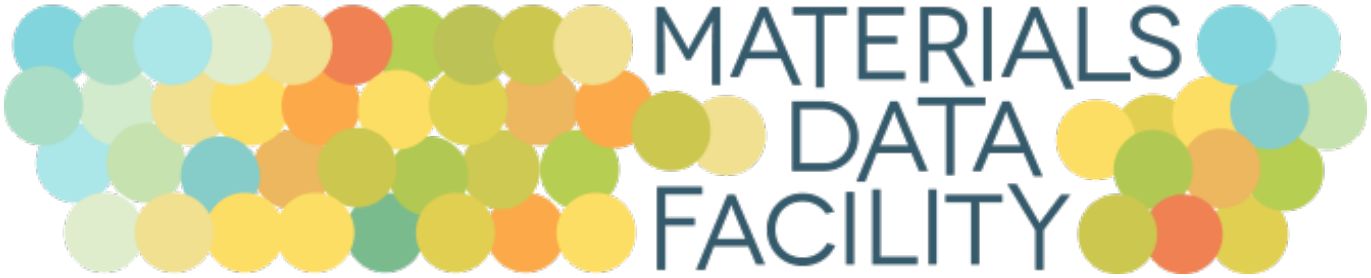


The Materials Data Facility (MDF) – Data Services to Advance Materials Research

Ben Blaiszik (blaiszik@uchicago.edu)
Ian Foster (foster@uchicago.edu)
Steve Tuecke, Kyle Chard,
Rachana Ananthakrishnan, Jim Pruyne (UC)
Michal Ondrejcek, Kenton McHenry,
John Towns (UIUC – NCSA)

materialsdatafacility.org

NIST National Institute of Standards and Technology U.S. Department of Commerce
CHiMaD Center for Hierarchical Materials Design
THE UNIVERSITY OF CHICAGO
ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
Argonne NATIONAL LABORATORY

Outline

- Overview
- Background on Globus platform (very quick)
- Key features
- Submission walk-through with use case

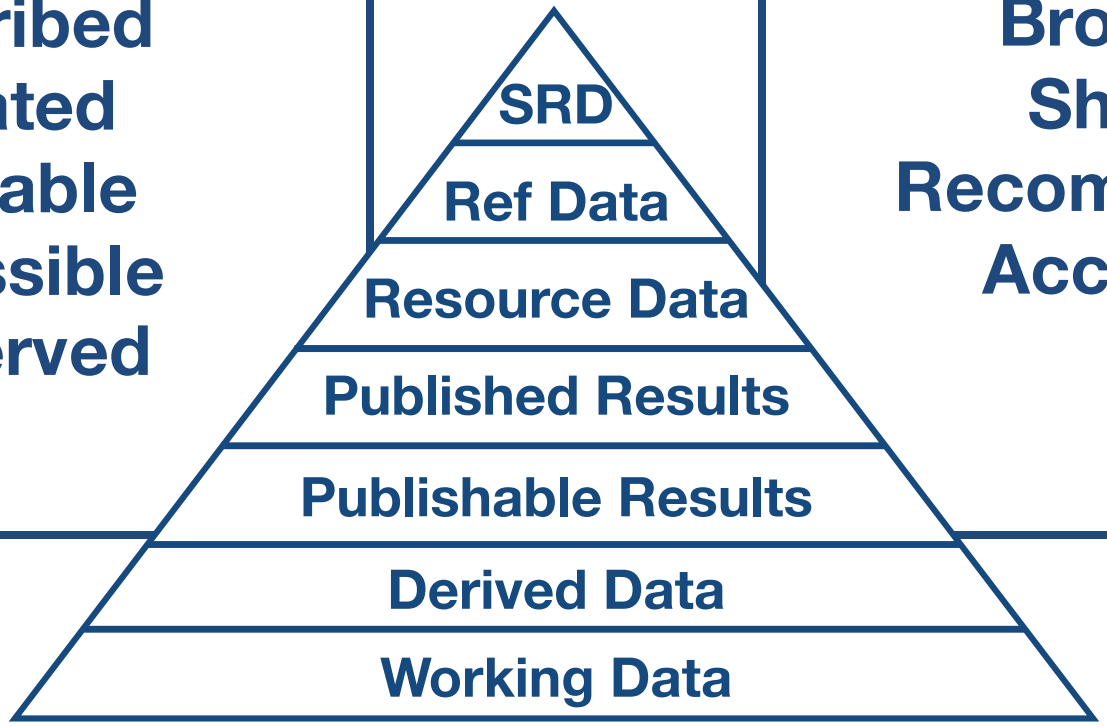
What is MDF?

We aim to make it simple for materials datasets and resources to be ...

Published
Identified
Described
Curated
Verifiable
Accessible
Preserved

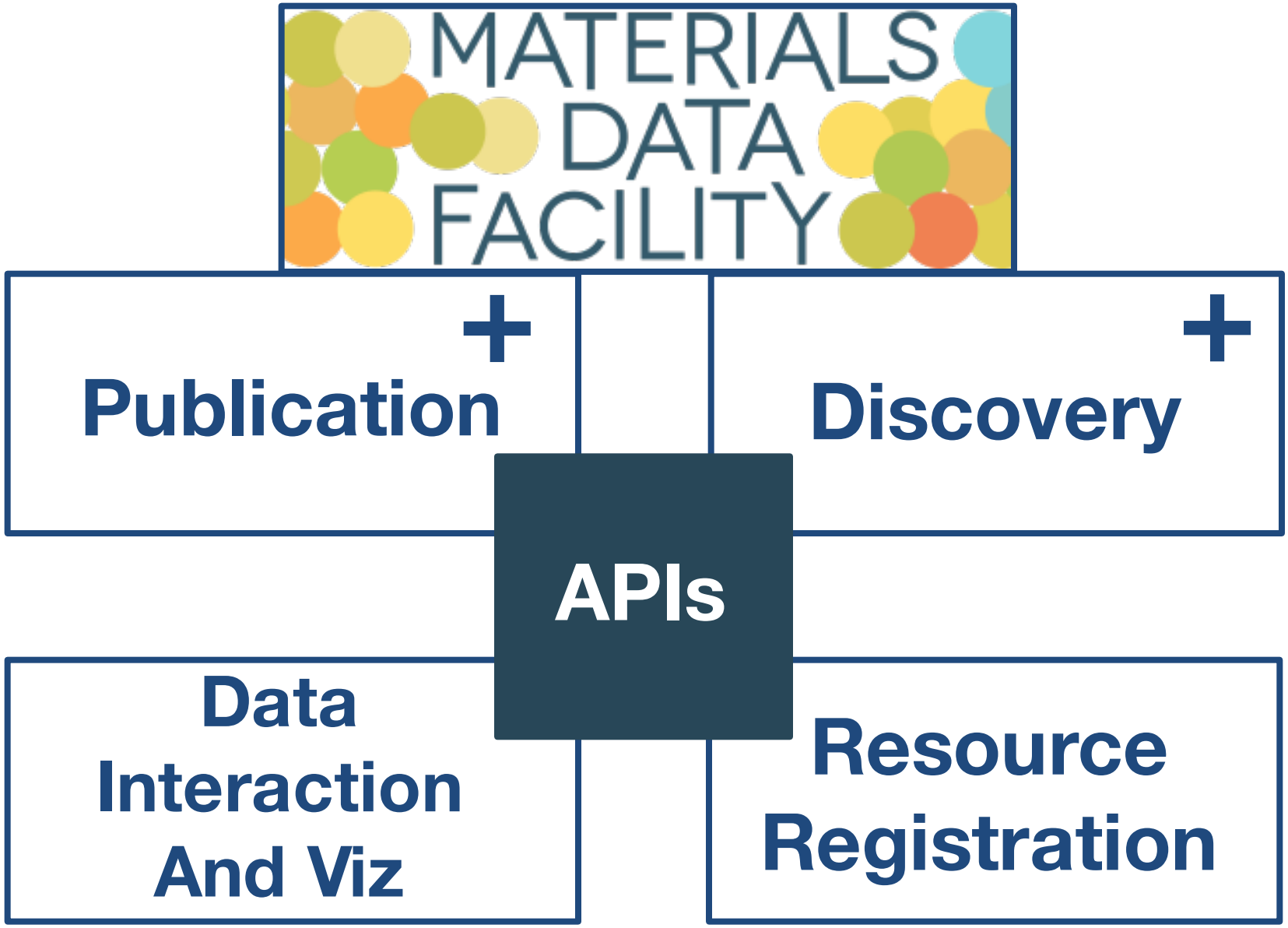
and

Discovered
Searched
Browsed
Shared
Recommended
Accessed



What infrastructure do we need to effectively support materials researchers?

Service Infrastructure



+ - Initial Foci

Publication

Some key features

- Identify and datasets with persistent identifiers
- Describe datasets with appropriate metadata, and provenance
- Curate metadata and data composition
- Verify dataset contents over time
- Preserve critical datasets in a state that increases transparency, replicability, and helps encourage reuse

Discovery

- Search and query datasets in modern ways – e.g. via indexed metadata rather than remembering file paths
- Discover distributed materials resources (more later)

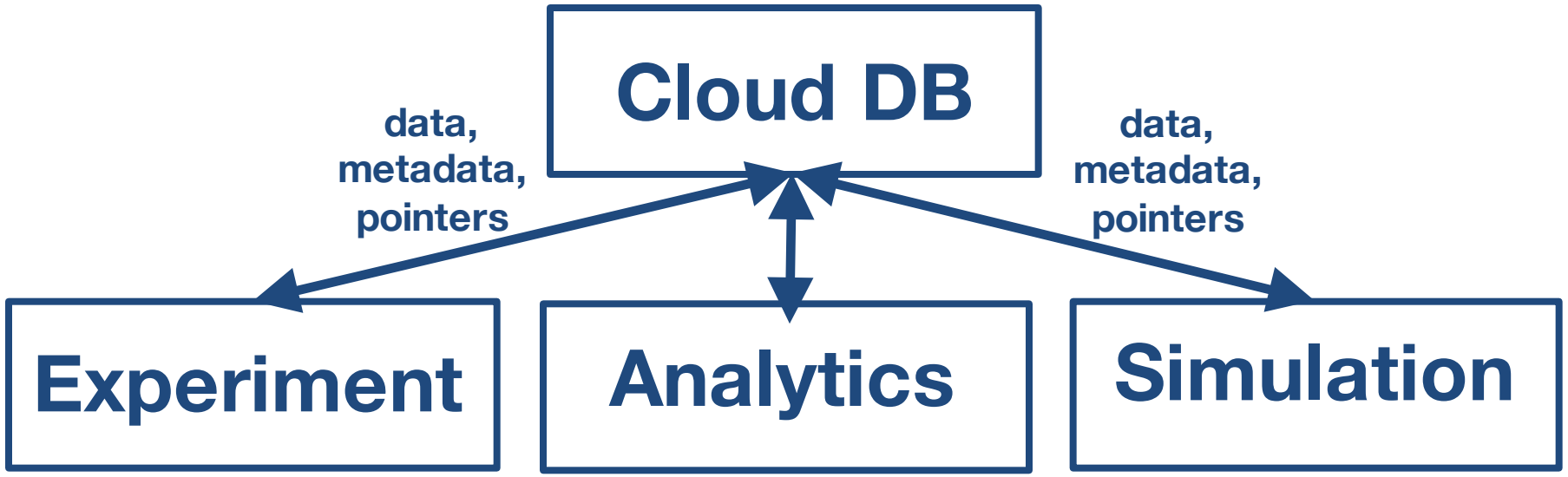
Resource Registration

- Find existing, widely distributed, materials resources
- Register new resources into the network
- Unified access and search methods for distributed resources

Data Interaction And Viz

See me for discussion

- Data-driven experiments using HPC resources and workflow technologies
- Real-time interaction with data regardless of data location (of course pending appropriate data access) and data size
- Machine learning across datasets and storage locations
- Automated discovery support



Understanding Incentives is Critical

Increasing Impact

- Increase paper citations[1]
- Add dataset citation capabilities

Meeting Award Requirements

- Simplify DMP compliance

Smoothing Dislocations

- Enable simple sharing among collaborators (near and far)
- Ease transitions between students
- Lessen need for *ad hoc* resource sharing (e.g. via group websites)

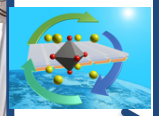
[1] Citation increase 30 (10.7717/peerj.175) - 60% (10.1371/journal.pone.0000308) [caveat bio research]

So where are we now?

Publication

Materials Data Publication/Discovery is Often a Challenge

Data Collection



Data Storage and Process

NATURE MATERIALS | ARTICLE

Reversible dilatancy in entangled single-wire materials

David Rodney, Benjamin Gadot, Oriol Riu Martinez, Sabine Rolland du Roscoat & Laurent Orgéas

Affiliations | Contributions | Corresponding author

Nature Materials (2015) | doi:10.1038/nmat4429

Received 25 May 2015 | Accepted 25 August 2015 | Published online 28 September 2015

Citation | Reprints | Rights & permissions | Article metrics

Abstract

Abstract • References • Author information • Supplementary information

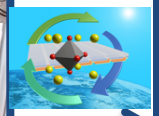
Designing structures that dilate rapidly in both tension and compression would benefit devices such as smart filters, actuators or fasteners. This property however requires an unusual Poisson ratio, or Poisson function at finite strains, which has to vary with applied strain and exceed the familiar bounds: less than 0 in tension and above 1/2 in compression. Here, by combining mechanical tests and discrete element simulations, we show that a simple three-dimensional architected material, made of a self-entangled single long coiled wire, behaves in between discrete and continuum media, with a large and reversible dilatancy in both tension and compression. This unusual behaviour arises from an interplay between the elongation of the coiled wire and rearrangements due to steric effects, which, unlike in traditional discrete media, are hysteretically reversible when the architecture is made of an elastic fibre.

Publication

Materials Data Publication/Discovery is Often a Challenge

Want to Publish → ← Want to Discover / Use

Data Collection

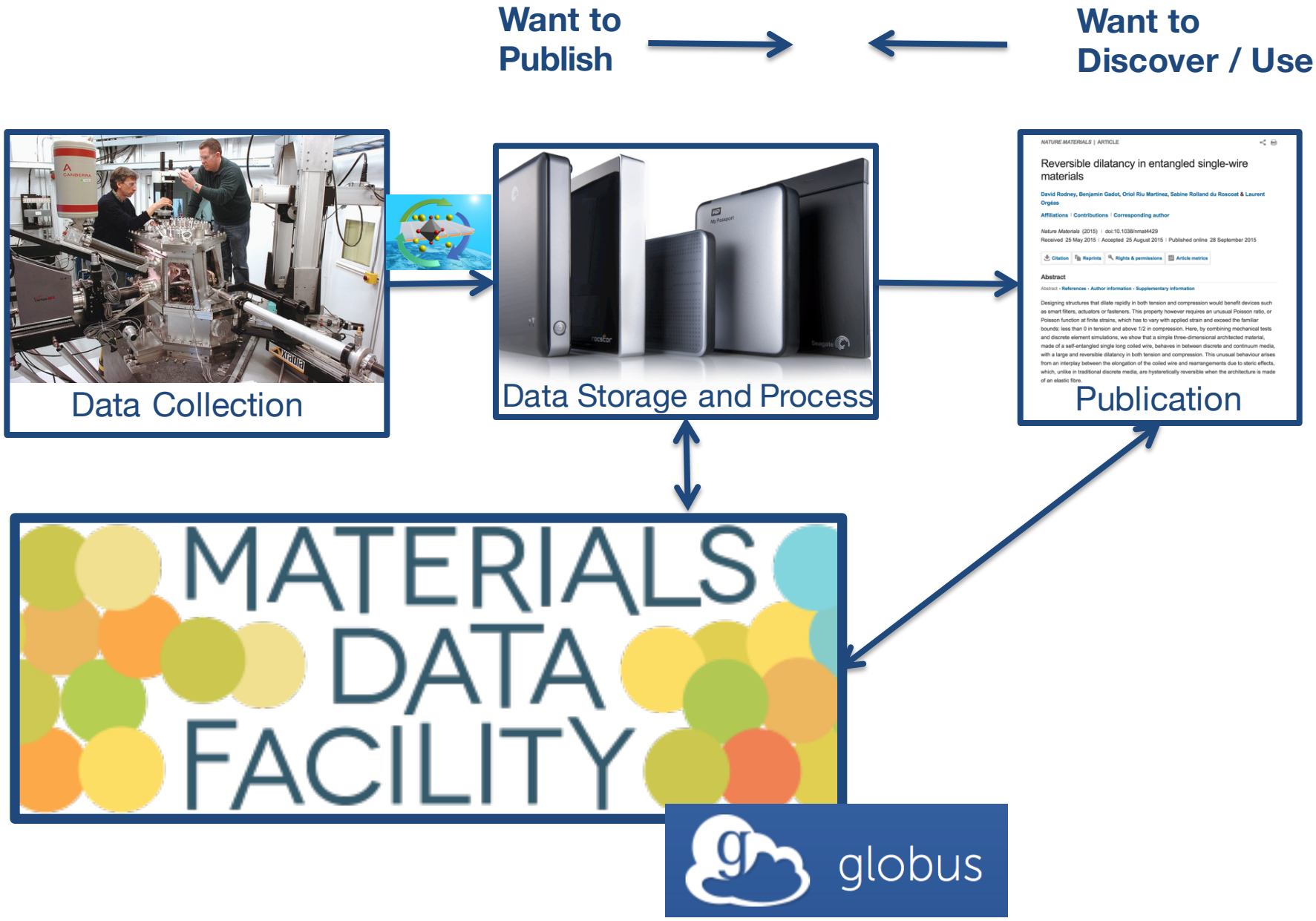


Data Storage and Process

Publication

- Need networked storage, sometimes many TB ?
- Need to uniquely identify data for search/cite ?
- Need custom metadata descriptions ?
- Need a data curation workflow
- Need automation capabilities

Materials Data Publication/Discovery is Often a Challenge



Globus Platform-as-a-Service

Identity management

- create and manage a unique identity linked to external identities for authentication

User groups

- Manage user group creation and administration flows
- Share data with user groups

Data publication

Data transfer

- High-performance data transfer from a web browser
- Optimize transfer settings and verify transfer integrity
- Add your laptop to the Globus cloud with Globus Connect Personal

Data sharing

- Share directly from your storage device (laptop or cluster)
- File and directory-level ACLs

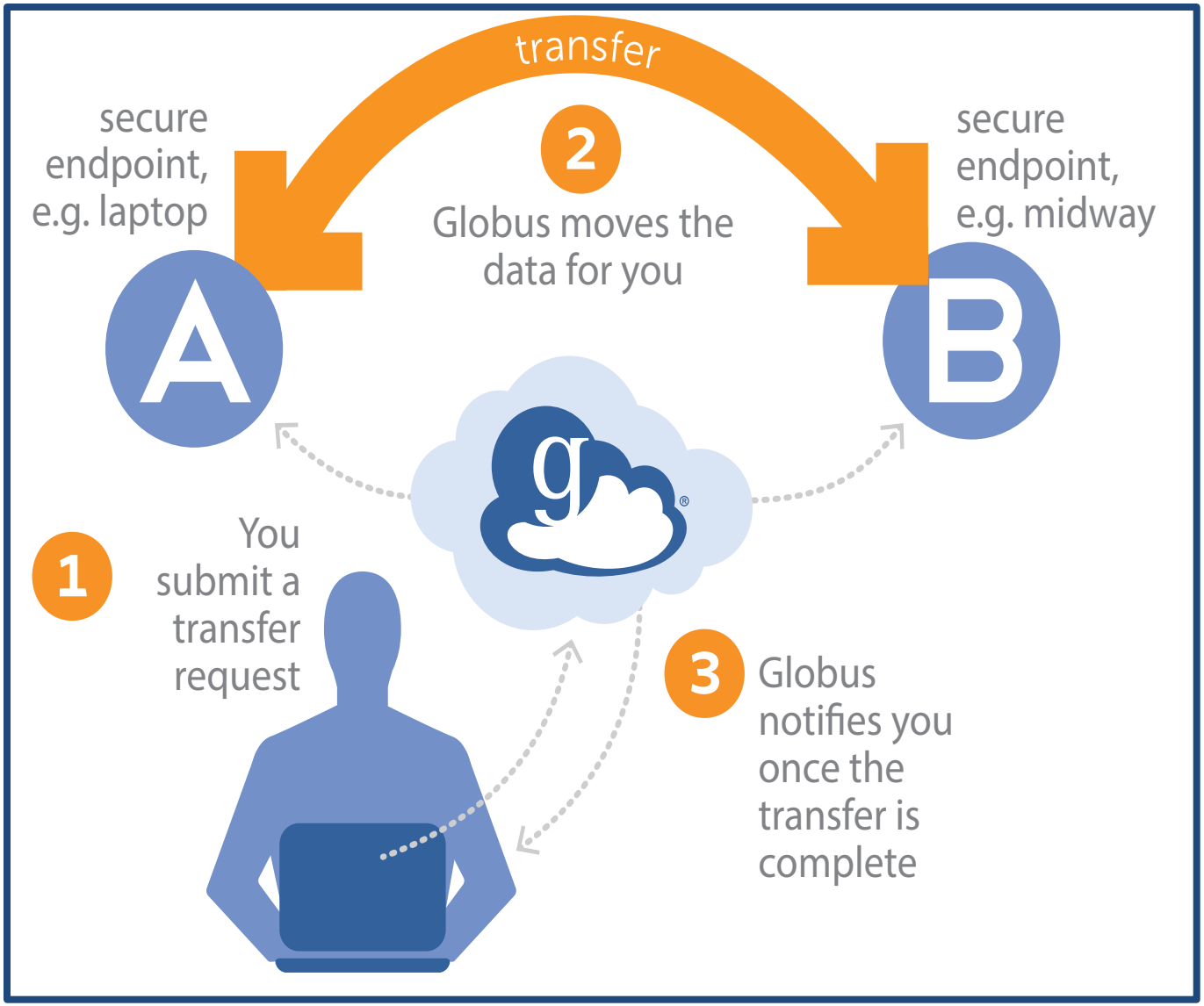
Globus Transfer Background

Endpoint

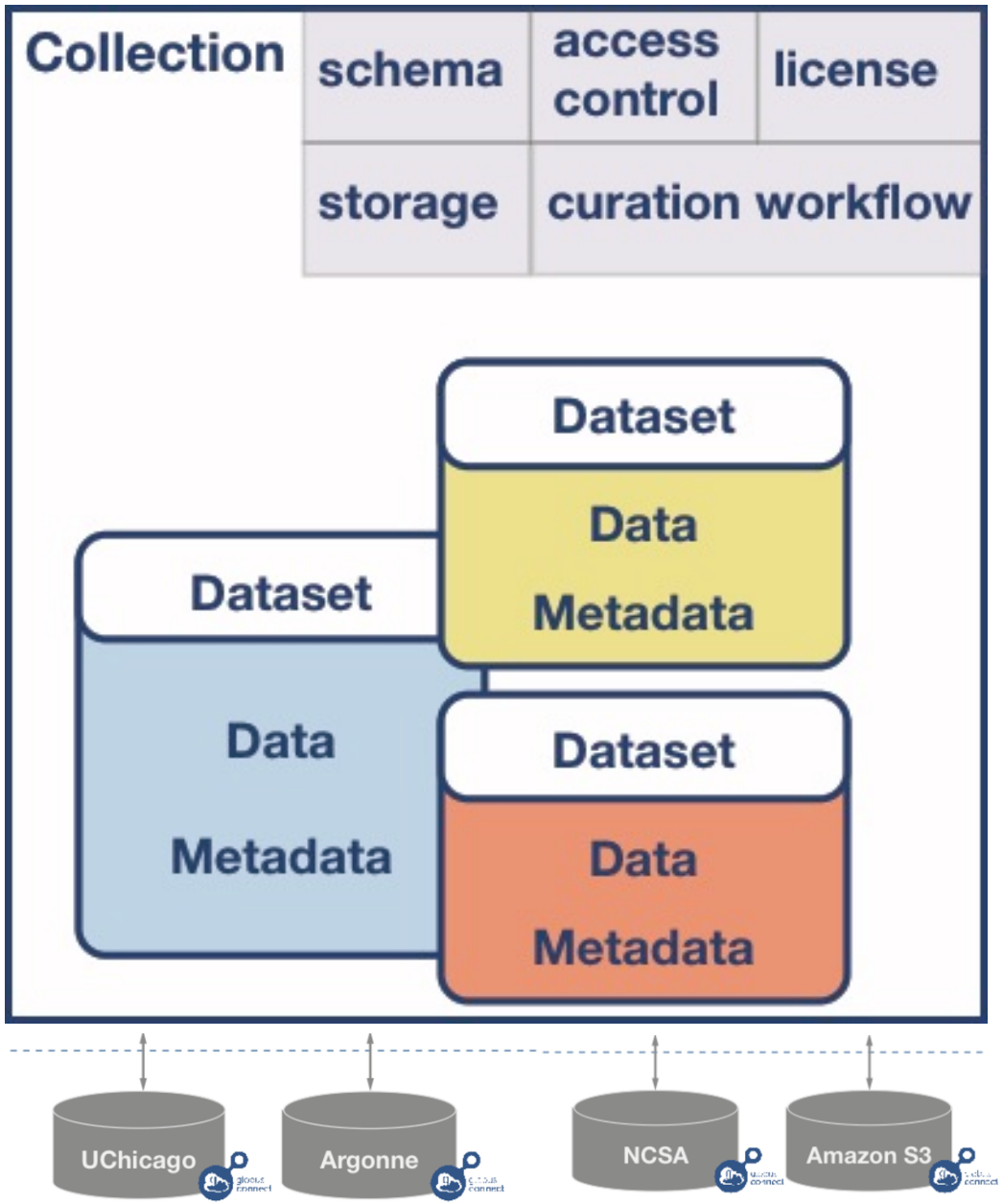
- E.g. laptop or server running a Globus client (e.g. Dropbox client)
- Enables advanced file transfer and sharing
- Currently GridFTP, future GridFTP + HTTP

Some Key Features

- REST API for automation and interoperability
- Web UI for convenience
- Optimizes and verifies transfers
- Handles auto-restarts
- Battle tested with big data



Collection Model



- **Collections might be a research group or a research topic...**
- **Collections have specified**
  - Mapping to storage endpoint
    - Currently handled as automatically created shared endpoints
  - Metadata schemas
  - Access control policies
  - Licenses
  - Curation workflows
- **Collections contain**
  - Datasets
    - Data
    - Metadata
- **Metadata Persistence**
  - Metadata log file with dataset
  - Metadata replicated in search index

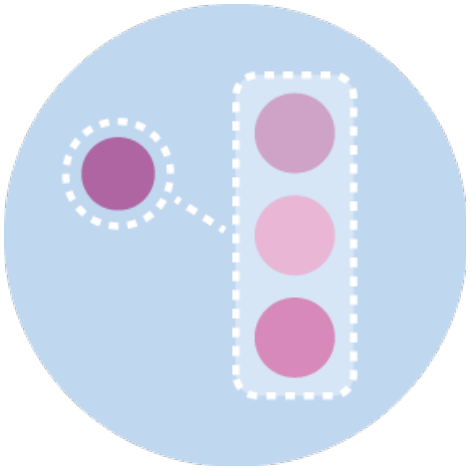
Publish Large Datasets

- Leverages Globus production capabilities for file transfer, user authentication, and groups

- 100 TB of reliable storage @ NCSA, and more storage at Argonne
  - Globus endpoint at ncsa#mdf
  - Expandable to PBs as necessary
  - Automated tape backup for reliability (in progress)

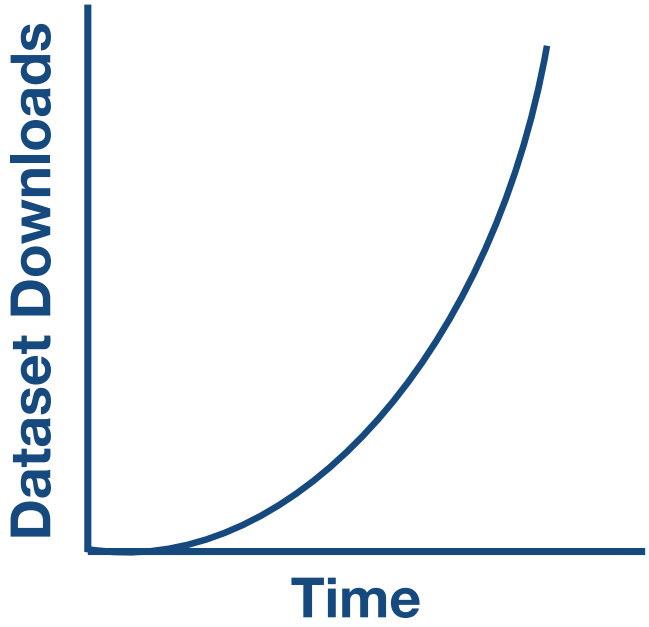
- Optionally use your own local or institutional storage

Uniquely Identify Datasets



- **Associate a unique identifier with a dataset**
  - DOI, Handle
- **Improve dataset discovery and citability**
  - Aligning incentives and understanding the culture will be critical to driving adoption

Future…



- Your work has been cited 153 times in the last year
- Researchers from 30 institutions have downloaded your datasets

Share Data with Flexible ACLs

- Share data publicly, with a set of users, or keep data private

Leverage Curation Workflows

- Collection administrators can specify the level of curation workflow required for a given collection e.g.
  - No curation
  - Curation of metadata only
  - Curation of metadata and files

Customize Metadata

- **Build a custom metadata schema for your specific research data**
- **Re-use existing metadata schemas**
- **Working in conjunction with NIST researchers to define these schemas**

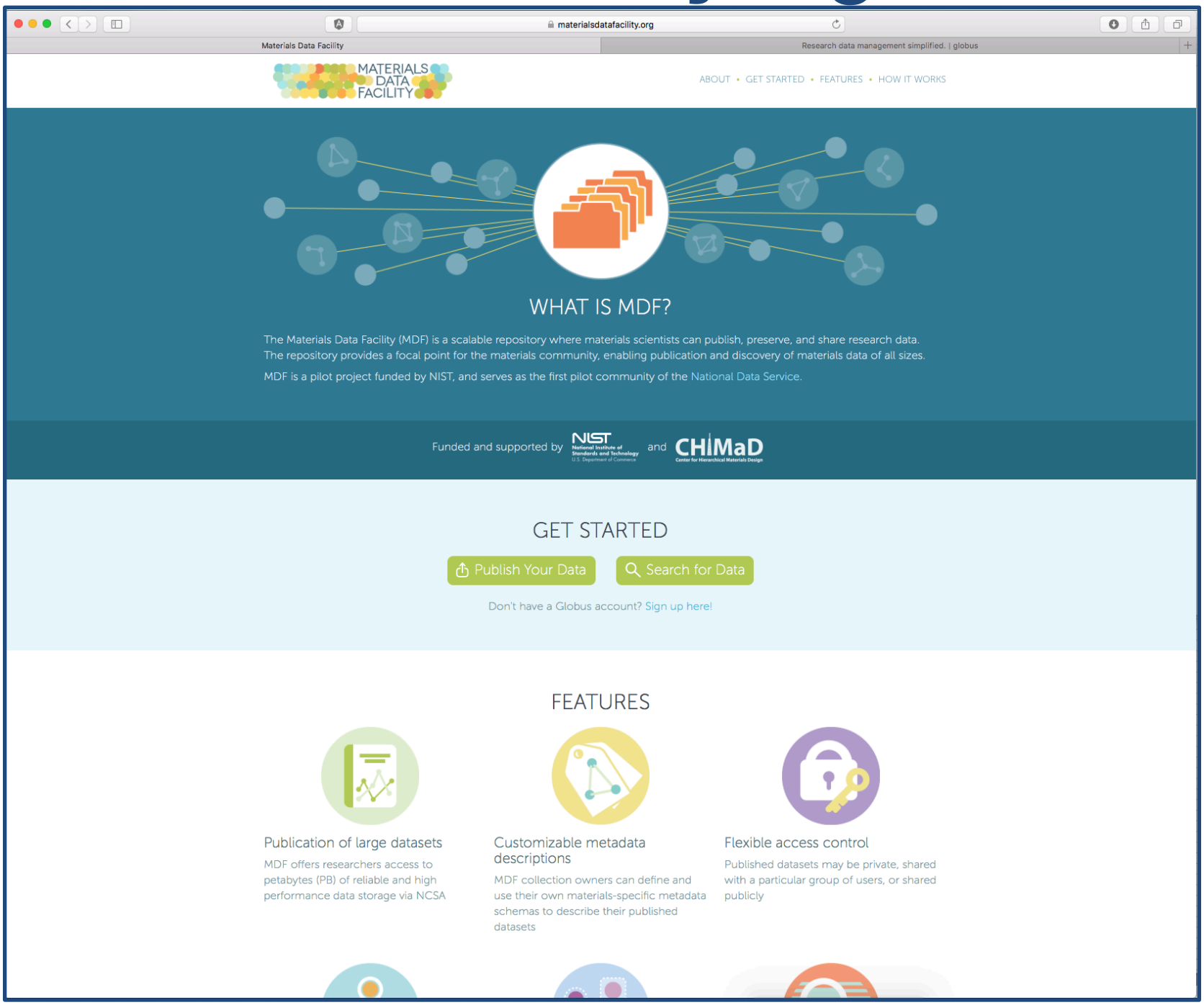
Future...

- **Can we build a system that allows schema:**
  - **Inheritance**
    - E.g. a schema “polymers” might inherit and expand upon the “base material” of NIST
  - **Versioning**
    - E.g. Understand contextually how to map fields between versions
  - **Dependence**
    - E.g. Allows the ability to build consensus around schemas

Discover Research Datasets

- Search on file metadata, custom metadata, and indexed file-level data
- Goal: Intuitive search (e.g. Google-style) with support for more complex range queries and faceting (e.g. Amazon-style)

MaterialsDataFacility.org



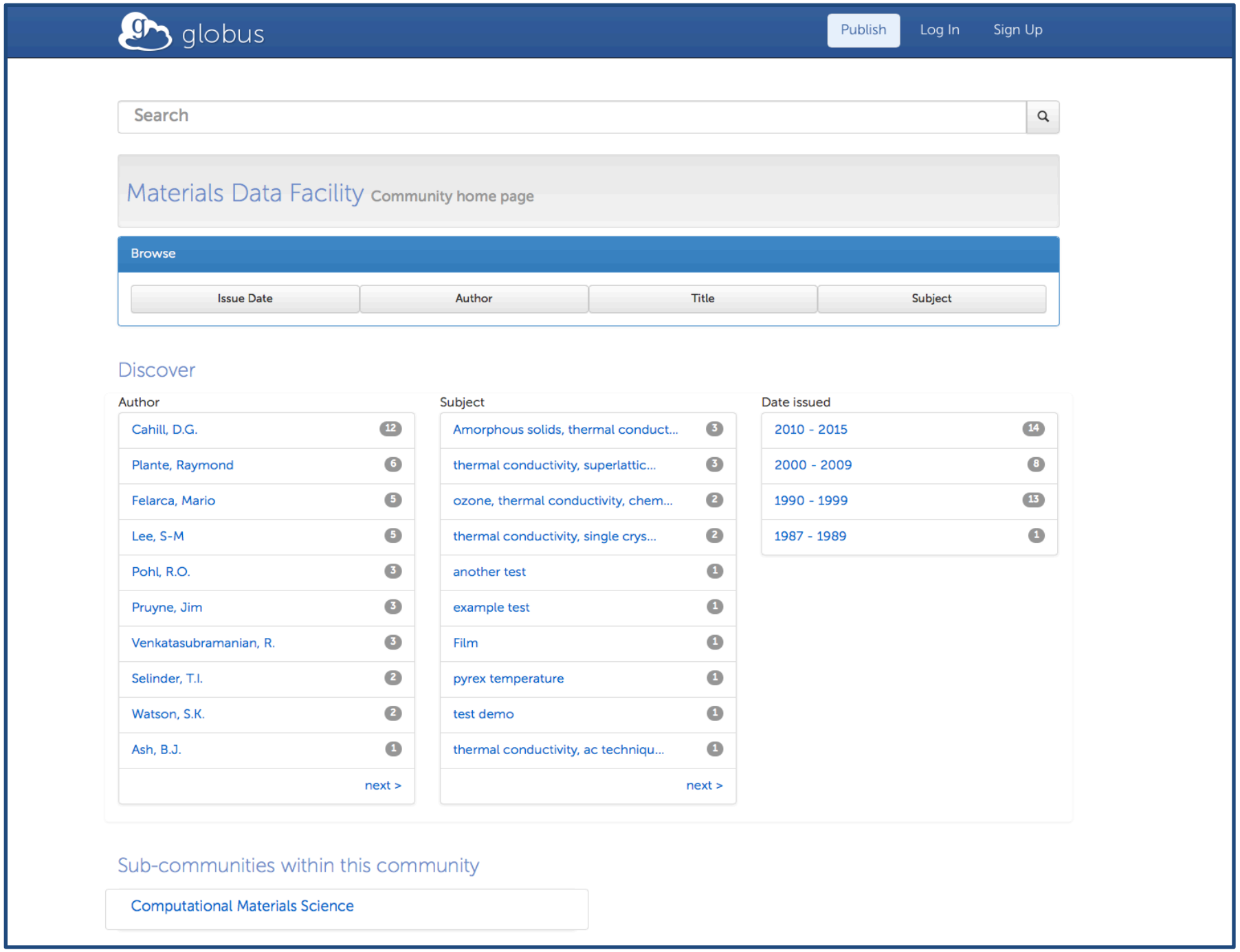
MDF Submission Walkthrough

Example Use Case

Publishing Big, Remote Data

- A research group is generating “big data” at a light source in Switzerland
- The researchers need to move o(50 TB) back to their home institution storage resources for analysis and archiving
- The researchers want to bundle multiple experimental runs into a single data publication with overarching metadata provenance
- The group PI wants to be able to verify the dataset and metadata contents before publication
- The research team wants a citable DOI so they can transparently share the raw and derived datasets with the community after publication
- The research team wants their data to be discoverable by both free text searching and custom metadata fields

MDF Collection Home



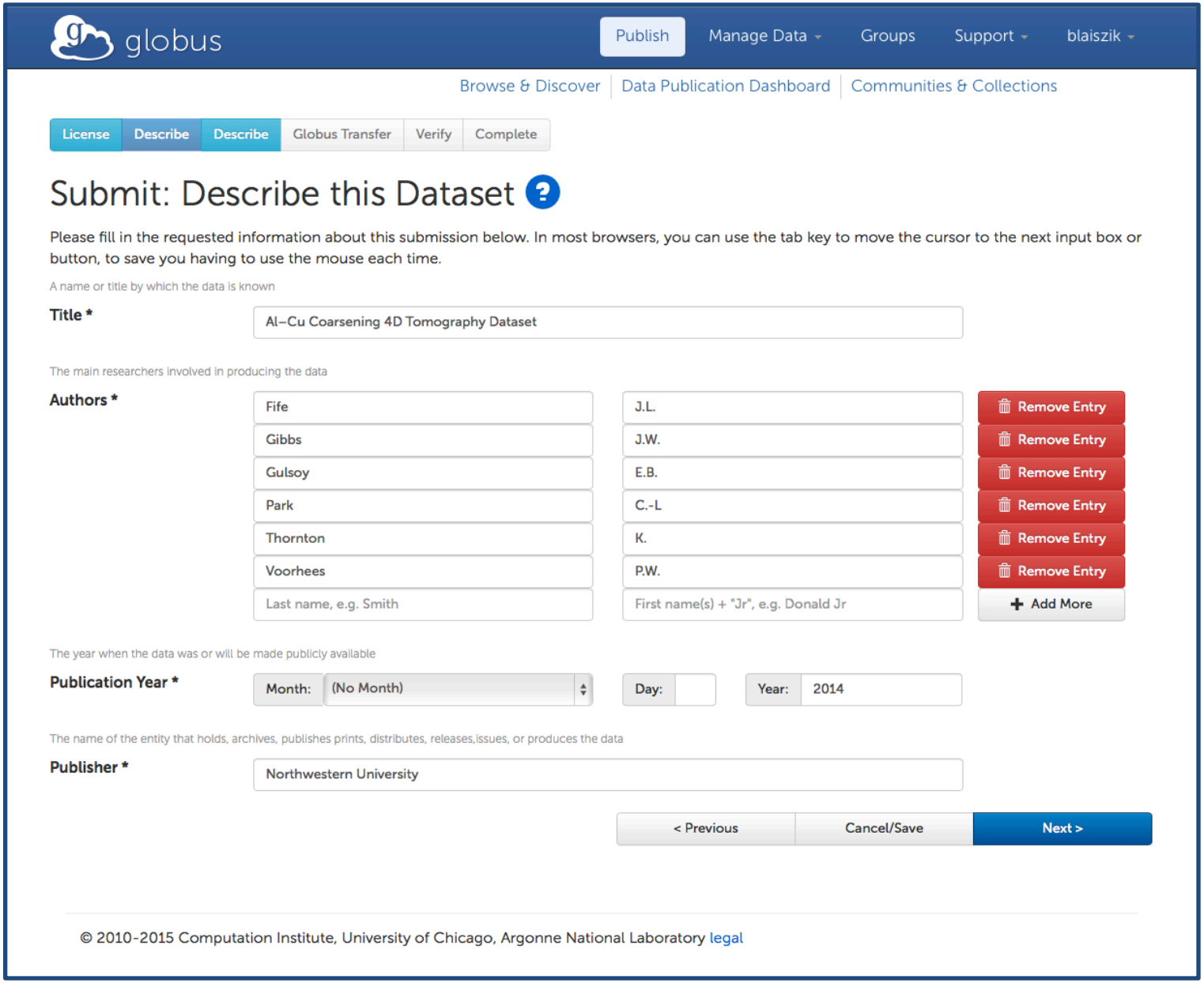
MDF Collections

Recall: Policies Set at the Collection Level

- Required metadata, schemas
- Data storage location
- Metadata curation policies

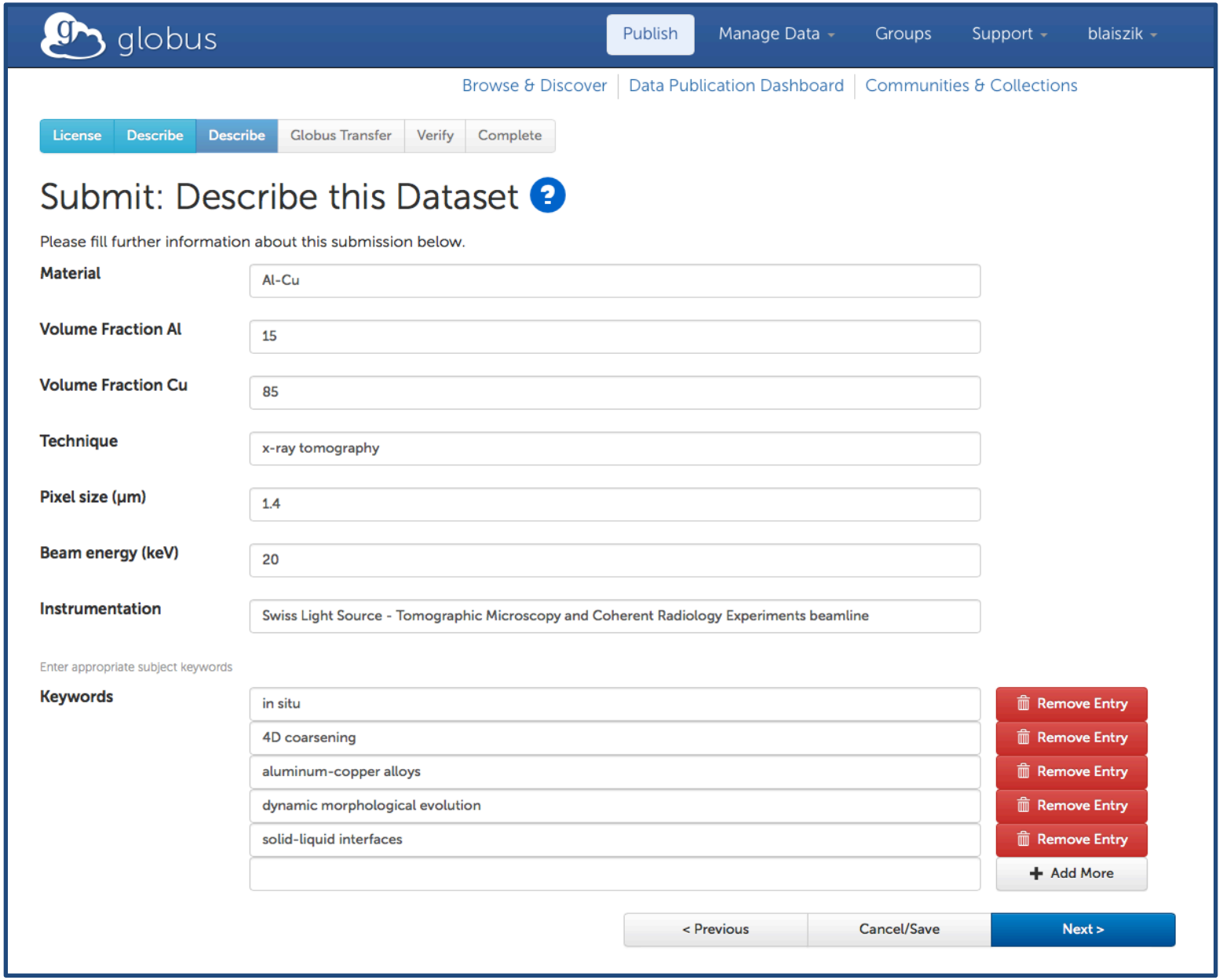
MDF Metadata Entry

- **Scientist or representative describes the data they are submitting**
- **For this collection Dublin Core and a custom metadata template are required**



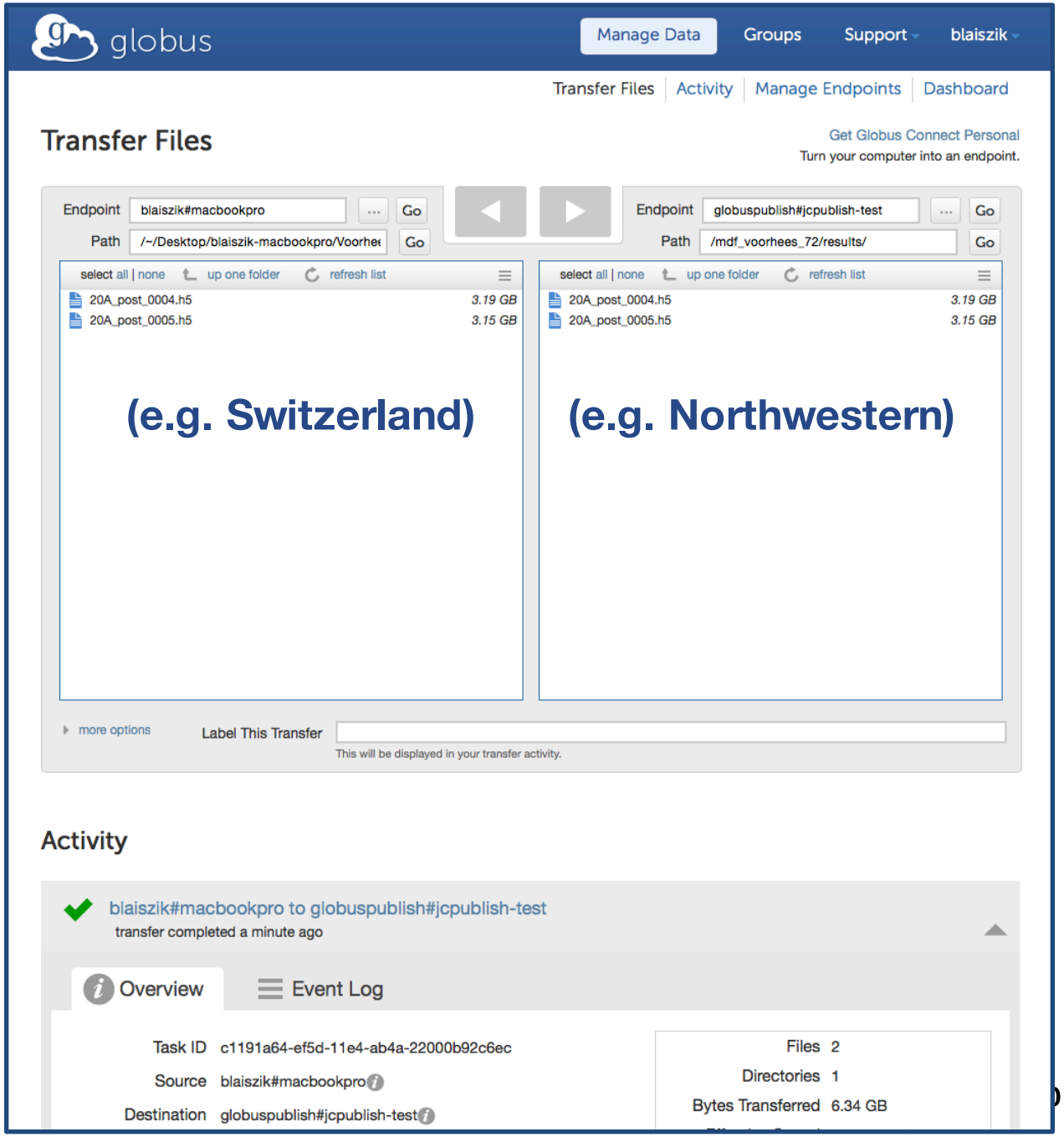
MDF Custom Metadata

- **Scientist or representative describes the data they are submitting**
- **For this collection Dublin Core and a custom metadata template are required**



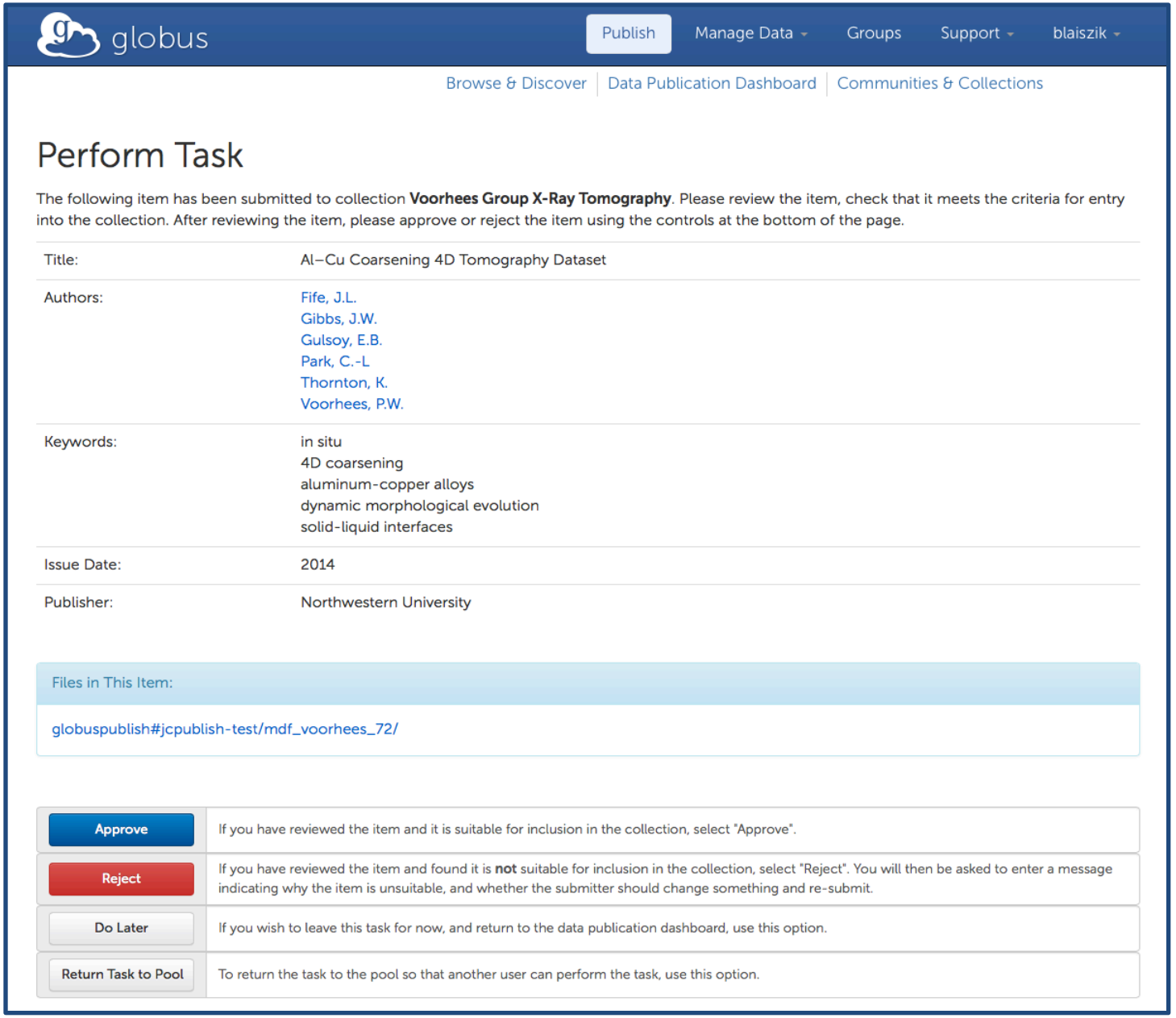
Dataset Assembly

- Shared endpoint is auto-created on collection-specified data store
- Scientist transfers dataset files to a unique publish endpoint
- Dataset may be assembled over any period of time
- When submission is finished, dataset will be rendered immutable via checksum

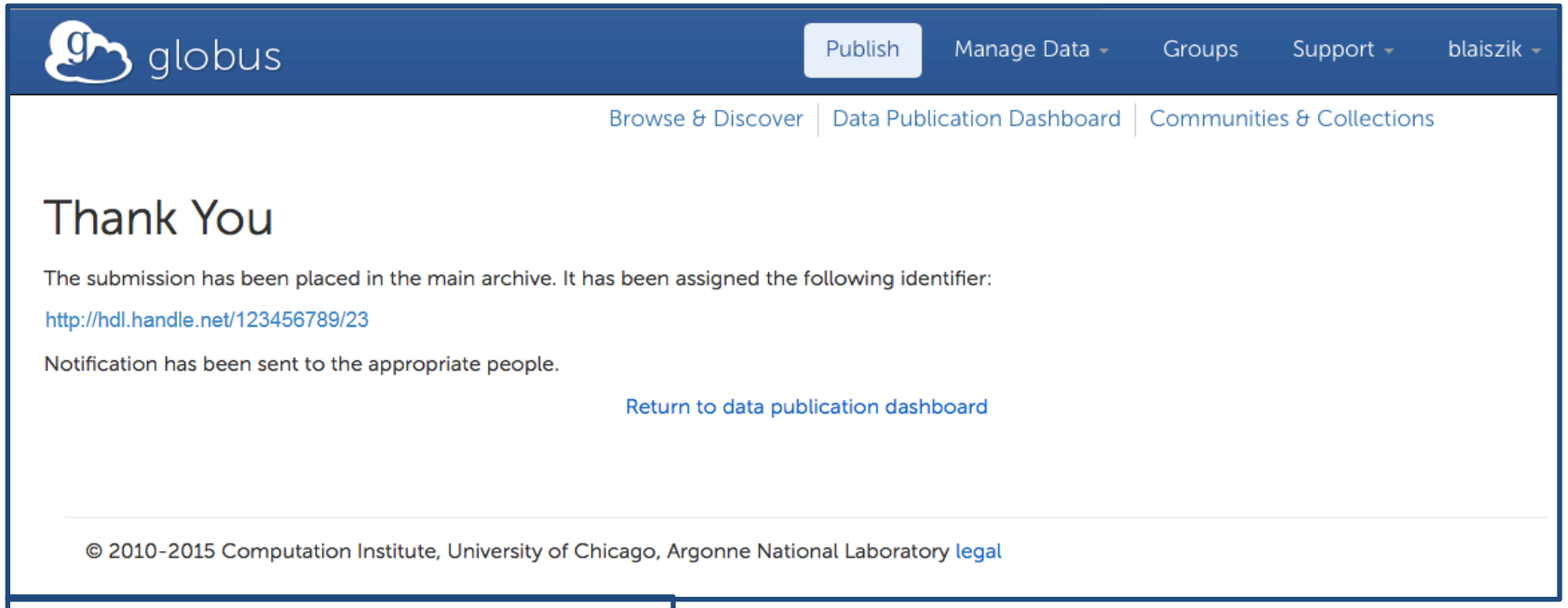


Dataset Curation

- **Optionally specified in collection configuration**
- **Can be approved or rejected (i.e. sent back to the submitter)**

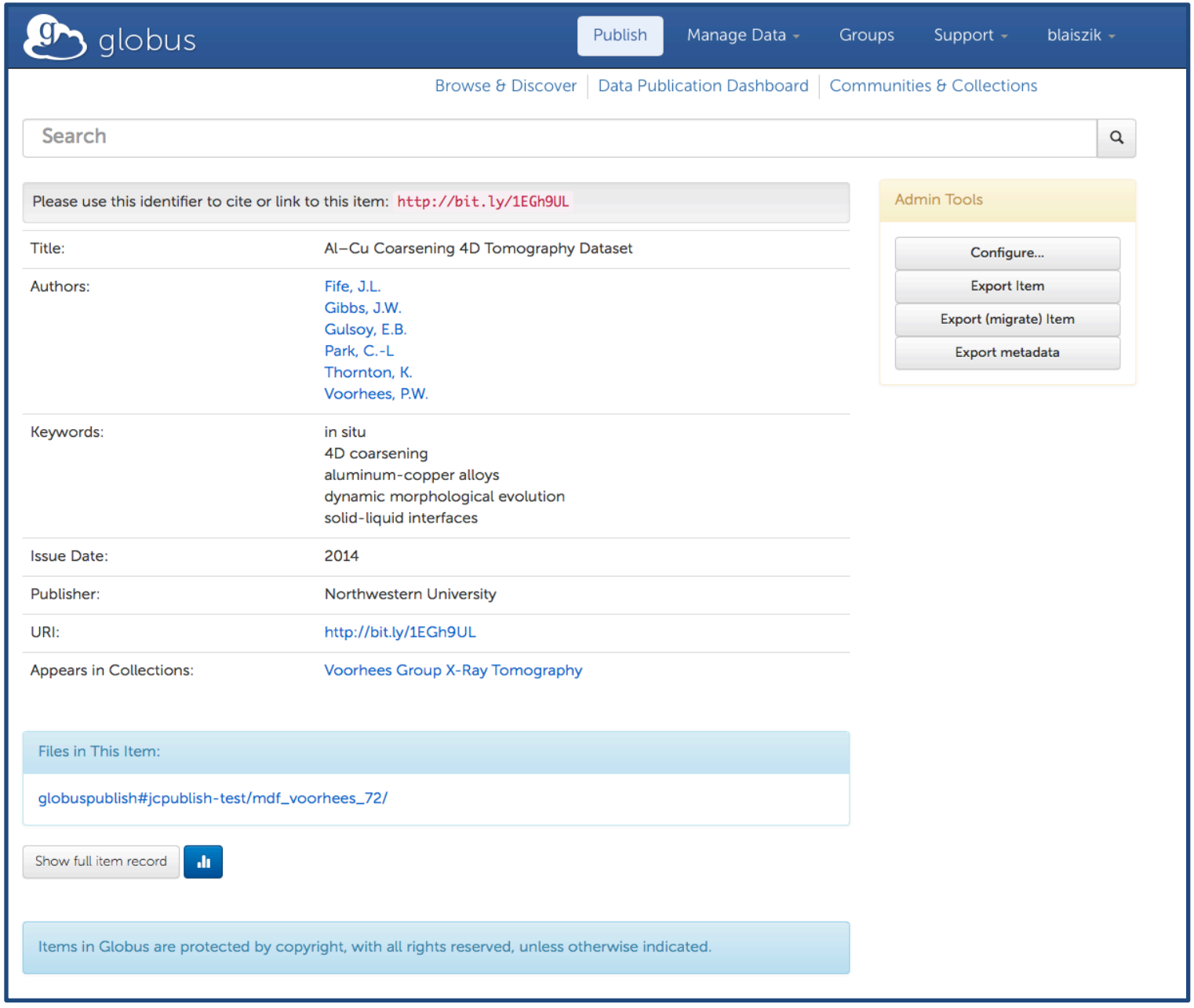


Mint a Permanent Identifier

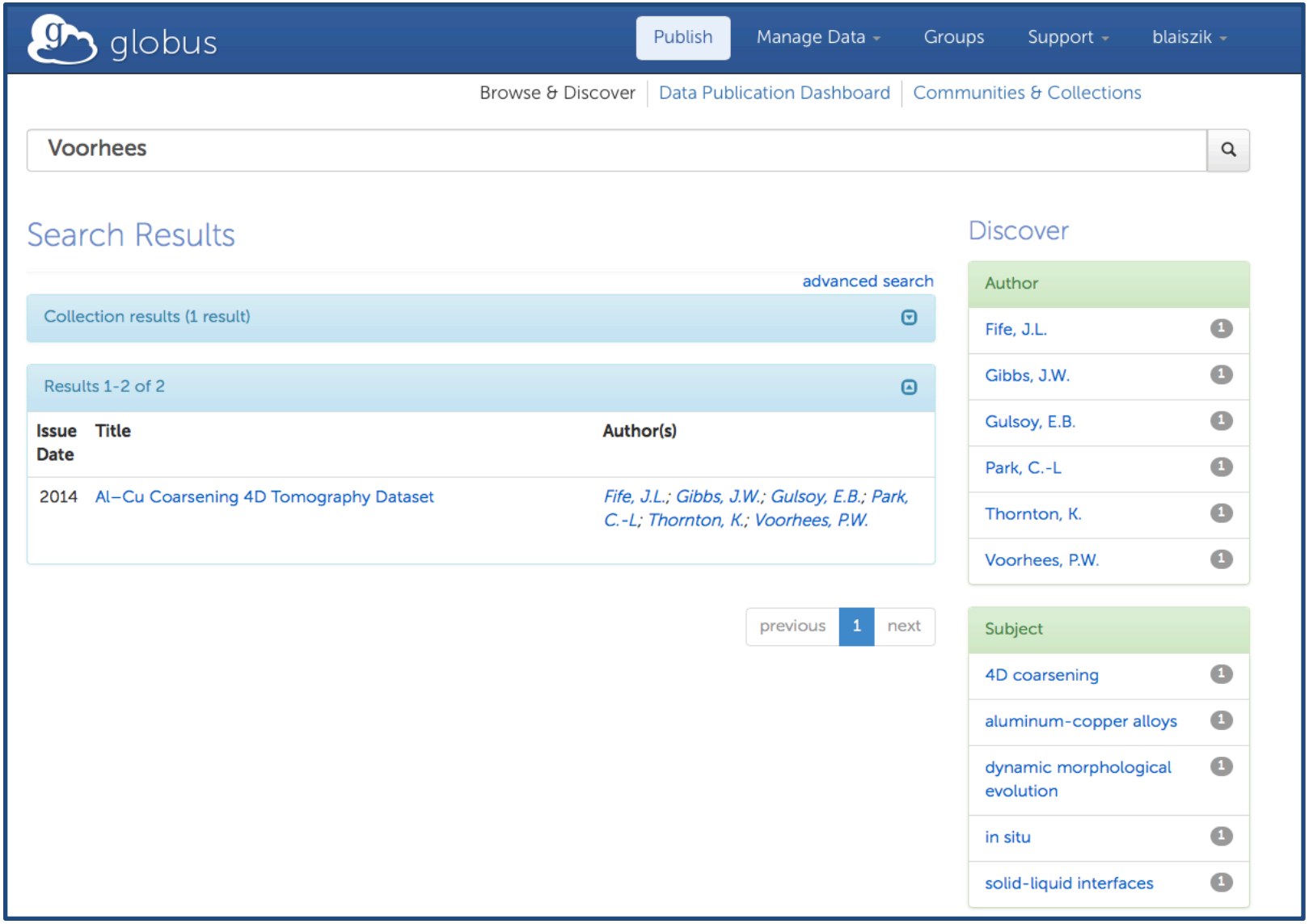


Can optionally be DOI or Handle

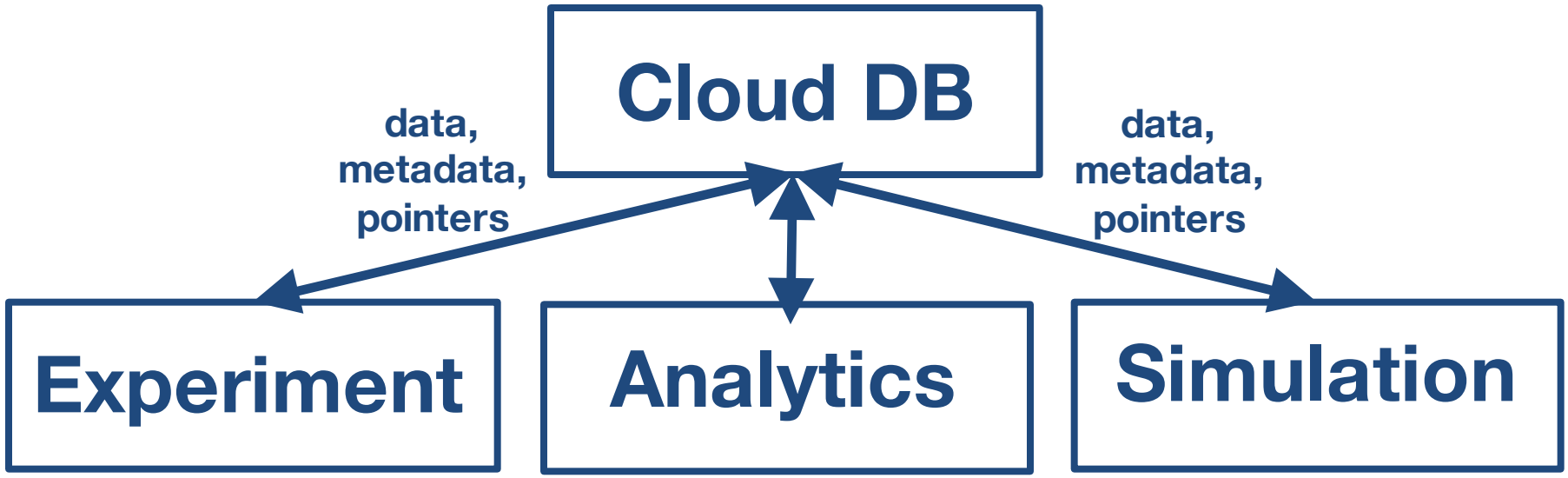
Dataset Record



Dataset Discovery



Data Interaction



Catalog

alpha software

Catalog → Datasets → Members
Datasets/Members have typed tags

- **Group** data based on features
  - Logical grouping to organize, search, etc.
  - Abstracts data location
- **Operate** on datasets as units
  - Copy, export, analyze, annotate, archive
- **Tag** datasets to reflect content
  - Capture metadata and provenance
    - i.e. all parameters for data file generation
  - Rich key:value pairing for advanced searching
- **Share/move** datasets
  - Fine-grained access control to data and metadata

iPhoto, ca. 2014

vs

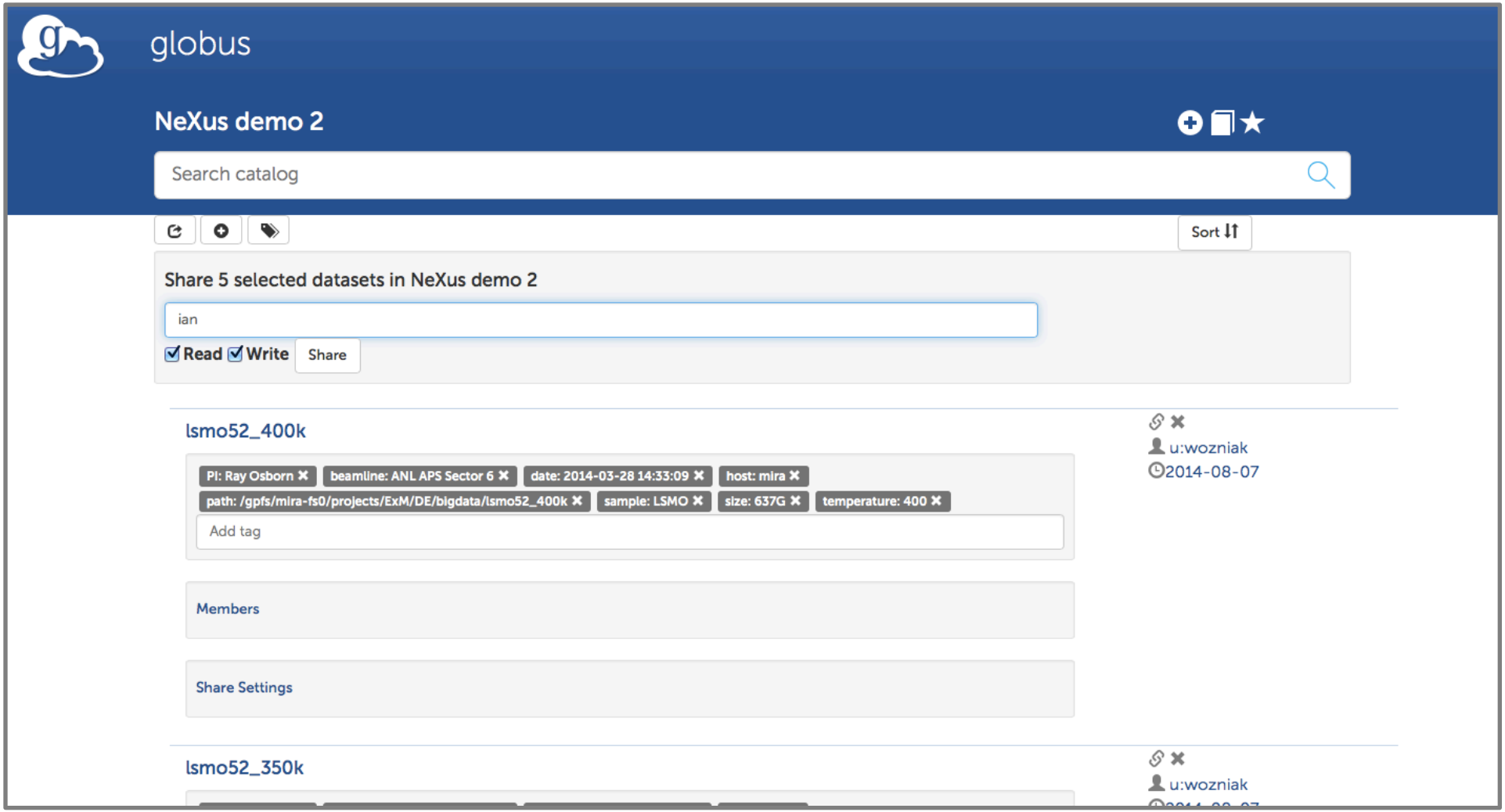
```
drwxr-xr-x  11 argonne  staff    374 Nov 11 10:41 .
drwxr-xr-x   5 argonne  staff    170 Nov 13 17:06 ..
-rw-r--r--   1 argonne  staff      0 Nov 11 10:41 __init__.py
drwxr-xr-x   4 argonne  staff    136 Nov 11 10:41 ca
-rw-r--r--   1 argonne  staff  12967 Nov 11 10:41 dataset_client.py
drwxr-xr-x  12 argonne  staff    408 Jan 22 15:43 examples
-rwxr-xr-x   1 argonne  staff   4670 Nov 11 10:41 goauth.py
-rw-r--r--   1 argonne  staff   2848 Nov 11 10:41 operators.py
-rw-r--r--   1 argonne  staff   7584 Nov 11 10:41 rest_client.py
drwxr-xr-x   4 argonne  staff    136 Nov 11 10:41 test
-rw-r--r--   1 argonne  staff   6975 Nov 11 10:41 verified_https.py
```

Standard file/directory structure

Catalog

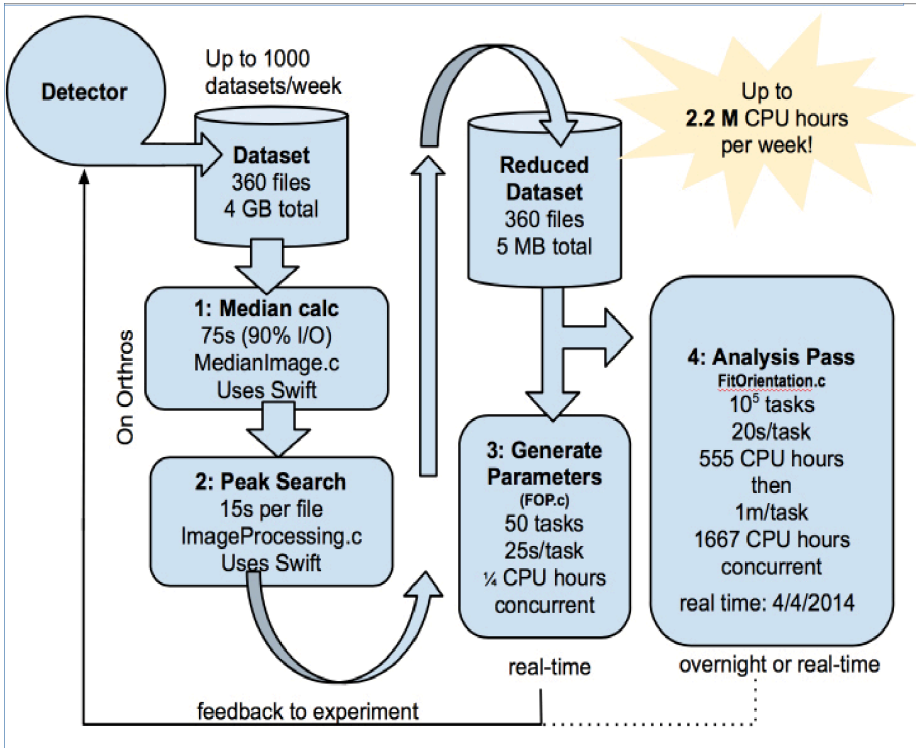
alpha software

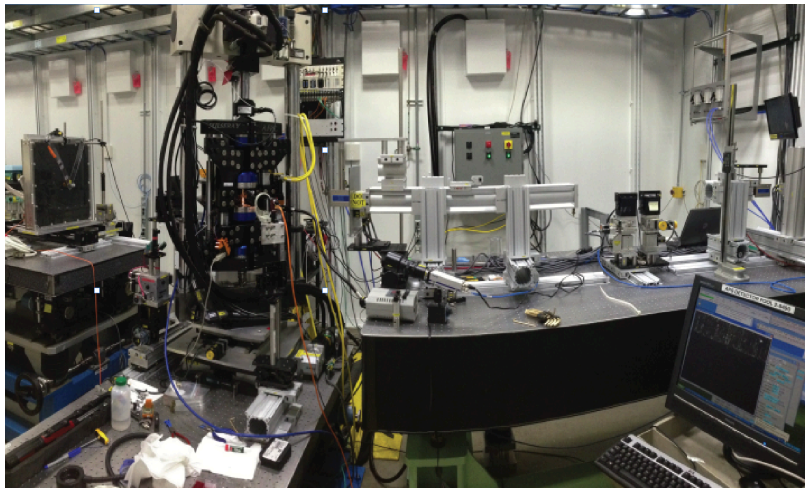
Catalog → Datasets → Members
Datasets/Members have typed tags

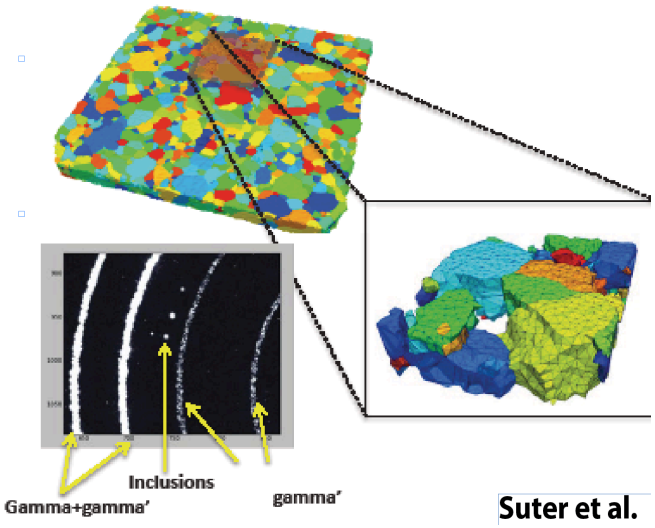


Workflows and Integrations

- HEDM experiments guided by *in operando* computational results
- parallel-enabled workflow to leverage HPC resources and dramatically reduce time from data gathering to actionable data analyses







Suter et al.

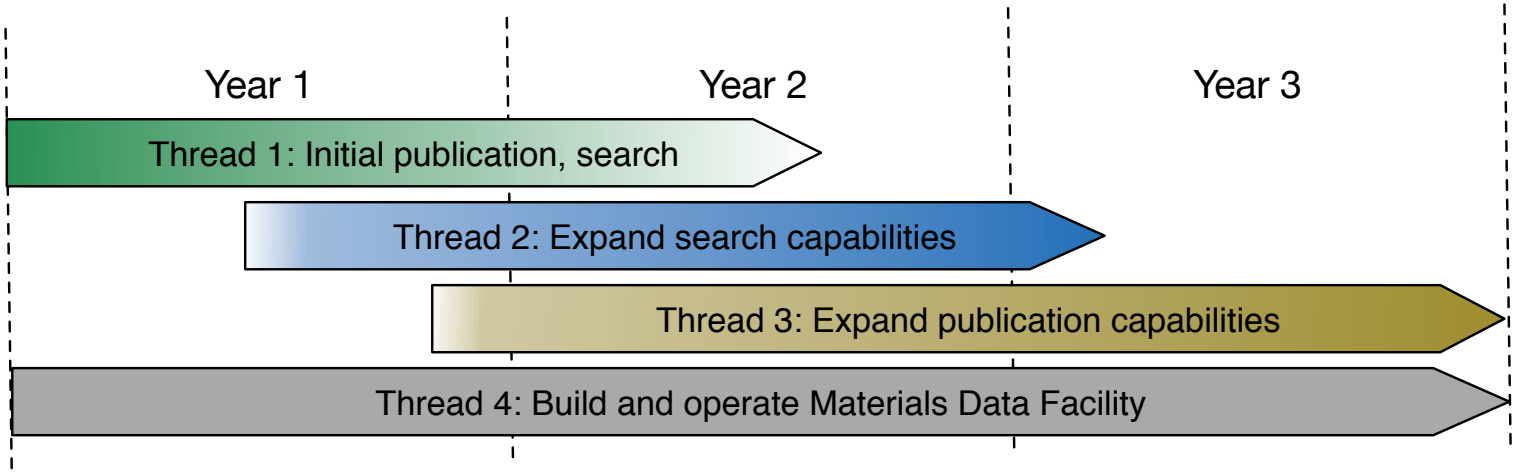
What's Available?

- Web interface to support data publication via Globus platform (identify management, user groups, optimized big data transfer)
- 100 TB of storage here at NCSA (scalable to 1 PB)
- Help with developing metadata schemas to describe your research datasets

What are we looking for?

- Early adopters, willing to get their hands dirty with the service and give honest feedback
- Key datasets of all sizes, shapes, raw or derived, that might help us better understand the process

First Steps



- Establish MDF with storage at Argonne and NCSA
- Identify datasets to pilot publication pipelines
- Engage with researches working with materials data to understand use cases and learn friction point
- Please talk to us if you have data you want to share, publish, discover, …
- Globus tutorials (identity, transfer, sharing): https://github.com/globusonline/globus-tutorials

Thanks to Our Sponsors!

U.S. DEPARTMENT OF ENERGY

THE UNIVERSITY OF CHICAGO