Globus Research Data Management: Introduction and Service Overview

Steve Tuecke
tuecke@uchicago.edu

Vas Vasiliadis
vas@uchicago.edu

SC15
Austin, TX | hpc transforms.

globus

Presentation material available at

globus.org/events/sc15

bit.ly/globus-sc15

Thank you to our sponsors!

U.S. DEPARTMENT OF ENERGY

NSF

ALFRED P. SLOAN FOUNDATION 1934

NATIONAL INSTITUTES OF HEALTH

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 Cluste

Campus Home Filesyster

All files and folders - Box

https://app.box.com/files

All Files

Security Notice: Heartbleed Update

A major security flaw was recently discovered with OpenSSL, an encryption standard that Box uses. We've patched the issue and your account is safe, but we are advising you reset your password now. Instructions here. More details are available on our help site and blog.

Globus Online

Updated Jan 21, 2014 by Andrew Keating 24 21

Account

Personal Account UPGRADE

Storage: 0B of 10.0GB

Business Solutions

Resources

Trash

Help

Success

Box Com

Lab Server

Personal Laptop

Desktop Workstation

XSEDE Resource

Public Cloud

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

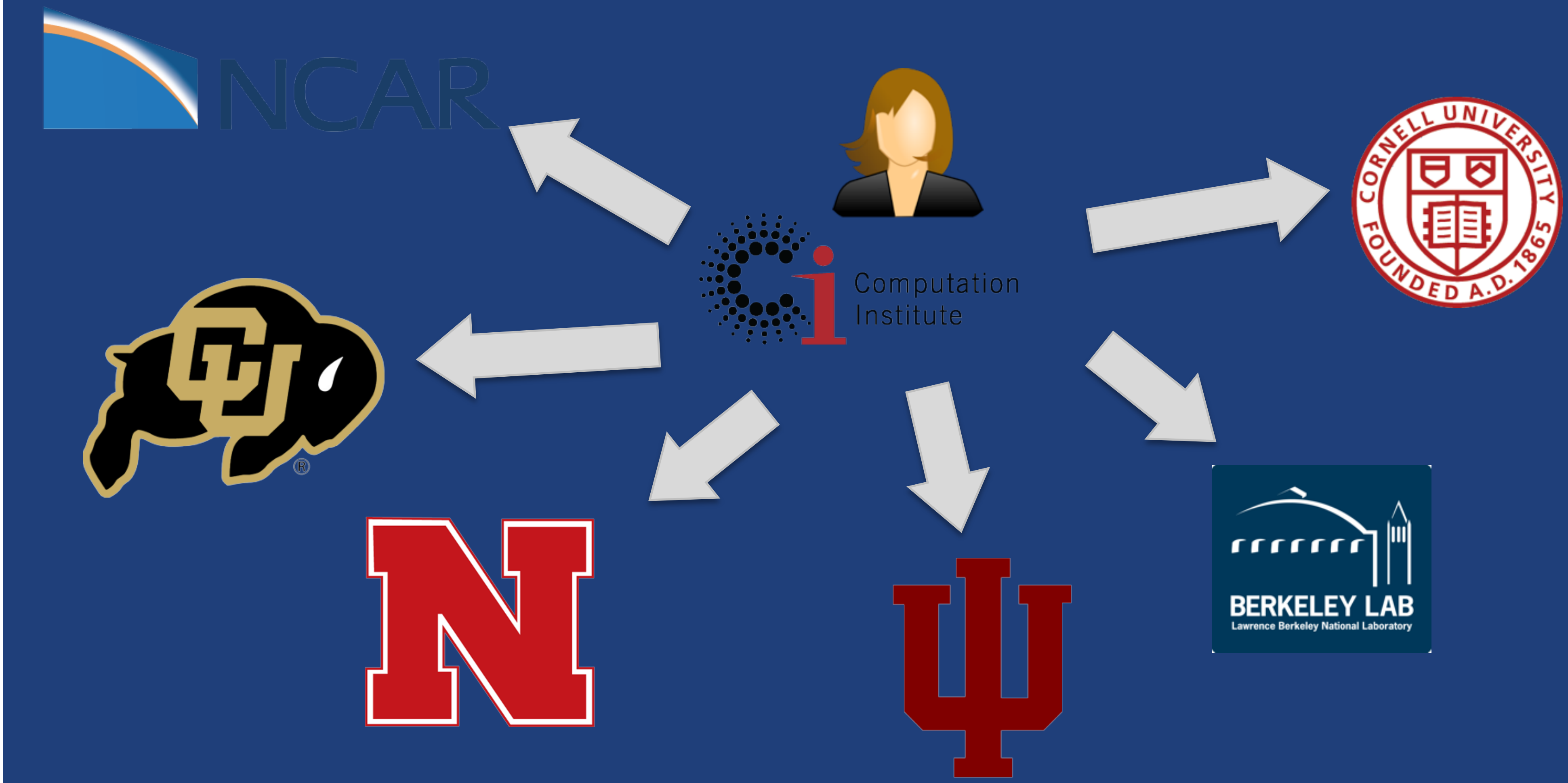
MRI

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
CORNELL UNIVERSITY FOUNDED A.D. 1865
BERKELEY LAB
Lawrence Berkeley National Laboratory

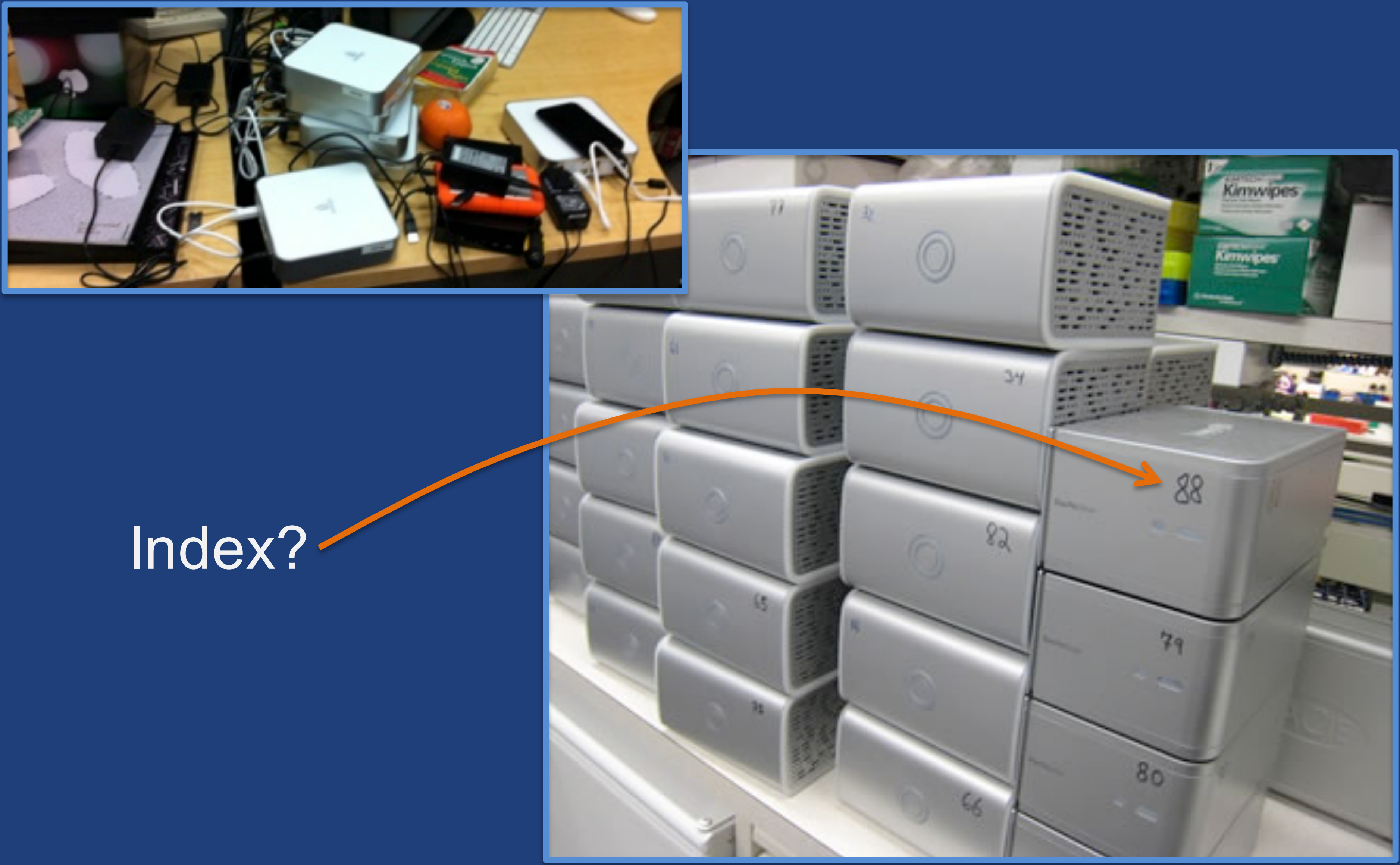
“I need to publish my data so others can find/use/validate/reproduce it.”

Reference Dataset

Scholarly Publication

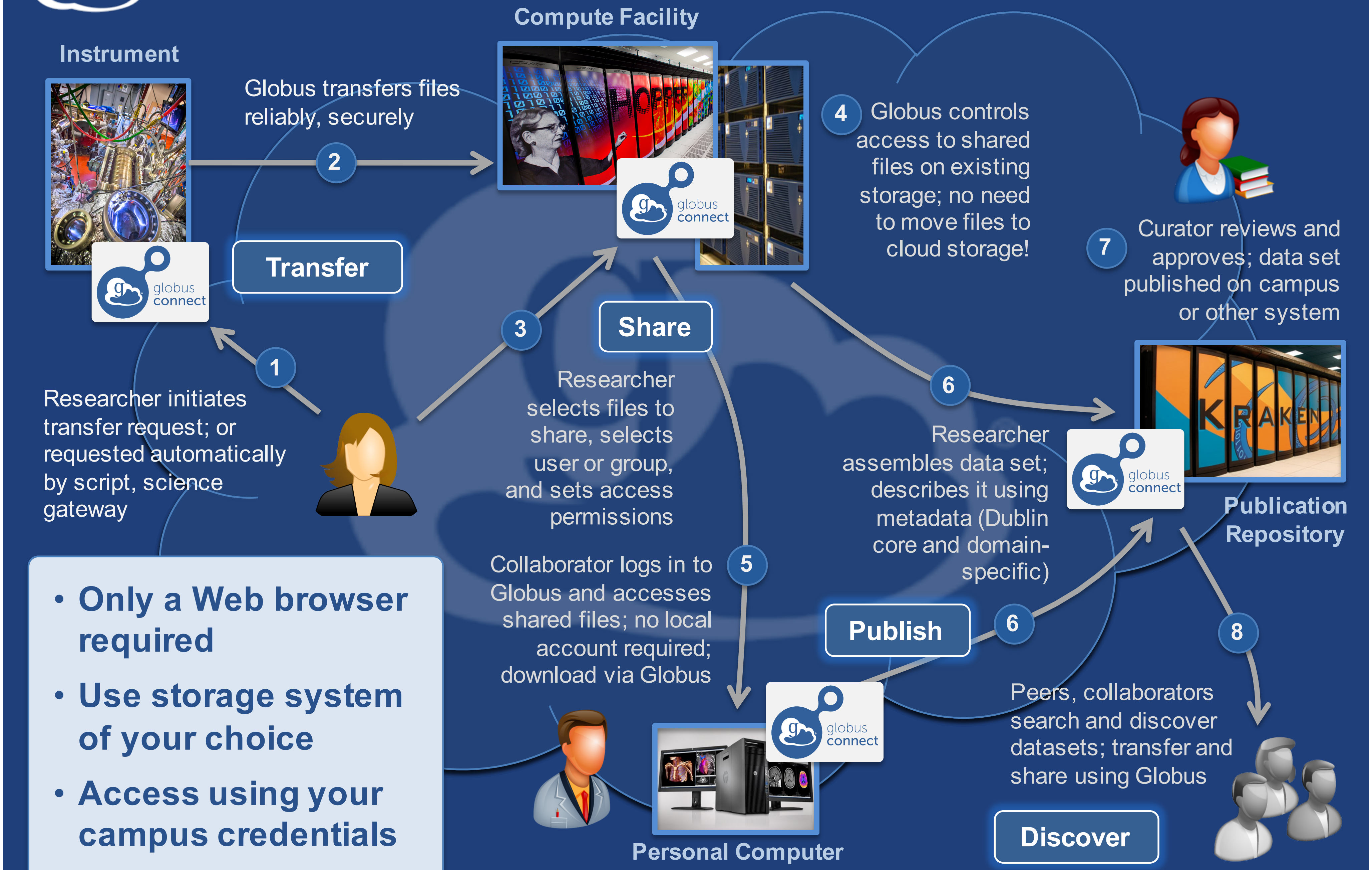
Research Community Collaboration

Research data management today



Index?

Globus and the research data lifecycle



Instrument

Compute Facility

Globus transfers files reliably, securely

Transfer

2

1 Researcher initiates transfer request; or requested automatically by script, science gateway

3 Share

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

Personal Computer

6 Publish

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

Publication Repository

8 Peers, collaborators search and discover datasets; transfer and share using Globus

Discover

- 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

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#sc15 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**

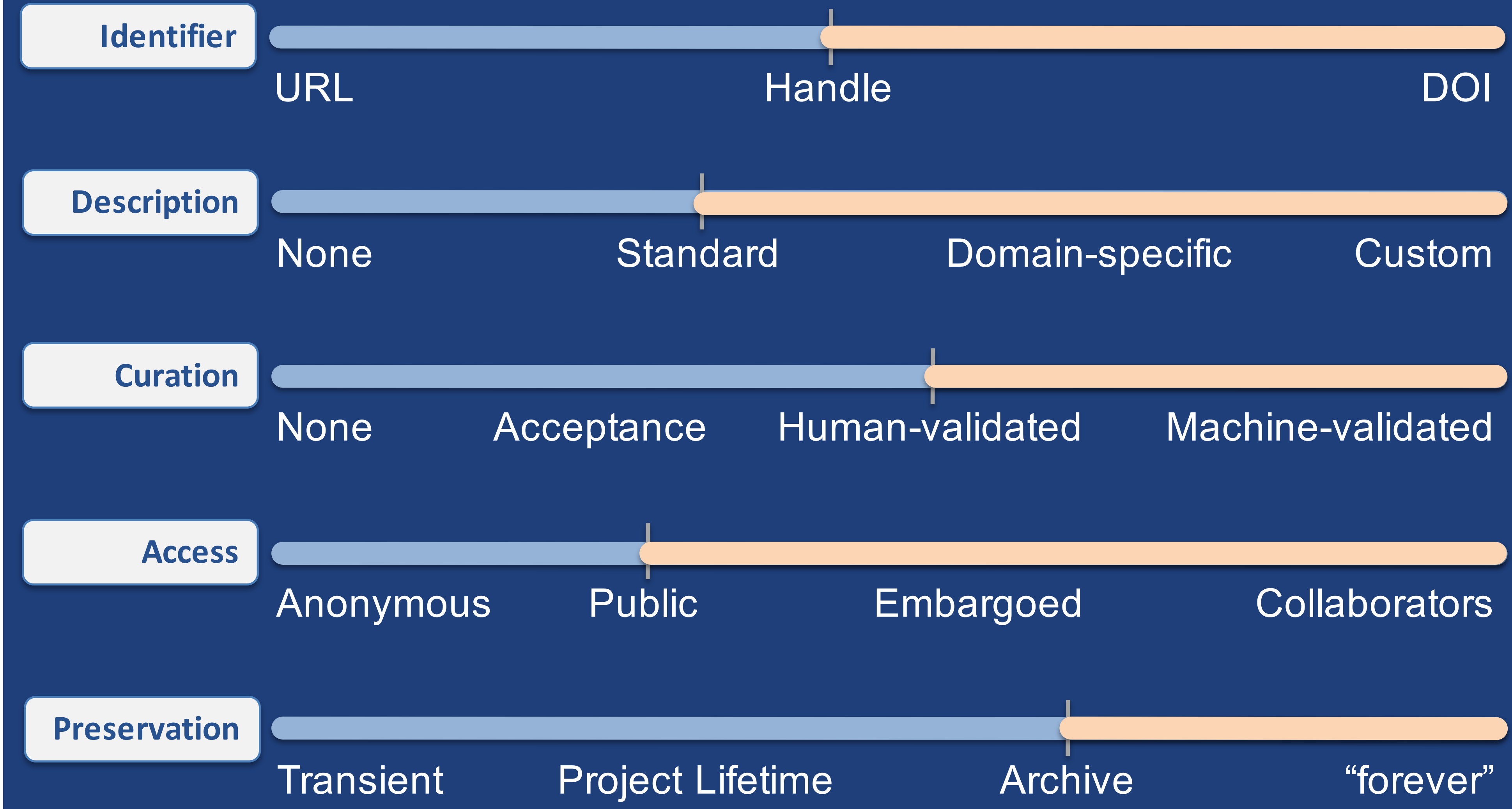
Group Management

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

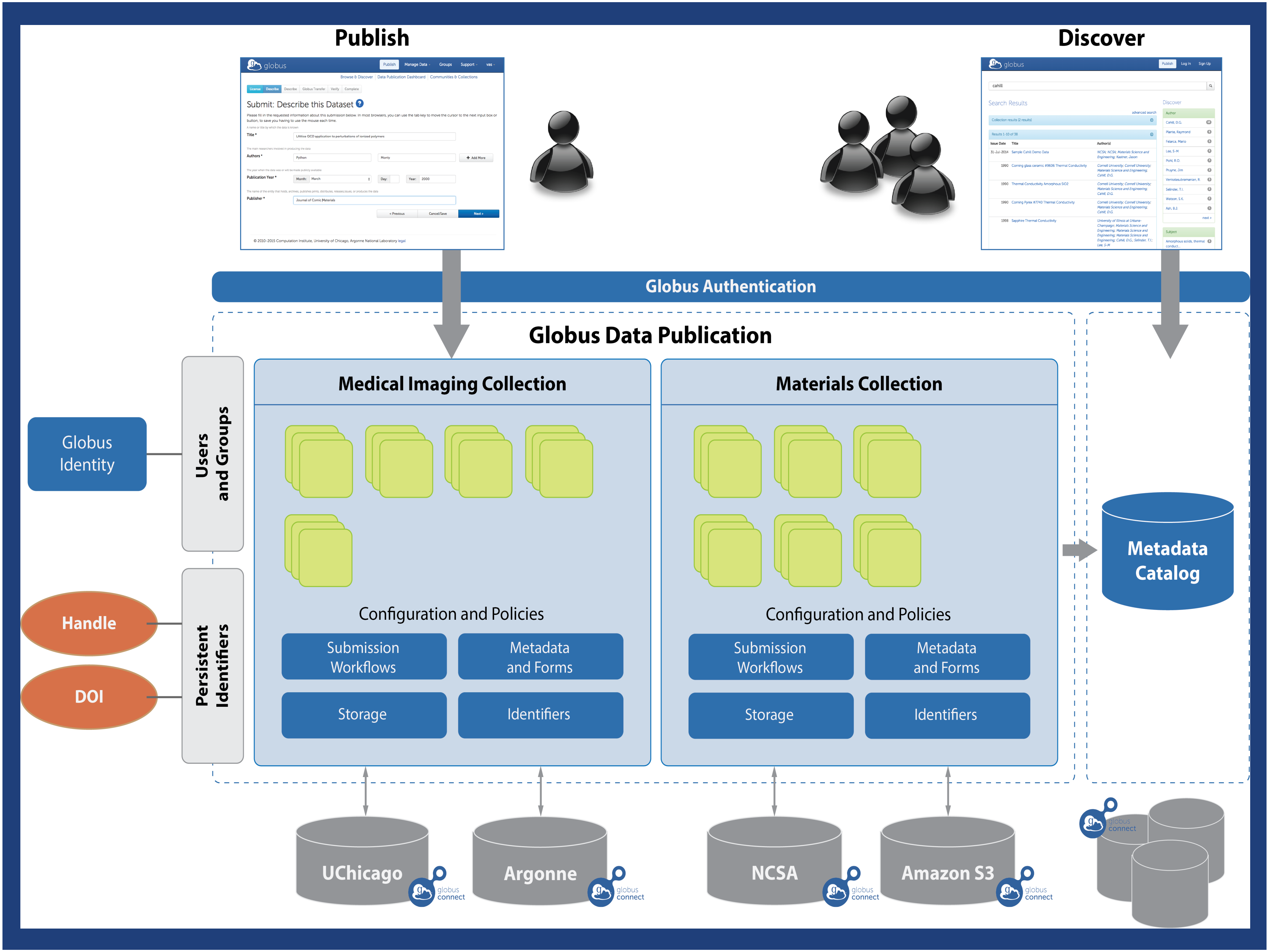
None — Acceptance — Human-validated — Machine-validated

Access

Anonymous — Public — Embargoed — Collaborators

Preservation

Transient — Project Lifetime — Archive — “forever”



globus — Publish | Manage Data | Groups | Support | vas

Browse & Discover | Data Publication Dashboard | Communities & Collections

License | Describe | Describe | Globus Transfer | Verify | Complete

Submit: Describe this Dataset ?

Please fill in the requested information about this submission below. In most browsers, you can use the tab key to move the cursor to the next input box or button, to save you having to use the mouse each time.

A name or title by which the data is known

Title * LAttice QCD application to perturbations of ionized polymers

The main researchers involved in producing the data

Authors * Python | Monty | + Add More

The year when the data was or will be made publicly available

Publication Year * Month: March | Day: | Year: 2000

The name of the entity that holds, archives, publishes prints, distributes, releases,issues, or produces the data

Publisher * Journal of Comic Materials

< Previous | Cancel/Save | Next >

© 2010-2015 Computation Institute, University of Chicago, Argonne National Laboratory legal

globus — Publish | Log In | Sign Up

cahill

Search Results

advanced search

Collection results (2 results)

Results 1-10 of 38

Issue Date	Title	Author(s)
31-Jul-2014	Sample Cahill Demo Data	NCSA; NCSA; Materials Science and Engineering; Kastner, Jason
1990	Corning glass ceramic #9606 Thermal Conductivity	Cornell University; Cornell University; Materials Science and Engineering; Cahill, D.G.
1990	Thermal Conductivity Amorphous SiO2	Cornell University; Cornell University; Materials Science and Engineering; Cahill, D.G.
1990	Corning Pyrex #7740 Thermal Conductivity	Cornell University; Cornell University; Materials Science and Engineering; Cahill, D.G.
1998	Sapphire Thermal Conductivity	University of Illinois at Urbana-Champaign; Materials Science and Engineering; Materials Science and Engineering; Materials Science and Engineering; Cahill, D.G.; Selinder, T.I.; Lee, S-M

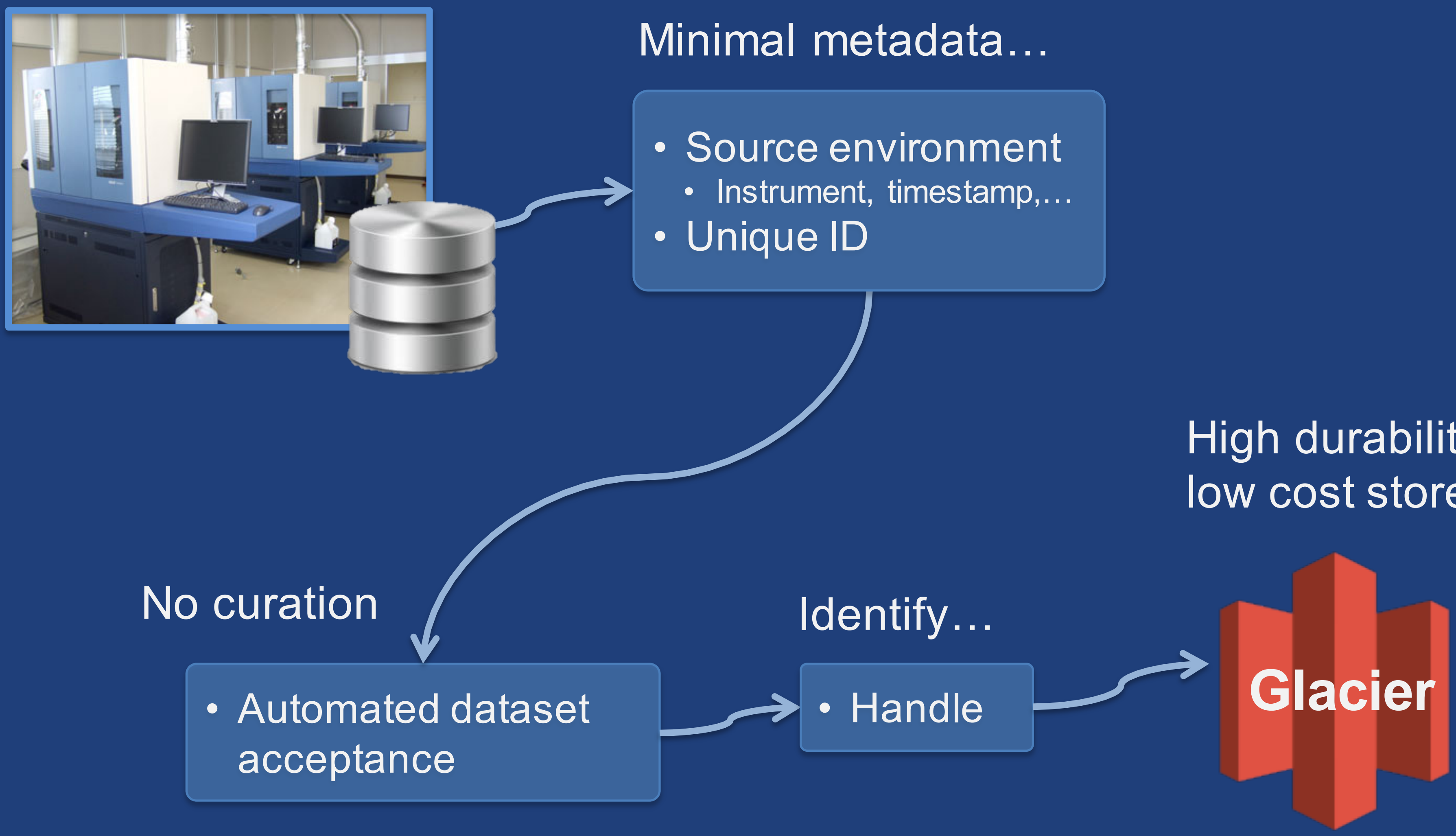
Discover

Author	
Cahill, D.G.	12
Plante, Raymond	6
Felarca, Mario	5
Lee, S-M	5
Pohl, R.O.	3
Pruyne, Jim	3
Venkatasubramanian, R.	3
Selinder, T.I.	2
Watson, S.K.	2
Ash, B.J.	1

next >

Subject	
Amorphous solids, thermal conduct...	3

Raw NGS output



Minimal metadata…

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

No curation

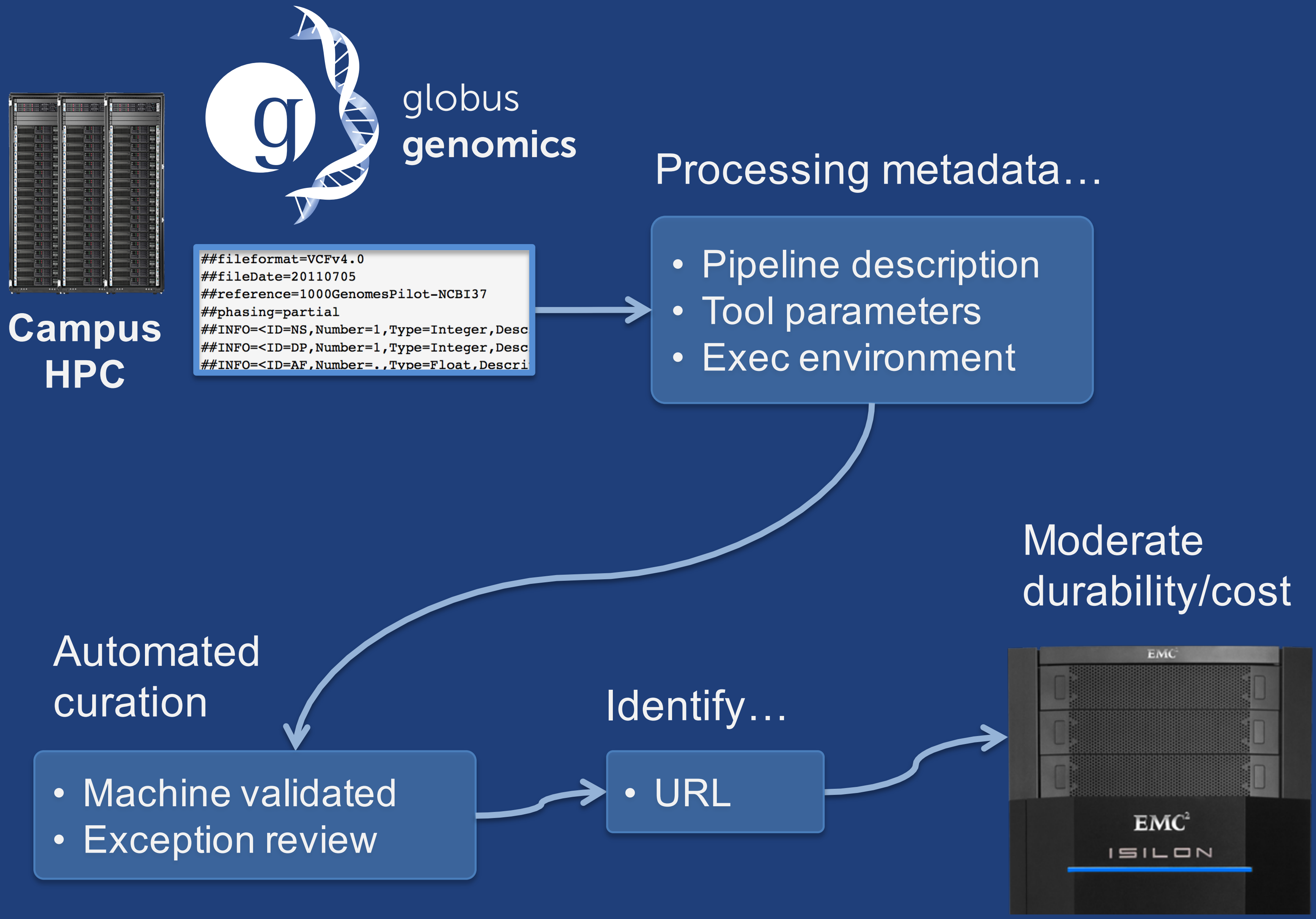
- Automated dataset acceptance

Identify…

- Handle

High durability, low cost store

Upstream analysis



Campus HPC

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Desc
##INFO=<ID=DP,Number=1,Type=Integer,Desc
##INFO=<ID=AF,Number=.,Type=Float,Descri
```

Processing metadata…

- Pipeline description
- Tool parameters
- Exec environment

Automated curation

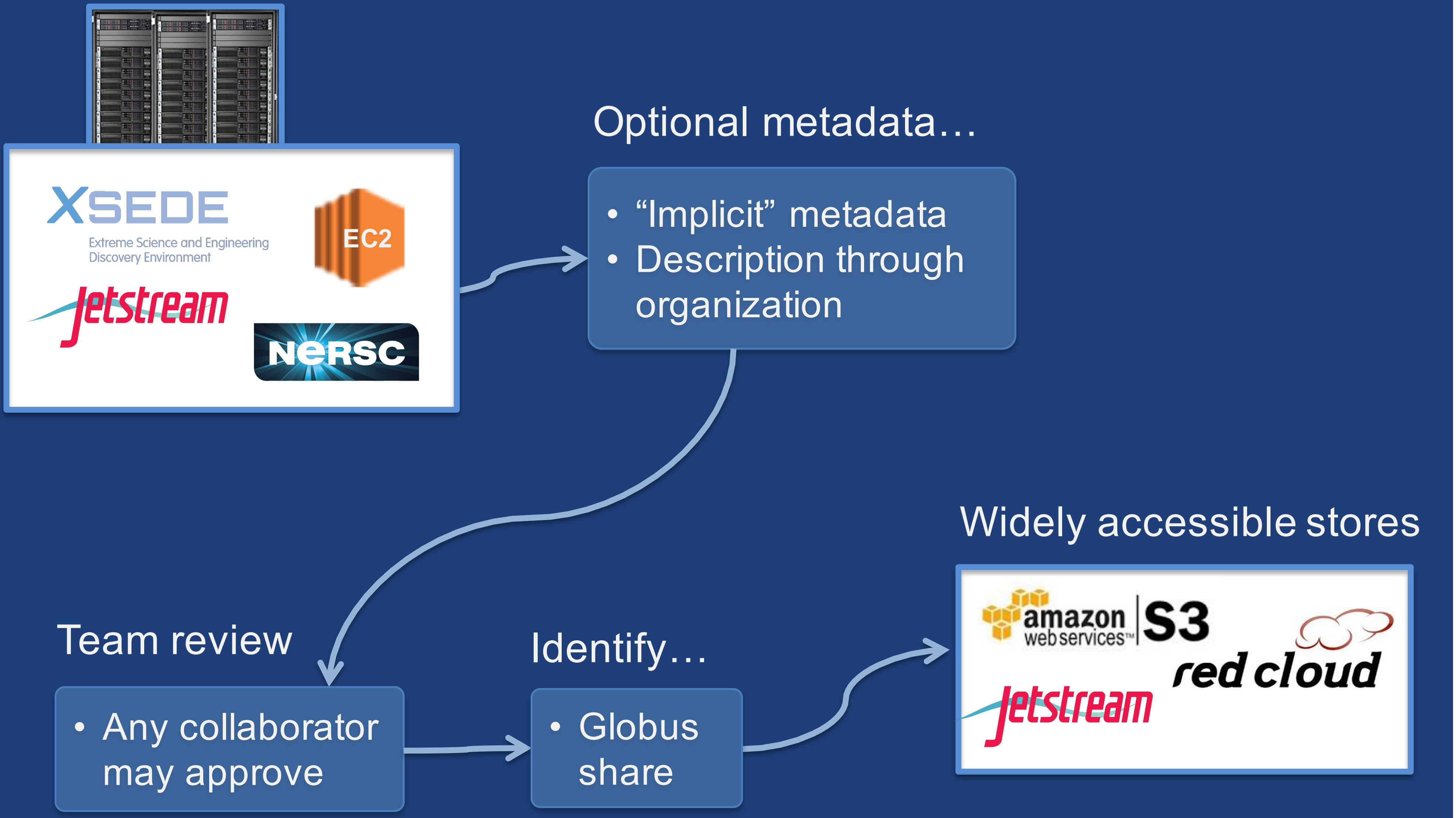
- Machine validated
- Exception review

Identify…

- URL

Moderate durability/cost

Downstream analysis



Optional metadata…

- “Implicit” metadata
- Description through organization

Team review

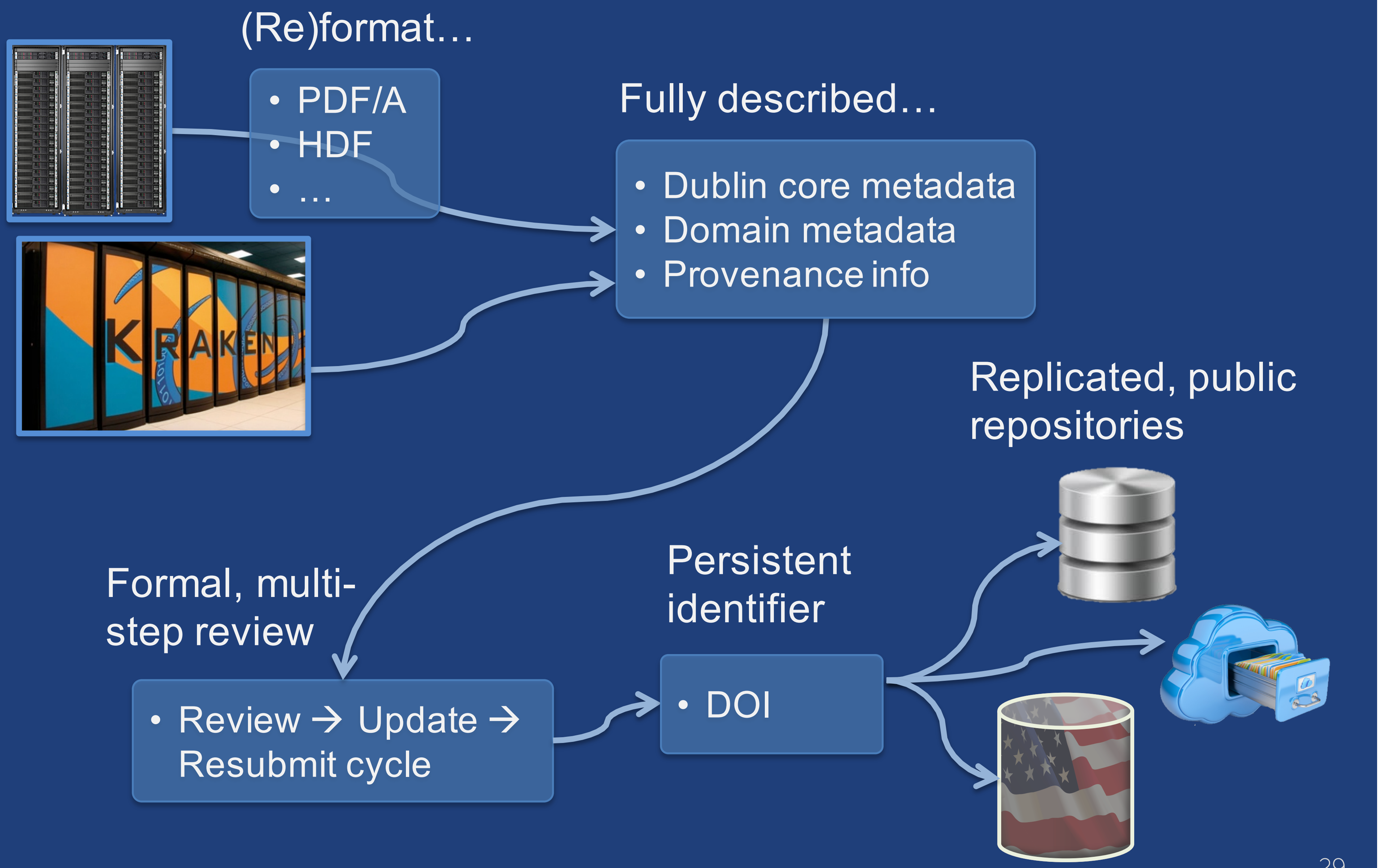
- Any collaborator may approve

Identify…

- Globus share

Widely accessible stores

Peer reviewed paper



(Re)format…

- PDF/A
- HDF
- …

Fully described…

- Dublin core metadata
- Domain metadata
- Provenance info

Formal, multi-step review

- Review → Update → Resubmit cycle

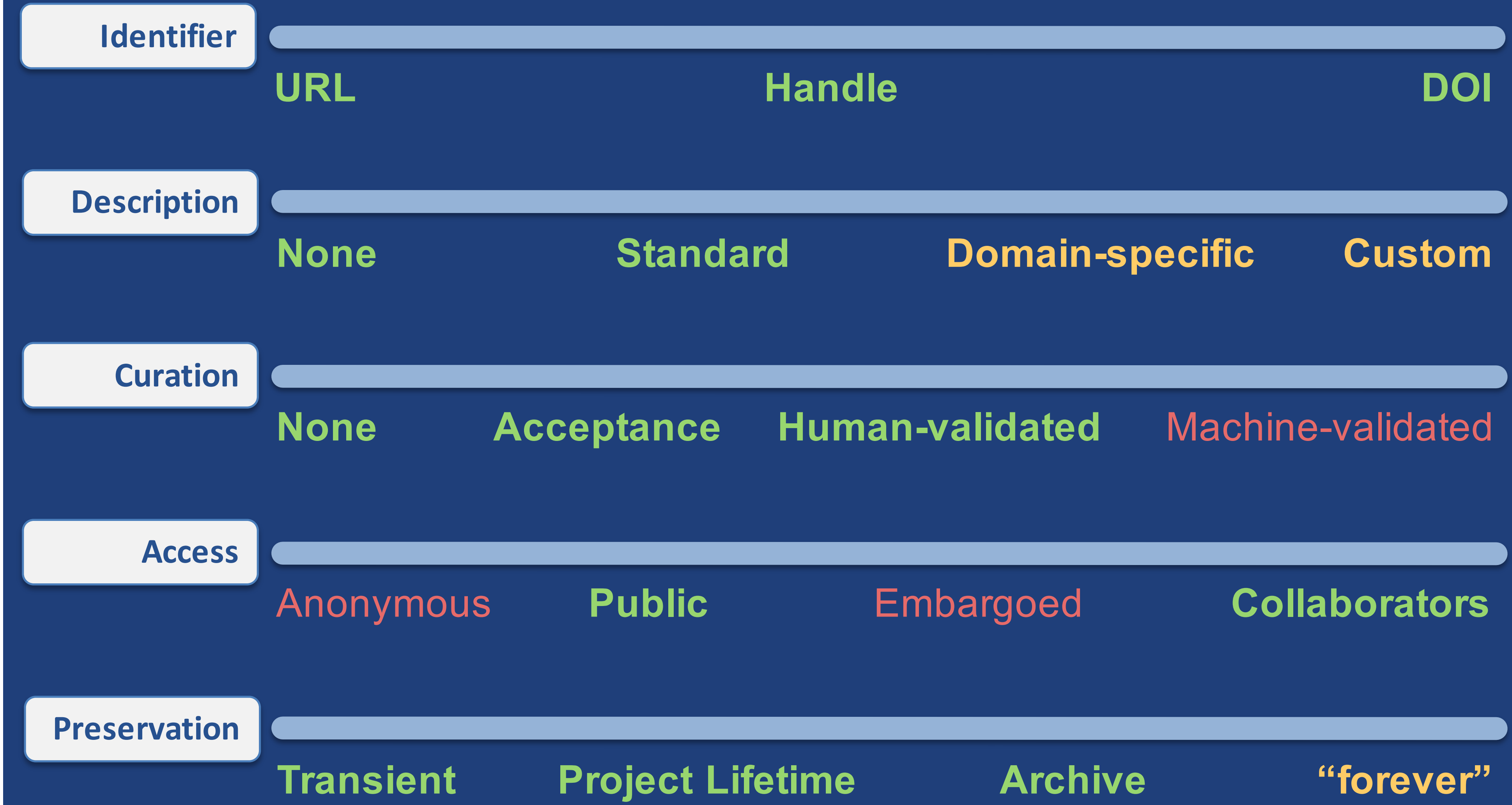
Persistent identifier

- DOI

Replicated, public repositories

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	**118 PB** transferred	**20 billion** files processed	**31,000** registered users
13 national labs use Globus	**10,000** active endpoints	**~350** active daily users	**99.95%** uptime
35+ 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

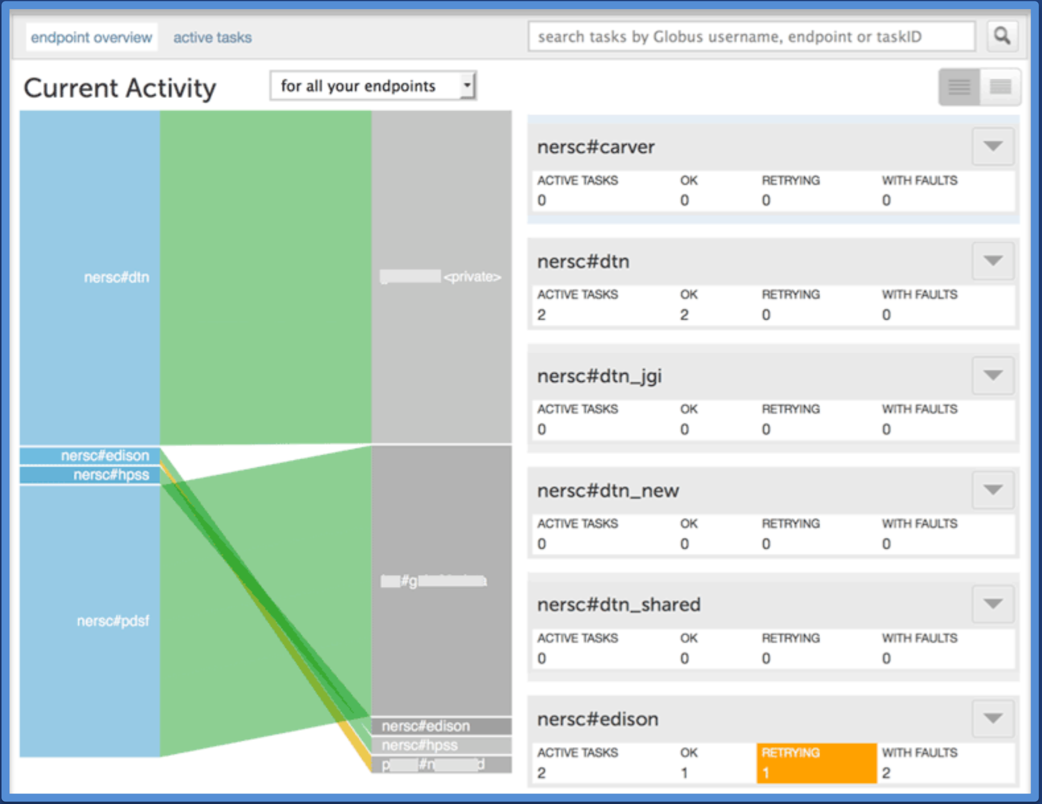
We are a non-profit, delivering a production-grade service to the non-profit research community

Our challenge:

Sustainability

Globus Provider Subscriptions

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



Current Activity

nersc#carver

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
0	0	0	0

nersc#dtn

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
2	2	0	0

nersc#dtn_jgi

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
0	0	0	0

nersc#dtn_new

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
0	0	0	0

nersc#dtn_shared

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
0	0	0	0

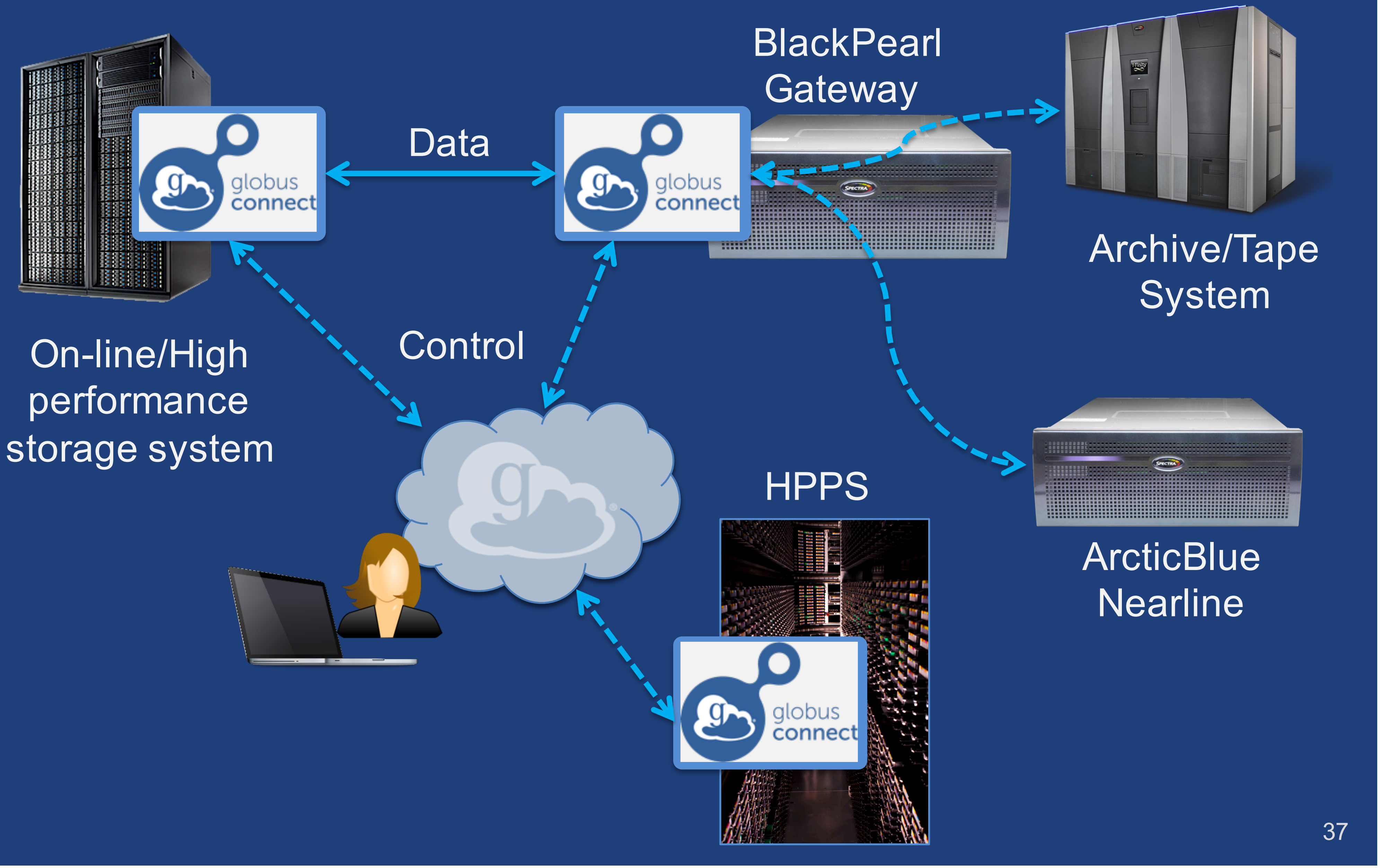
nersc#edison

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
2	1	1	2

- **Alternate Identity Provider (InCommon is standard)**
- **Mass Storage System optimization**

globus.org/provider-plans

Bridging the storage hierarchy



Demonstration:
Globus management console

Demonstration:

Bridging to Cloud Storage

- Amazon S3: supported

- Ceph: coming soon