Globus Research Data Management: Introduction and Service Overview

Kyle Chard
chard@uchicago.edu

Ben Blaiszik
blaiszik@uchicago.edu

Thank you to our sponsors!

U.S. DEPARTMENT OF
ENERGY

THE UNIVERSITY OF
CHICAGO

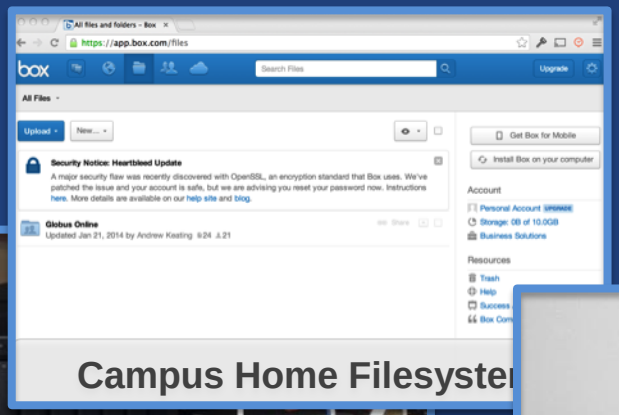
Argonne
NATIONAL LABORATORY

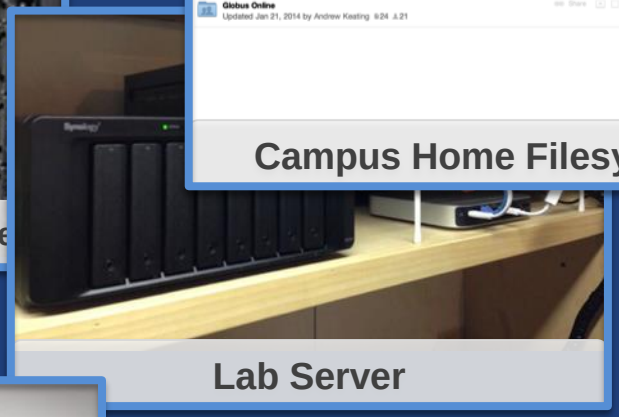
Agenda

- **Research data management challenges**
- **Globus: a high-level flyover**
- **File Transfer and Sharing: Accelerating and streamlining collaboration**
- **Data Publication: Enhancing reproducibility and discoverability**
- **New features: Globus Auth and HTTP**
- **Leveraging the Globus platform**

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.”

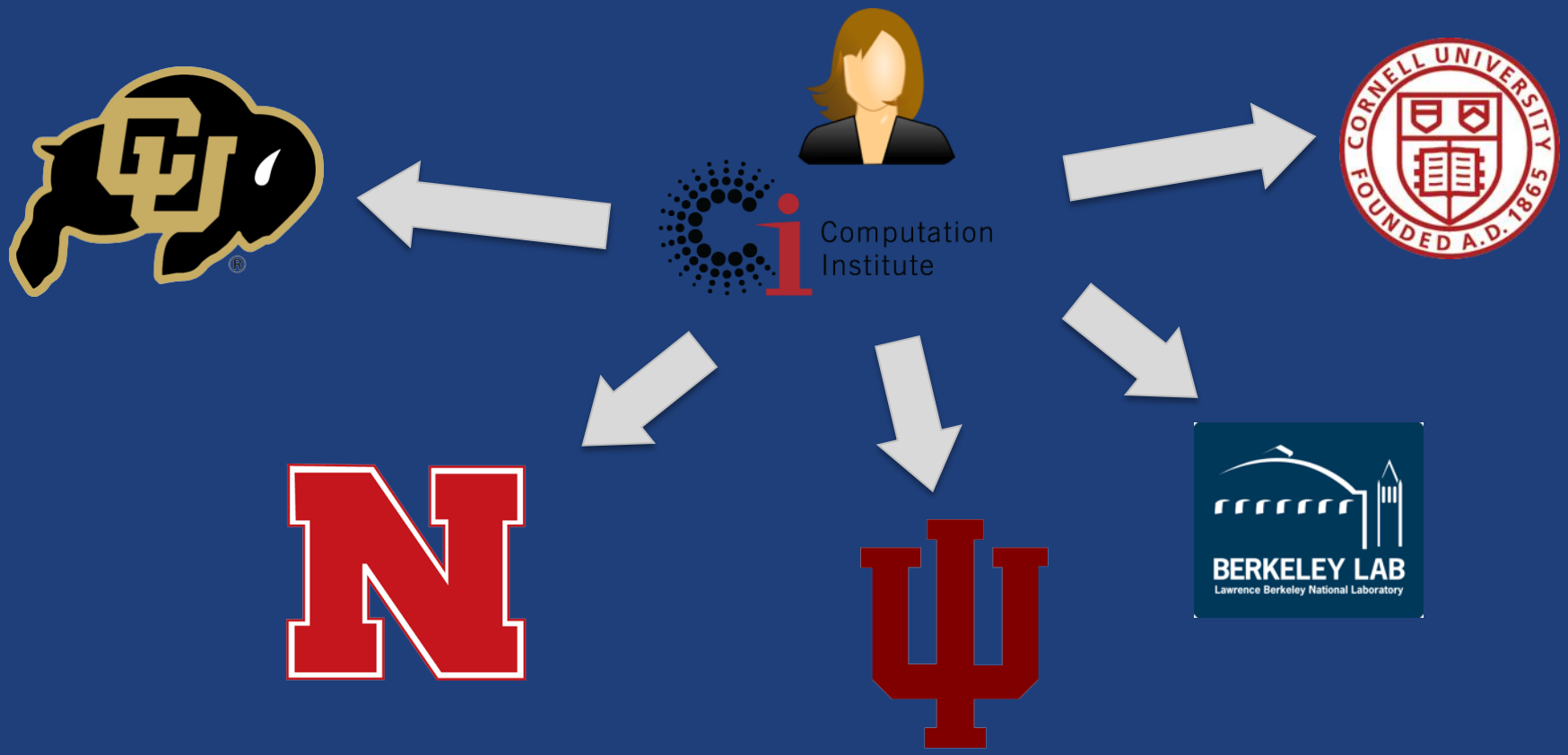
MRI

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

Reference Dataset

Scholarly Publication

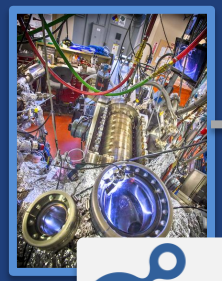
Research Community Collaboration

Research data management today

Index?

Globus and the research data lifecycle

Instrument



Globus transfers files reliably, securely

2

Compute Facility

4 Globus controls access to shared files on existing storage; no need to move files to cloud storage!

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

Transfer

globus connect

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

Collaborator logs in to Globus and accesses shared files; no local account required; download via Globus

5

Personal Computer

6

Researcher assembles data set; describes it using metadata (Dublin core and domain-specific)

Publish

6

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 kyle#ncsa-tutorial 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**

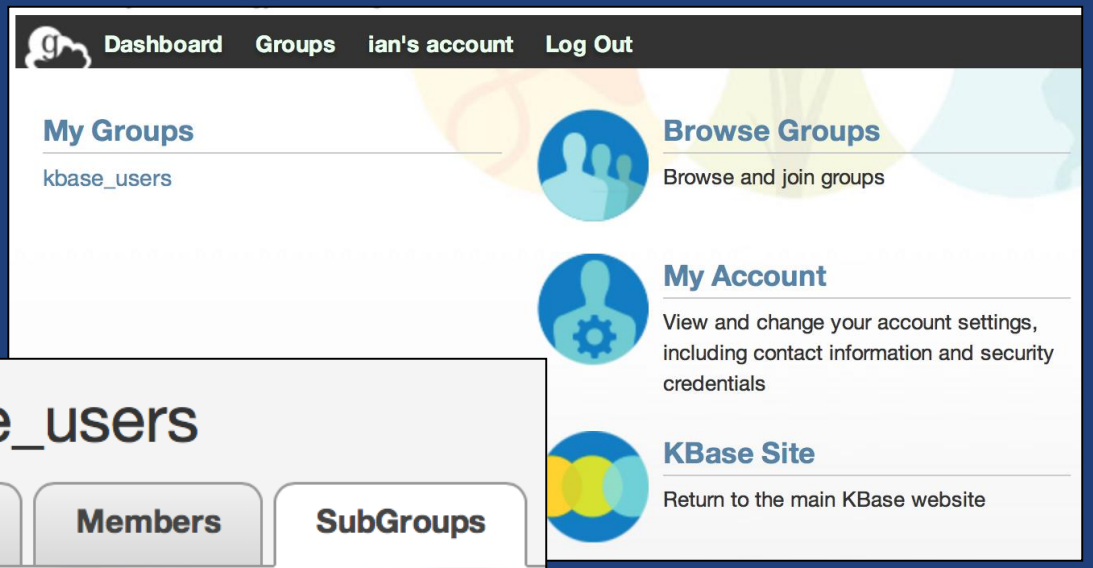
Group Management

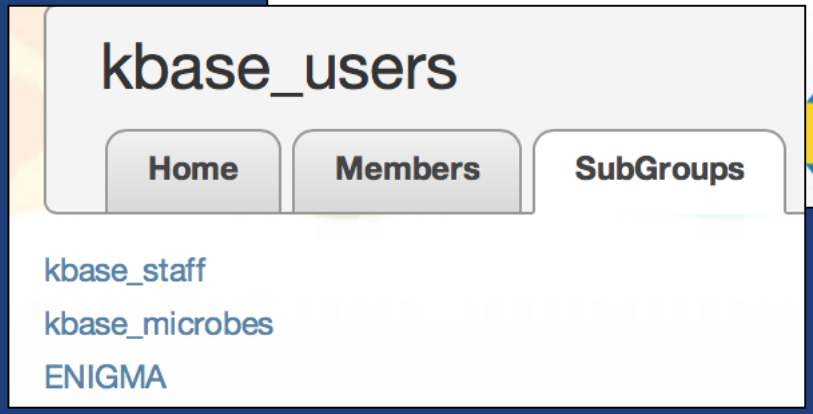
Group Management

- **User-managed group creation and management**
- **Flexible control over roles, policies, workflows, and visibility**
- **Groups can be used in authorization decisions**

Example: kBase

- Every kBase user added to kbase_users
- Subgroups also created
- Groups used for access control



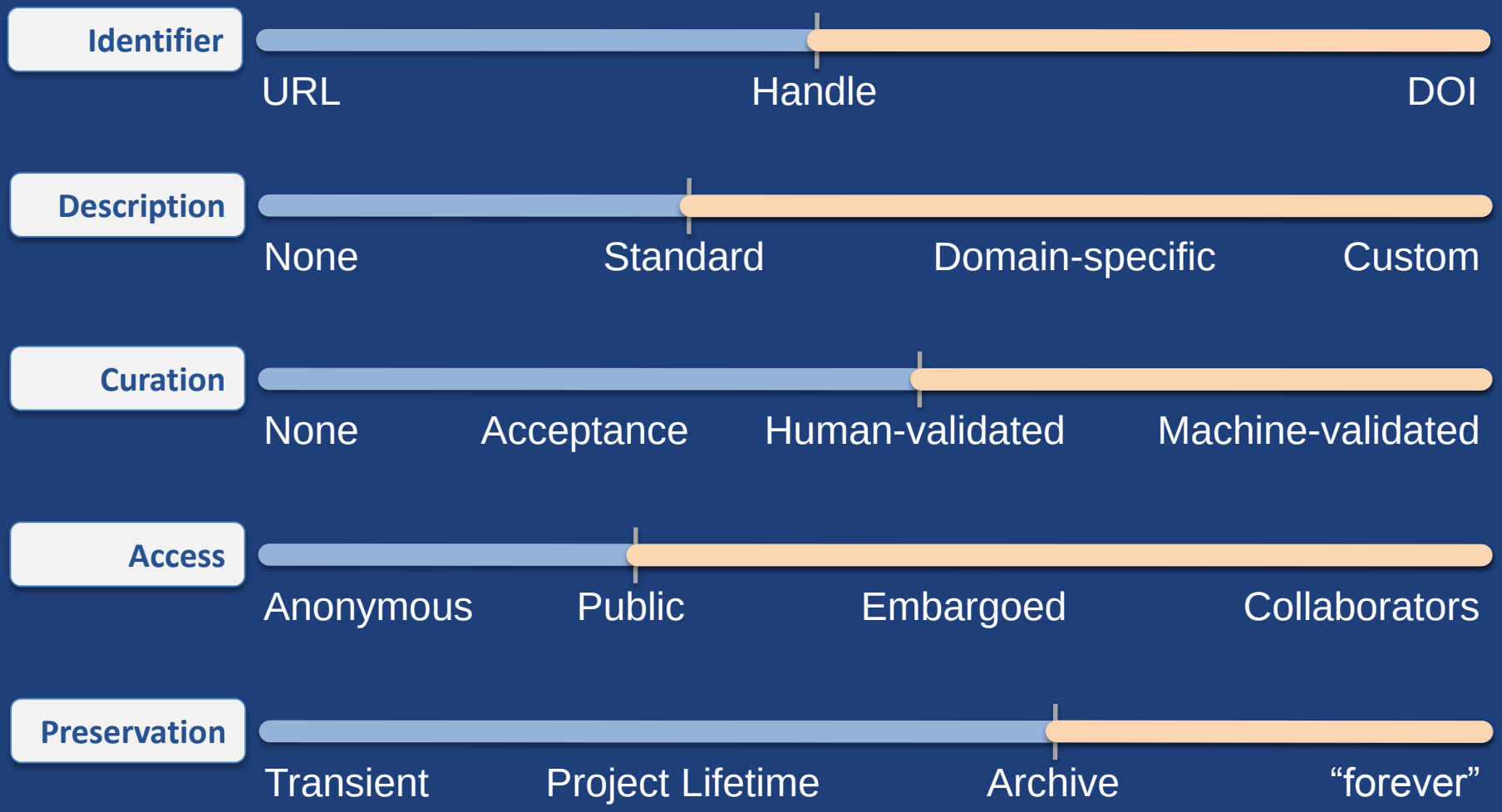


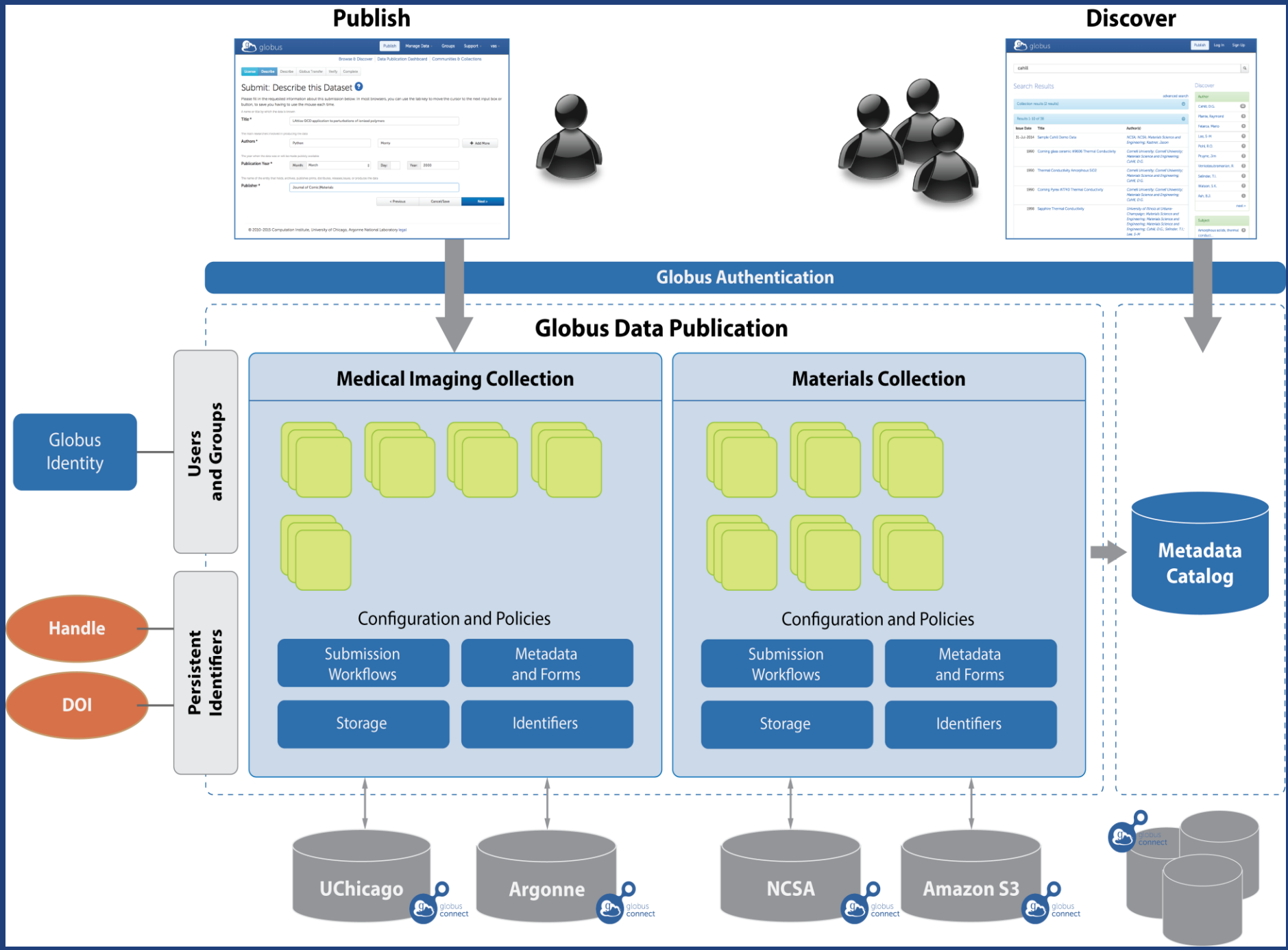
Exercise 3: Create/configure group

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
Discover
Globus Authentication
Globus Data Publication
Medical Imaging Collection
Materials Collection
Configuration and Policies
Submission Workflows
Metadata and Forms
Storage
Identifiers
Users and Groups
Persistent Identifiers
Globus Identity
Handle
DOI
Metadata Catalog
UChicago
Argonne
NCSA
Amazon S3
globus connect

Yet another publication system?

- **Software as a Service**
- **Self service management**
  - Identifiers, policies, submission and curation workflows, storage, metadata, access control
- **Remote storage**
- **Supports arbitrarily large datasets**
- **Powerful search**

Exercise 4: Publish a dataset

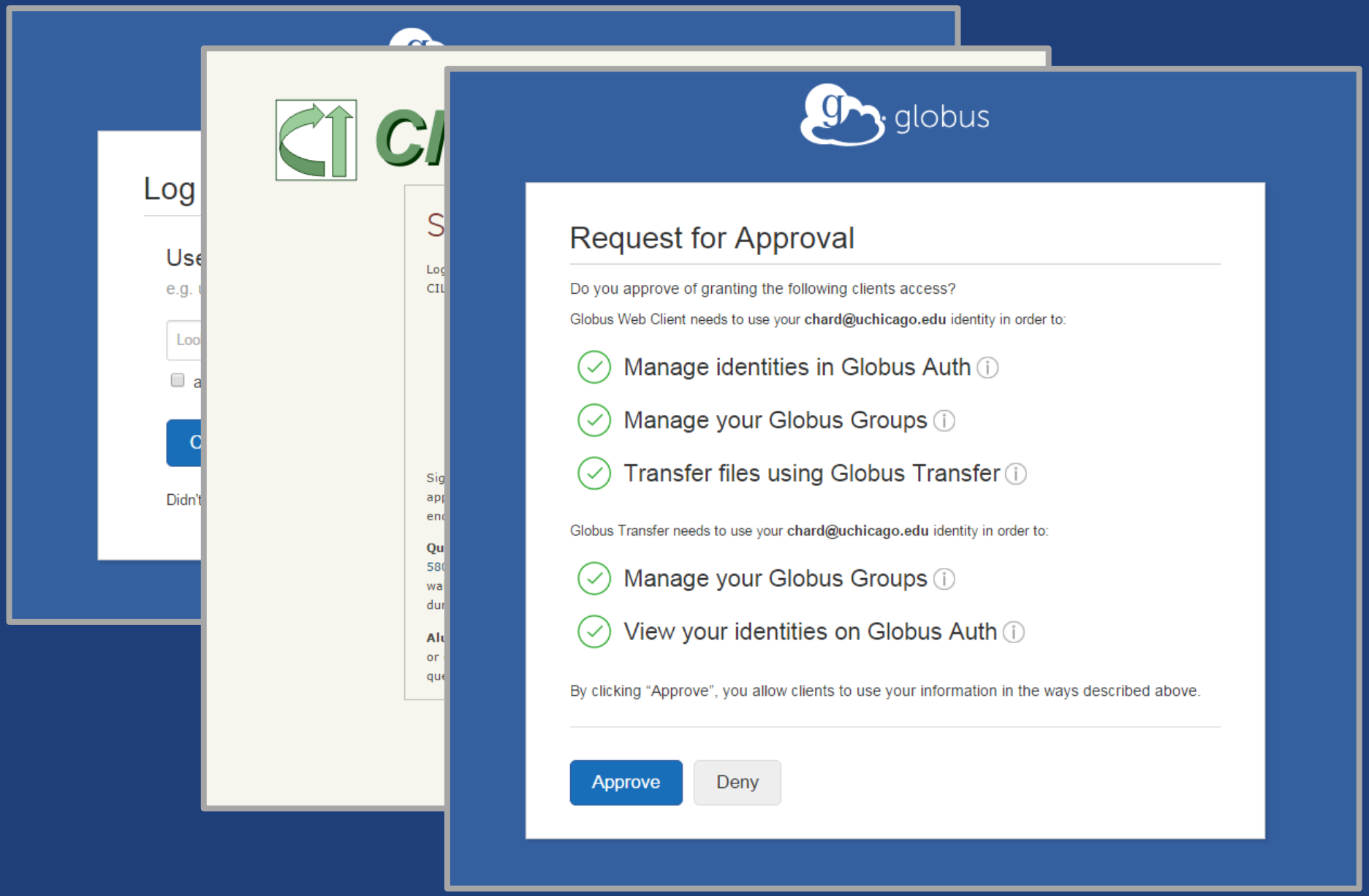
1. **Go to trial.publish.globus.org**
2. **Log in, click “Submit a New Dataset”**
3. **Select either of the Open Trial collections and continue**
   1. Accept the license terms
   2. Enter required metadata to describe the dataset
   3. Assemble data set from the kye#ncsa-tutorial endpoint (or your own laptop if you installed Globus Connect Personal)
   4. Complete the workflow and submit
4. **Curators (a.k.a. presenters) will “review” your submission and publish**
5. **Search for your published dataset and browse the data**

Globus today…

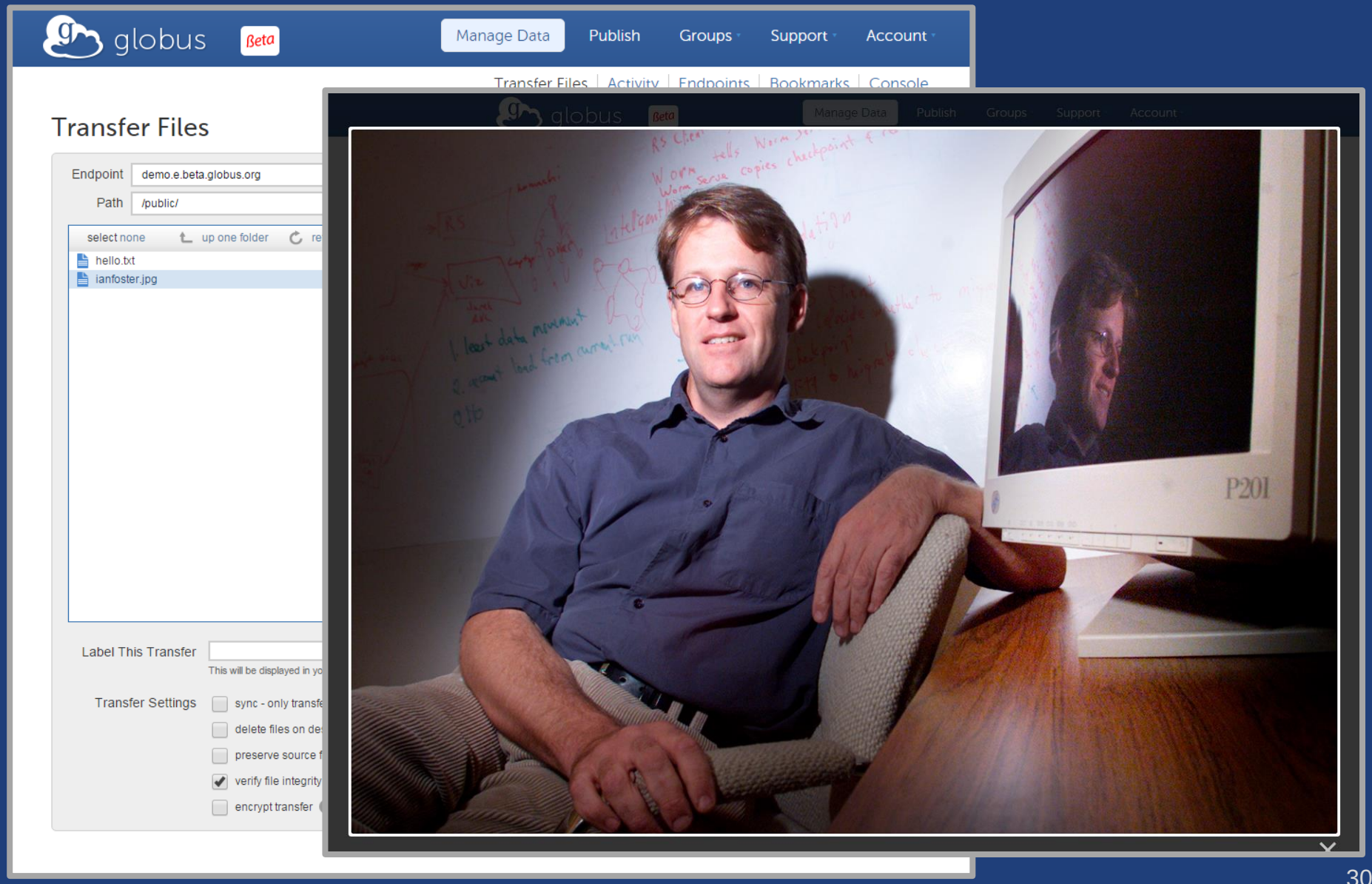
4 major services	**122 PB** transferred	**20 billion** files processed	**31,000** registered users
13 national labs use Globus	**10,000** active endpoints	**~350** active daily users	**99.95%** uptime
35+ institutional subscribers	**1 PB** largest single transfer to date	**3 months** longest continuously managed transfer	**130** federated campus identities

New Features: Globus Auth and HTTP

New features: Globus Auth

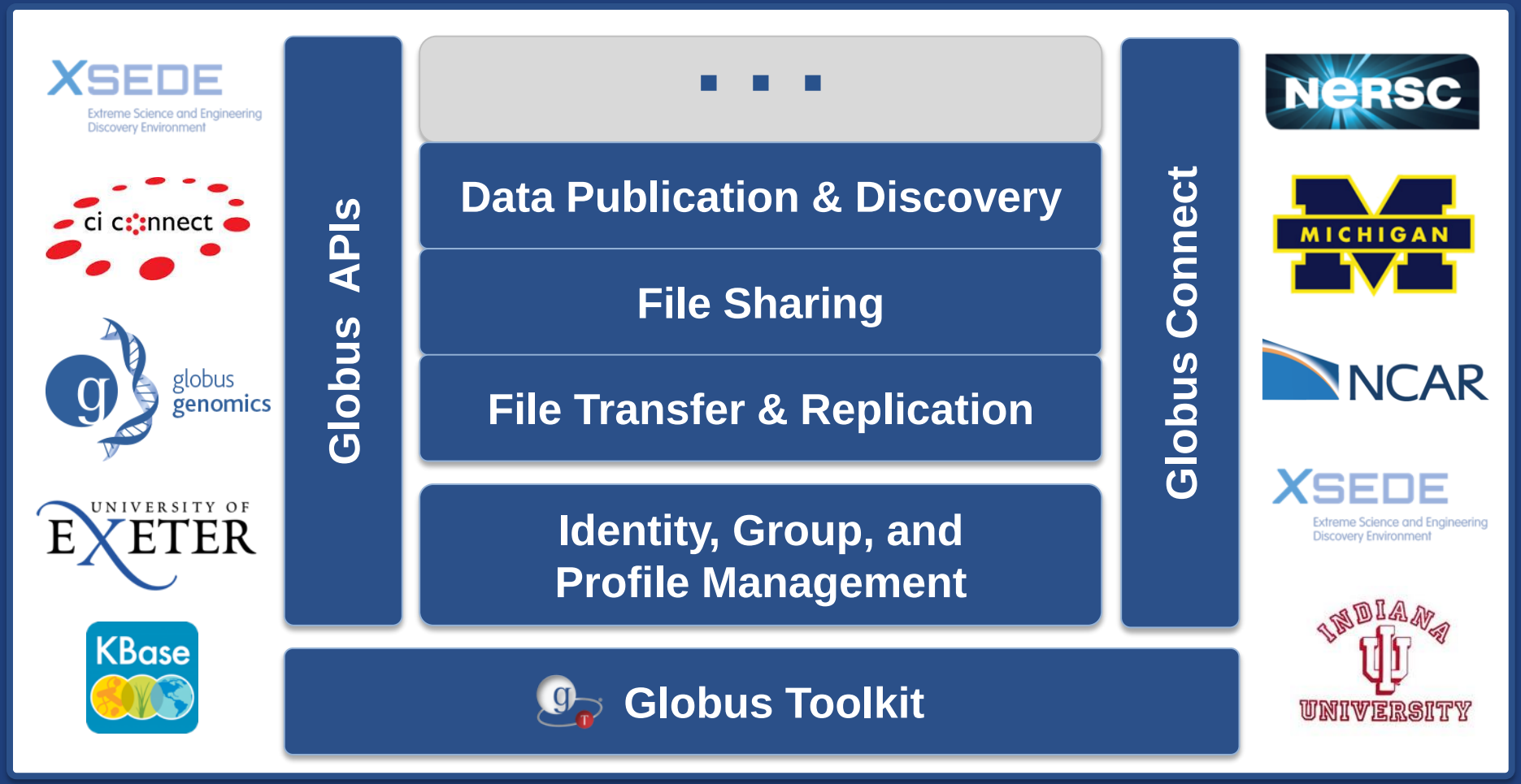


New features: HTTP



Leveraging the Globus Platform

Globus Platform-as-a-Service



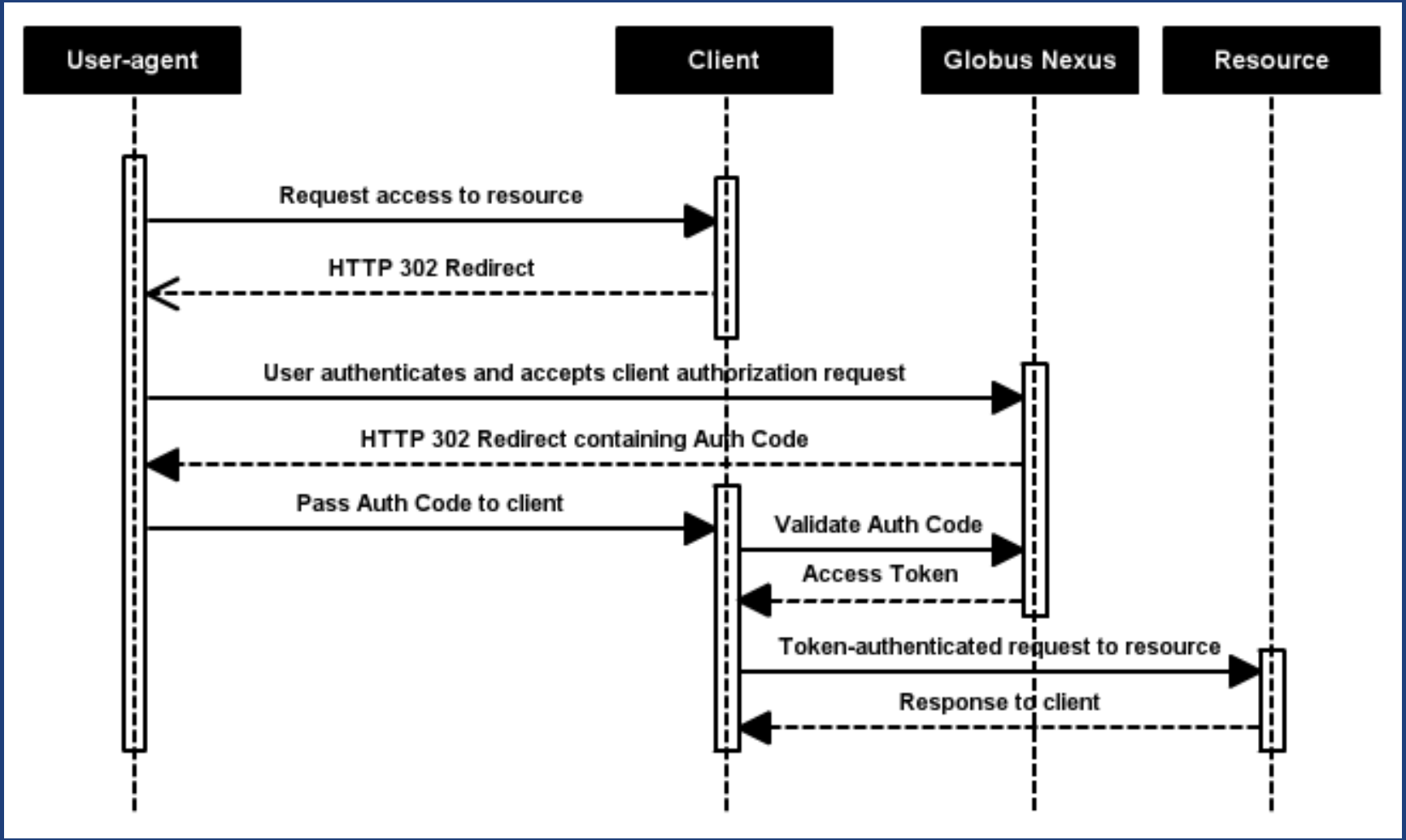
APIs and Clients

- **Globus Nexus (identity, profile, groups)**
  - API: http://globusonline.github.io/nexus-docs/api.html
  - Clients:
    - https://github.com/globusonline/python-nexus-client
    - https://github.com/globusonline/java-nexus-client
- **Globus transfer**
  - API: https://docs.globus.org/api/transfer/
  - Clients:
    - https://github.com/globusonline/transfer-api-client-python
    - https://github.com/globusonline/transfer-api-client-java
- **Globus publication (coming soon)**

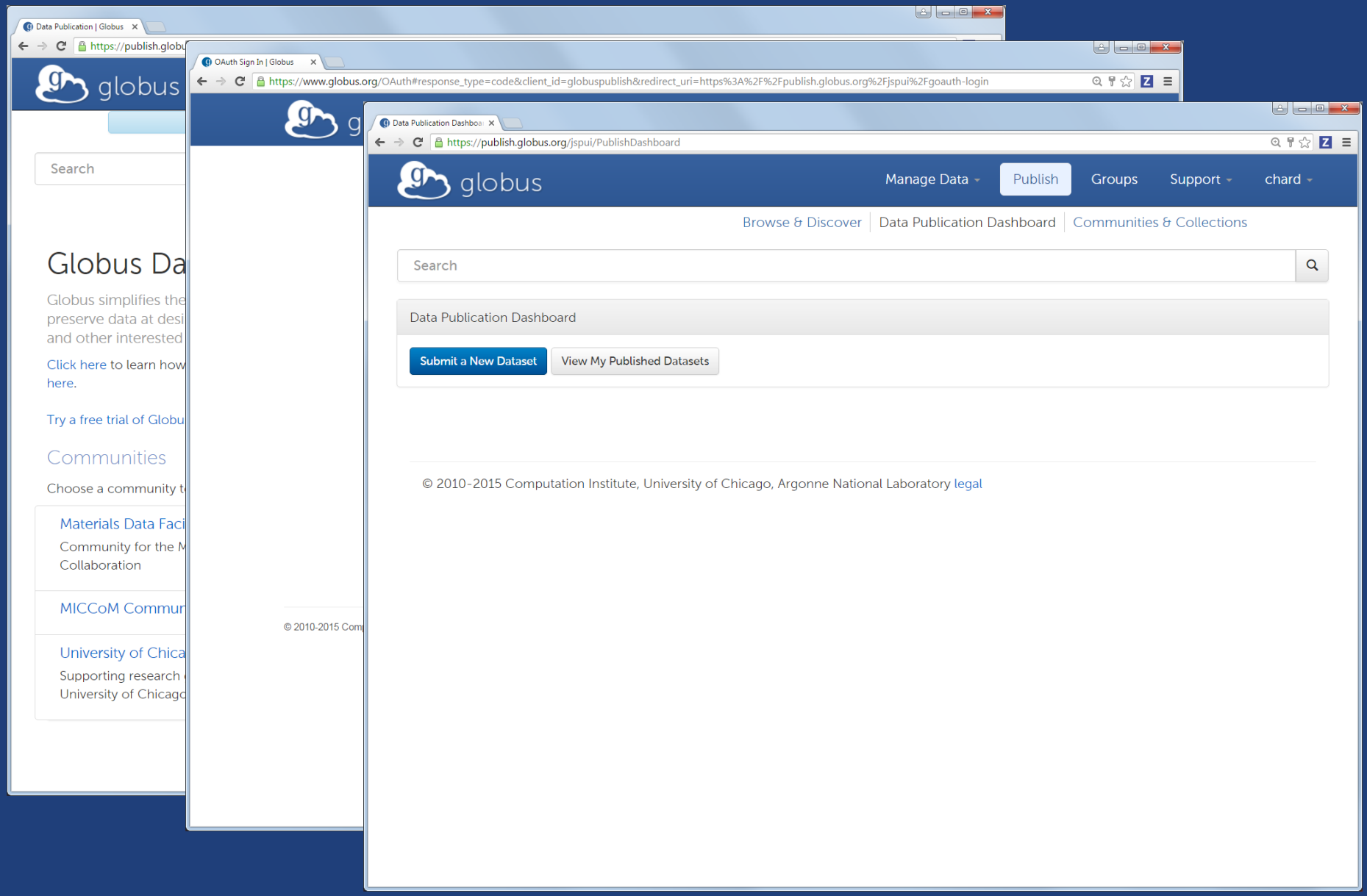
Developing a 3rd party application

- **Register client (URL and client username/password)**
- **Implement an OAuth client (or reuse an existing one)**
- **Store the resulting access token and make calls to Globus APIs on behalf ot the authenticated user**

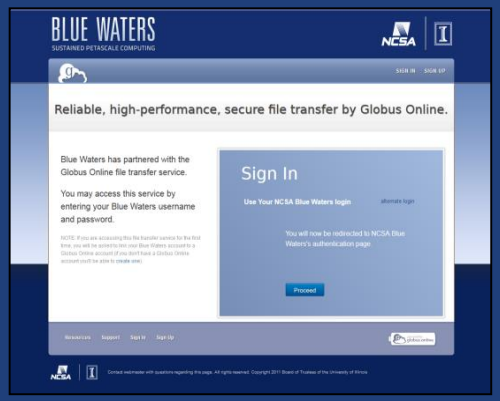
Globus OAuth workflow

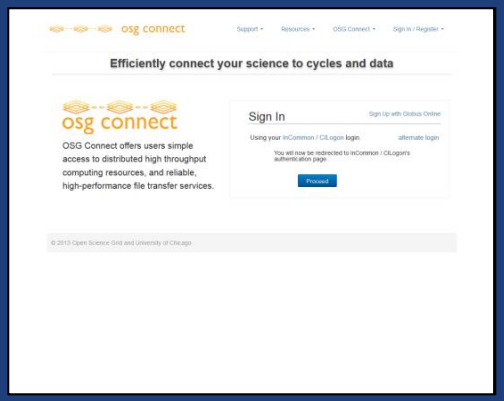


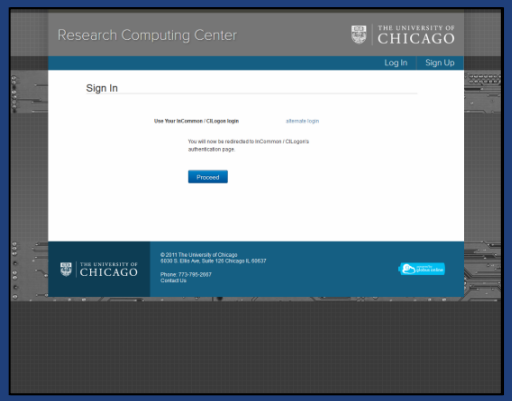
Developing a 3rd party application

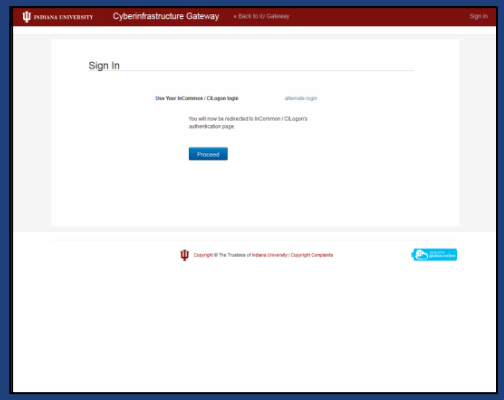


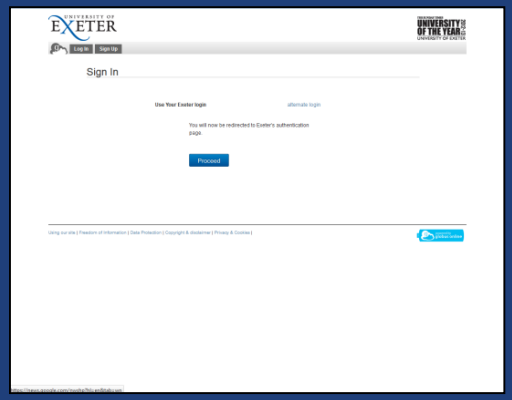
Branded sites

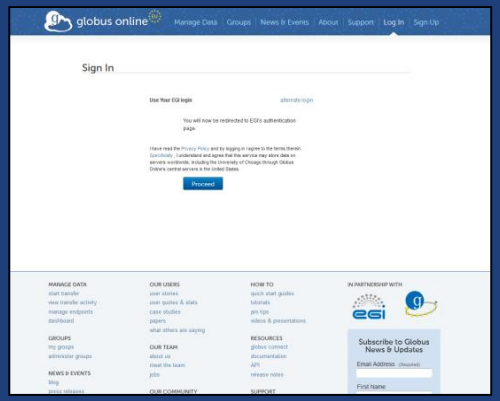


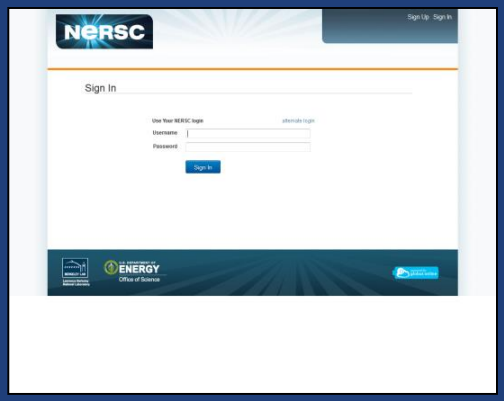












Resouces

- Signup: **globus.org/signup**
- Enable your resource: **globus.org/globus-connect-server**
- Need help? **support.globus.org**
- Follow us: **@globusonline**

Thank you!