

Globus: Simplifying Research Data Management via Software-as-a-Service



Vas Vasiliadis

University of Chicago, Argonne National Laboratory

vas@uchicago.edu

4th Workshop on Promoting Open Science Data, March 1, 2017





My background

- 2010 – Present
 - Computation Institute, University of Chicago and Argonne National Laboratory
 - Masters Program in Computer Science: Lecturer in Cloud Computing and Product Management
 - Globus : Chief Customer Officer
- Prior Experience
 - Innovation consulting
 - Large scale systems integration
 - Real-time data distribution systems development
- Education
 - MBA, University of Michigan, Ann Arbor, MI, USA
 - BSEE, University of the Witwatersrand, South Africa



Agenda

- **Motivation**
- **Globus Software-as-a-Service**
- **Globus Platform-as-a-Service**
- **Building the research app ecosystem**
- **A word on sustainability**



Research data management

Index?



How do we...
...move?
...share?
...discover?
...reproduce?



Bridge to national cyberinfrastructure

Move datasets to supercomputer,
national facility



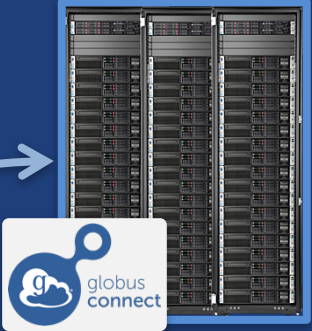
Move results to campus (...)



Bridge to instruments



- Clear staging store
- Data cleansing
- Pre-processing
- ...



Analysis store

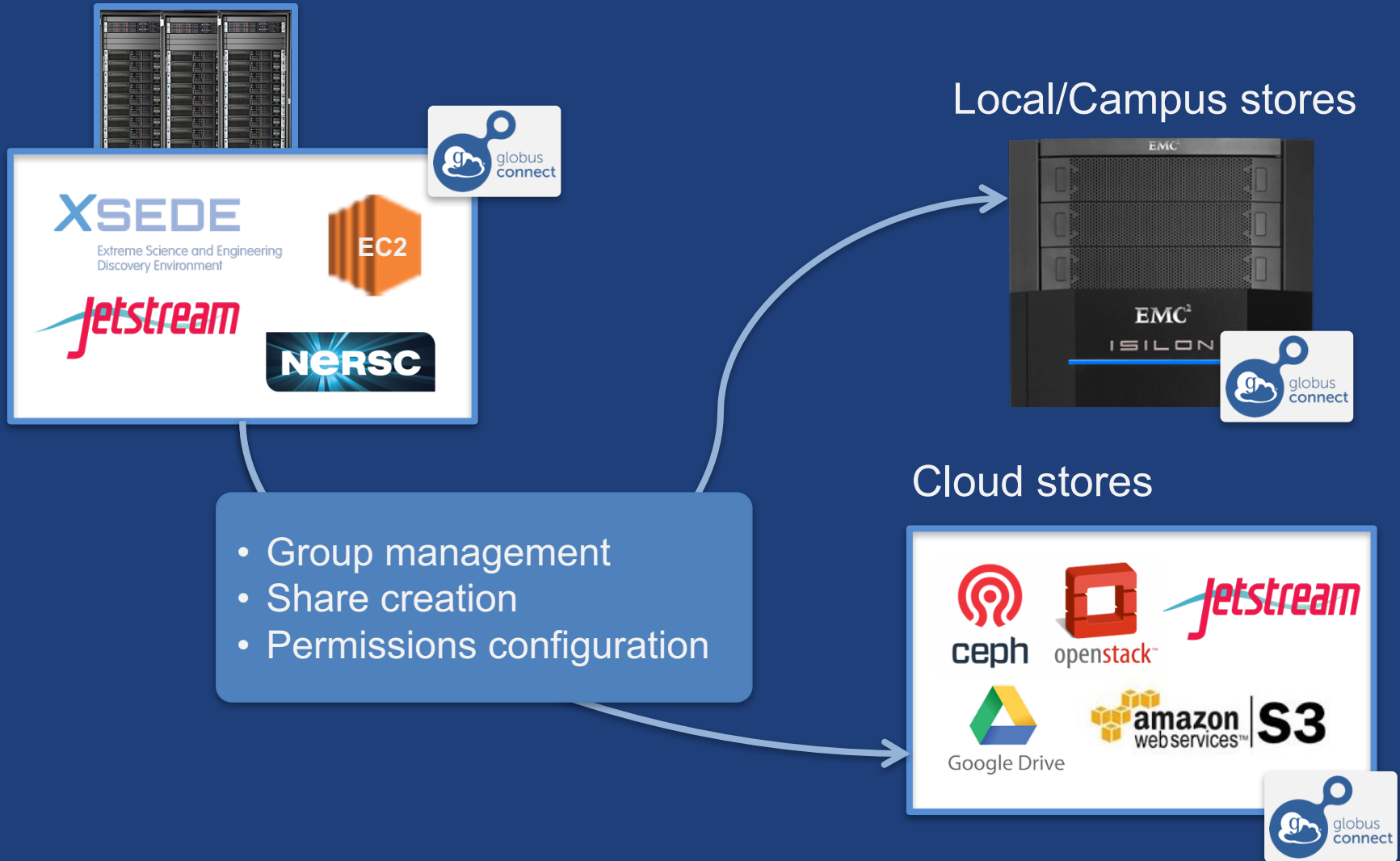
Raw Source Data



High durability,
low cost store

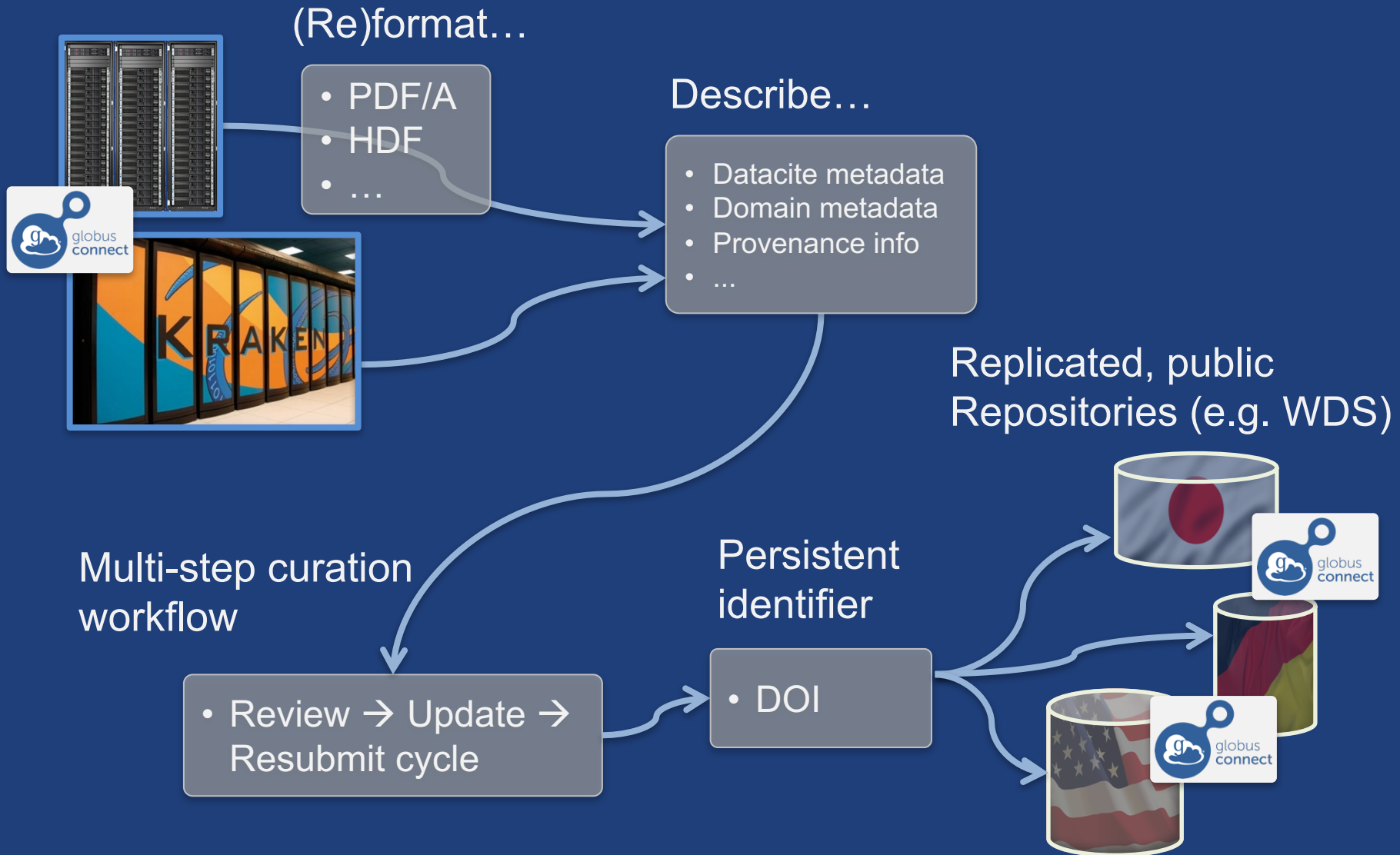


Bridge to cloud resources





Bridge to community/public





Globus enables...

Collaboration

...within and beyond campus
boundaries





Globus delivers...

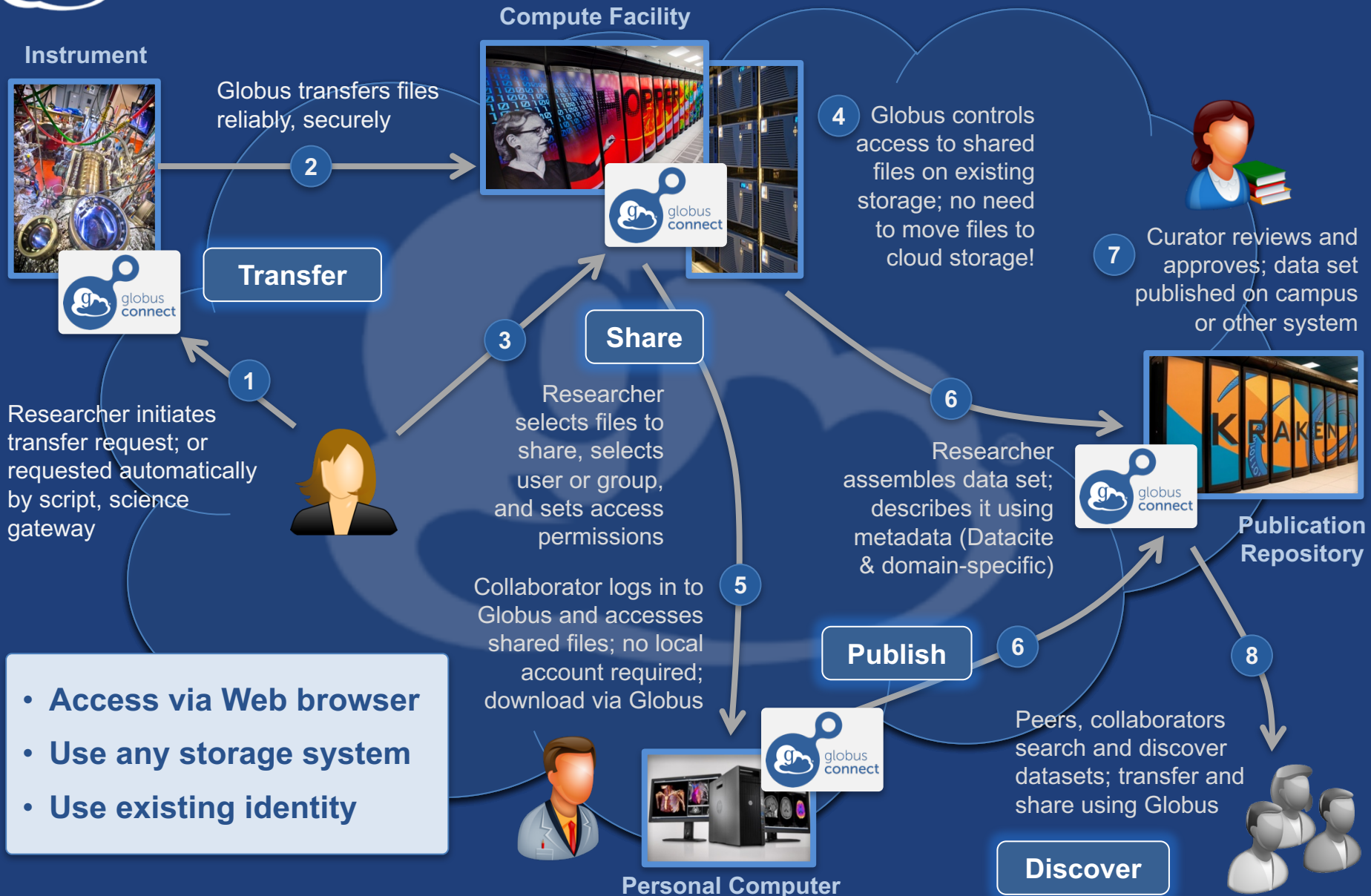
Big data transfer, sharing,
publication, and discovery...

...directly from your own
storage systems...

...via software-as-a-service



Globus and the research data lifecycle





Why use Globus?

- **Ease of use: consistent UI across systems**
- **“Fire-and-forget” file transfer**
- **No/low-overhead external collaboration**
- **Secure access w/multi-tier security model**
- **Maximized WAN throughput**
- **Rapid deployment w/standard packages**
- **Highly automatable: CLI, RESTful API**



Focus on User Experience



...for your photos

Google



...for your office docs

NETFLIX

...for your entertainment

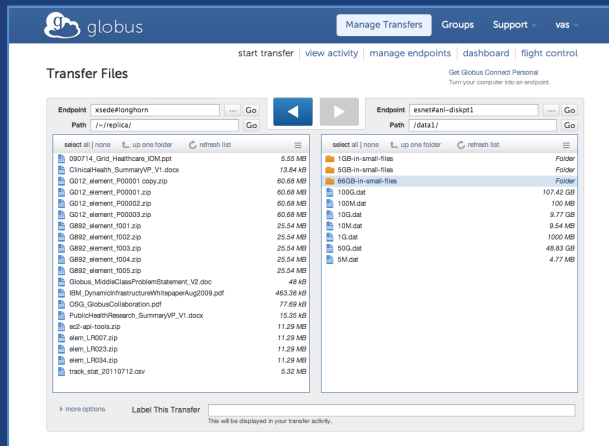


globus

...for your research data



Use(r)-appropriate interfaces



Web

CLI



Globus service

```
laptop:~ ssh vas@cli.globusonline.org
$ Welcome to globusonline.org, vas. Type 'help' for help.
$ endpoint-modify vas#ebs --organization="University of Chicago"
$
```

```
GET /endpoint/go%23ep1
PUT /endpoint/vas#my_endpt
200 OK
X-Transfer-API-Version: 0.10
Content-Type: application/json
...
```

Rest API

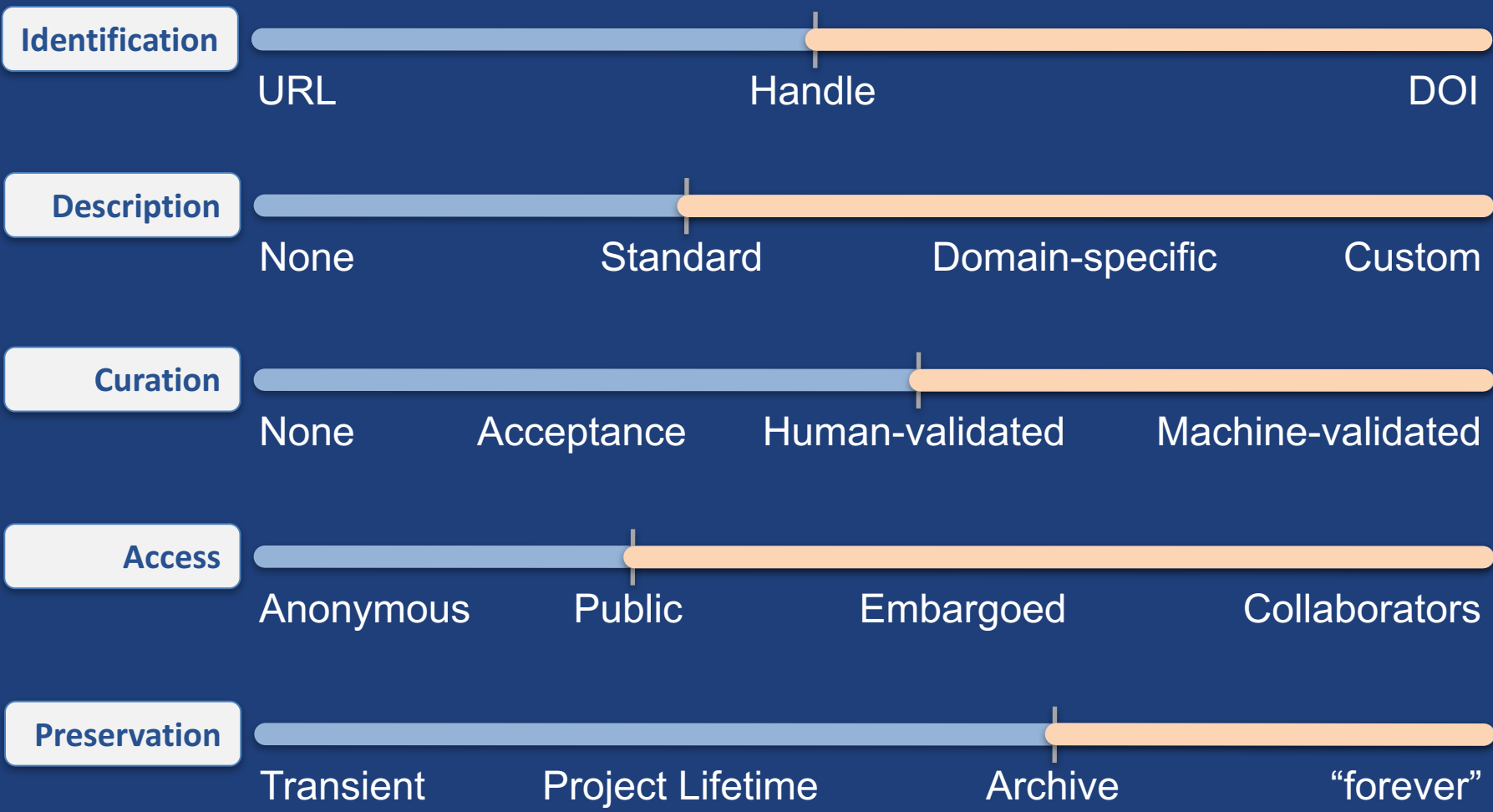


Demonstration

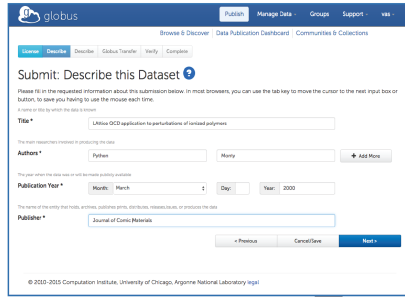
- Identity Federation
- Transfer and Sharing



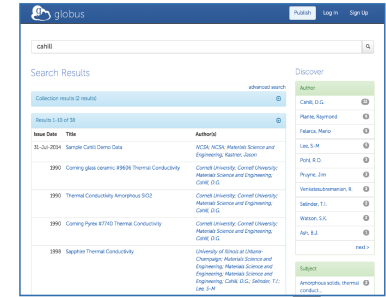
Globus data publication framework



Publish



Discover



Globus Authentication

Globus Data Publication

Users and Groups

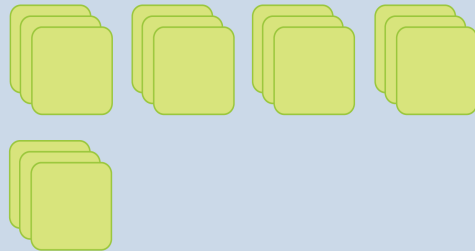
Persistent Identifiers

Globus Identity

Handle

DOI

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

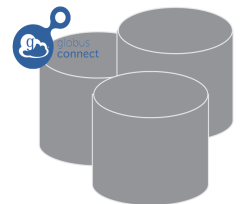
Metadata Catalog

UChicago

Argonne

NCSA

Amazon S3





Demonstration

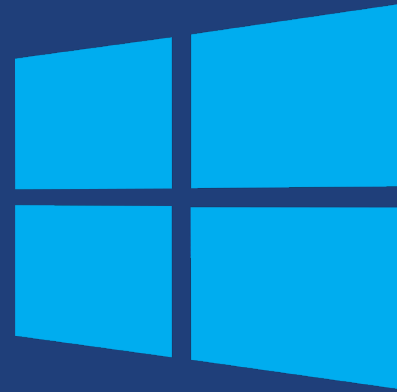
- Data Publication



Globus Connect...
Makes your storage
system a Globus
endpoint



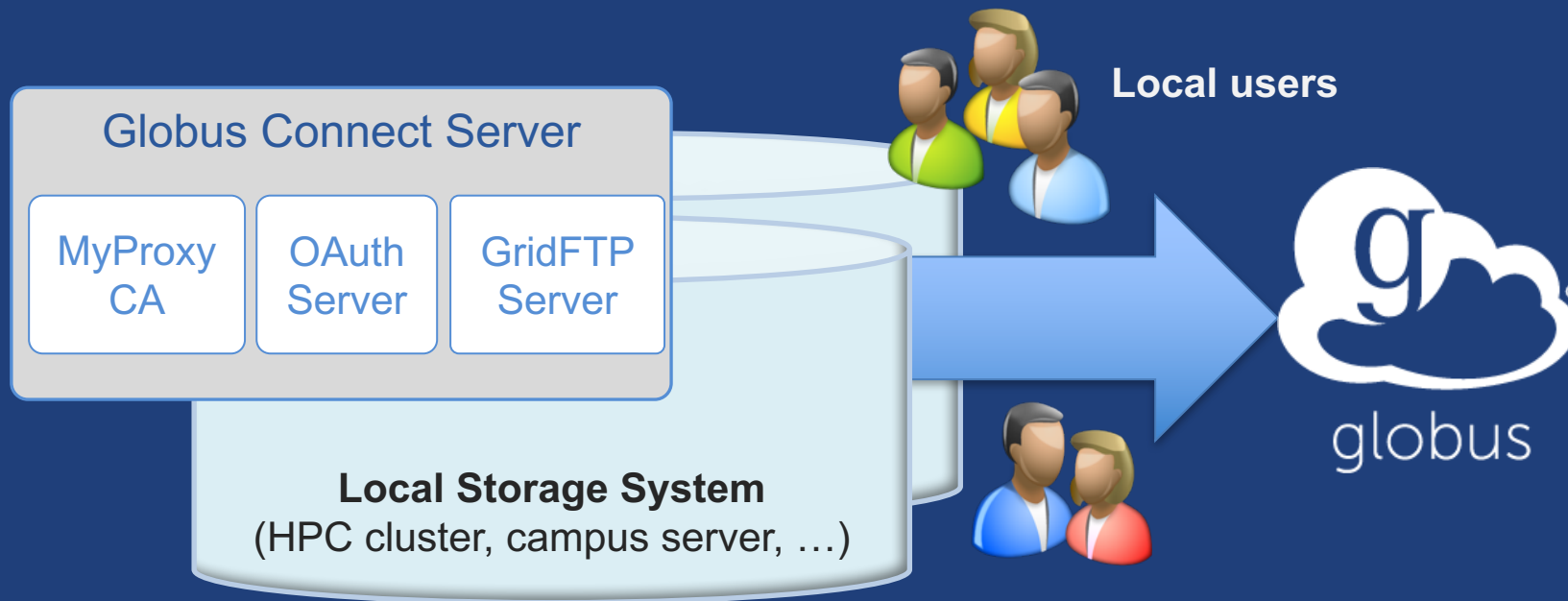
Globus Connect Personal



- **Installers do not require admin access**
- **Zero configuration; auto updating**
- **Handles NATs**



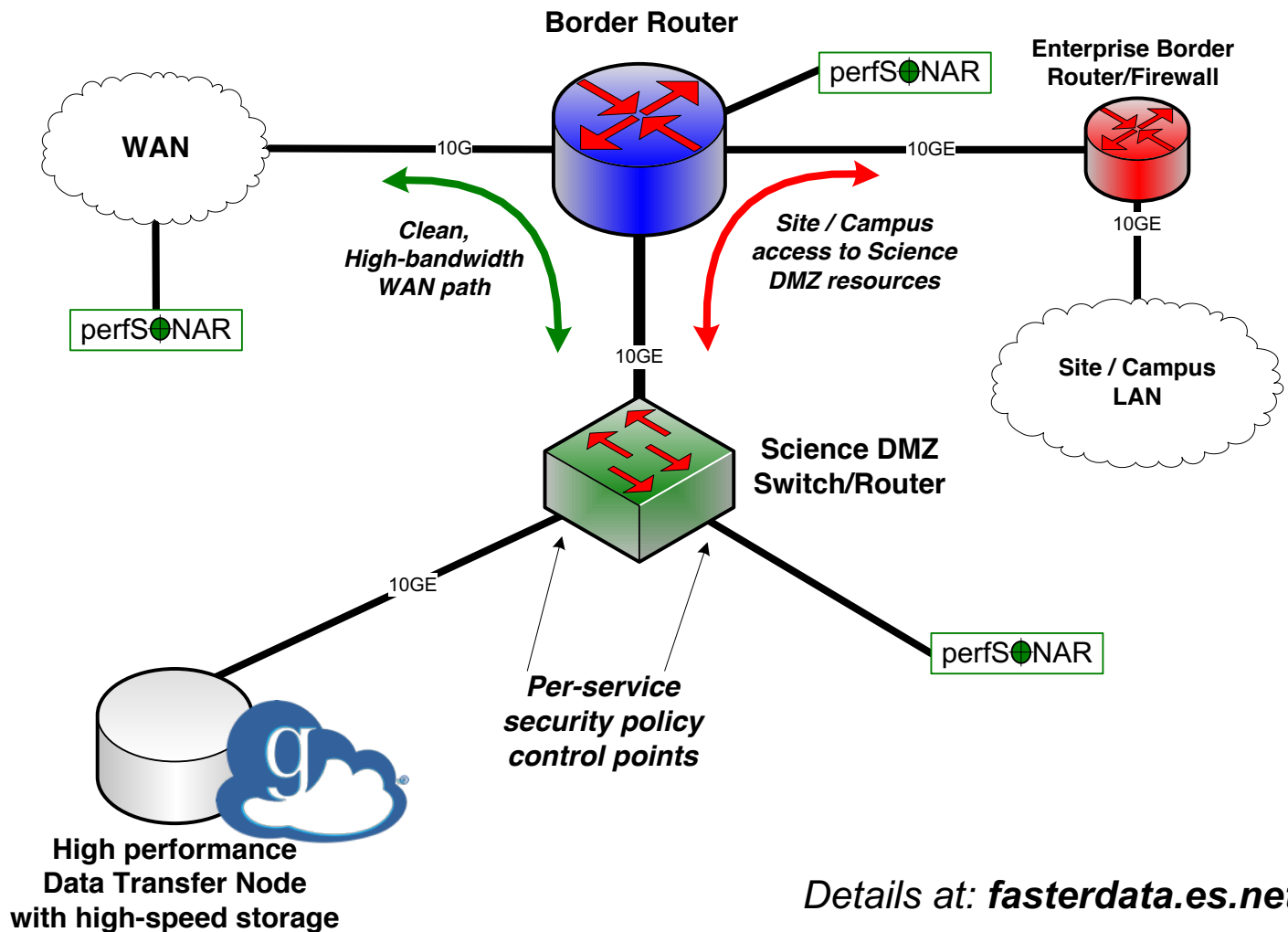
Globus Connect Server



- Enable access for all users with local accounts
- Create endpoint on practically any filesystem
- Native packages: RPMs and DEBs



Best-practice deployment



Details at: fasterdata.es.net



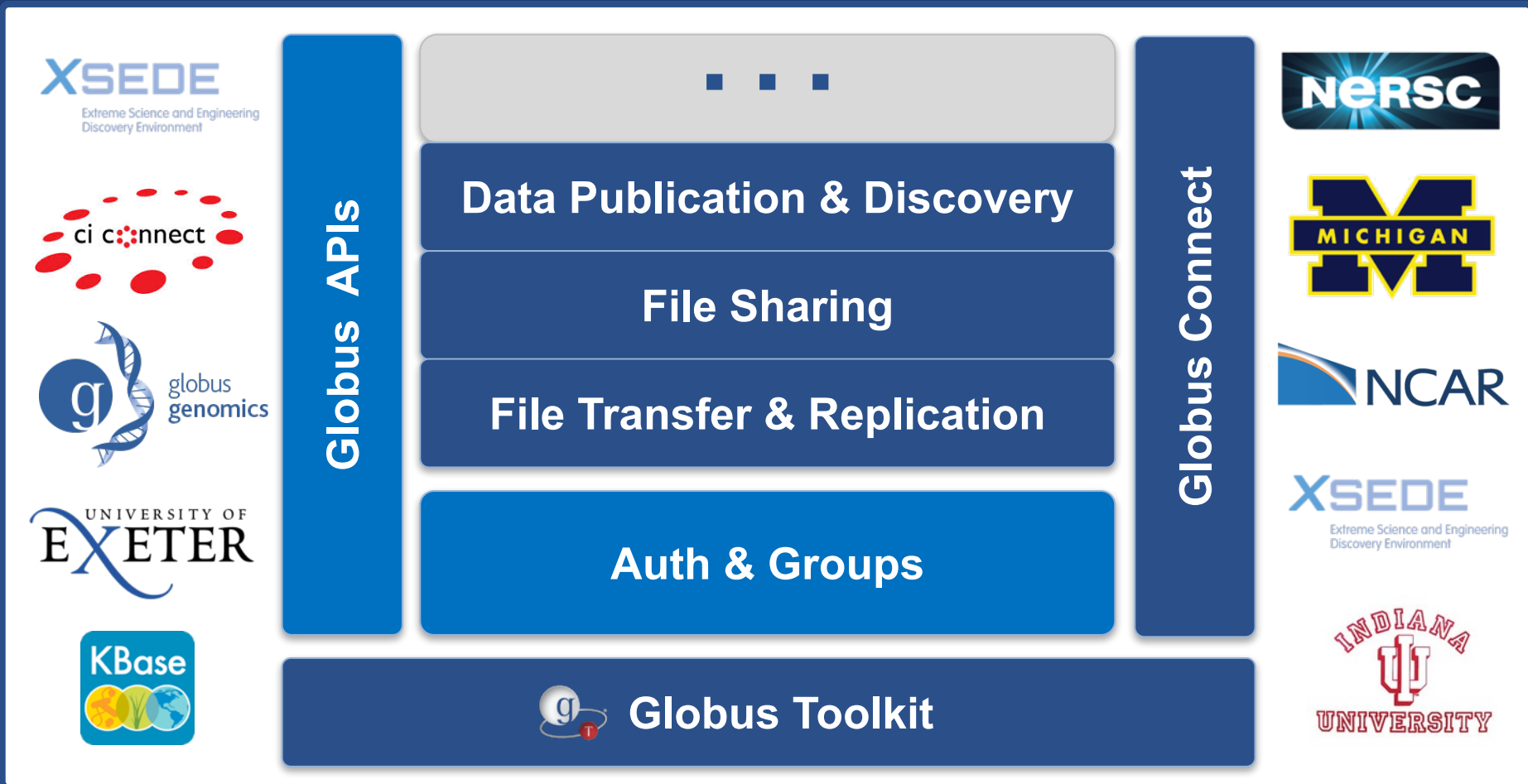
Storage connectors

- **Standard storage connectors (Posix)**
 - Linux, Windows, MacOS
 - Lustre, GPFS, OrangeFS, etc.
- **Premium storage connectors**
 - HPSS
 - HDFS
 - S3
 - Ceph RadosGW (S3 API)
 - Spectra Logic BlackPearl
 - Google Drive (coming soon)





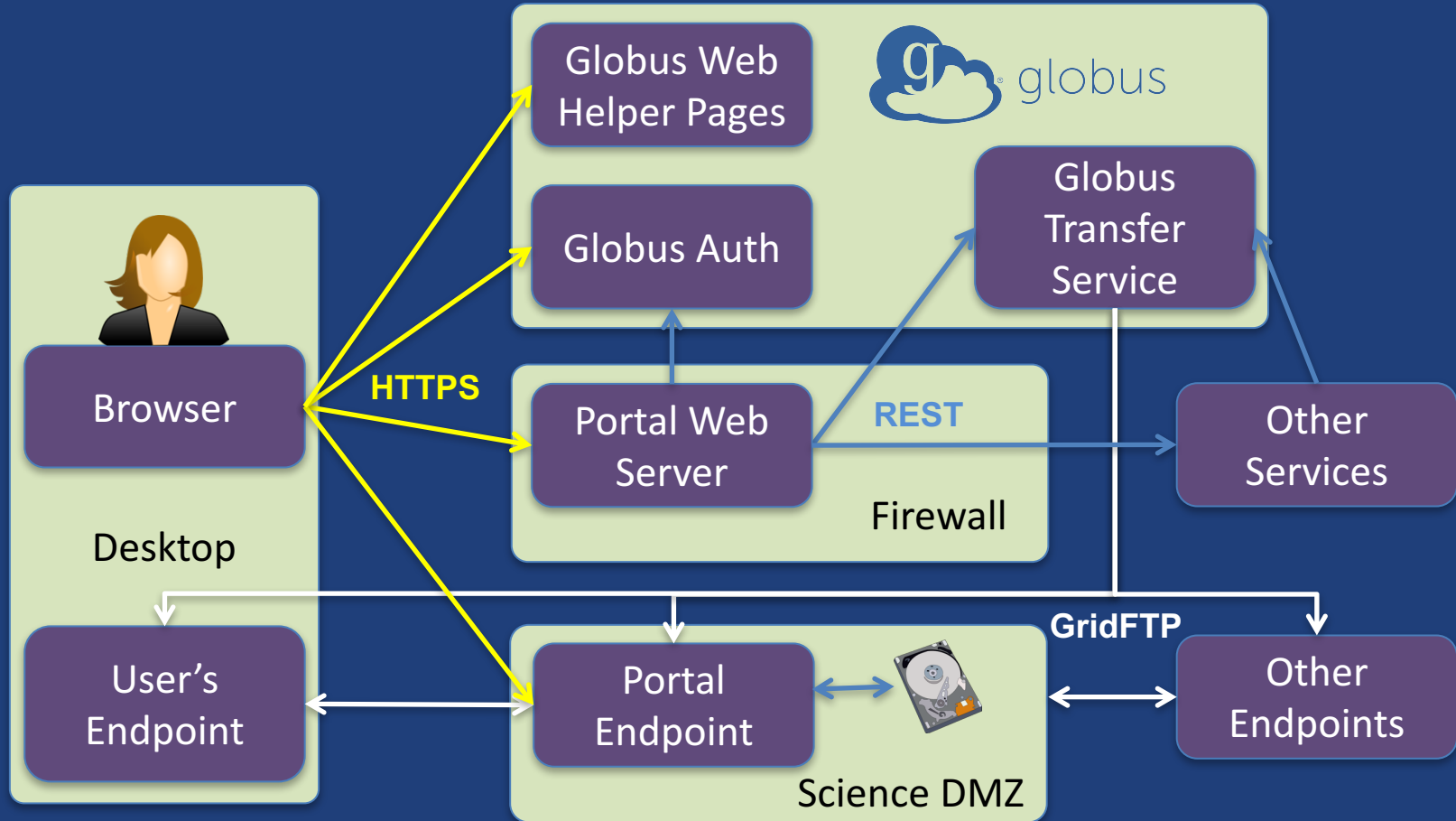
Globus Platform-as-a-Service



API (and other) documentation: docs.globus.org



Building research apps with the Globus Platform-as-a-Service



- **Move portal storage into Science DMZ, with Globus endpoint**
- **Leave Portal Web server behind firewall**
- **Globus handles the security and data heavy lifting**



Data App: NCAR RDA

UCAR NCAR Closures/Emergencies Locations/Directions Find Pe

Hello [twacke@uchicago.edu](#) [dashboard](#) [sign out](#)

NCAR UCAR Research Data Archive Computational & Information Systems Lab *weather • data • climate*

Go to Dataset:

Home Find Data Ancillary Services About/Contact Data Citation Web Services For Staff

NCEP Climate Forecast System Version 2 (CFSv2) Monthly Products
ds094.2

For assistance, contact [Bob Dattore \(303-497-1825\)](#).

Description Data Access

Mouse over the table headings for detailed descriptions

Data Description		Data File Downloads		Customizable Data Requests	Other Access Methods	NCAR-Only Access	
		Web Server Holdings	Globus Transfer Service (GridFTP)	Subsetting	THREDDS Data Server	Central File System (GLADE) Holdings	Tape Archive (HPSS) Holdings
Union of Available Products		Web File Listing	Request Globus Invitation	Get a Subset	TDS Access	GLADE File Listing	HPSS File Listing
P R O D	Diurnal monthly means	Web File Listing		Get a Subset		GLADE File Listing	HPSS File Listing
	Regular monthly means	Web File Listing		Get a Subset		GLADE File Listing	HPSS File Listing



Analysis App: Wellcome Sanger

Sanger Imputation Service **Beta**

Home

About

Instructions ▾

Resources

Status

Sanger Imputation Service

This is a free genotype **imputation** and **phasing** service provided by the [Wellcome Trust Sanger Institute](#). You can upload GWAS data in VCF or 23andMe format and receive imputed and phased genomes back. Click [here](#) to learn more and [follow us on Twitter](#).

Before you start

Be sure to [read through the instructions](#).

You will need to set up a free account with [Globus](#) and have [Globus Connect](#) running at your institute or on your computer to transfer files to and from the service.

Ready to start?

If you are ready to upload your data, please fill in the details below to **register an imputation and/or phasing job**. If you need more information, see the [about](#) page.

What is this [?](#)

➔ Next

News

[@sangerimpute](#)

11/05/2016

Thanks to [EAGLE](#), we can now return **phased data**. The HRC panel has been updated to r1.1 to fix a [known issue](#). See [ChangeLog](#) for more details.

15/02/2016

Globus API changed, please see [updated instructions](#).

17/12/2015

New status page and reworked internals. See [ChangeLog](#).

09/11/2015

Pipeline updated to add some features requested by users. See [ChangeLog](#).

[See older news...](#)



Globus PaaS: National Resource Access

XSEDE
Extreme Science and Engineering
Discovery Environment

globus Account ▾

Jetstream Web App would like to:

✓ Access all Jetstream resources

By clicking "Allow", you allow Jetstream Web information and services. You can rescind this

Allow

Deny

globus Globus Account Log In

compute | **calcul**
canada | canada

Compute Canada has partnered with Globus to offer this high performance file transfer service.

Calcul Canada s'est associé à Globus pour vous offrir ce service de transfert de fichier à haute performance.

Log in to use Compute Canada Globus Web App

Use your existing organizational login
e.g. university, national lab, facility, project, Google or [Globus ID](#)
(Your Globus username and password used prior to February 13, 2016 is now Globus ID)


WestGrid ▾

Continue

Didn't find your organization? Then use Globus ID to [sign up](#).



Globus PaaS: Identity Management



KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase


[Home](#) [About](#) [News](#) [Developer Zone](#) [KBase Labs](#) [Contact Us](#)

The new **Systems Biology Knowledgebase (KBase)** is a collaborative effort designed to accelerate our understanding of microbes, microbial communities, and plants. It will be a community-driven, extensible and scalable open-source software framework and application system. KBase will offer free and open access to data, models and simulations, enabling scientists and researchers to build new knowledge and share their findings.

[Collaborate with us](#) [Get Started](#) [Develop with us](#)

What can KBase do?

- ✔ Combine heterogenous data types
- ✔ Offer standardized access to bioinformatic and modeling analyses
- ✔ Use evidence-supported annotations of genome structure and genetic function
- ✔ Discover new associations and network structures in community and molecular networks
- ✔ Map genotype to complex organismal traits
- ✔ Design and refine experiments using models of metabolism, regulation and community function
- ✔ Enable sharing of data, hypotheses, and newly-generated knowledge



Latest News

[KBase at International Plant and Animal Genome XXI](#)
Posted by salazar Jan 09, 2013

[KBase Team at Argonne for November Build](#)
Posted by nharris Nov 30, 2012

[November Build at Argonne](#)
Posted by salazar Nov 23, 2012

[view news](#)

Upcoming Events

2013-01-12
[International Plant and Animal Genome XXI \(PAG 2013\)](#)

2013-02-18
[BERAC Presentations](#)

2013-02-24
[DOE/NIFA Plant Feedstocks Genomics for Bioenergy](#)

2013-02-25
[Proposed: Genomic Science Contractors-Grantees Meeting](#)



Globus PaaS developer resources

globus-sdk-python 0.2.5 documentation

Table Of Contents

- Globus SDK for Python (Beta)
- Installation
- Basic Usage
- API Documentation
- License

Next topic: High Level API

This Page: Show Source

Quick search: Go

Globus SDK for Python (Beta)

This SDK provides a convenient Pythonic interface to REST APIs is available at <https://docs.globus.org>.

Two interfaces are provided - a low level interface, sup methods for common API resources.

Source code is available at <https://github.com/globus/globus-sdk-python>

Python SDK

Installation

The Globus SDK requires Python 2.6 or higher.

The simplest way to install the Globus SDK installations:

```
pip install globus-sdk
```

This will install the Globus SDK and its dependencies. Bleeding edge versions of the Globus SDK can be installed with:

```
git checkout https://github.com/globus/globus-sdk-python
cd globus-sdk-python
python setup.py install
```

Basic Usage

Modern Research Data Portal

Modern Research Data Portal

It's how research data management is done!

LOGIN | SIGN UP

Globus Transfer API

API reference for transfer and sharing functions

Sample Application

Requirements

- You need to be in the tutorial environment
- Installed Globus Python SDK

Jupyter Notebook

```
In [15]: from __future__ import print_function # for python 2
tutorial_endpoint_1 = "ddb59aef-6d04-11e5-ba46-22000b92c6ec" # endpoint "Globus Online"
tutorial_endpoint_2 = "ddb59af0-6d04-11e5-ba46-22000b92c6ec" # endpoint "Globus Connect Online"
tutorial_users_group = "50b6a29c-63ac-11e4-8062-22000ab68755" # group "Tutorial Users"
```

Configuration

First you will need to configure the client with an OAuth2 access token. For the purpose of this tutorial, you can use a token from the Globus website. Click the "Jupyter Notebook" option and copy the resulting text below, or click on "Globus CLI" and

```
In [16]: transfer_token = None # if None, tries to get token from ~/.globus.cfg file
```

docs.globus.org/api

github.com/globus



Demonstration

- Command Line Interface
- Globus PaaS



Globus Milestones

November 18, 2010
Globus Online
launches

270 users



April 26, 2011
Globus Connect solves
the "last mile" problem

1,000 users

June 4, 2011



January 23, 2012
NCSA selects Globus
Online for Blue Waters



June 20, 2012
Globus Online wins
R&D 100 Award



January 22, 2013
Globus Online accepted as
XSEDE production service

10,000 users

June 13, 2013



100PB moved



November 18, 2013
Globus file sharing
announced at SC13

June 26, 2015



March 12, 2015
Management console provides admins with
dynamic real-time view into transfer activity

February 2014
NERSC and NCAR become first
paying Globus subscribers



200PB moved

50,000 users

February 11, 2016
Globus Auth débuts, providing immediate
access to 100,000's of users with existing credentials

October 14, 2016

December 3, 2016



Our vision for the future



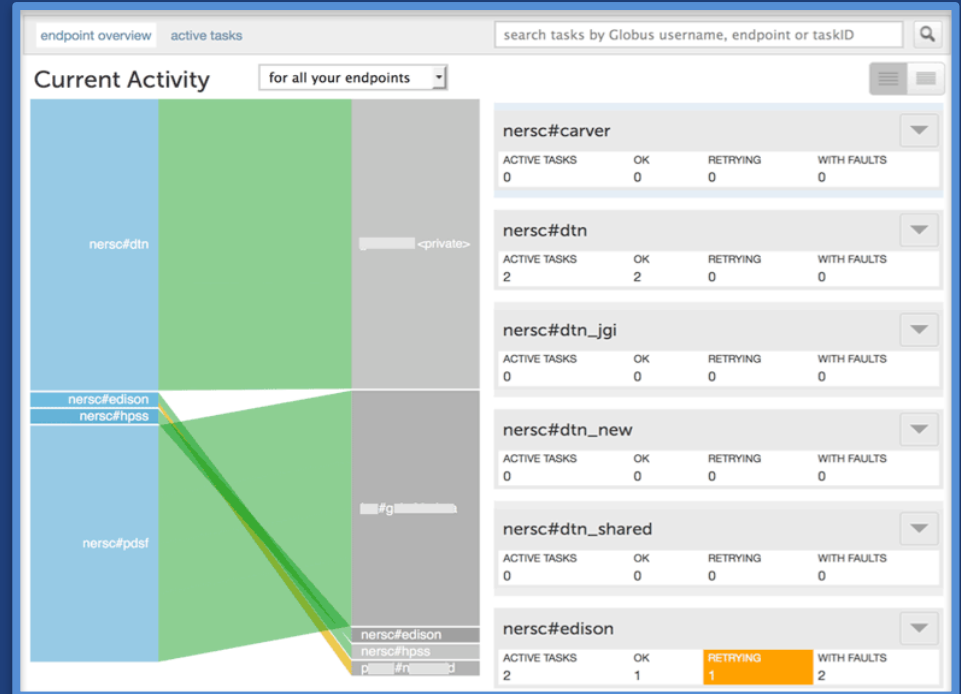




Globus Subscriptions

- **Standard Subscription**

- Shared endpoints
- Data publication
- HTTPS support*
- Management console
- Usage reporting
- Priority support
- Application integration



- **Branded Web Site**

- **Premium Storage Connectors**

- Amazon S3, Ceph, HPSS, Spectra, Google Drive*, ...

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

* When available



Demonstration

- Management Console



Thank you to our sponsors



U.S. DEPARTMENT OF
ENERGY



THE UNIVERSITY OF
CHICAGO

NIST

**National Institute of
Standards and Technology**
U.S. Department of Commerce



Argonne
NATIONAL LABORATORY



powered by
amazon
web services



Thank you to our users...

5

major services

245 PB

transferred

40 Bn

files processed

53,000

registered users

13

national labs
use Globus

10,000

active endpoints

10,000

active users/year

99.5%

uptime

60+

institutional
subscribers

1 PB

largest single
transfer to date

3 months

longest
continuously
managed transfer

130

federated
campus identities



...and thank YOU, our subscribers!



NEW YORK UNIVERSITY



Berkeley
UNIVERSITY OF CALIFORNIA



JOHNS HOPKINS
UNIVERSITY

CORNELL
UNIVERSITY



UF | UNIVERSITY of
FLORIDA

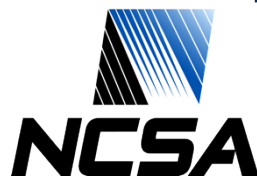


MICHIGAN STATE
UNIVERSITY

Yale



THE UNIVERSITY OF
CHICAGO



syngenta

NIST



VirginiaTech
Invent the Future





Engage with us



to the GlobusWorld Tour!

a series of Globus tutorials and developer workshops across on the success of the [workshop held at GlobusWorld 2016](#). are made possible by the various hosting institutions that the meeting space and other financial support.

rkshops are currently scheduled:

4 2016 - LRNI, Berkeley, CA

Why Attend?

- Learn how the Globus platform simplifies development of web applications for researchers
- Experiment with new Globus services and APIs
- Exchange ideas with peers on ways technologies knowledge of Globus features





Join the Globus community

- Access the service: globus.org/login
- Create endpoints: globus.org/app/endpoints
- Learn more: docs.globus.org
- Engage: www.globus.org/mailing-lists
- Need help? support@globus.org
- Subscribe: globus.org/subscriptions
- Follow us: [@globusonline](https://twitter.com/globusonline)



Thank You.

vas@uchicago.edu