

Globus Platform Services for Data Publication

Greg Nawrocki – greg@globus.org

University of Chicago & Argonne National Lab

GeoDaRRS – August 7, 2018





Outline

- **Globus Overview**
- **Globus Data Publication v1**
- **Lessons learned**
- **Globus Data Publication v2 – Globus Automate**
- **Example application**
- **Summary**



Globus SaaS: Research data lifecycle

Instrument



Globus transfers files quickly, reliably, securely

2

Transfer

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



Publication Repository

1 Researcher initiates transfer request; or requested automatically by script, science gateway

1



3 Researcher selects files to share, selects user or group, and sets access permissions

3

Share

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

5



Personal Computer



6 Publish

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

6

8 Peers, collaborators search and discover datasets; transfer and share using Globus

8



Discover

- Use a Web browser
- Access any storage
- Use an existing identity



Globus Data Publication V1

- Cloud-based web app
- BYO storage & in-place publication
- User-managed collections
- Select pre-defined schema
- Handle, DOI persistent identifiers
- Adoption since 2015:
 - >2000 users, >600 datasets

The image displays three screenshots of the Globus Data Publication web application. The top screenshot shows the main dashboard with a search bar, navigation menu, and a 'MDF CONNECT' section. The middle screenshot shows a search results page for 'Find and Share Canadian Research Data' with a list of datasets and a 'Deposit Data' button. The bottom screenshot shows a 'HOW TO GET STARTED' section with steps for describing and submitting data.

MDF CONNECT

It has never been easier to share your data with the community. Deposit data once, send to partner services.

Tell your research story.

[Become a Contributor](#)

HOW TO GET STARTED

2 - Describe Your Data
Describe your dataset using the MDF Connect form, and add any additional descriptions to a README or README.md file in the base directory.

3 - Submit Data
Select where you want your dataset deposited and let us handle the rest.

[Become a Contributor](#)



Many variations of data publication...



Citable Data

Standard metadata,
persistent identifiers,
durable storage



Institutional Data

Many domains,
custom metadata,
locally managed storage



Community Data

Agreed schema,
larger datasets,
fine grained metadata



Includes active data management



Active Research Data

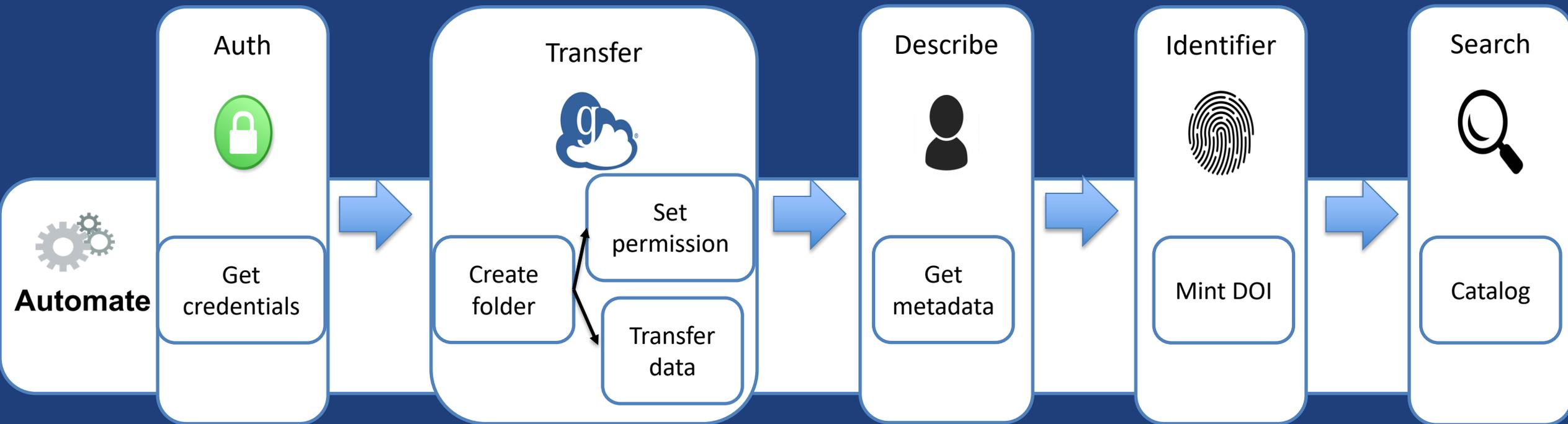
Less standard and evolving schema, organize data independent of storage, support active collaboration, location agnostics identifiers

Enable automation.



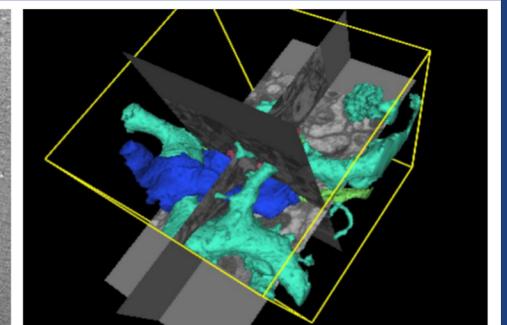
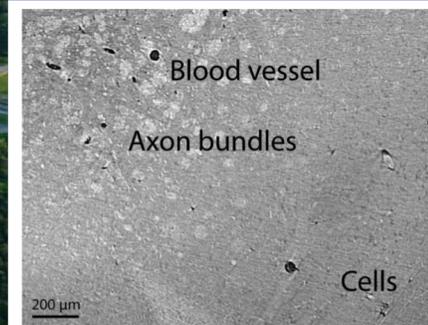
Globus Automation Platform

- Decompose Globus Publish v1 into platform services
- Allow for flexible re-composition and adaptation of services
- Enable extension and enhancement



Automation example

- UChicago's Kasthuri Lab study brain aging and disease
 - Construct connectomes -- mapping of neuron connections



g Neuroanatomy reconstruction pipeline

APS



1. Imaging



2. Acquisition



3. Pre-processing



ALCF



4. Preview & Centre



5. User validation & input



6. Reconstruction



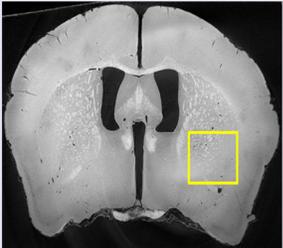
JLSE



7. Publication



UChicago



8. Visualization



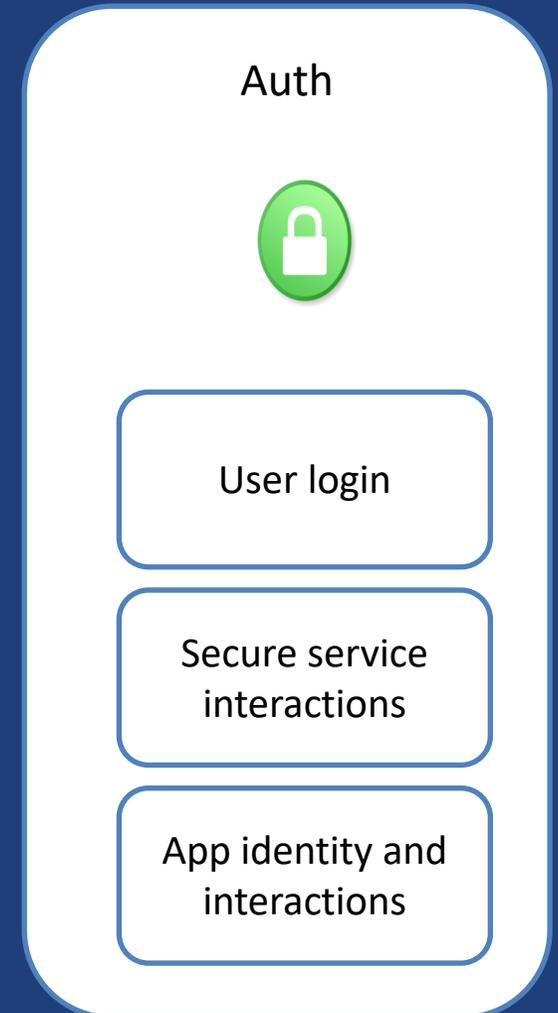
9. Science!





Globus Auth

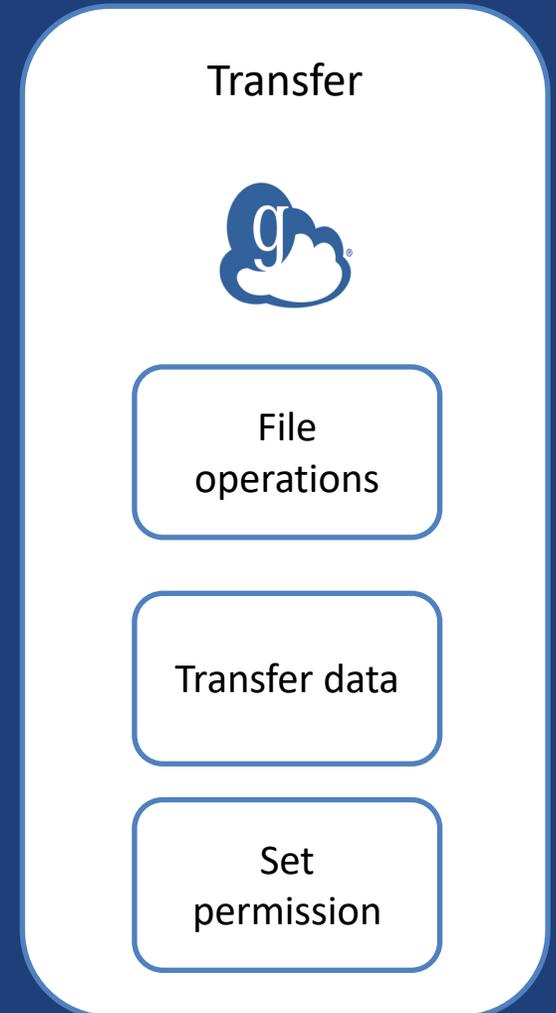
- **Foundational Identity and Access Management service**
- **Protects REST API communications**
- **Enables login for diverse app ecosystem, with no new identity required**
- **Employs least privileges security model**





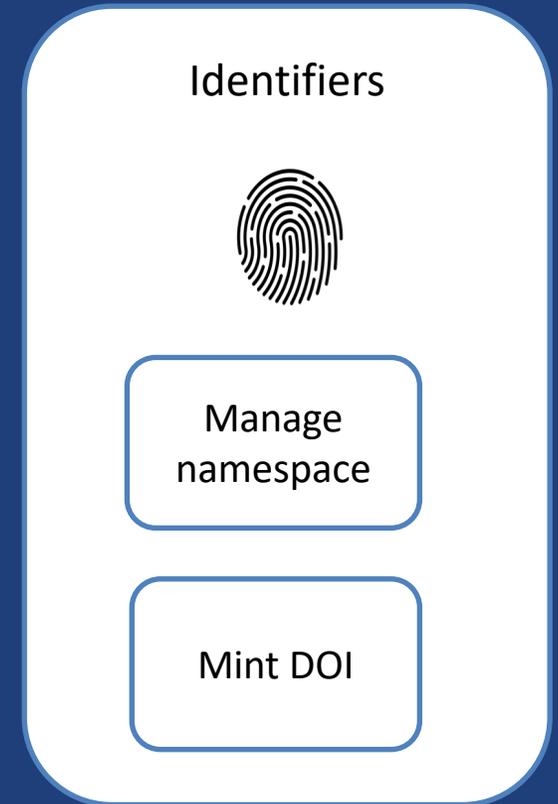
Globus Transfer and Sharing

- **File management**
 - Uniform interface for file operations
- **Transfer**
 - Managed, secure, high-performant data transfer
- **Sharing**
 - Fine grained sharing from existing storage system
 - Share with specific users, groups or public
- **Leverages existing security solutions at institutions**
- **APIs for integration with applications**



Globus Identifiers

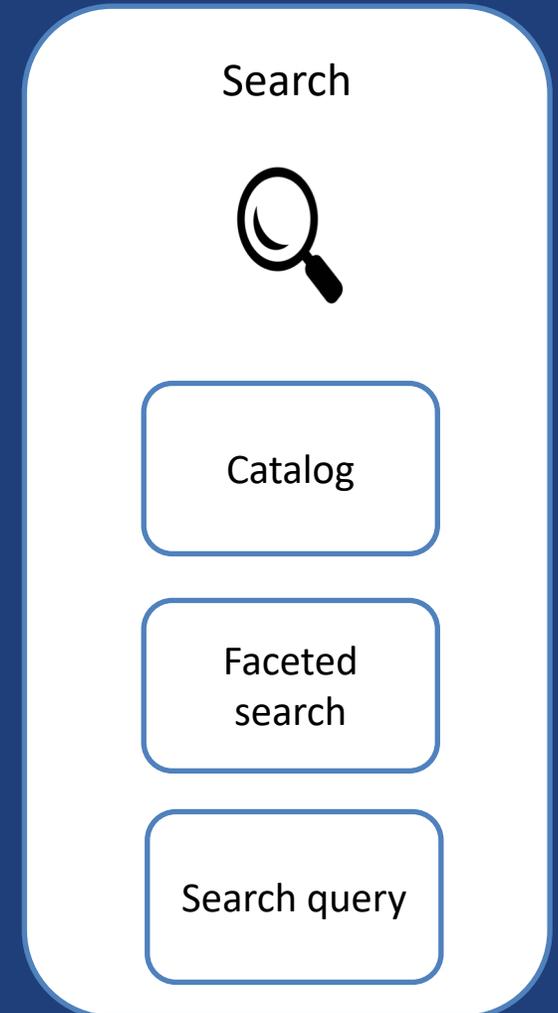
- **Issue persistent identifiers**
 - DOI, ARK, Handle, Globus
- **Within a namespace**
 - Control which identities and groups can create identifiers in your namespace
- **Each identifier has:**
 - Link to data
 - Landing page
 - **Visibility**
 - **Checksum**
 - **(Extensible) Metadata**
 - **Replaces / Replaced-by**





Globus Search

- **Hosted, scalable service for research data discovery**
 - Schema agnostic
 - Fine grain access control
 - Plain text search
 - Faceted search
 - Rich query language

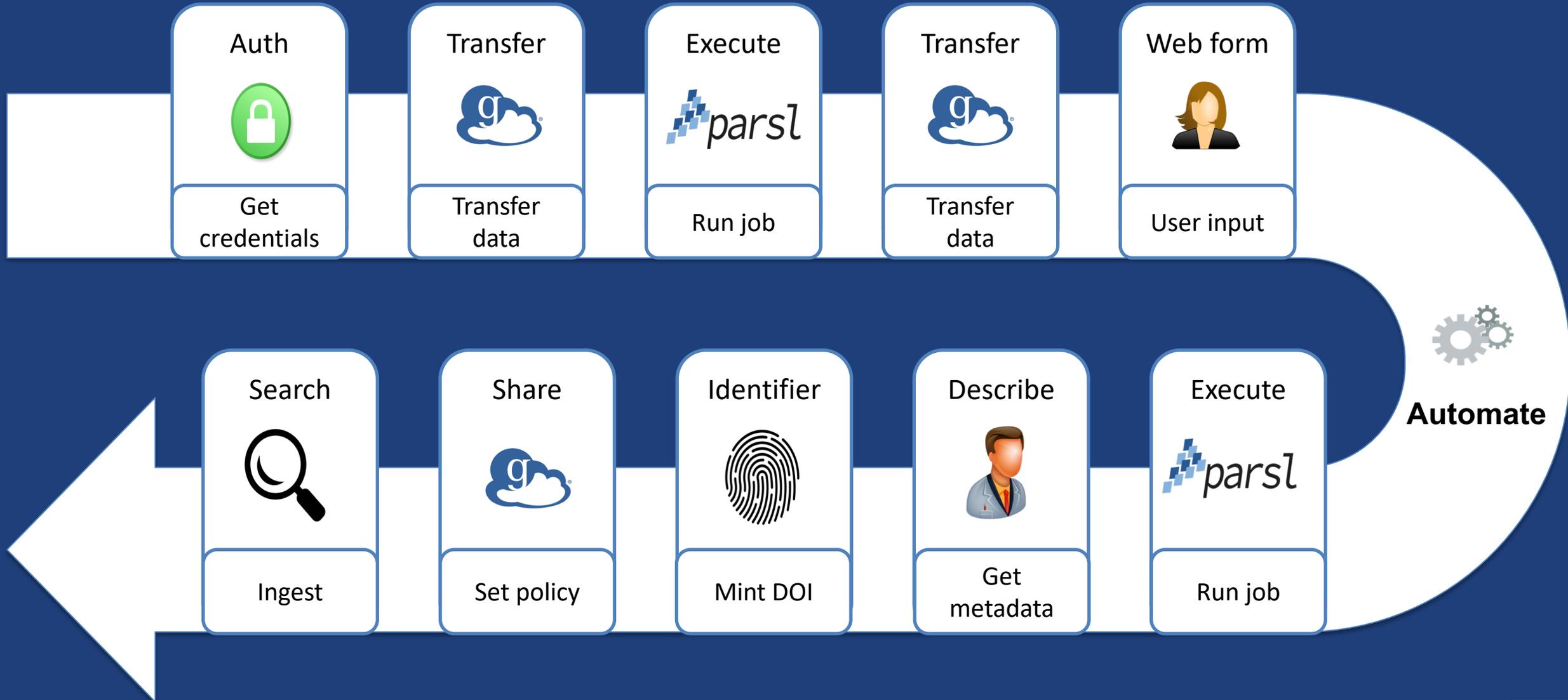




Globus Automate

- **Composition and execution service for automating research data management**
- **Higher level language and authoring tools**
 - Describe flow or state machine
- **Pluggable API to integrate any actions**
 - E.g. Automated validation, metadata extraction
- **Flexible invocation of actions**
 - E.g. User driven, event driven

Neuroanatomy reconstruction pipeline





Summary

Data publication solutions can be built using an automation platform.

Globus provides a customizable and extensible research data automation platform.



Thanks to...



THE UNIVERSITY OF
CHICAGO



U.S. DEPARTMENT OF
ENERGY

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



Argonne 
NATIONAL LABORATORY

 powered by
amazon
web services

Rachana Ananthakrishnan, Ben Blaiszak, Kyle Chard,
Ryan Chard, Brendan McCollam, Jim Pruyne, Stephen
Rosen, Steve Tuecke, & Ian Foster