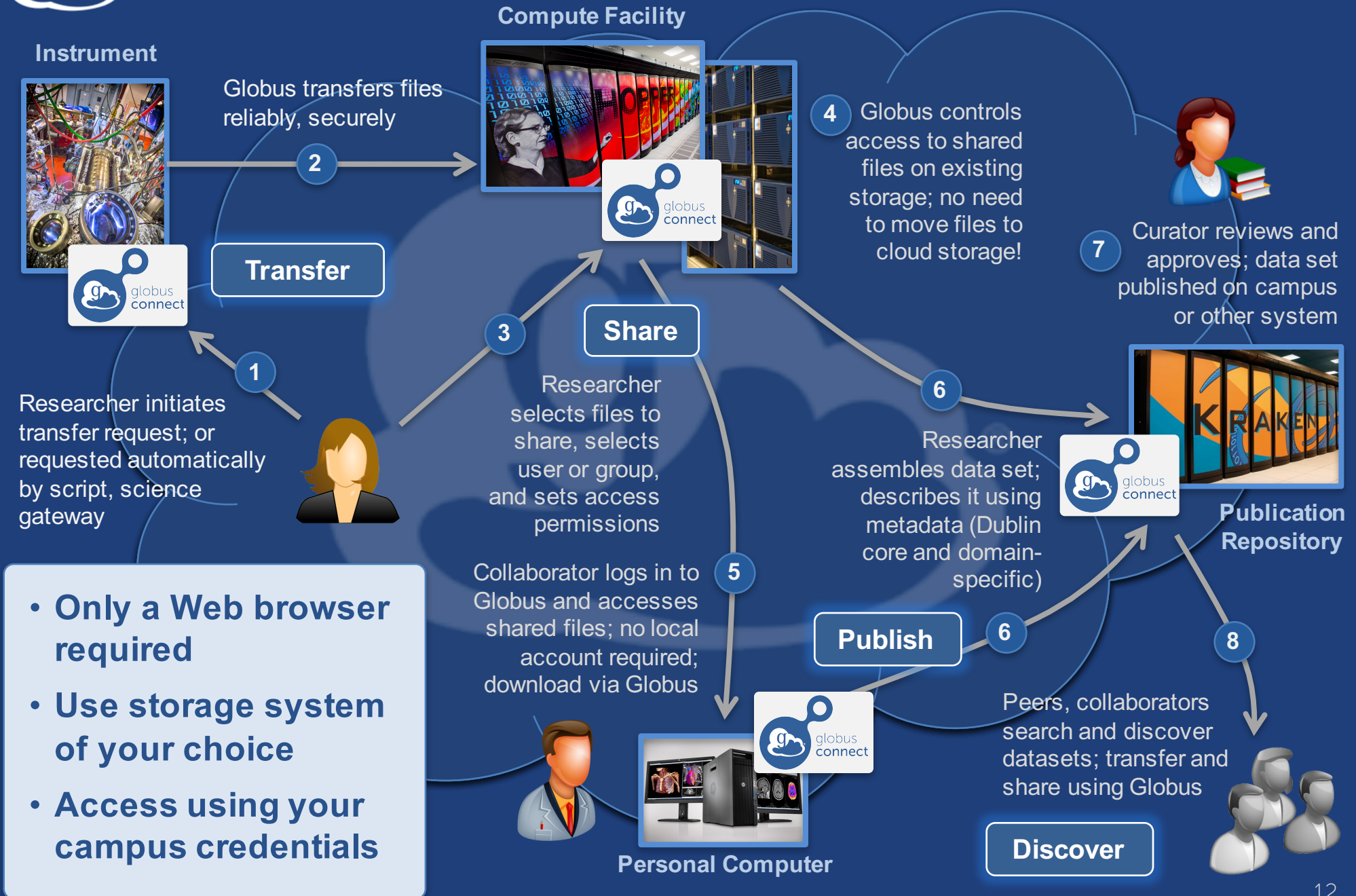
Research
Community
Collaboration



Index?



Globus and the research data lifecycle





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



Focus: User Experience
+
High performance
(but invisible) network



Accessing Globus and Moving Data



Exercise: Log in & transfer files

1. Go to: www.globus.org/login
2. Select your institution from the list and click “Continue”
3. Authenticate with your institution’s identity system
4. Install Globus Connect Personal
5. Move file(s) from esnet#???-diskpt1 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

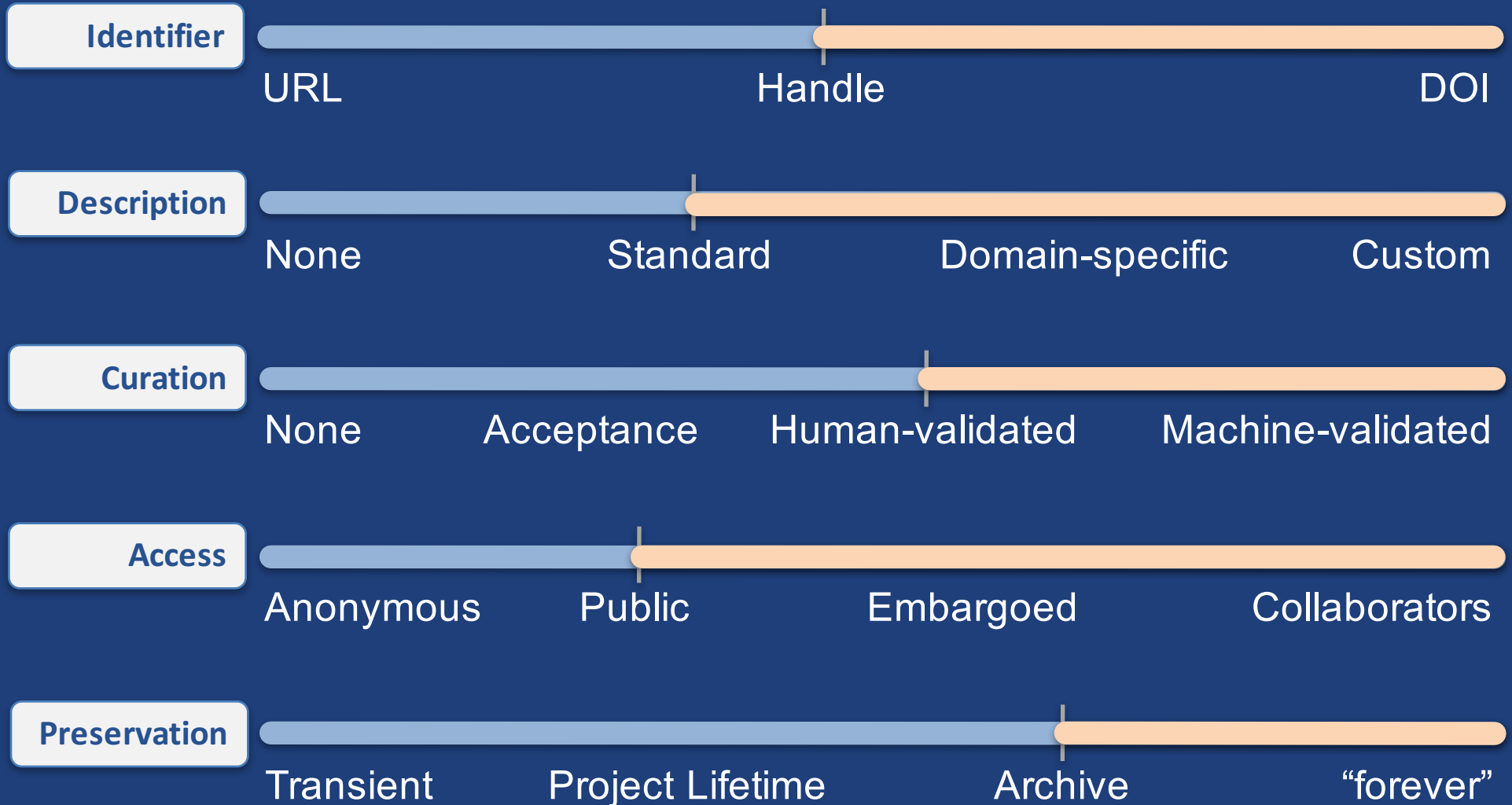
5. Confirm your neighbor can access your shared endpoint



Enhancing reproducibility and discoverability



Globus data publication framework



Publish

The screenshot shows the 'Submit: Describe this Dataset' form in the Globus interface. It includes fields for Title, Authors, Publication Year, and Publisher. The form is titled 'Submit: Describe this Dataset' and has a 'Publish' button at the top right. The form is divided into sections for 'Title', 'Authors', 'Publication Year', and 'Publisher'. The 'Title' field contains 'Linux GCD application for perturbations of central pathways'. The 'Authors' field contains 'Python' and 'Harty'. The 'Publication Year' field contains '2000'. The 'Publisher' field contains 'Journal of Comic Materials'. There are 'Previous', 'Cancel/Save', and 'Next' buttons at the bottom.

Discover

The screenshot shows the 'Discover' interface with search results for 'cath'. The results are displayed in a table with columns for 'Collection results (2 of 2)', 'Title', 'Author', and 'Subject'. The table lists several datasets, including 'Sample Cath Data', 'Thermal Conductivity', and 'Thermal Conductivity Anomalous SOC'. The 'Author' column lists various institutions and researchers, such as 'NCS, NCS, Materials Science and Engineering, James, John' and 'Cornell University, Cornell University, Materials Science and Engineering, Cath, D.G.'. The 'Subject' column lists 'Thermal Conductivity' and 'Thermal Conductivity Anomalous SOC'.

Globus Authentication

Globus Data Publication

Medical Imaging Collection



Configuration and Policies

Submission
Workflows

Metadata
and Forms

Storage

Identifiers

Materials Collection



Configuration and Policies

Submission
Workflows

Metadata
and Forms

Storage

Identifiers

Metadata Catalog

UChicago

Argonne

NCSA

Amazon S3

Globus
Identity

Users
and Groups

Handle

DOI

Persistent
Identifiers



Raw NGS output



Minimal metadata...

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

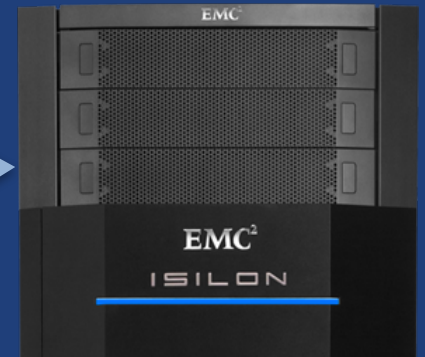
No curation

- Automated dataset acceptance

Identify...

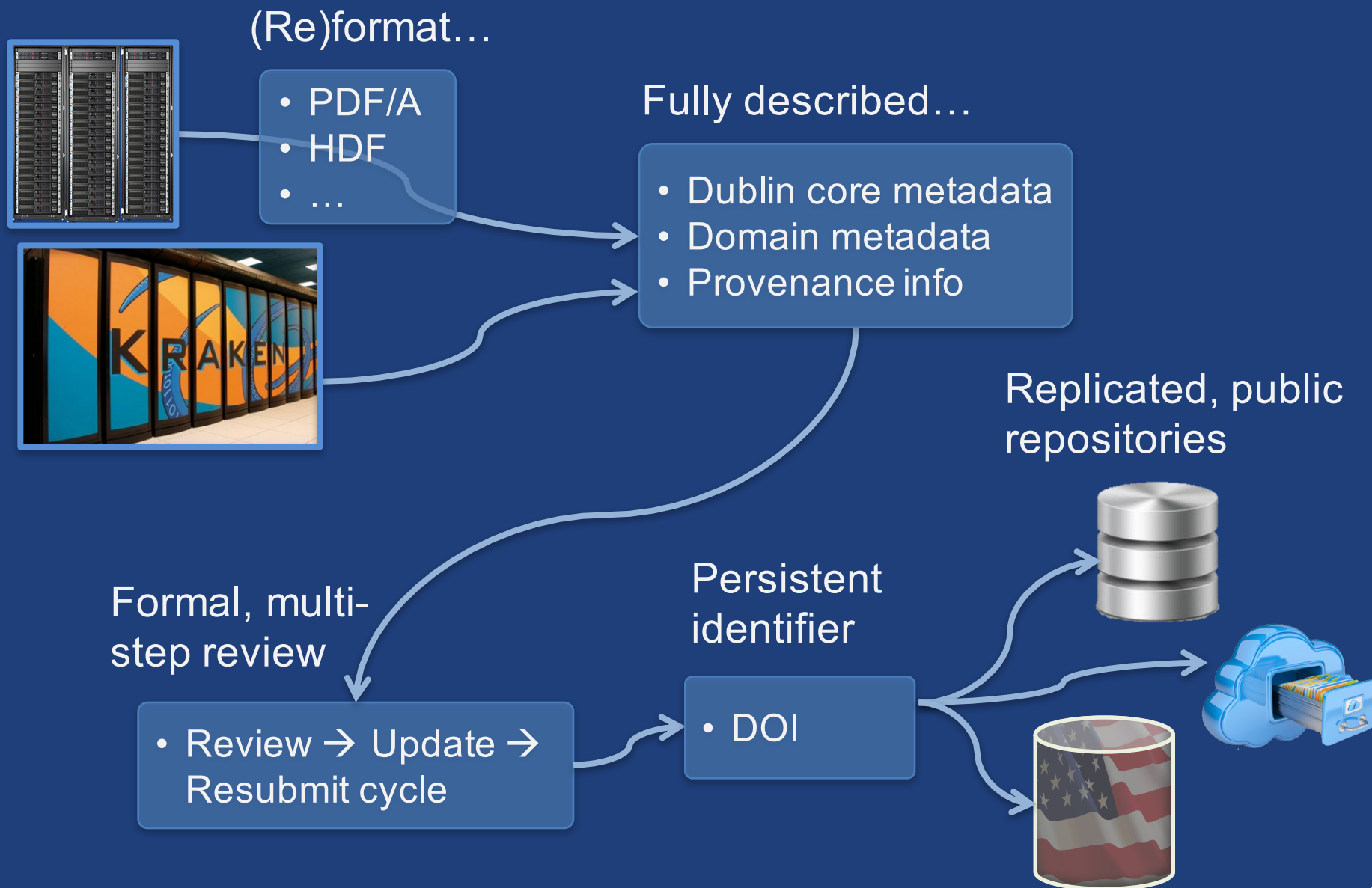
- Handle

Moderate durability,
Moderate cost store





Peer reviewed paper





Globus publication - Initial release



Supported in GA release



Consulting support



Planned

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”



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#sc15` 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



Globus: today and tomorrow



Globus today...

4

major services

155 PB

transferred

26 billion

files processed

39,000

registered users

13

national labs
use Globus

20,000

active endpoints

~450

active daily users

99.9%

uptime

60+

institutional
subscribers

1 PB

largest single
transfer to date

3 months

longest
continuously
managed transfer

130

federated
campus identities



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 Subscriptions

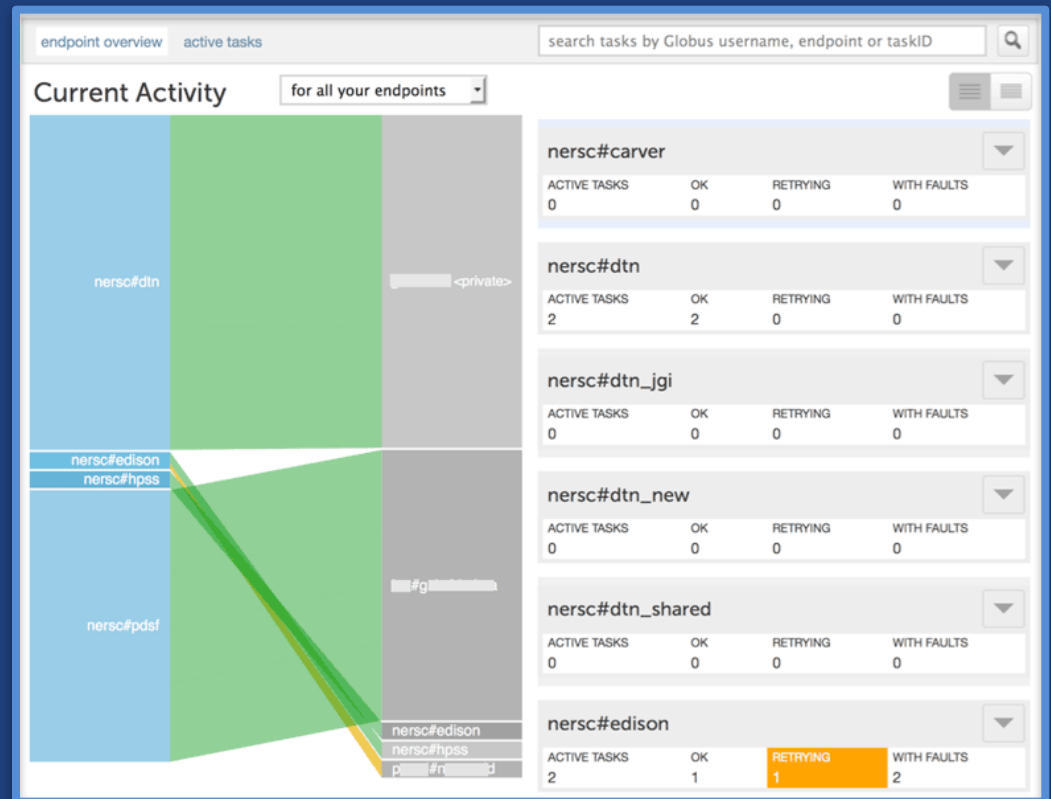
- **Globus Provider Plan**

- Shared endpoints
- Data publication
- Amazon S3 endpoints
- Management console
- Usage reporting
- Priority support
- Application integration

- **Branded Web Site**

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

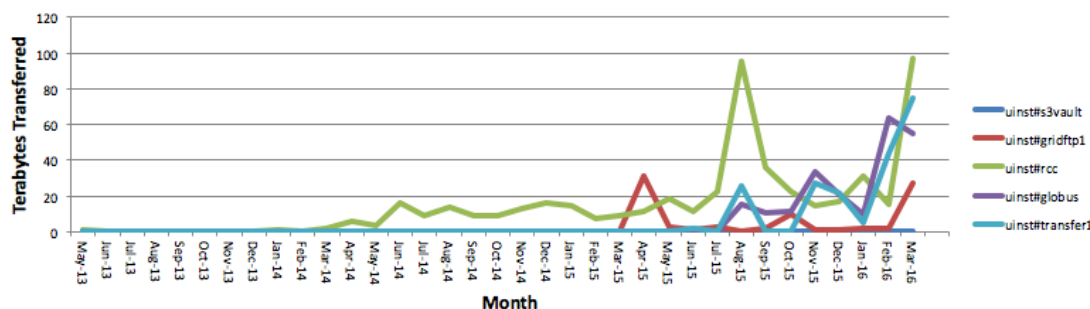
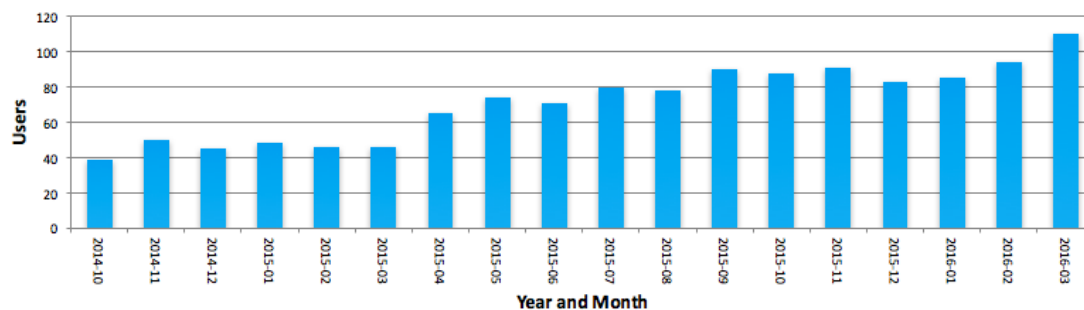
- **Mass Storage System optimization**



globus.org/provider-plans



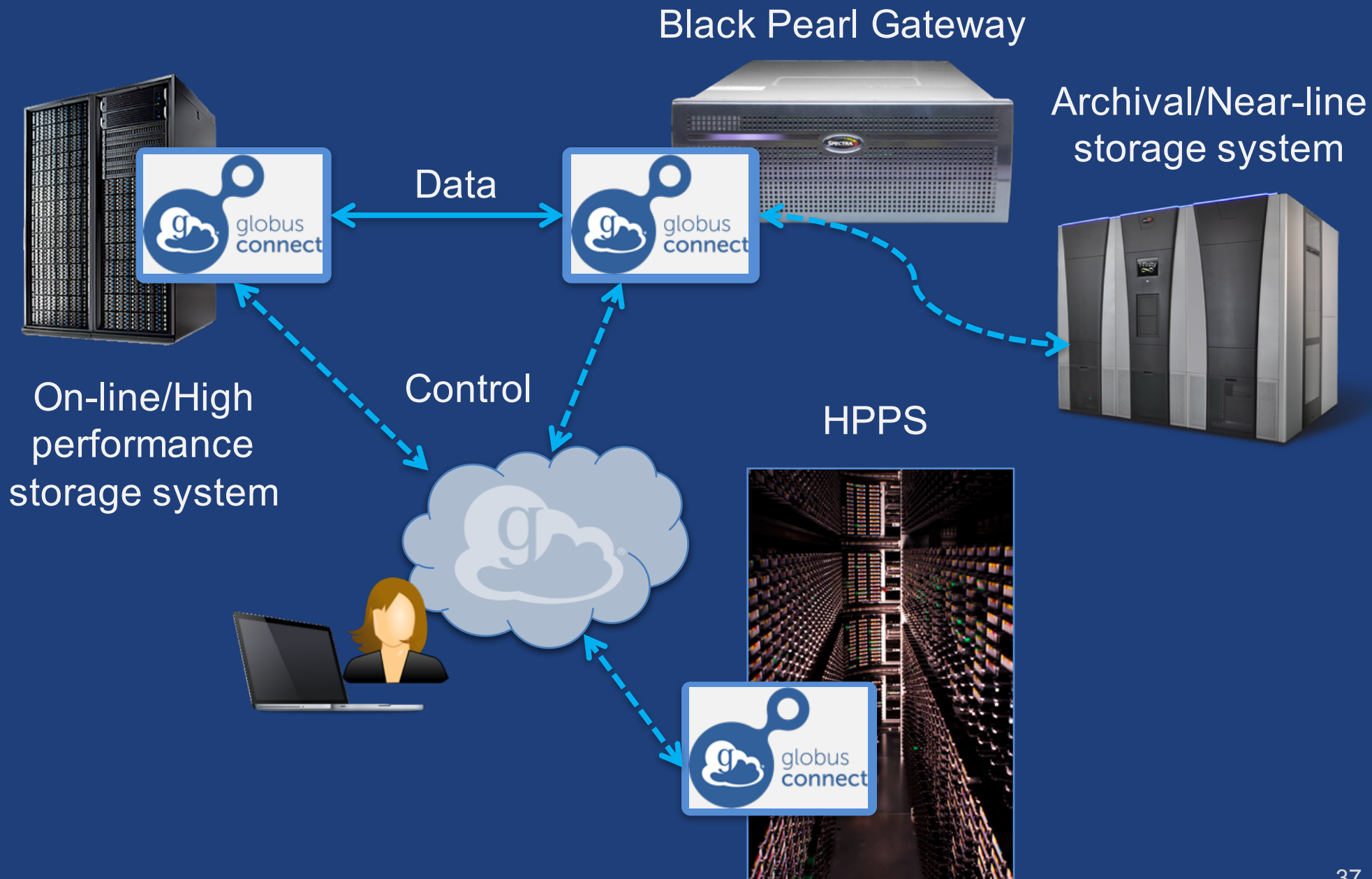
Demonstration: Globus management console



- **Monthly summaries**
- **Complete history of all endpoint activity**
- **Available on endpoint “Globus Usage Reports”**



Bridging the storage hierarchy





Demonstration:

Bridging to Cloud Storage

- Amazon S3: supported
- Ceph: coming soon



Enable your storage system

- Access Globus: **globus.org/login**
- Create endpoint: **globus.org/globus-connect-server**
- Learn more: **docs.globus.org**
- Need help? **support.globus.org**
- Subscribe to help us make Globus self-sustaining:
globus.org/provider-plans
- Follow us: **[@globusonline](https://twitter.com/globusonline)**