Building the Model Research Data Portal

Ian Foster
foster@uchicago.edu

globusworld 2016
APRIL 20-22
CHICAGO

Developer Workshop: Building the Modern Research Data Portal

New high-speed networks make it possible, in principle, to transfer and share research data at tremendous speeds and scales–but have also proved challenging to use in practice. Two new technologies now allow us to translate this potential into reality: Science DMZ architectures provide frictionless end-to-end network paths; and Globus APIs allow programmers to create powerful research data portals that leverage these paths for data distribu synchronization, and other useful purposes.

Introduction, Concepts, and Components

IMPERIAL 2

Led by: **TBD**

We will introduce the Modern Research Data Portal and set the context for how Globus and the ScienceDMZ combine to deliver unique data management capabilities. This will include:

- Overview of use cases: Common patterns like data publication/distribution, orchestration of data flows, etc.
- Overview of the Globus platform: Architecture and brief overview of available services
- Introduction to the Globus Auth API: Authenticating and authorizing a client
- Introduction to the Globus Transfer API: Make your first call and move data with Globus
- Introduction to the Python SDK for using Globus Auth and Transfer

Come to Chicago in April to learn more!

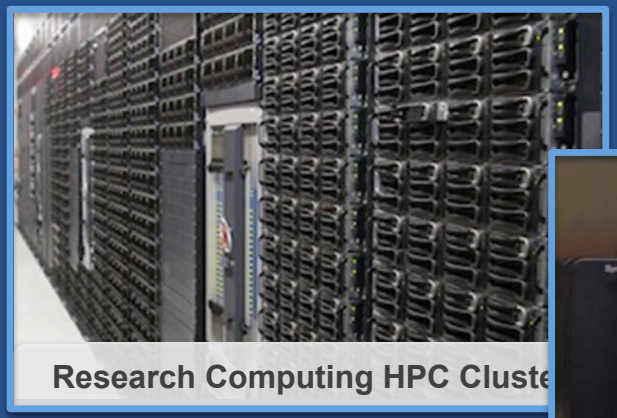
Thank you to our sponsors!

U.S. DEPARTMENT OF
ENERGY

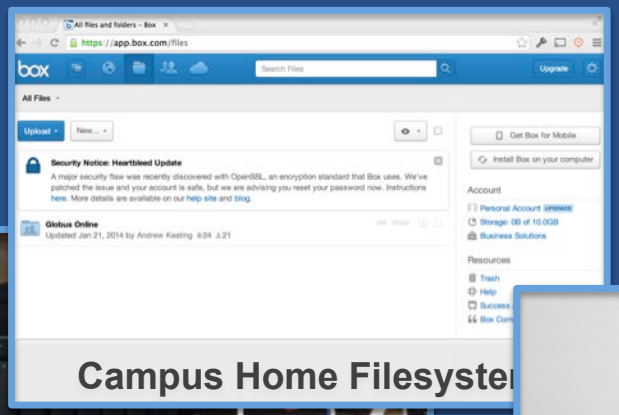
THE UNIVERSITY OF
CHICAGO

Argonne
NATIONAL LABORATORY

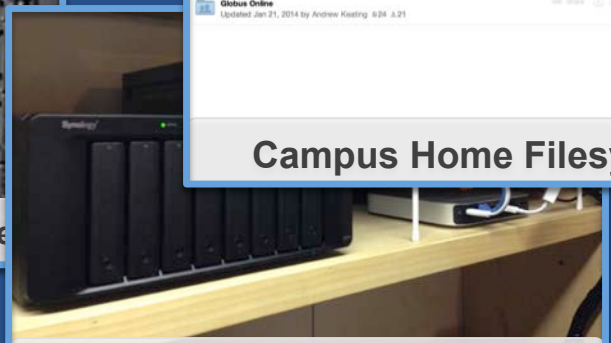
“I need to easily, quickly, & reliably move portions of my data to other locations.”



Research Computing HPC Cluste



Campus Home Filesyste



Lab Server

Personal Laptop

Desktop Workstation

XSEDE Resource

amazon web services S3

red cloud

Windows Azure

Google Cloud Platform

Public Cloud

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

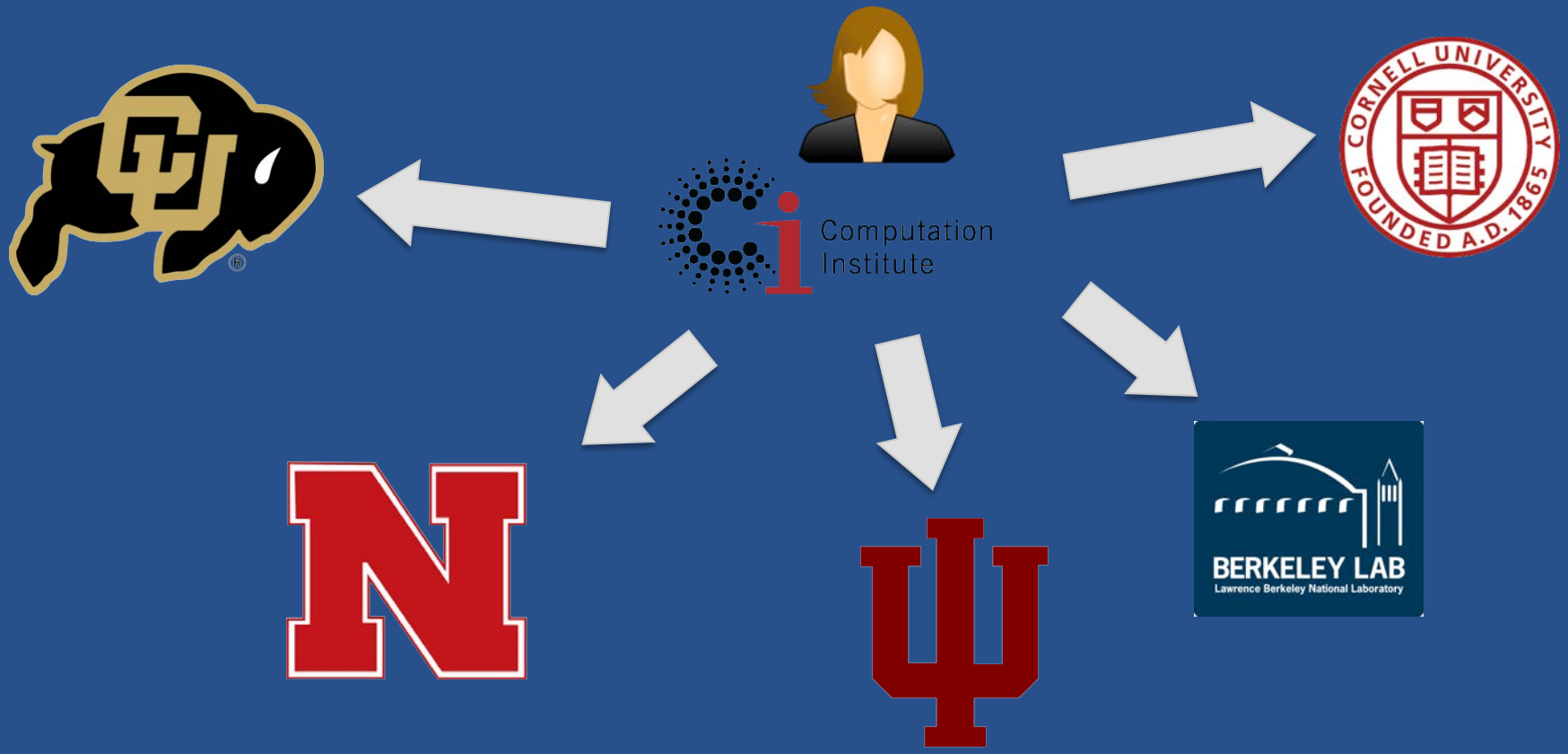
MRI

Advanced Light Source

Next Gen Sequencer

Light Sheet Microscope

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



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

Reference Dataset

Scholarly Publication

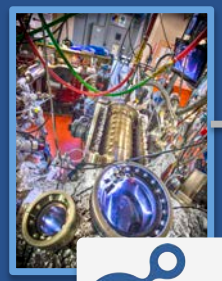
Research Community Collaboration

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

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

6

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

Publication Repository

5

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

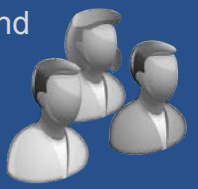
Personal Computer

Publish

6

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

Lowering collaboration overhead

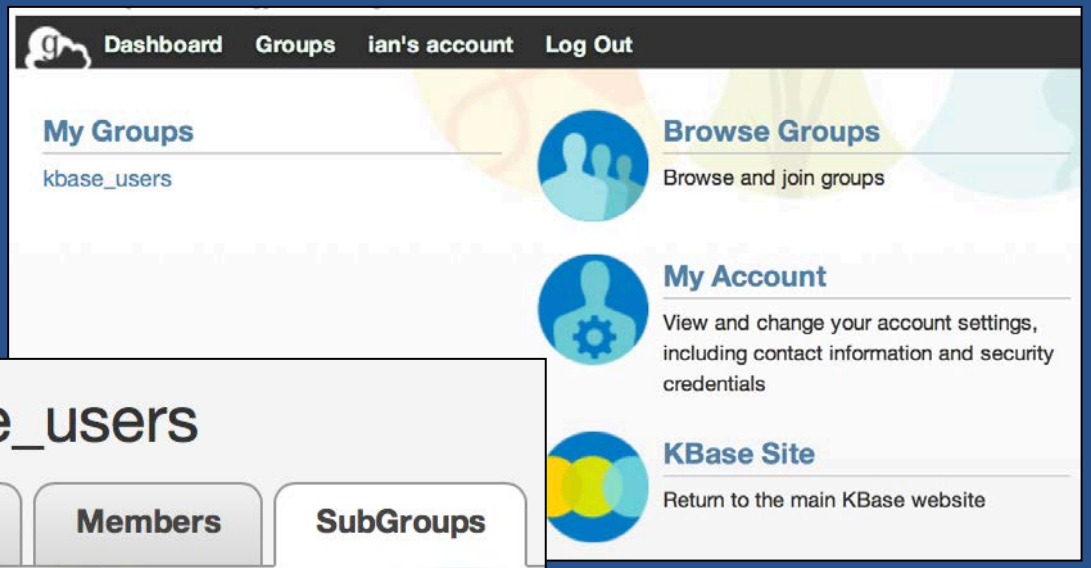
- **Grant collaborators access to data on systems without requiring local accounts**
- **No need to replicate or move data to separate system/cloud just for sharing**
- **Researchers manage “virtual” ACLs…**
- **Respect local system access controls**

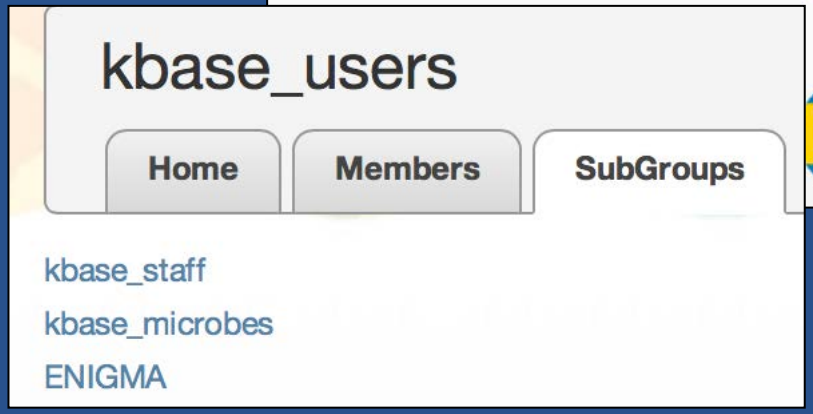
Group Management

- **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



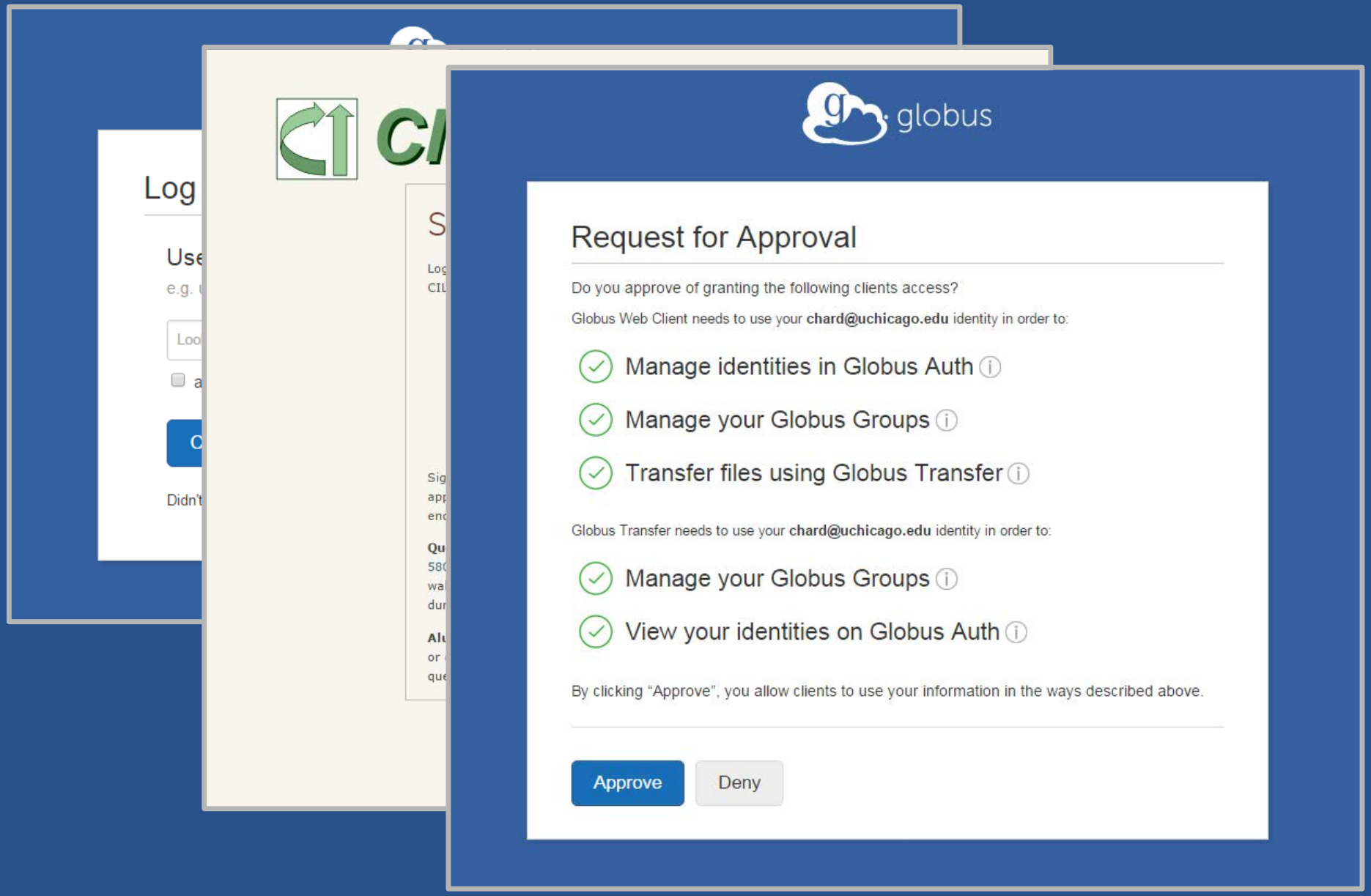


Globus today…

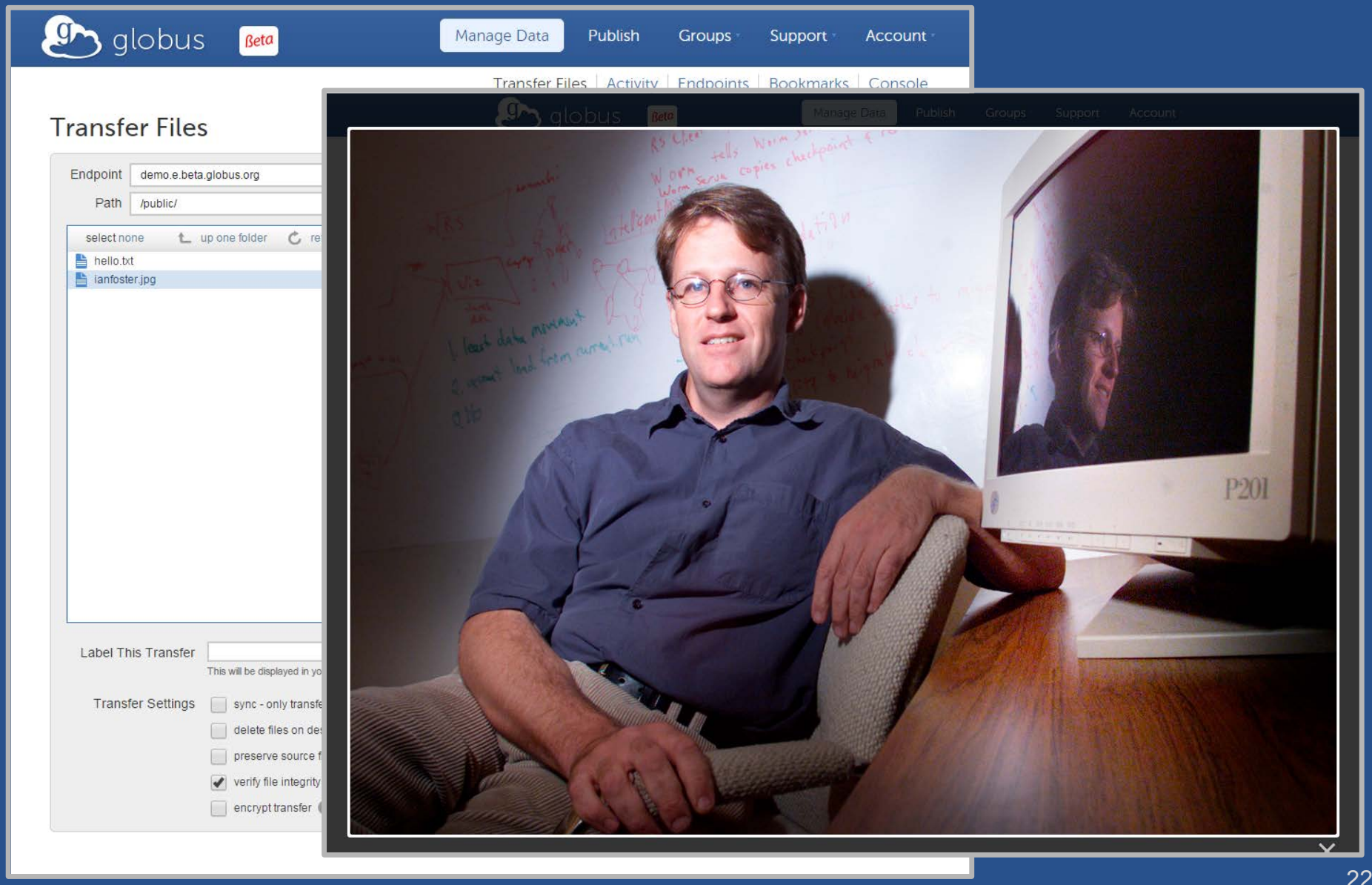
4 major services	**137 PB** transferred	**20 billion** files processed	**31,000** registered users
13 national labs use Globus	**10,000** active endpoints	**~400** active daily users	**99.95%** uptime
35+ institutional subscribers	**1 PB** largest single transfer to date	**3 months** longest continuously managed transfer	**130** federated campus identities

New Features: Globus Auth and HTTP

New features: Globus Auth

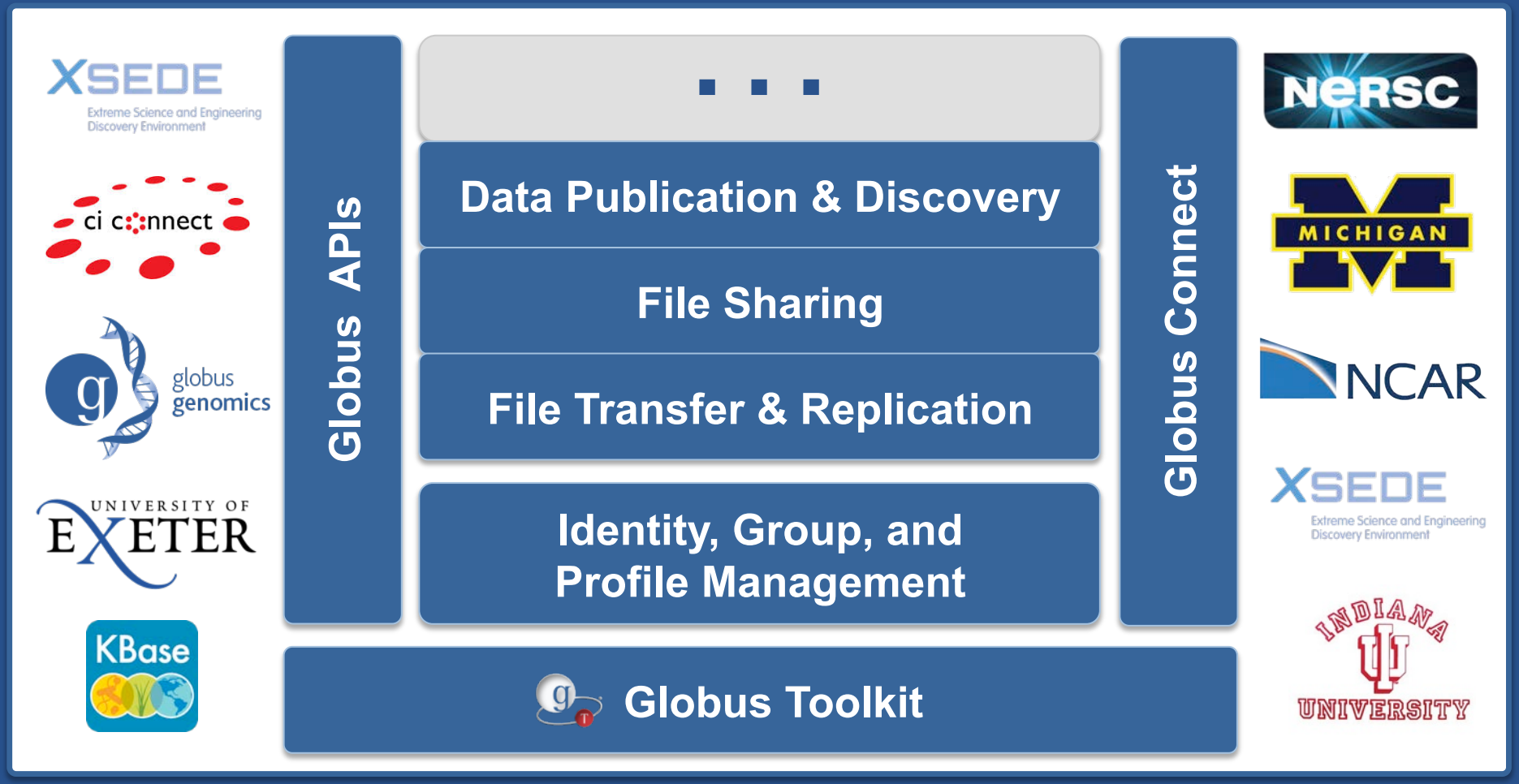


New features: HTTP



Leveraging the Globus Platform

Globus Platform-as-a-Service



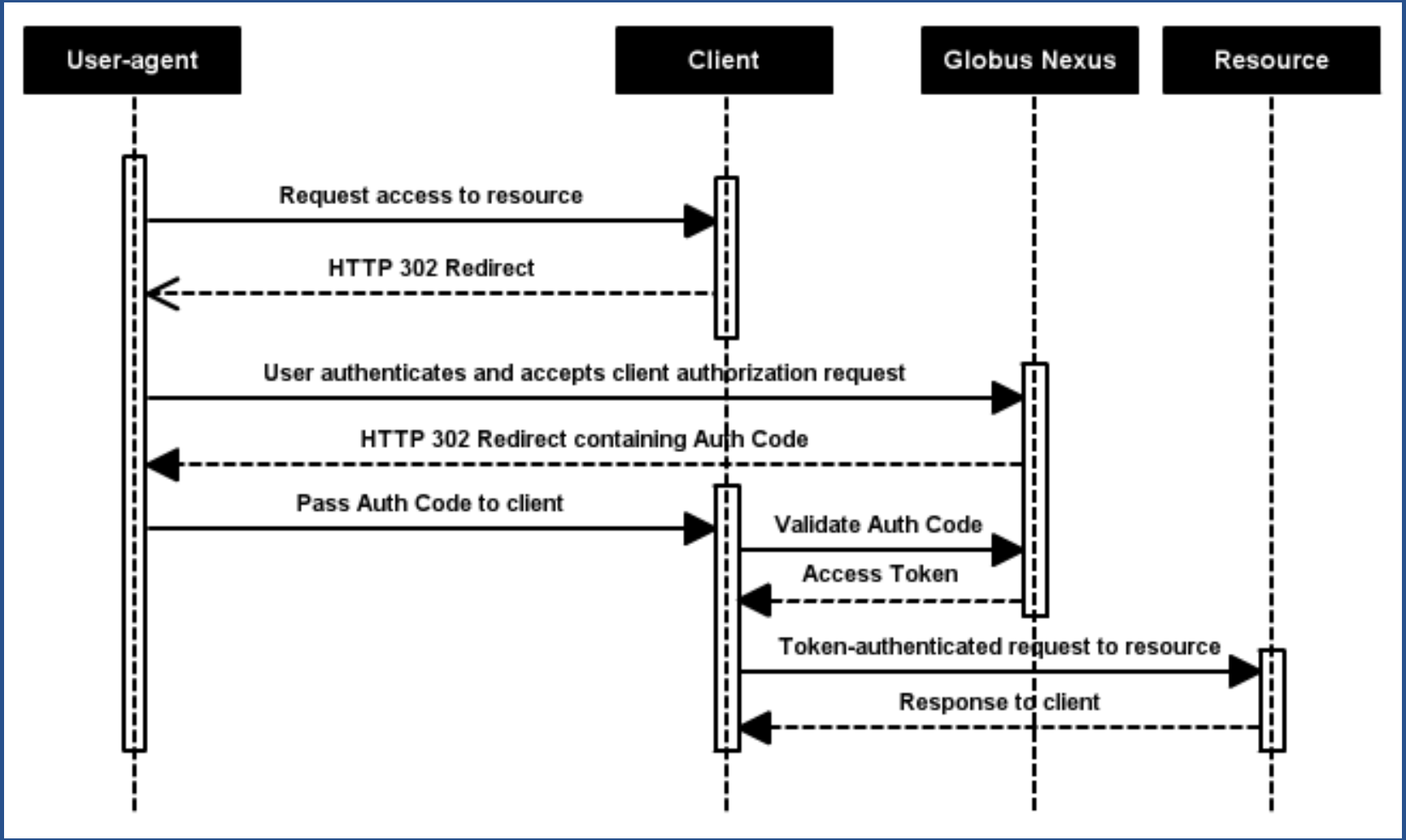
APIs and Clients

- **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)**

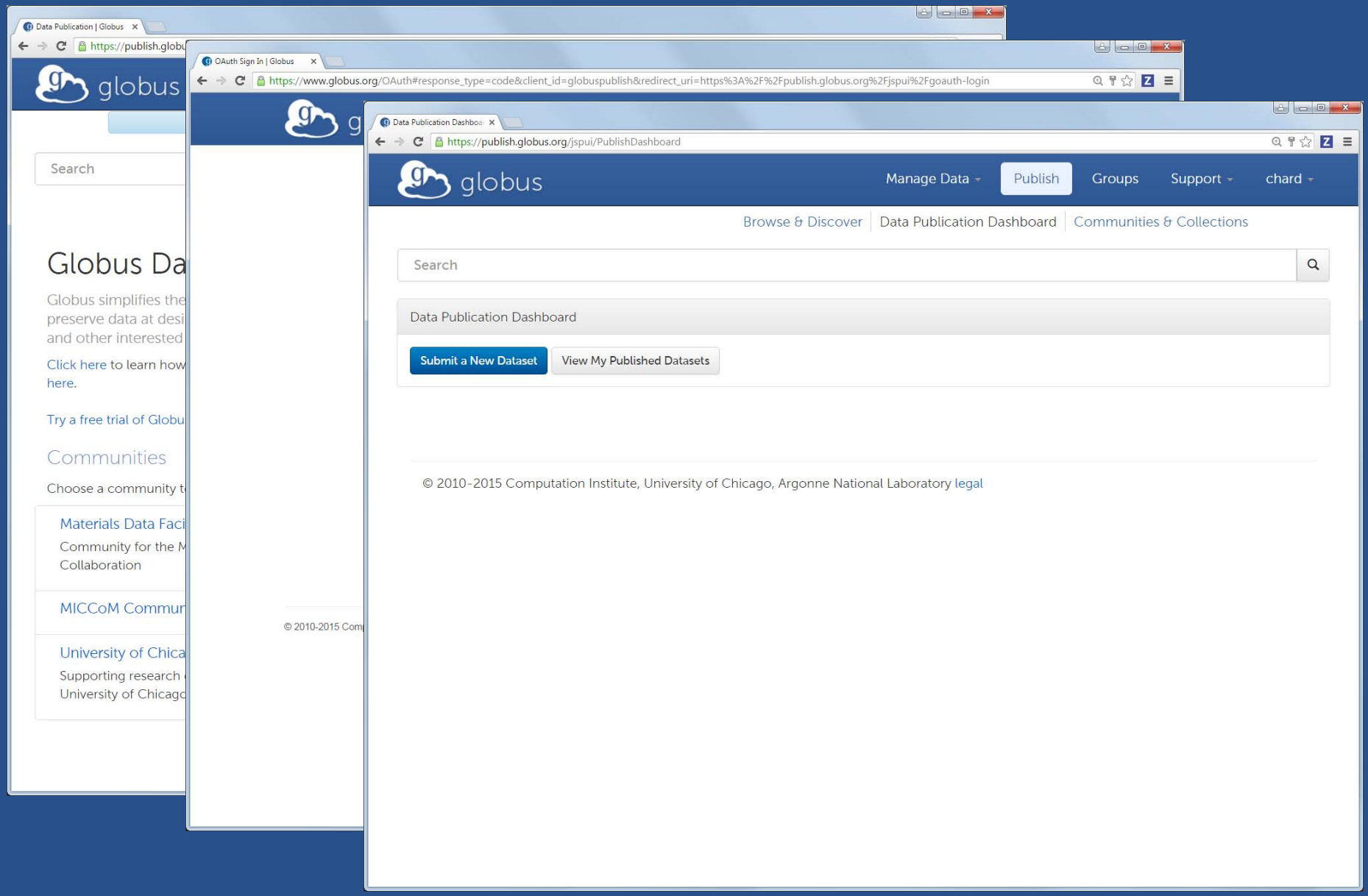
Developing a 3rd party application

- **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 of the authenticated user**

Globus OAuth workflow

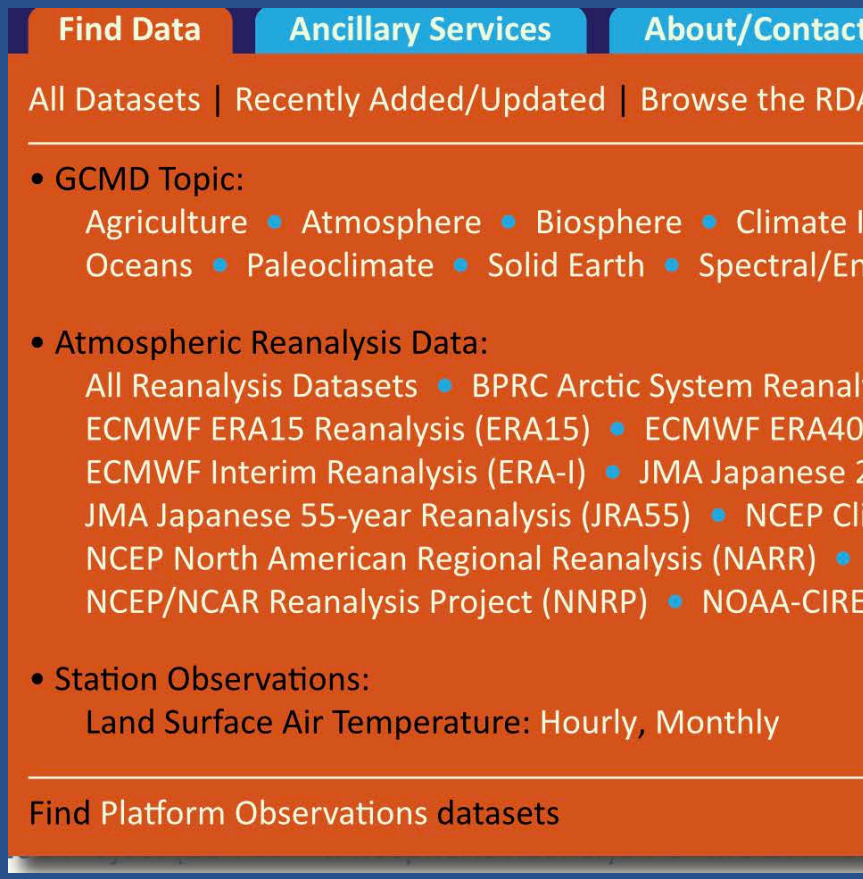


Developing a 3rd party application



Globus PaaS at NCAR

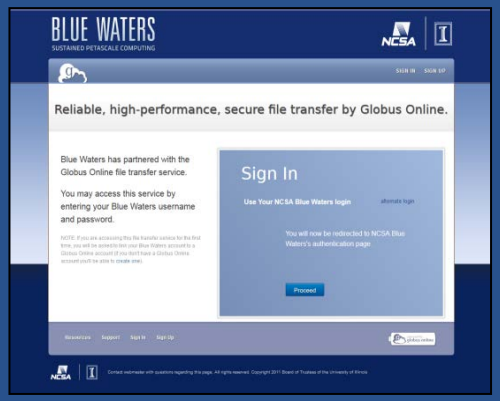
- **Research Data Archive at NCAR**
- **Integrate Globus for data downloads**
- **Shared endpoint with subfolder per request**
- **Single sign on via streamlined account provisioning**

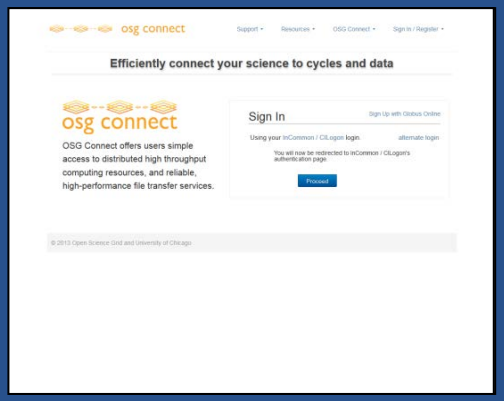


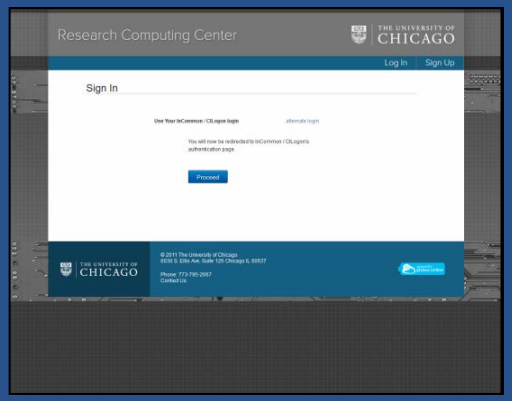
CISL Research Data Archive

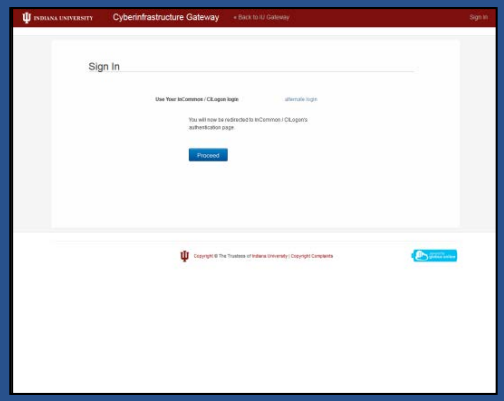
Managed by NCAR's Data Support Section
Data for Atmospheric and Geosciences Research

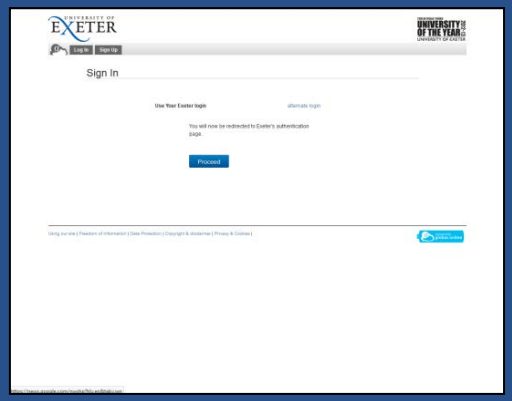
Branded sites

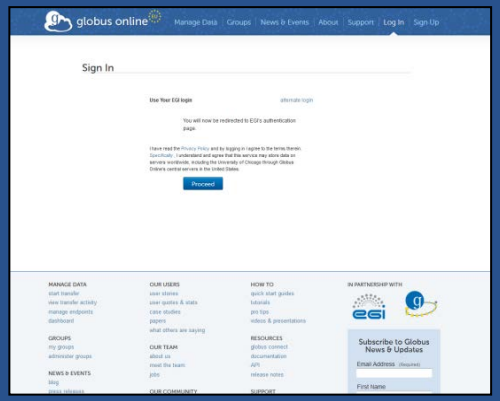


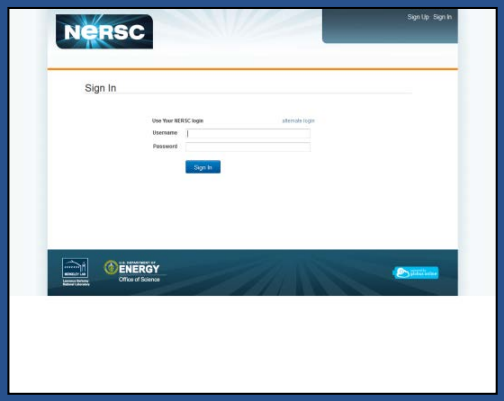












Thank you!