

Globus: The Platform for Research Data Management

Vas Vasiliadis — vas@uchicago.edu





Globus is ...

a non-profit service
developed and operated by



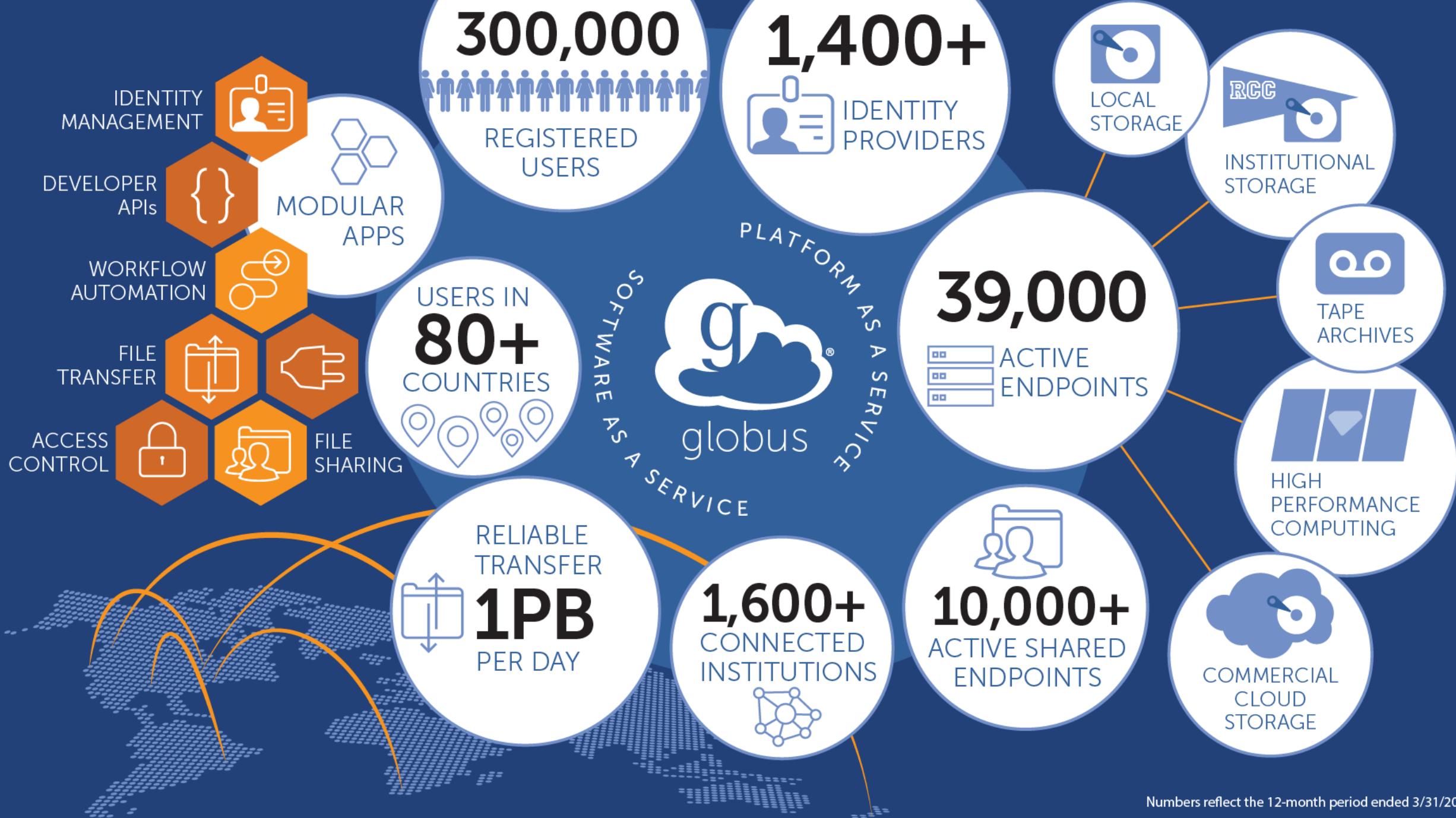
THE UNIVERSITY OF
CHICAGO



Our team comprises ...
professional software developers
and business operators with
extensive experience in industry
and academia



Our mission is to...
increase the efficiency and
effectiveness of researchers
engaged in data-driven
science and scholarship
through *sustainable* software





Development is funded by...



U. S. DEPARTMENT OF
ENERGY



THE UNIVERSITY OF
CHICAGO



powered by
amazon
web services



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

Argonne
NATIONAL LABORATORY



Operations are funded by subscribers



Yale

SIMONS FOUNDATION



RIT

Duke
UNIVERSITY



Dartmouth



THE ROCKEFELLER UNIVERSITY



University of
Pittsburgh



Northeastern
University



Washington
University in St. Louis



UNIVERSITY OF
BIRMINGHAM



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

CORNELL
UNIVERSITY



NCSA

MICHIGAN STATE
UNIVERSITY



Tufts
UNIVERSITY

Washington
University in St. Louis



Sustainability enabled by freemium model

- Basic file transfer free to non-profit research institutions
- Annual subscriptions provide advanced capabilities to researchers and visibility/control to administrators



80% US Institutions



20% International Institutions



13% Federal Agencies/
National Facilities



46% R1 Universities



13% Independent
Research Institutes



4% Hospital/Health
Care Systems



20% Other
Universities



4% Commercial



Globus is known for reliable, secure file transfer

Activity List ✓ RDA to ALCF noverify
transfer completed

Overview Event Log

Task Label	RDA to ALCF noverify
Source	NCAR RDA Dataset Archive ⓘ
Destination	DME PerfTest - Argonne ⓘ
Task ID	20ebf766-a46d-11eb-8a95-d70d98a40c8d
Owner	Vas Vasiliadis (vas@globusid.org)
Condition	SUCCEEDED
Requested	2021-04-23 02:50 pm
Completed	2021-04-23 02:53 pm
Duration	2 minutes 47 seconds
Transfer Settings	<ul style="list-style-type: none">transfer is not encryptedoverwriting all files on destination

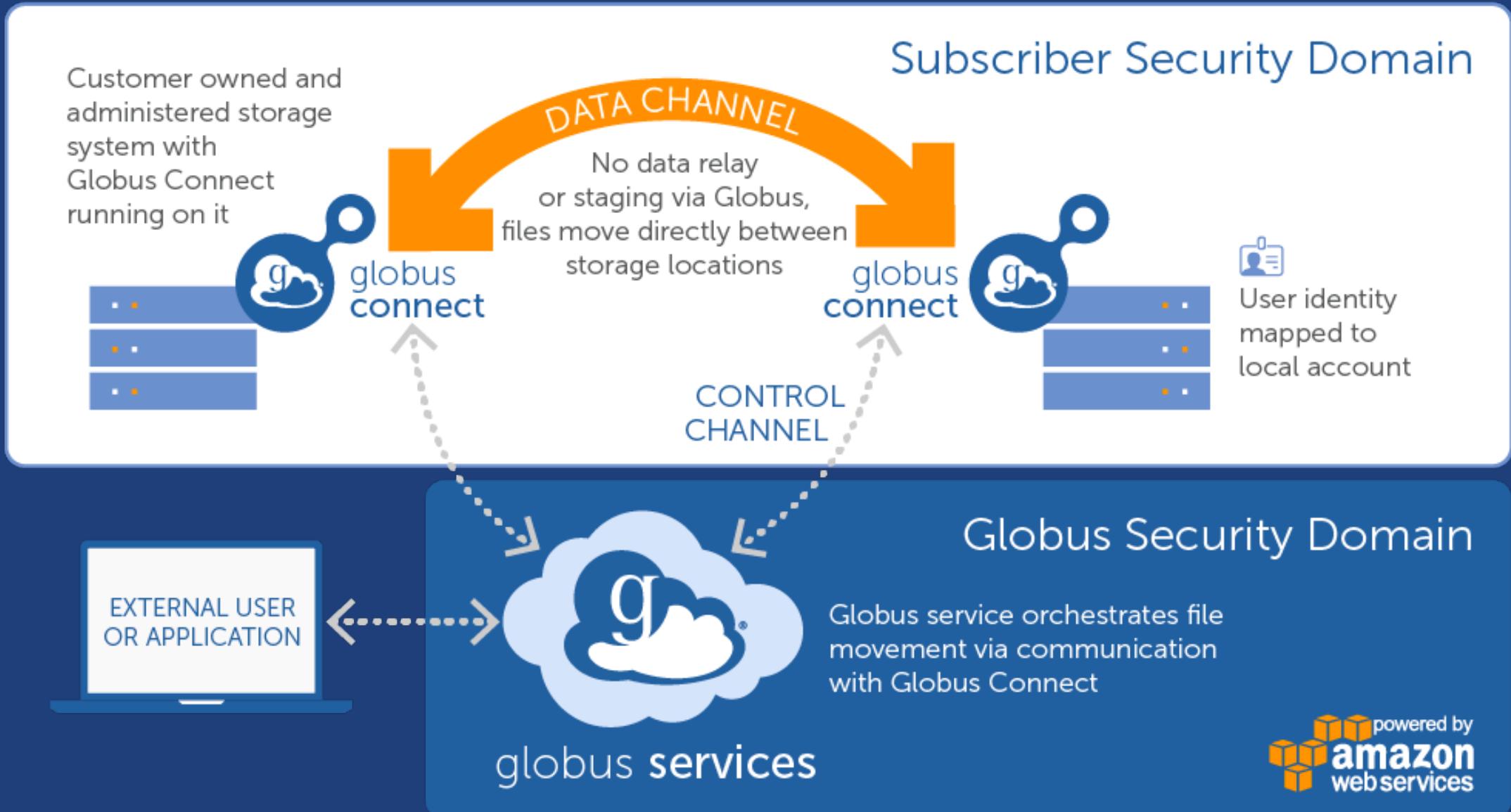
72.8Gbps

6151 Files
2 Directories
1.51 TB Bytes Transferred
9.10 GB/s Effective Speed
0 Skipped files on sync
0 Skipped files on error

[View debug data](#)



Conceptual Architecture: Hybrid SaaS





Globus Connect unifies data access across diverse landscape of storage systems...



Research Computing HPC



Personal Systems



Desktop Workstations



Archives



Instruments



Public Cloud Storage



Currently supported systems



Google Cloud

Microsoft Azure
Blob Storage



IBM Cloud



Google Drive



iRODS®

Quantum.™ ACTIVE SCALE.™



HPSS



wasabi
hot cloud storage

lustre™



We enable a “data appropriate” storage strategy

- **Access frequency
(data temperature)**
- **Access modality(ies)**
 - Ad hoc, via web browser
 - Scripted, via CLI tools
 - Programmatic, via APIs
- **Uniform interface and user experience**



Google
Cloud Storage



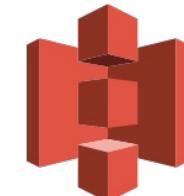
Google Drive



Microsoft Azure
Blob Storage



Microsoft OneDrive



amazon
S3



Uniform interface, consistent user experience

The screenshot displays a uniform user interface for managing multiple cloud storage collections. At the top, there are four collection cards:

- Vas Google Cloud Storage Collection**: Path: /
- Vas Google Drive Collection**: Path: /My Drive/
- Vas Box Collection**: Path: /
- Vas AWS S3 Collection**: Path: /

Each card includes search, refresh, and three-dot menu icons. Below the cards are two main sections, each containing a list of files and folders:

Left Section (Google Cloud Storage and Google Drive):

- globus-archive** (12/31/1969, 5:59 PM)
- uchicago-public** (12/31/1969, 5:59 PM)
- 211207_SciDAS_Panel_Enabling_F_and_A_Va...** (12/7/2021, 9:05 AM, 831.77 KB)
- CC Spring 2022** (6/5/2021, 1:02 PM)
- globus** (10/7/2019, 12:11 AM)
- migration** (1/11/2021, 7:18 PM)
- MPCS 51083-1 Spring 2021** (3/6/2021, 1:40 PM)
- MPCS 51083-1 Winter 2022** (1/10/2022, 10:26 PM)
- MPCS 51240 - Product Management** (9/24/2020, 3:56 PM)
- Product Management - Interim Feedback** (1/31/2022, 12:11 AM, 0 B)

Right Section (Box and AWS S3):

- globus-box-archive** (12/17/2020, 1:06 AM)
- uchicago-perftest** (3/24/2021, 9:41 PM)
- vvarchive** (12/17/2020, 10:45 AM)
- globus-vault** (12/31/1969, 5:59 PM)
- mpcs-practicum** (12/31/1969, 5:59 PM)
- personal-vault** (12/31/1969, 5:59 PM)

Both sections feature a toolbar at the top with "Start" buttons, transfer/timer options, and three-dot menus. On the far right, there are additional three-dot menu icons.



Move without (worrying about) limits

- API request rates
- Data volume
- Third-party tools cannot circumvent these...
- ...but you can use Globus to “fire-and-forget”
- → it will (eventually) be done
- File size ...see “Data appropriate” above :-)

A screenshot of a computer screen displaying an error message from Globus. The message is as follows:

3/9/2022, 08:33 PM endpoint too busy

Error (transfer)
Endpoint: Vas Google Drive Collection (d6d62391-fdda-4ba5-ac78-6523f806ea79)
Server: m-422a8b.d8b83.36fe.data.globus.org:443
File: /My%20Drive/migration/uchicago-perftest/cc32-16p32-16/test1185
Command: STOR /My Drive/migration/uchicago-perftest/cc32-16p32-16/test1185
Message: Fatal FTP response

Details: 451-GlobusError: v=1 c=TOO_BUSY\r\n451-GlobusError: v=1 c=INTERNAL_ERROR\r\n451-\r\n451-GD-Method: "PATCH"\r\n451-GD-URI: "https://www.googleapis.com/upload/drive/v3/files/11h9NbppR7qG7YalZDqyWi-8w9rw9gGj"\r\n451-GD-Response-Code: "403"\r\n451-GD-Response: "User rate limit exceeded."\r\n451 End.\r\n



We simplify secure sharing with collaborators...



UCSF



On-premises
stores



XSEDE

Extreme Science and Engineering
Discovery Environment



NCSA



NIH

compute | calcul
canada | canada

Project repositories,
replication stores



Public repositories

g

Public / private cloud stores

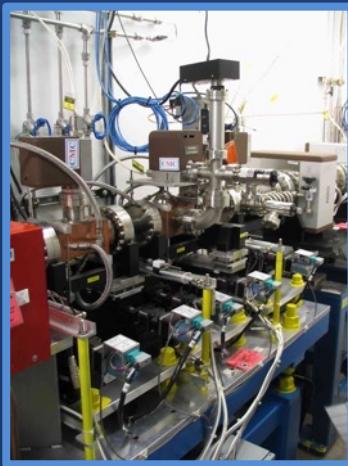




...help researchers manage instrument data...



Next-Gen Sequencer



Advanced Light Source



MRI



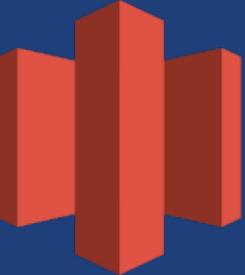
Cryo-EM



Light Sheet Microscope



Analysis
store



High-durability,
low-cost store



Remote visualization

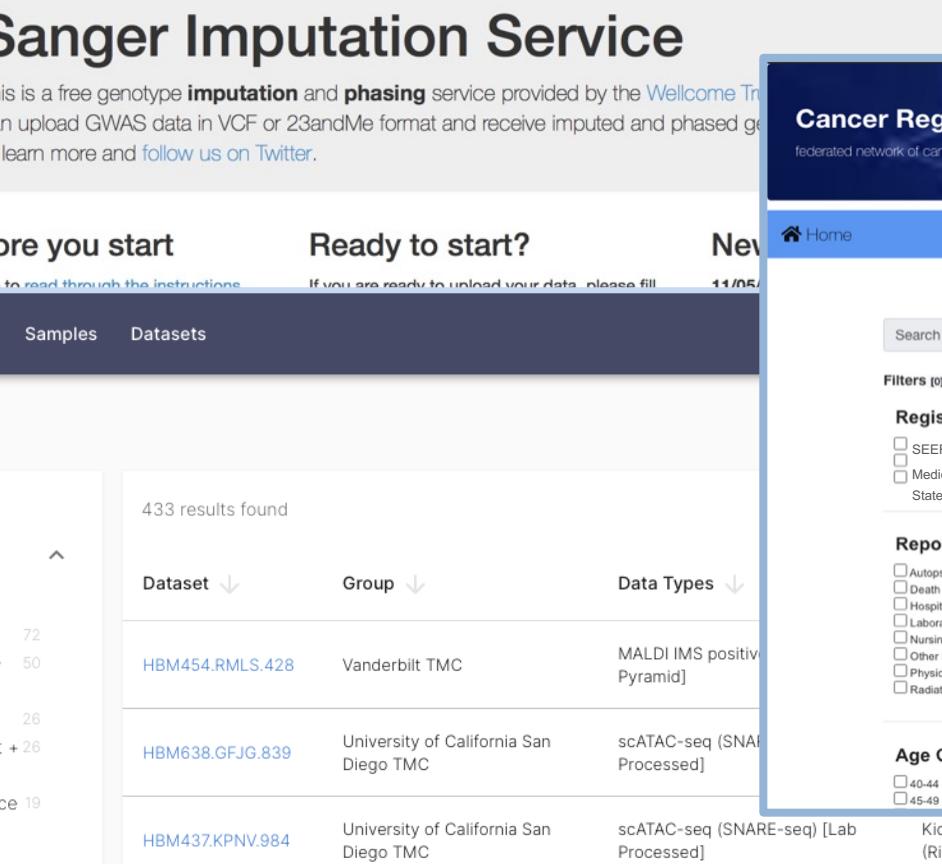


Personal system





...and build data-centric applications



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 genotypes. Please refer to our [Instructions](#) to learn more and [follow us on Twitter](#).



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

Ready to start?
If you are ready to upload your data, please fill out the [form](#).

New!

HuBMAP Donors Samples Datasets

Datasets

Dataset Metadata

433 results found

Dataset	Group	Data Types
HBM454.RMLS.428	Vanderbilt TMC	MALDI IMS positive [Pyramid]
HBM638.GFJG.839	University of California San Diego TMC	scATAC-seq (SNARe-seq) [Processed]
HBM437.KPNV.984	University of California San Diego TMC	scATAC-seq (SNARe-seq) [Lab Processed]
HBM595.QDQQ.996	University of California San Diego TMC	scrRNA-seq (SNARe-seq) [Lab Processed]

Cancer Registry Records for Research (CR3)
federated network of cancer registry data

Home Make a Request Sign Out braumann@uchicago.edu

Search All Advanced

Filters (0) [Clear](#)

Registry

- SEER Registry (173,646)
 Medical Center Registry (108,515)
 State Registry (151,989)

Reporting Source

- Autopsy only (201)
 Death certificate only (3,853)
 Hospital inpatient/outpatient or clinic (388,460)
 Laboratory only (hospital or private) (10,707)
 Nursing/convalescent home/hospice (3,408)
 Other hospital outpatient unit or surge... (1,754)
 Physicians office/private medical pract... (17,023)
 Radiation treatment or medical oncology... (8,744)

Age Group at Diagnosis

- 40-44 (16,291)
 45-49 (25,314)

Diagnosis per Year

Age Group at Diagnosis

Registry

Age Group at Diagnosis

Diagnosis per Year

AJCC Stage/Best CS



Use(r)-appropriate interfaces

Globus service

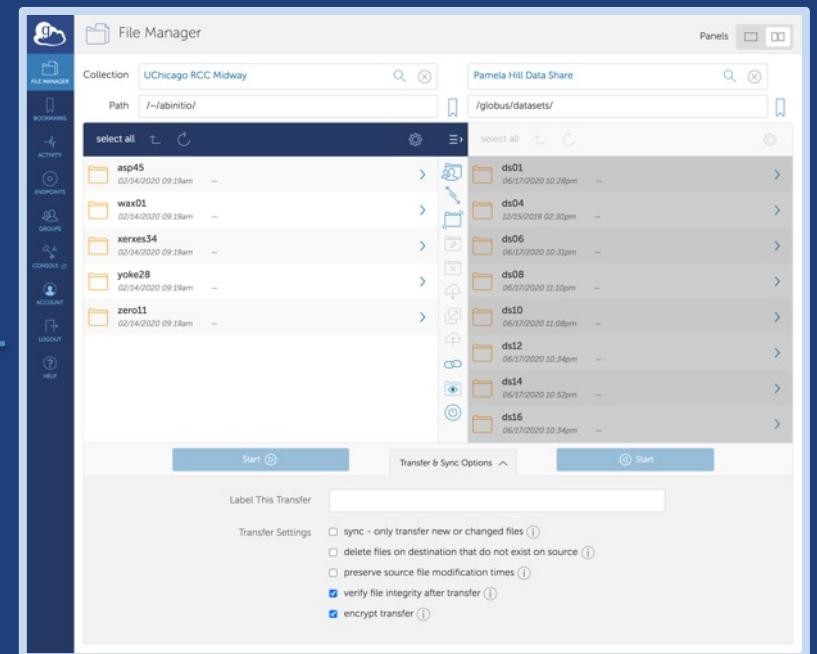


Platform
(RESTful APIs)

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

CLI

```
Usage: globus [OPTIONS] COMMAND [ARGS]...
Options:
-v, --verbose           Control level of output
-h, --help              Show this message and exit.
-F, --format [unix|json|text] Output format for stdout. Defaults to text
--jmespath, --jq TEXT   A JMESPath expression to apply to json
                       output. Takes precedence over any specified
                       '--format' and forces the format to be json
                       processed by this expression
--map-http-status TEXT Map HTTP statuses to any of these exit codes:
                       0,1,50-99. e.g. "404=50,403=51"
Commands:
bookmark      Manage endpoint bookmarks
config        Manage your Globus config file. (Advanced Users)
delete        Submit a delete task (asynchronous)
endpoint     Manage Globus endpoint definitions
get-identities Lookup Globus Auth Identities
list-commands List all CLI Commands
login         Log into Globus to get credentials for the Globus CLI
logout        Logout of the Globus CLI
ls            List endpoint directory contents
mkdir        Make a directory on an endpoint
rename       Rename a file or directory on an endpoint
rm           Delete a single path; wait for it to complete
session      Manage your CLI auth session
task         Manage asynchronous tasks
transfer    Submit a transfer task (asynchronous)
update      Update the Globus CLI to its latest version 2020
version     Show the version and exit
whoami      Show the currently logged-in primary identity.
```



Web



- Custom portals
- Science Gateways
- Unique workflows
- Any research application...





Globus core security features

- **Access Control**
 - Identities provided and managed by institution
 - Institution controls all access policies
 - Globus is identity broker; no access to/storage of user credentials
- **Data remain at institutions, not stored by Globus**
- **Integrity checks of transferred data**
- **High availability and redundancy**
- **Encryption of user files and Globus control data**





Globus High Assurance features



- **Additional authentication assurance**
 - Authenticate with a specific identity within session
 - Reauthenticate after specified time period
- **Session/device isolation**
 - Authentication context is per application, per session
- **Enforces encryption of all user data in transit**
- **Audit logging**



Globus for protected data management

Security controls

- NIST 800-53
- 800-171 Low+



Restricted data handling

- PHI, PII, CUI
- Compliant data sharing

BAA w/UChicago

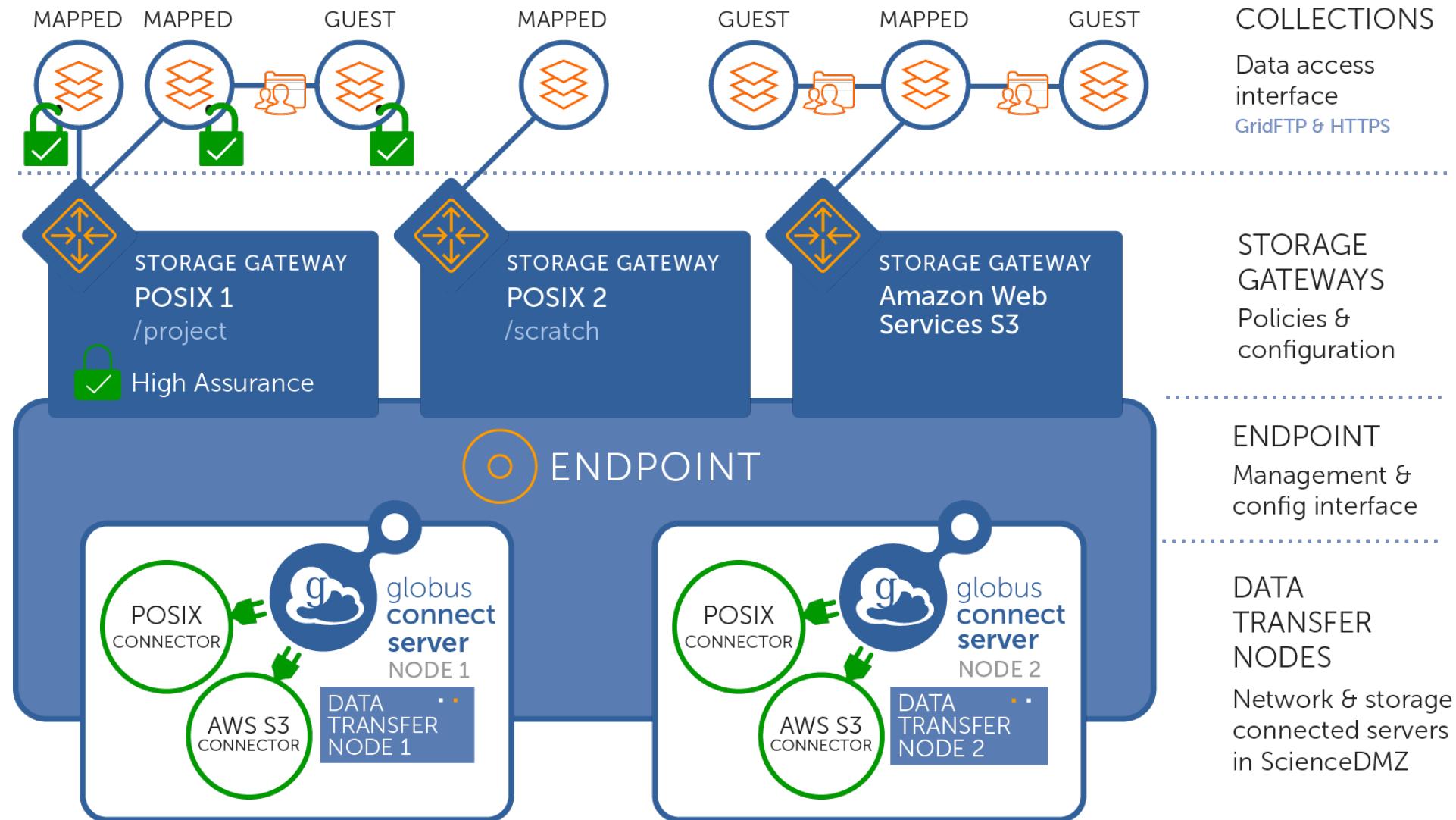
- UChicago BAA with Amazon



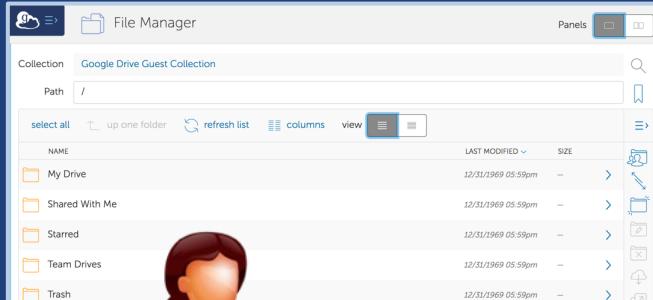
Let's take a look....



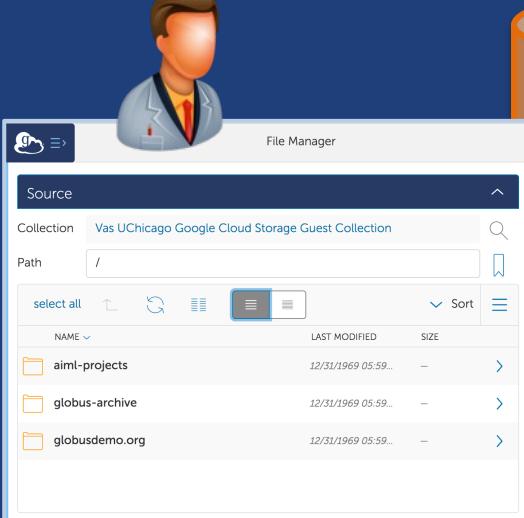
Globus Connect Server architecture



g Cloud Connectors/gateways for cloud storage



Globus Collection



Globus Collection

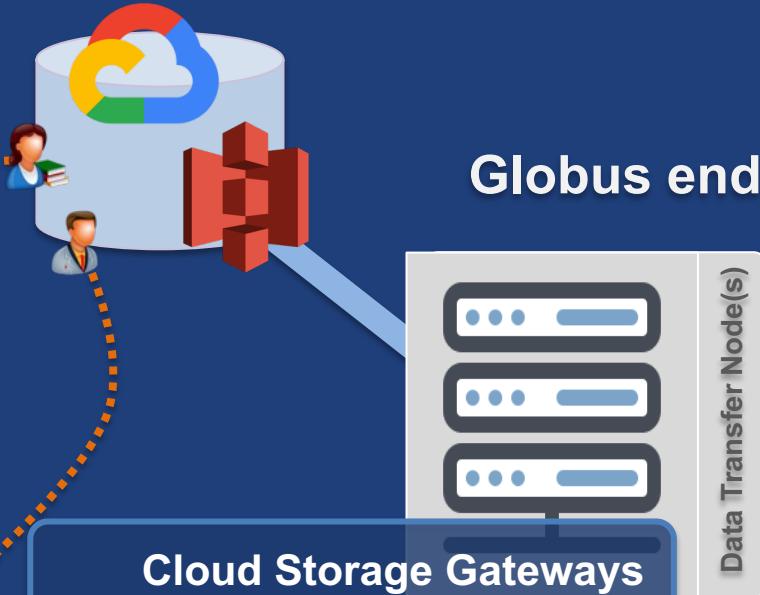
Cloud Storage Gateways
(using Globus connectors)



aws | s3

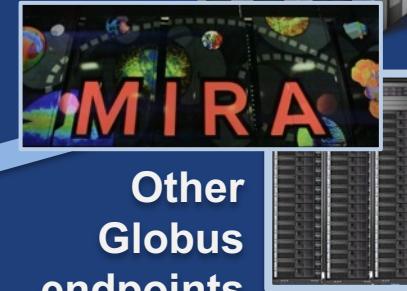


Google Cloud



Data

Globus endpoint



Control





Deployment varies by cloud provider, but has common elements

- **Provider configuration**
 - Register an application
 - Get app credentials
- **Storage gateway creation**
 - Use app credentials
 - Specify access policy
- **Mapped collection creation**
 - User account \leftrightarrow Globus identity mapping
- **Guest collections shield users from complexity**

The screenshot shows the 'Client ID for Web application' section of the Google Cloud Platform API & Services Credentials page. It includes fields for 'Name' (set to 'Globus Connect Server v5.4 Gateways'), a note about authorized domains, and a list of 'Authorized JavaScript origins' with two entries: 'https://75869.7567.data.globus.org' and 'https://d65b7.ac45.data.globus.org'. A '+ ADD URI' button is also visible.



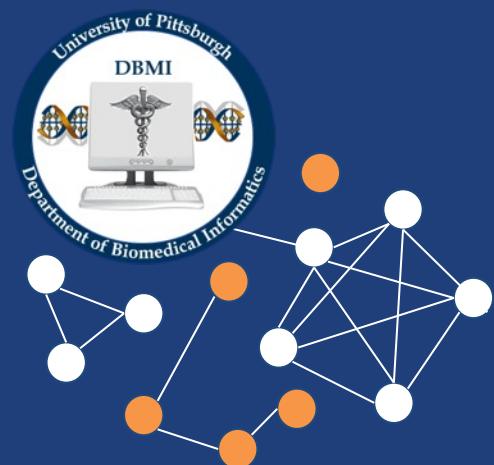
Globus Search and Portal Framework Example: Advanced Photon Source

acdc.alcf.anl.gov



Cancer Registry Records for Research (CR3)

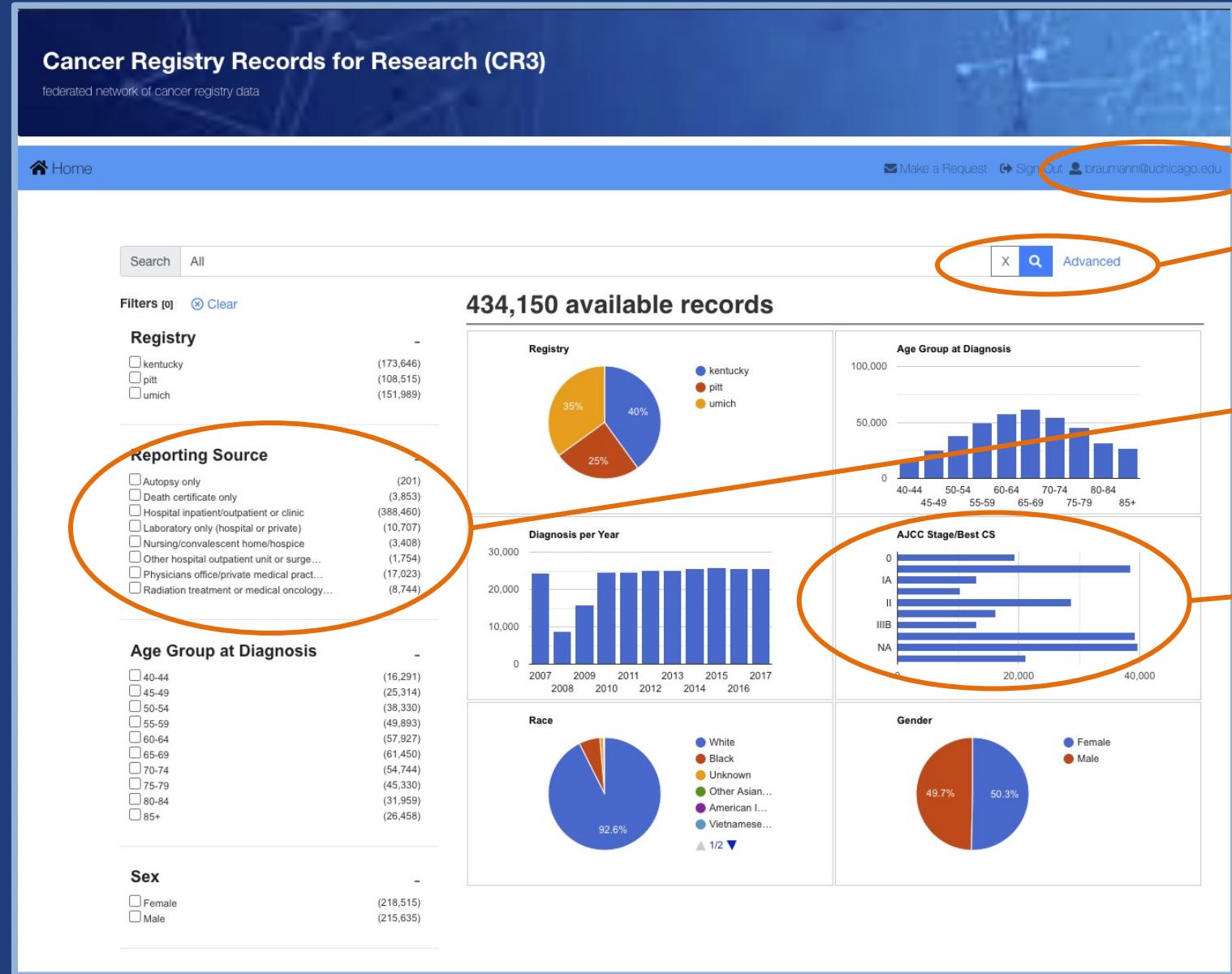
- **Create network of federated cancer registries**
 - Deploy similar infrastructure at other cancer registries
 - Enable queries across multiple registries
- **Federation via Globus: network scale \leftrightarrow local control**
 - Data owners input/export data sets, apply QC, set access policies
 - Registry data remain at the institution where they were generated
 - Identities are provided/authenticated by the institution, not Globus
 - System scale depends on data owners providing storage resources



University of
Pittsburgh

UPMC LIFE
CHANGING
MEDICINE

THE UNIVERSITY OF
CHICAGO



Federated logon using Globus Auth with Pitt/UPMC as identity providers

Google-like text search with facets for filtering

Variable facets based on source registry index

Dynamically updating charts as facets change

Developed using a framework based on the Globus Modern Research Data Portal* design pattern (docs.globus.org/mrdp)



Current Focus...



Automating data management at scale



Timer Service

Scheduled and recurring transfers
(*a.k.a. Globus cron*)

Command Line Interface

Ad hoc scripting and integration



Globus Flows service

Comprehensive task (data and compute) orchestration with human in the loop interactions



The Globus Timer service

- Scheduled/recurring file transfers
- Supports all Globus transfer and sync options
- Service with a command line interface
- Example: NIH – hpc.nih.gov/storage/globus_cron.html





Automating Cryo-EM Flows

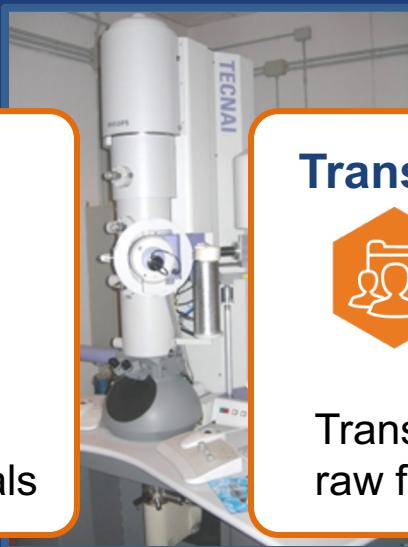
Globus
Flows



Auth



Get
credentials



Transfer



Transfer
raw files

funcX



Launch
analysis job

Carbon!



Correct,
classify, ...

The screenshot shows the cryoSPARC software interface. At the top, there's a navigation bar with 'Projects' and 'Workspaces'. Below it is a 'Statistics' section showing 'This week' and 'This month' data: Projects (5, 4), Workspaces (14, 13), Jobs (119, 86), and Completed jobs (102, 76). A 'Change Log' section details recent updates. To the right, there's a 'My Recent Jobs' list with entries for P40_JB, P40_JT, P40_JI, and P40_JS. The main area contains various tabs and buttons for different analysis steps like 'Ab initio Reconstruction', 'Select 2D classes', '2D Classification Streaming', etc. On the far right, there's a 'Discussion Forum' and a 'RELION' logo.

This is a screenshot of the ALCF Community Data Co-Op / Serial Crystallography Search interface. The top navigation bar includes 'ALCF Community Data Co-Op', 'Serial Crystallography Search', and other options. The main content area shows a search results table with columns for 'Run' (e.g., S9, S8, S10, SSX, S15, S14, S18, S12, S16, S17) and 'Date Processed' (e.g., 2020, 2021). A sidebar on the left lists 'Run' categories and 'Date Processed' filters. The bottom of the screen shows a progress bar and some status messages.

Search



Search,
discover,
reuse

Share



Set access
controls

Transfer



Move final
files to repo

Search



Index
ingest

funcX



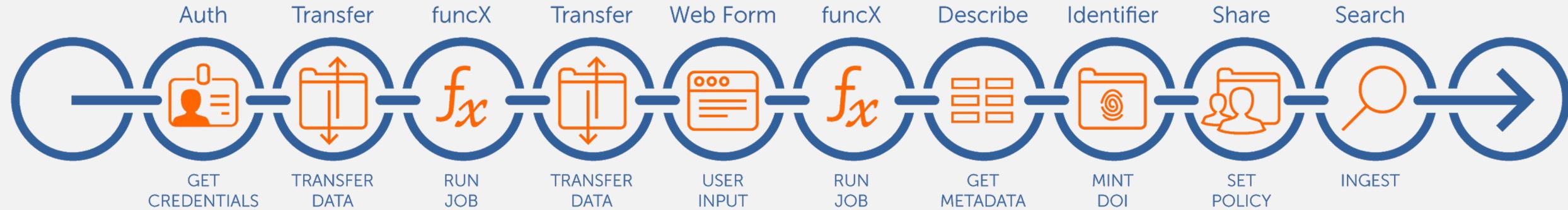
Extract
metadata



Managed automation of tasks

- **Flows:** A platform service for defining, applying, and sharing distributed research automation flows
- Flows comprise **Actions**
- **Action Providers:** Called by Flows to perform tasks
- **Triggers***: Start flows based on events

* In development





docs.globus.org
globus.org/connectors
globus.org/subscriptions
outreach@globus.org
support@globus.org