Globus Platform Services for Data Publication

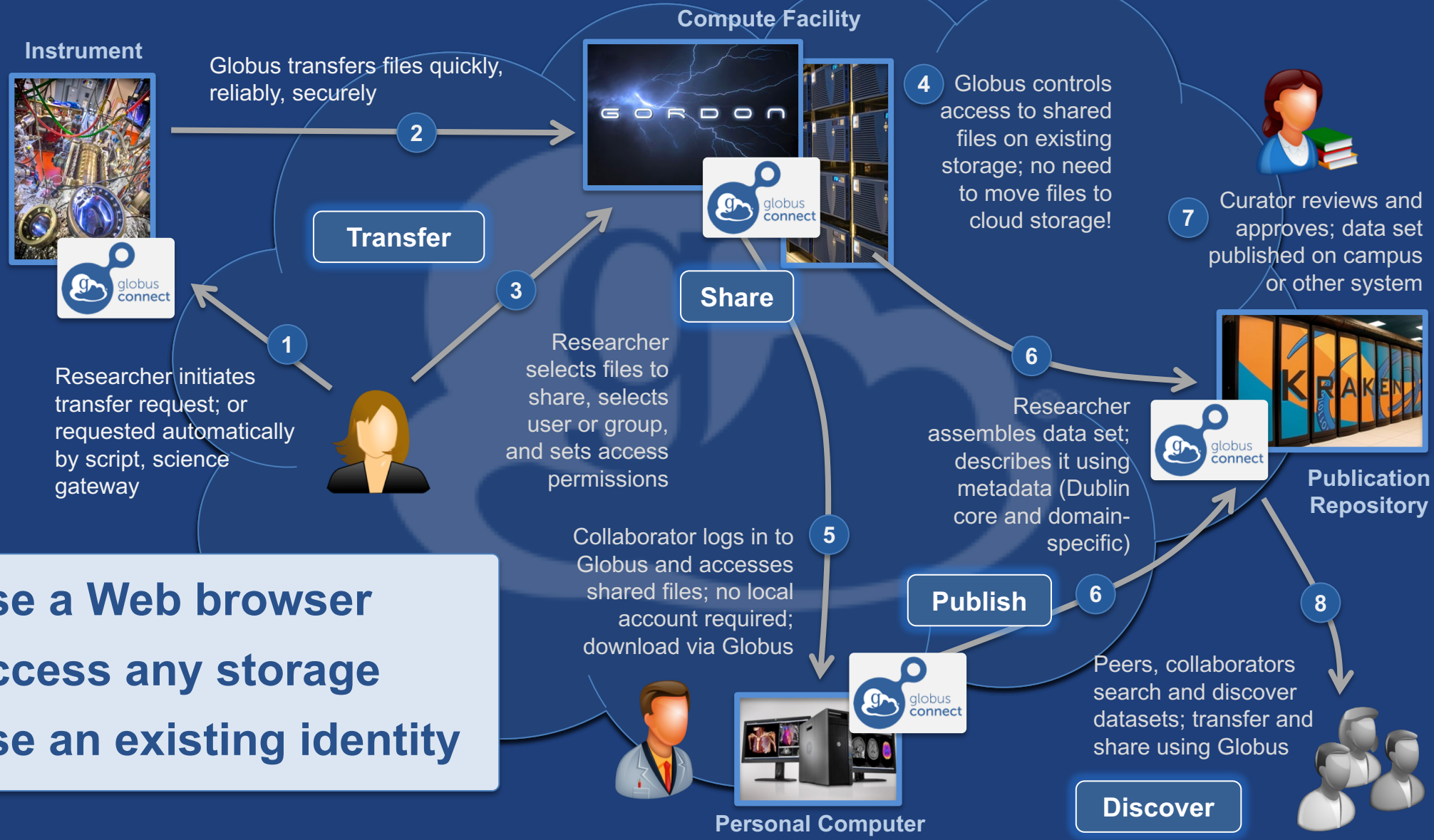
Greg Nawrocki – greg@globus.org
University of Chicago & Argonne National Lab

GeoDaRRS – August 7, 2018

Outline

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

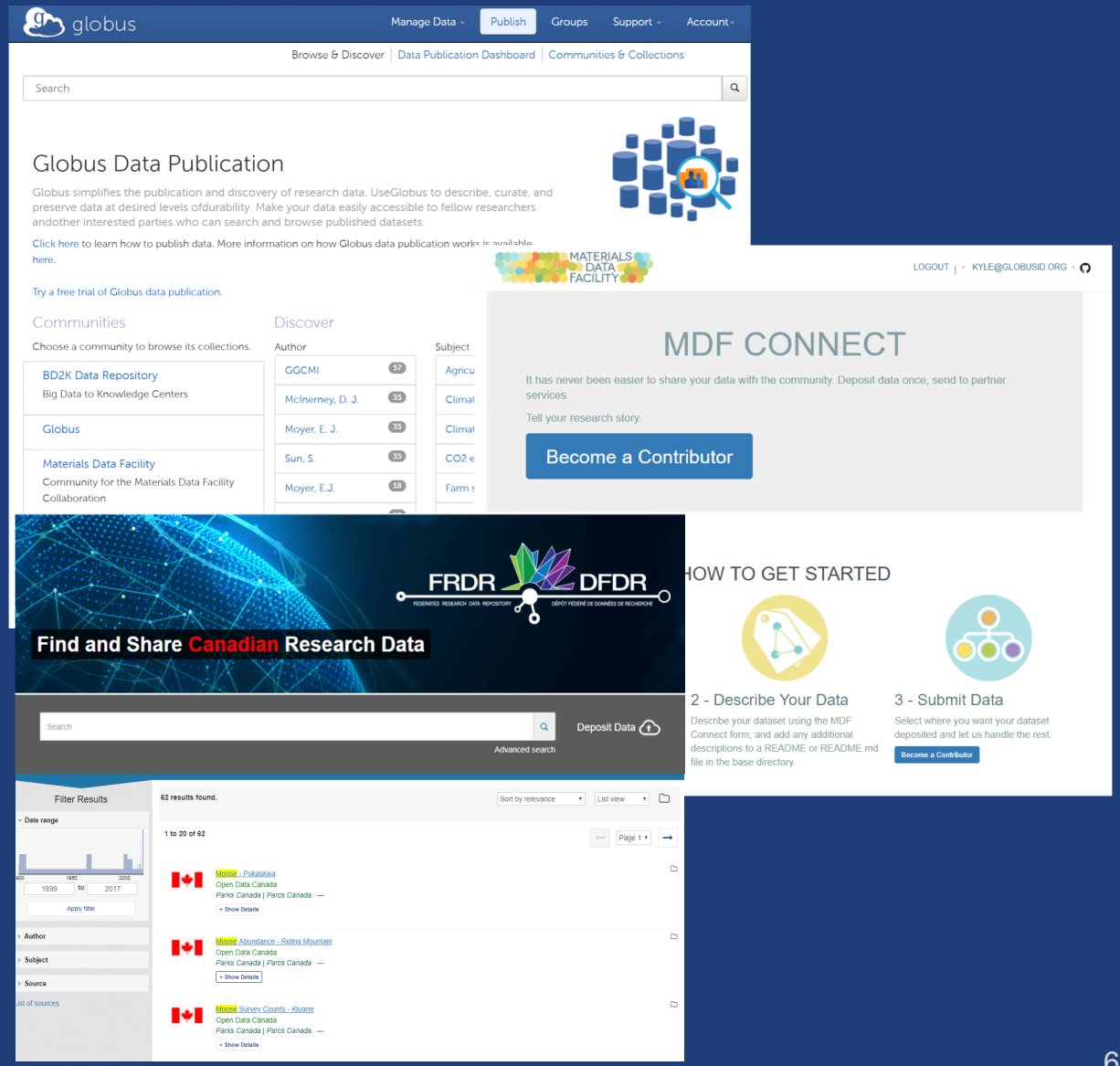
Globus SaaS: Research data lifecycle



- **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



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

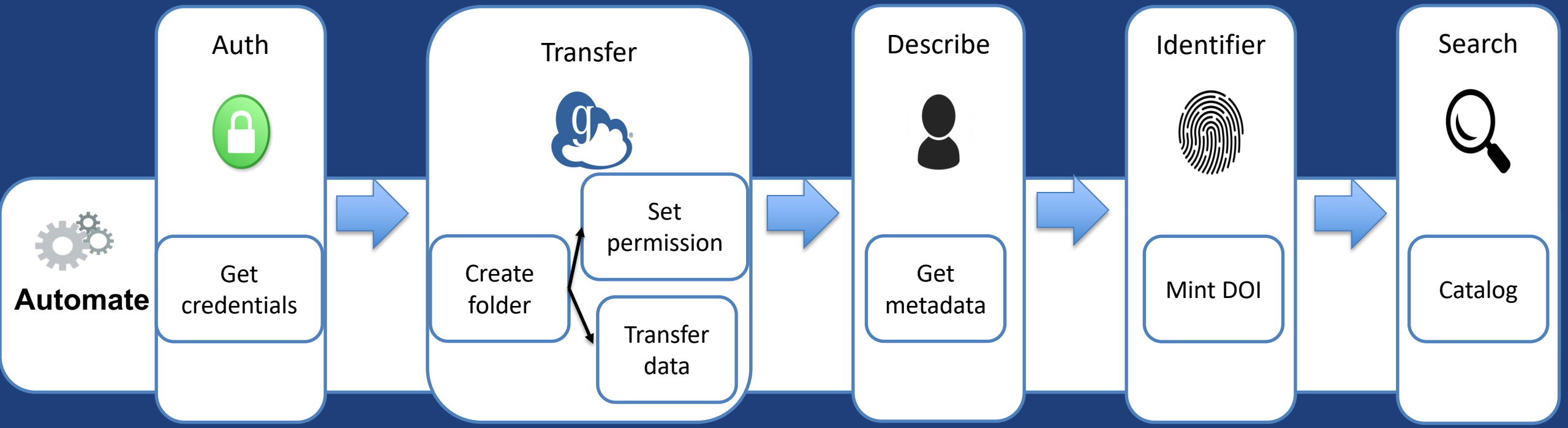
Active Research Data

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

Enable automation.

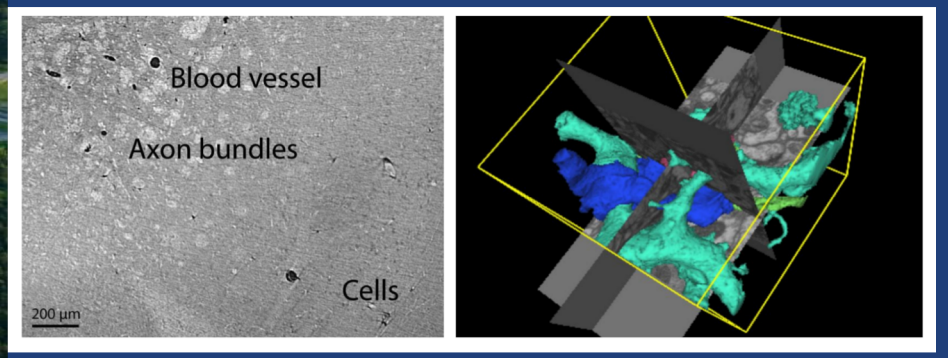
Globus Automation Platform

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



Automation example

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



Neuroanatomy reconstruction pipeline

APS

1. Imaging

2. Acquisition

3. Pre-processing

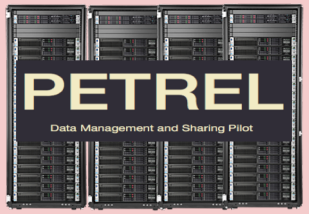
ALCF

4. Preview & Centre

5. User validation & input

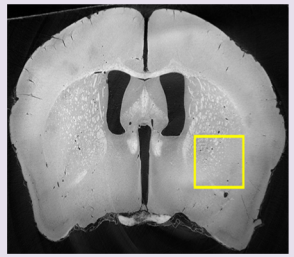
6. Reconstruction

JLSE



7. Publication

UChicago

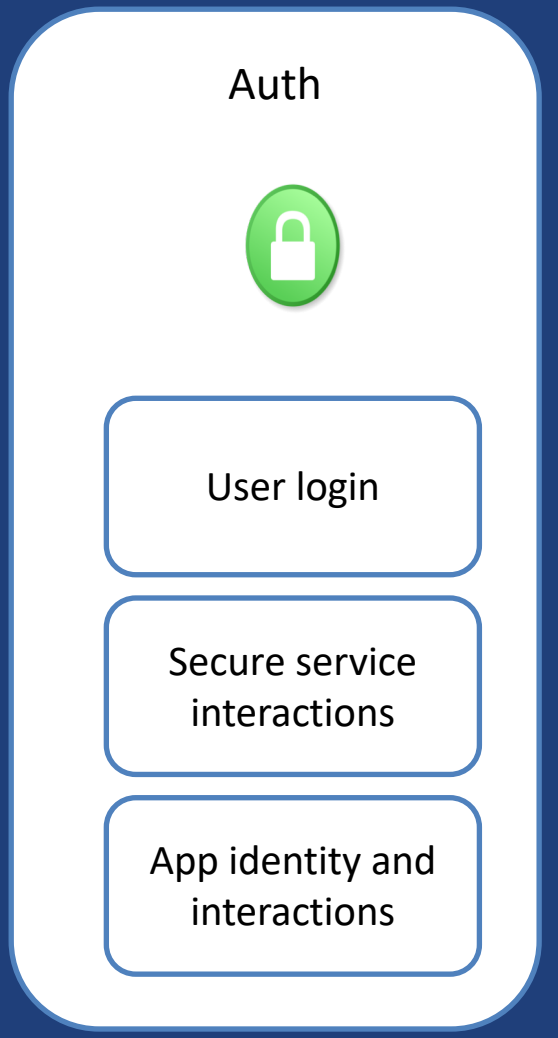


8. Visualization

9. Science!

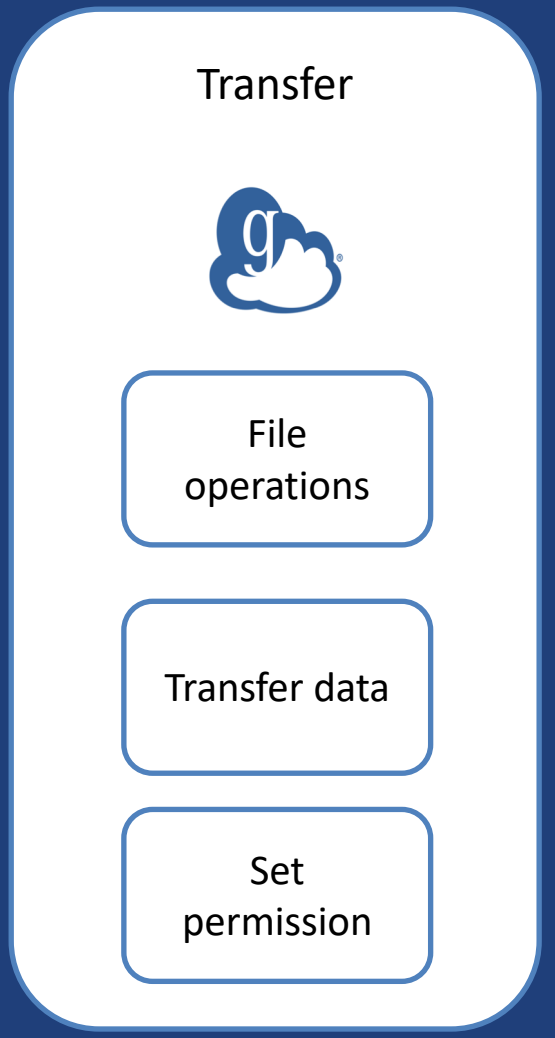
Globus Auth

- **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**



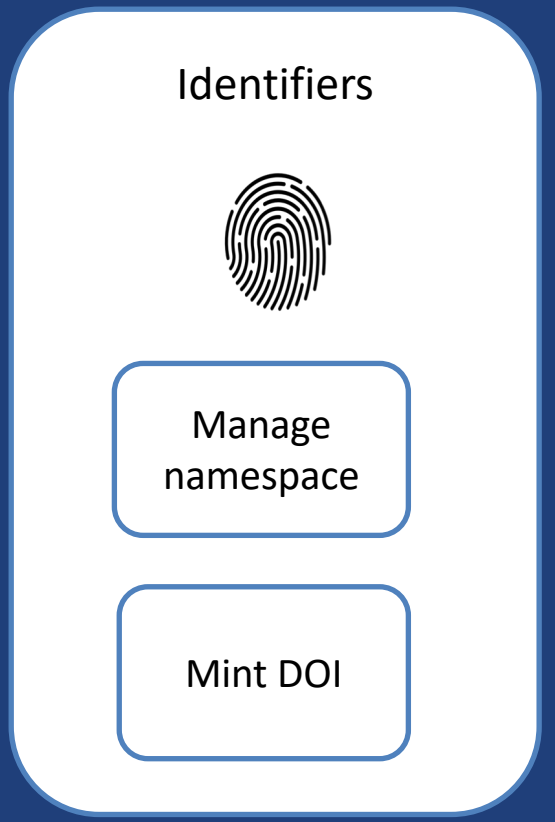
Globus Transfer and Sharing

- **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**



DataCite
FIND, ACCESS, AND REUSE DATA

doi

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

Auth	Transfer	Execute	Transfer	Web form
Get credentials	Transfer data	Run job	Transfer data	User input

Automate

Search	Share	Identifier	Describe	Execute
Ingest	Set policy	Mint DOI	Get metadata	Run job

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

Argonne NATIONAL LABORATORY

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

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