# Globus Research Data Management: Introduction and Service Overview

XSEDE

Extreme Science and Engineering
Discovery Environment

Steve Tuecke
Vas Vasiliadis

globus

Presentations and other useful information available at

**globus.org/events/xsede15/tutorial**

# 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 scenarios and challenges**
- **Introduction to Globus**
- **Demonstrations and Exercises**
  - Accessing Globus and Transferring Files
  - File sharing and Group Management
  - Data publication and discovery
- **Globus: today and tomorrow**

# Research data management scenarios and challenges

# “I need to easily, quickly, & reliably move or mirror portions of my data to other places.”

Research Computing HPC Cluste

Lab Server

All files and folders – Box
https://app.box.com/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

Campus Home Filesyste

Personal Laptop

Desktop Workstation

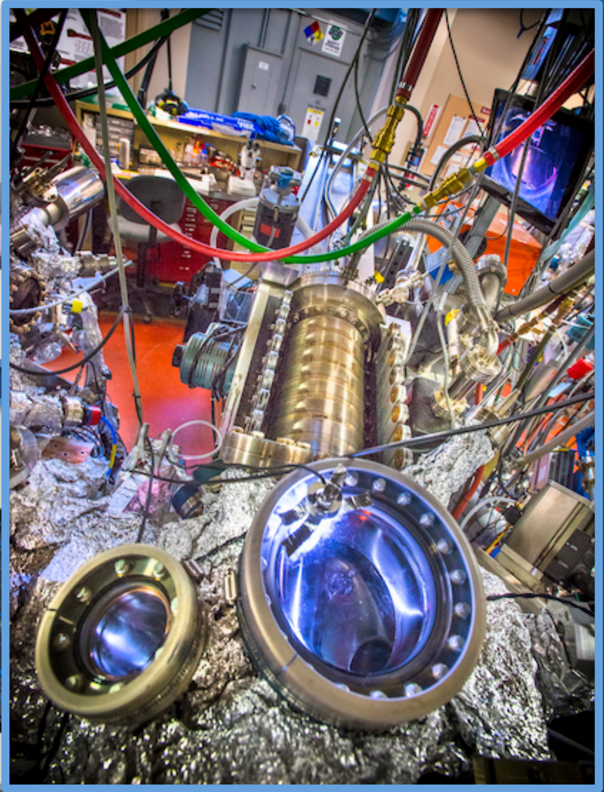
XSEDE Resource

Public Cloud

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

MRI

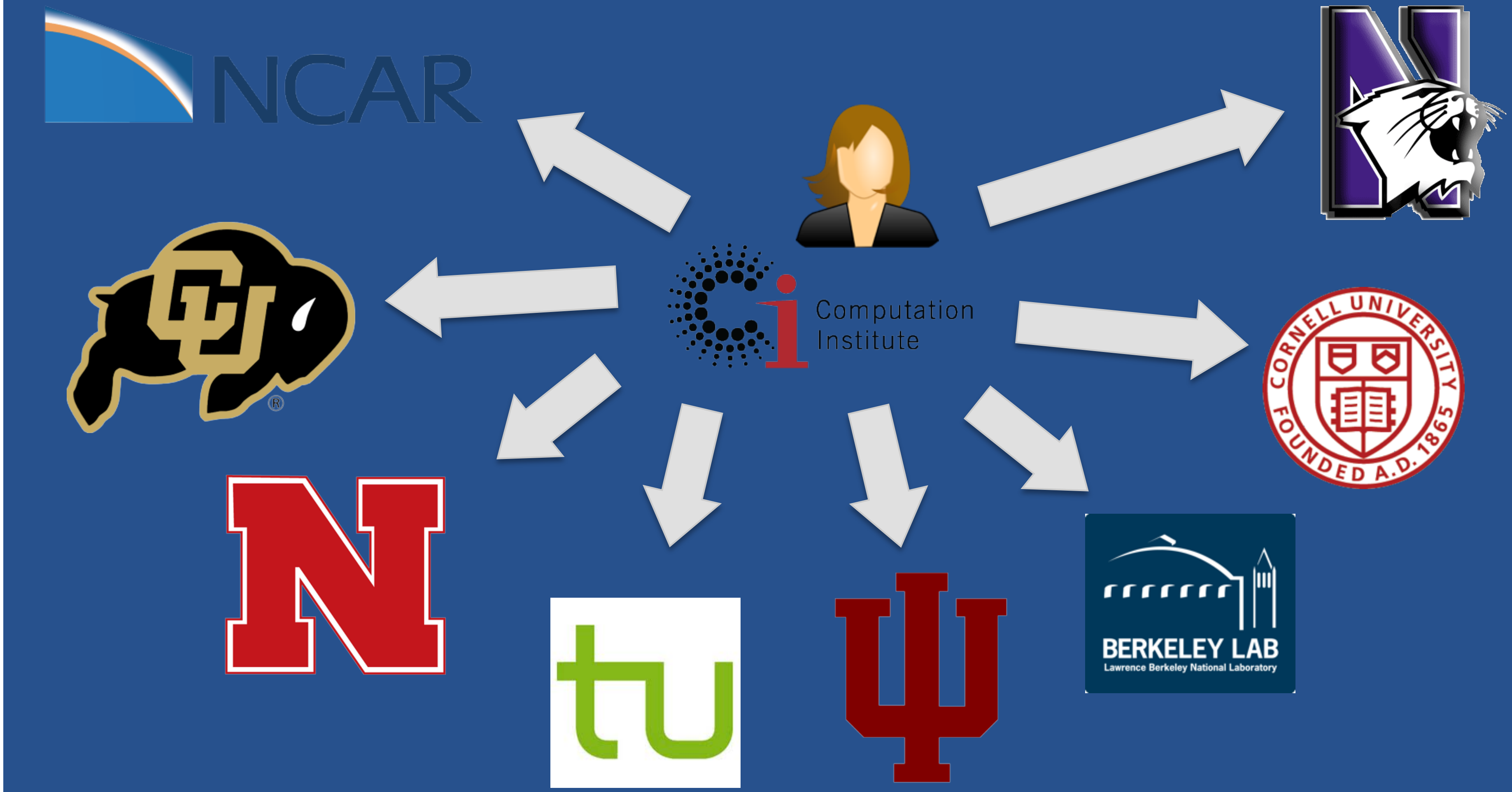
Advanced Light Source



Next Gen Sequencer

Light Sheet Microscope

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



# “I need a good place to store / backup / archive my (big) research data, at a reasonable price.”

Campus Store

Mass Store

Public Cloud Archive

# "I need to publish my data so that others can find it and use it."

Reference Dataset

Scholarly Publication

Active Research Collaboration

# Globus introduction and demonstration

Globus is…

Research data management…

…delivered via SaaS

Globus delivers…

Big data transfer, sharing, publication, and discovery…

…directly from your own storage systems

# It’s about the user experience…

flickr …for your photos

Gmail by Google …for your e-mail

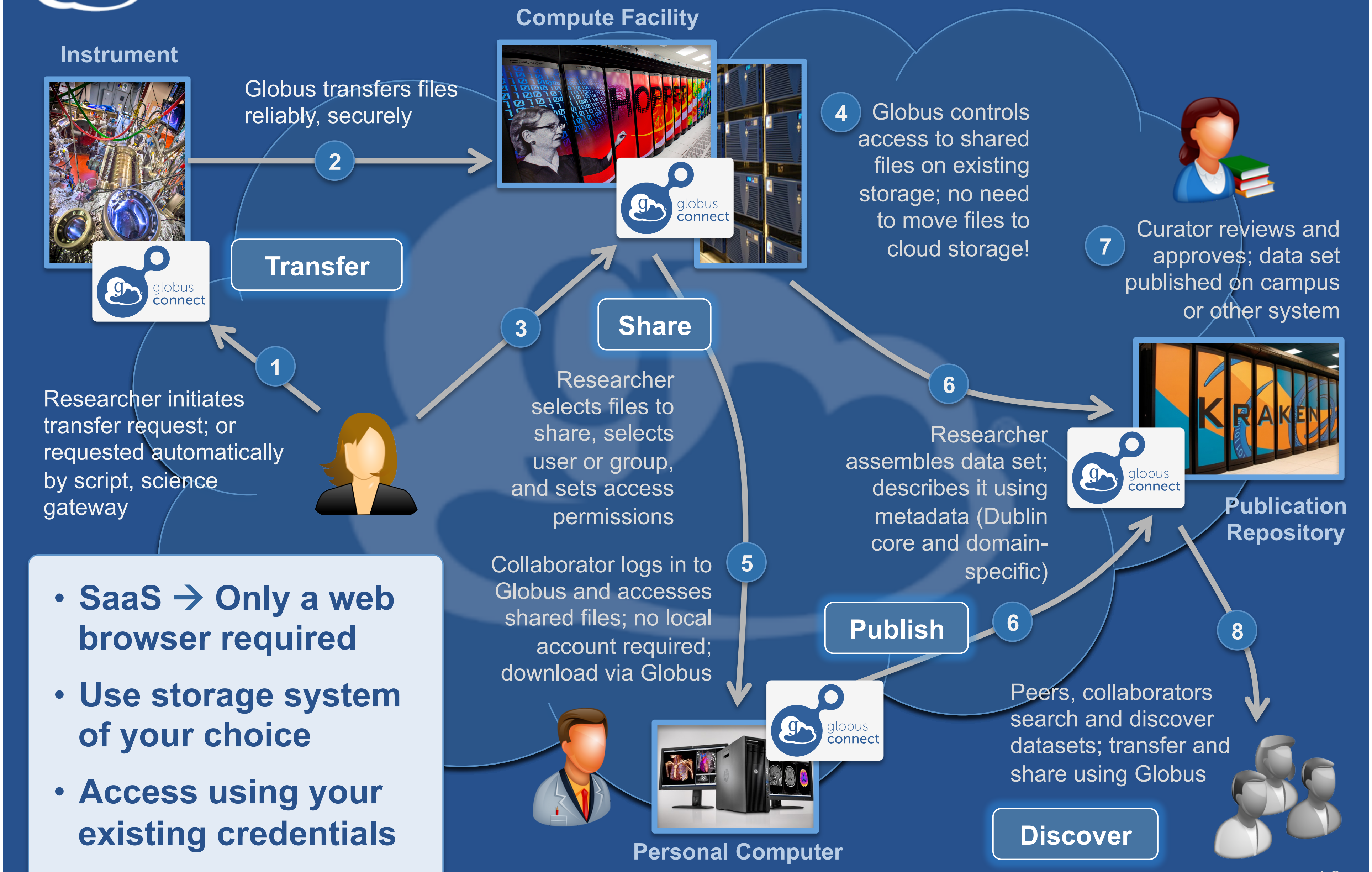
NETFLIX …for your entertainment

globus …for your research data

# Globus is SaaS

- **Web, command line, and REST interfaces**
- **Reduced IT operational costs**
- **New features automatically available**
- **Consolidated support & troubleshooting**
- **Easy to add your laptop, server, cluster, supercomputer, etc. with Globus Connect**

# 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 existing credentials

# Globus and XSEDE

- **Globus endpoints available on all XSEDE storage resources**
- **Accessed using your XSEDE user portal username/password (see demo)**
- **First service to pass XSEDE acceptance test**
  - Available for file transfer
  - File sharing in final stages of approval

# Demonstration:

# - Accessing Globus

# - File Transfer

# Exercise: 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 file(s) from esnet#lbl-diskpt1 to your laptop

# Demonstration:

# - File Sharing

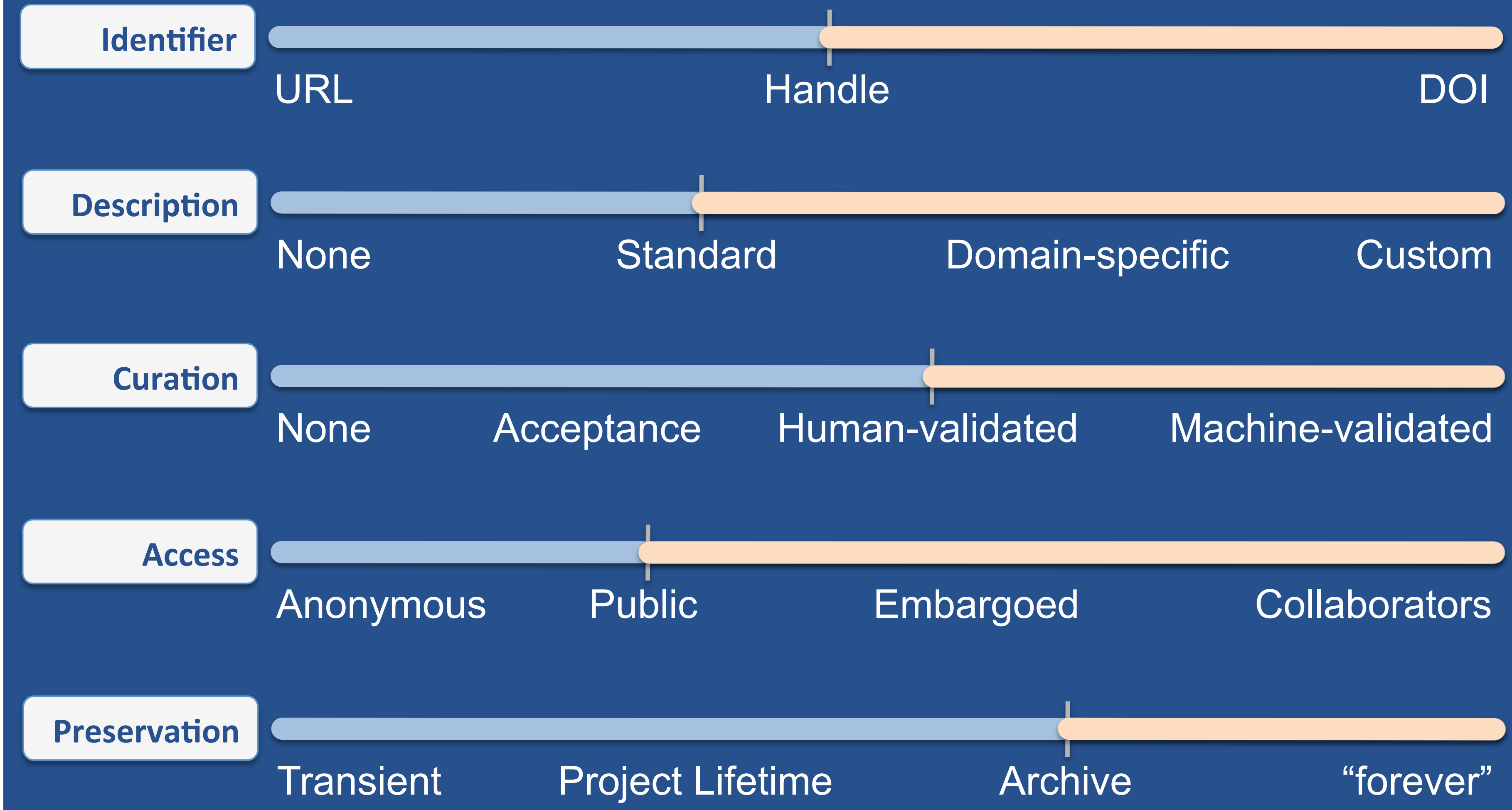
# - Group Management

# Exercise: File sharing

- **Join the “Tutorial Users” group**
  - Go to “Groups” and search for “tutorial”
- **Enable sharing in Globus Connect Personal configuration**
  - e.g. on Mac OS: Preferences → Access → Shareable
- **Create a shared endpoint on your laptop**
- **Grant your neighbor permissions on your shared endpoint**
- **Access your neighbor’s shared endpoint**

# Globus Data Publication and Discovery

# Globus data publication framework



**Identifier**

URL Handle DOI

**Description**

None Standard Domain-specific Custom

**Curation**

None Acceptance Human-validated Machine-validated

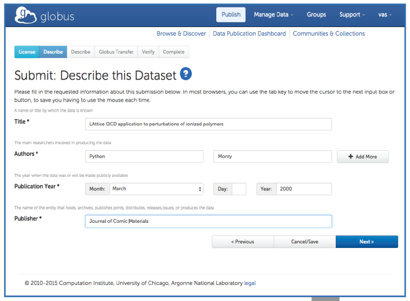
**Access**

Anonymous Public Embargoed Collaborators

**Preservation**

Transient Project Lifetime Archive “forever”

## Publish



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

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

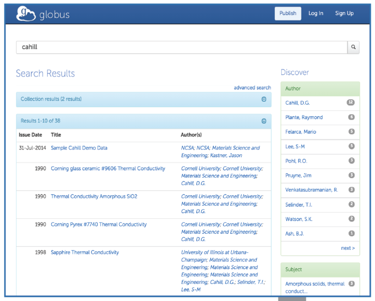
Publication Year * Month: March Day: Year: 2000

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

Publisher * Journal of Comic Materials

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

## Discover



cahill

Search Results

Collection results (2 results)

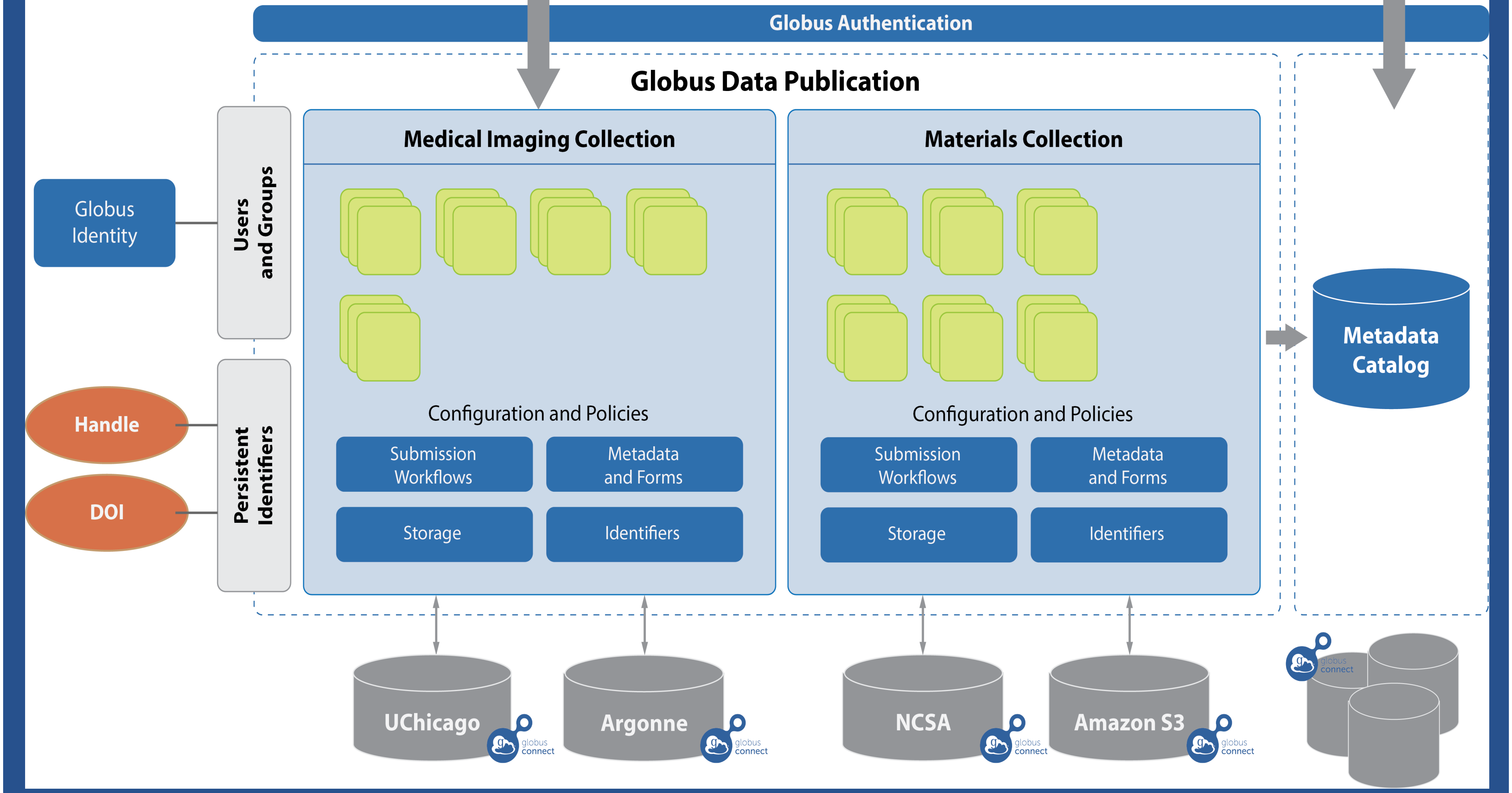
Results 1-10 of 38

| 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 |

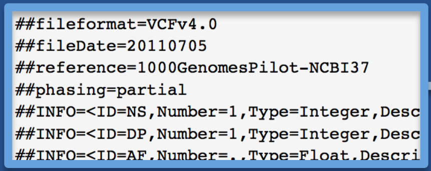
| Subject | |
|---|---|
| Amorphous solids, thermal conduct... | 3 |



# Scenario 1: Genomics analysis

**Campus HPC**

globus **genomics**



```
##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

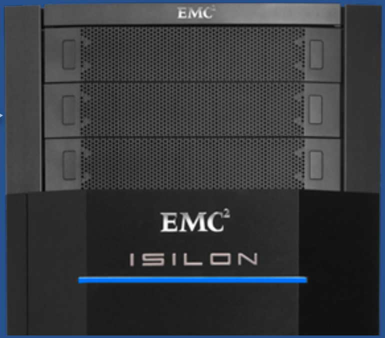
Automated curation

- Machine validated
- Exception review

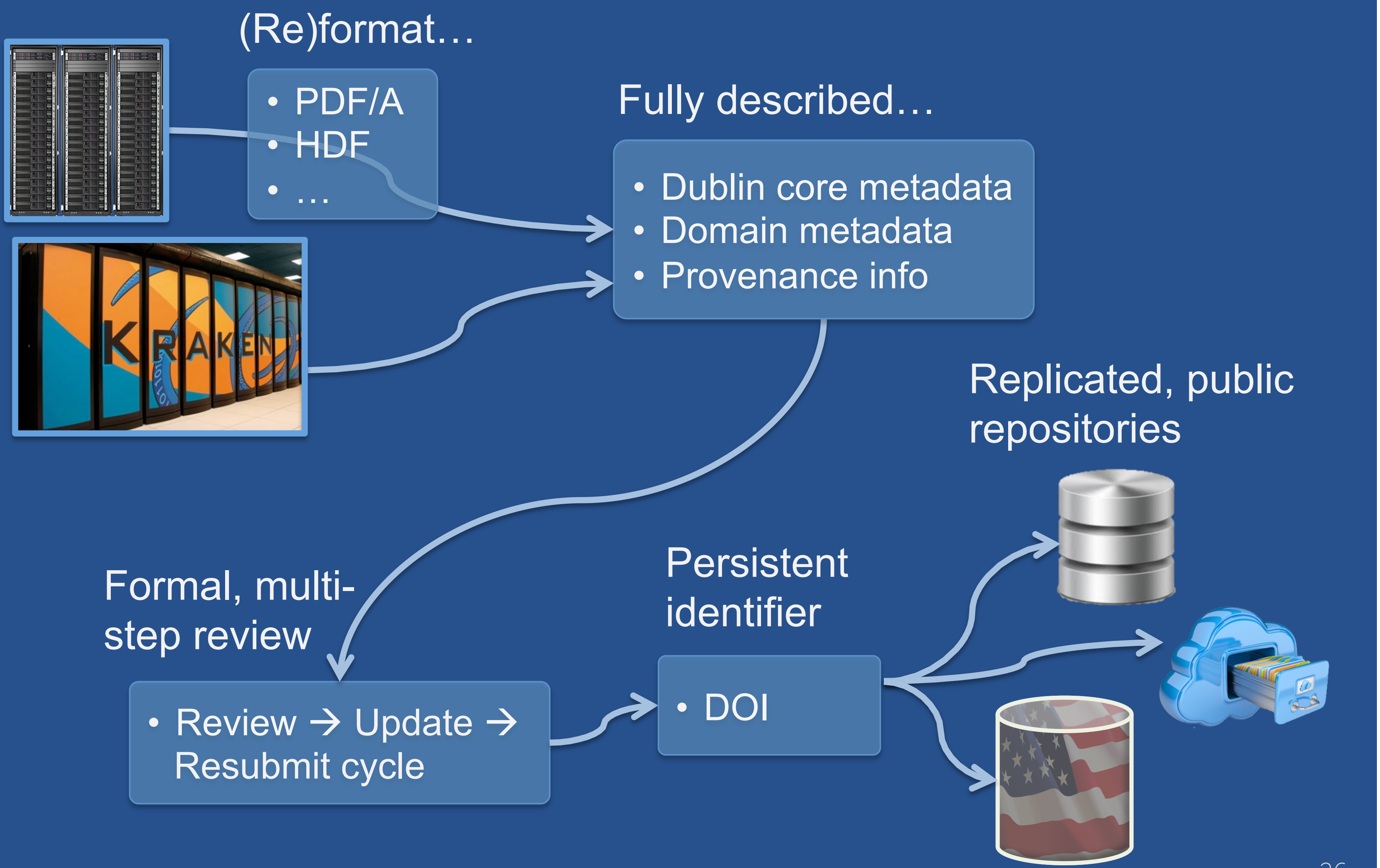
Identify…

- URL

Moderate durability/cost



# Scenario 2: 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

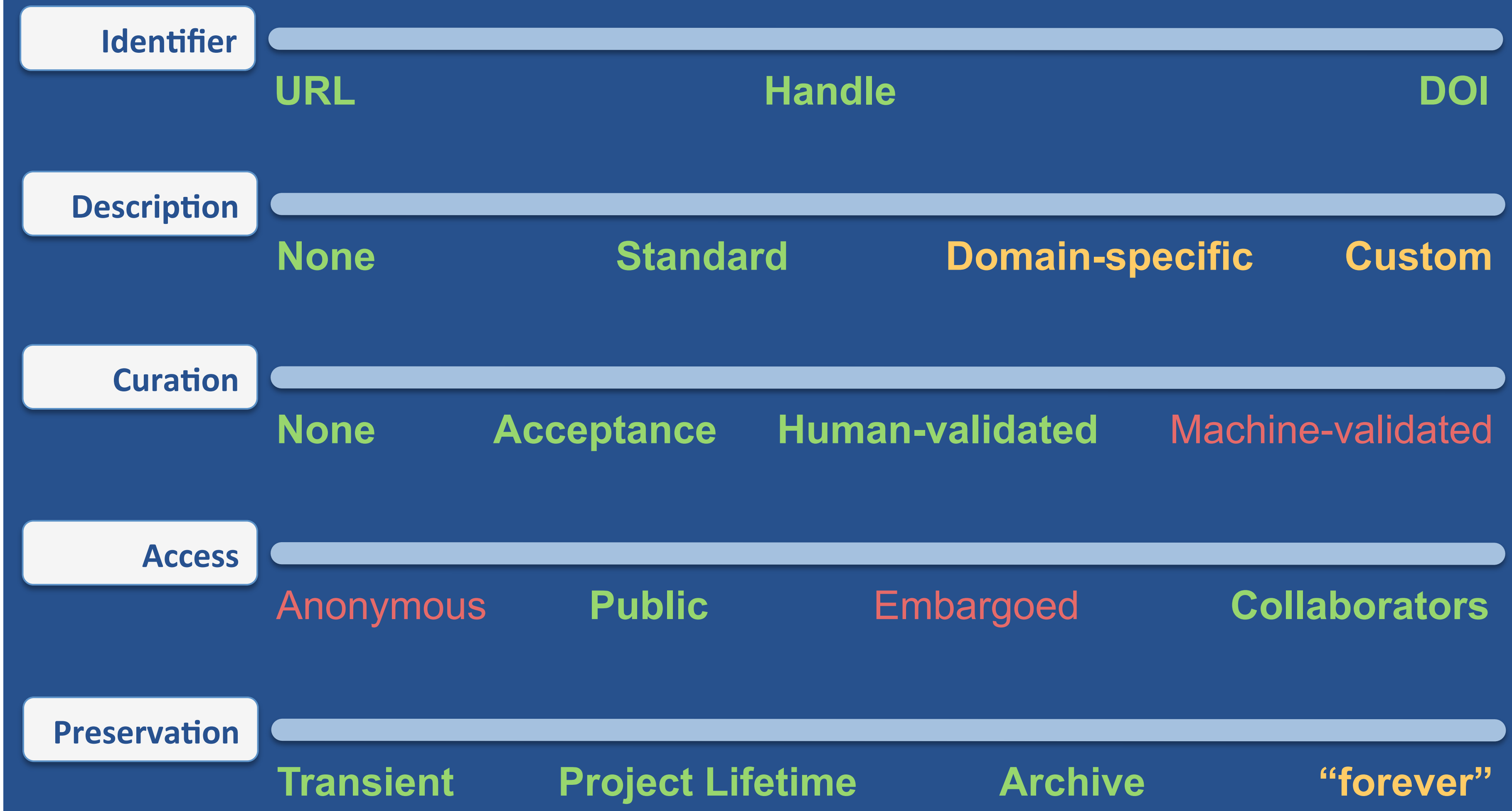
Persistent identifier

- DOI

Replicated, public repositories

# Globus publication – Current 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”

# Demonstration:
# Data Publication
# and Discovery

# 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

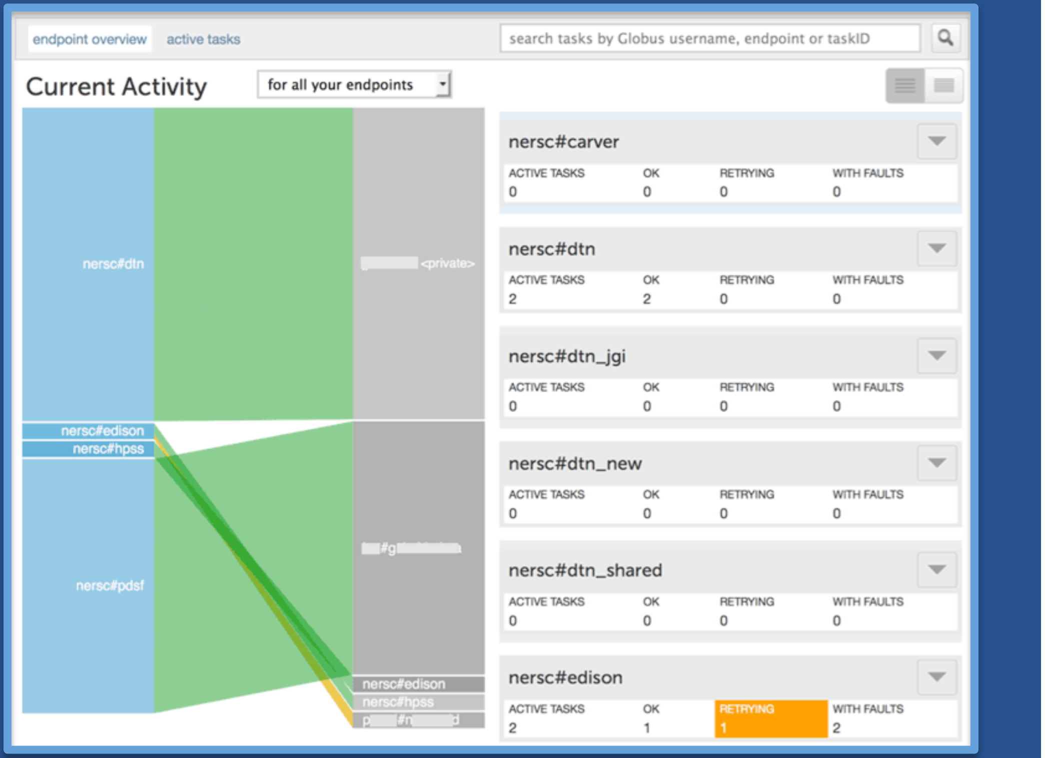
We are a non-profit, delivering a production-grade service to the non-profit research community

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**



Current Activity — for all your endpoints

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

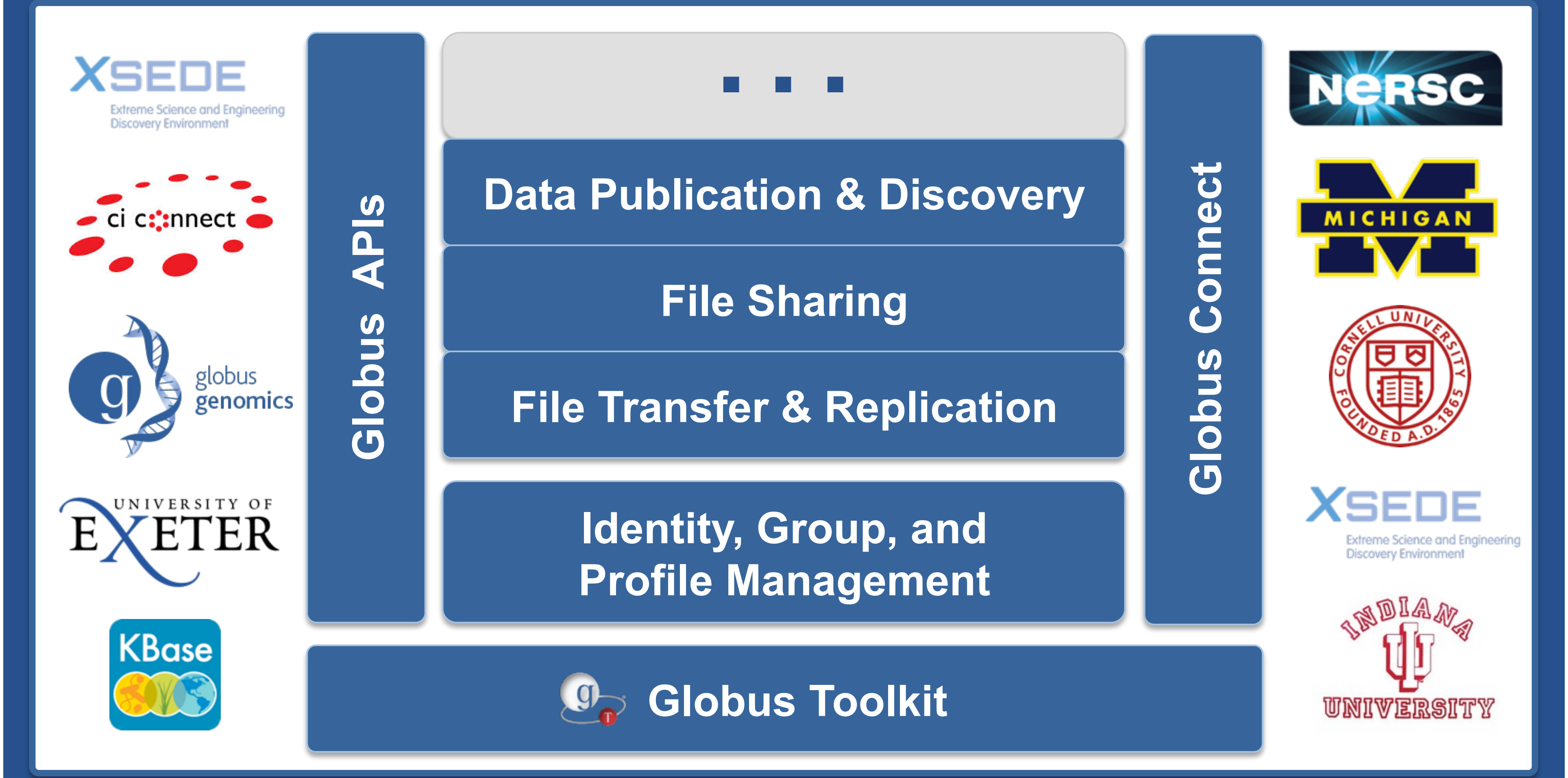
| ACTIVE TASKS | OK | RETRYING | WITH FAULTS |
|---|---|---|---|
| 2 | 1 | 1 | 2 |

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

**globus.org/provider-plans**

# Demonstration:
# Globus management console

# Globus Platform-as-a-Service



■ ■ ■

Data Publication & Discovery

File Sharing

File Transfer & Replication

Identity, Group, and Profile Management

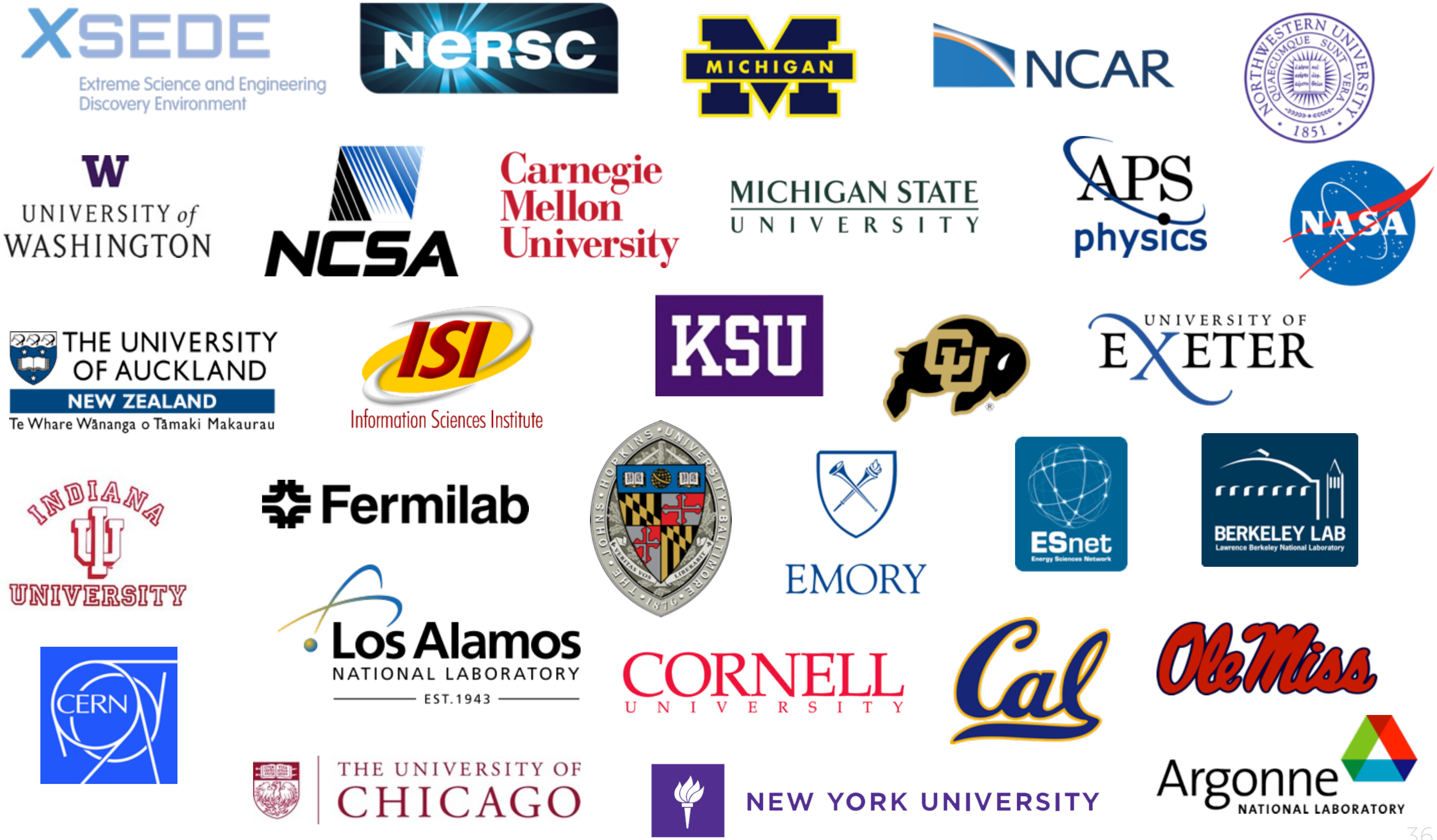
Globus APIs

Globus Connect

Globus Toolkit

XSEDE
Extreme Science and Engineering Discovery Environment
ci connect
globus genomics
UNIVERSITY OF EXETER
KBase
NERSC
MICHIGAN
CORNELL UNIVERSITY FOUNDED A.D. 1865
XSEDE
Extreme Science and Engineering Discovery Environment
INDIANA UNIVERSITY

# Some early Globus adopters


XSEDE
Extreme Science and Engineering Discovery Environment
NeRSC
MICHIGAN
NCAR
NORTHWESTERN UNIVERSITY QUAECUMQUE SUNT VERA 1851
W UNIVERSITY of WASHINGTON
NCSA
Carnegie Mellon University
MICHIGAN STATE UNIVERSITY
APS physics
NASA
THE UNIVERSITY OF AUCKLAND NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau
ISI
Information Sciences Institute
KSU
CU
UNIVERSITY OF EXETER
INDIANA UNIVERSITY
Fermilab
THE JOHNS HOPKINS UNIVERSITY BALTIMORE 1876
EMORY
ESnet
Energy Sciences Network
BERKELEY LAB
Lawrence Berkeley National Laboratory
CERN
Los Alamos NATIONAL LABORATORY EST.1943
CORNELL UNIVERSITY
Cal
Ole Miss
THE UNIVERSITY OF CHICAGO
NEW YORK UNIVERSITY
Argonne NATIONAL LABORATORY

# End: Introduction and Service Overview