Materials Data Facility - Data Services to Advance Materials Science Research

Ben Blaiszik (blaiszik@uchicago.edu),
Ian Foster (foster@uchicago.edu)
Kyle Chard, Rachana Ananthakrishnan, Steven Tuecke
Michael Ondrejcek, Kenton McHenry, John Towns

materialsdatafacility.org
globus.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
  - MDF
  - Globus quick introduction
- Data Publication
  - Key MDF features
  - Publication walk-through
  - Early lessons learned
- Materials Resource Registration
- Next Steps

Lessons Learned this Week

- Finding good materials datasets from individual groups is hard
- Finding well-described materials datasets is hard
- Combining materials datasets across groups is harder
- Securing rights and permissions to those datasets for mining and sharing is hard

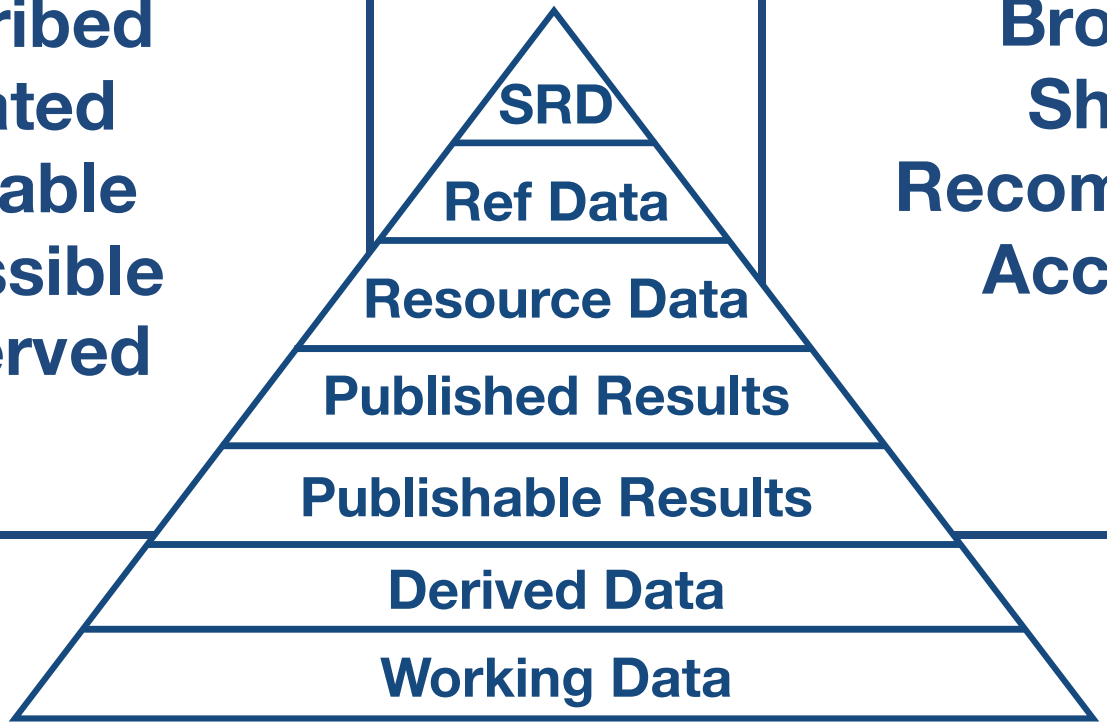
What is MDF?

We are developing services to make it more simple for materials datasets and resources to be ...

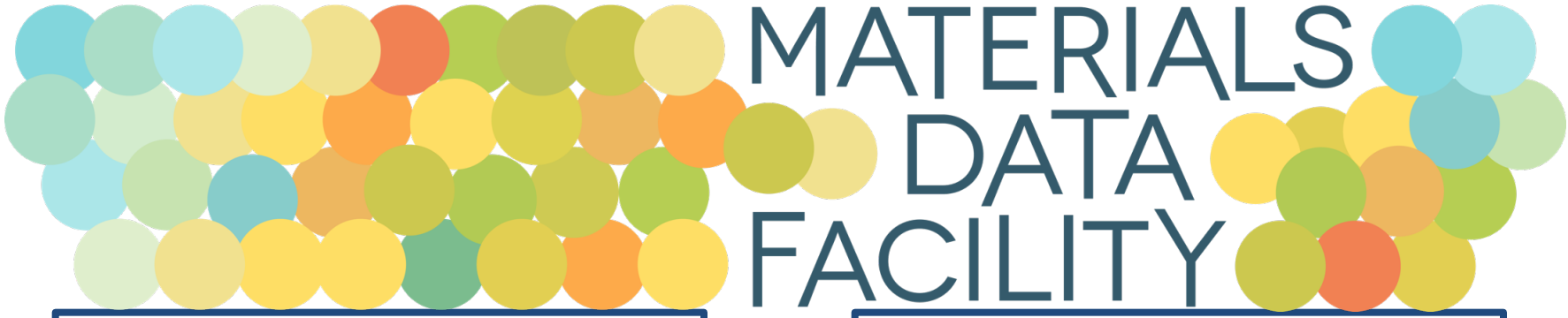
Published
Identified
Described
Curated
Verifiable
Accessible
Preserved

and

Discovered
Searched
Browsed
Shared
Recommended
Accessed



Data Service Infrastructure



Publication +

Discovery +

APIs

Compute for data interaction and viz

Resource Registration

+ - Initial Foci

Publication

Opened to external users in mid Feb 2016

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

Discovery

Coming late 2016-ish

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

Future...

Spotlight for all data you have access to regardless of location

Resource Registration

Coming Q2 2016 via collaboration w/ NIST

- Find existing, widely distributed, materials resources
- MDF will run an instance of MRR, currently populating before making widely available

[Dima, Youssef, et al. @ NIST]

Globus Background

Globus Platform-as-a-Service (PaaS)

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

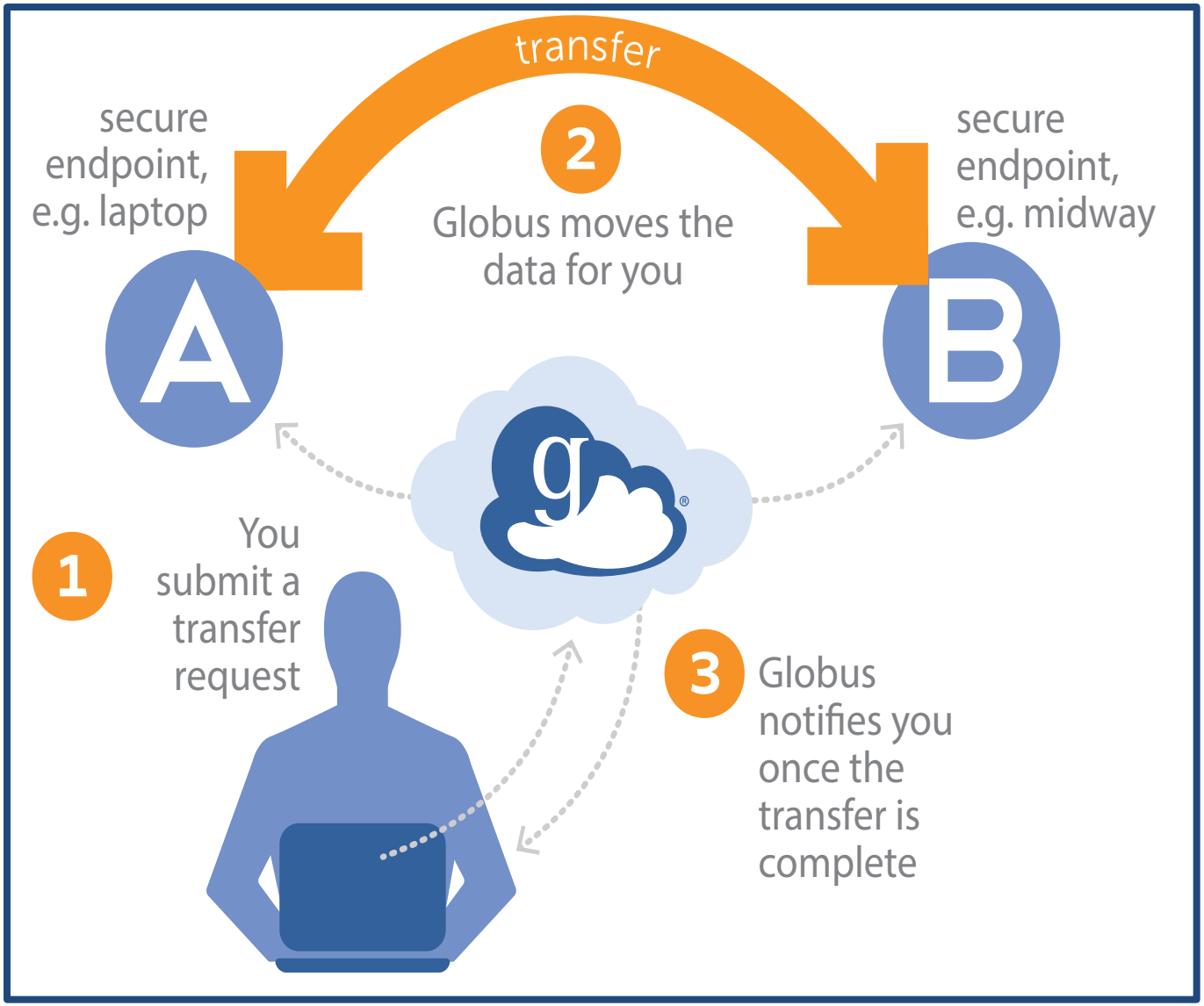
Globus Background

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



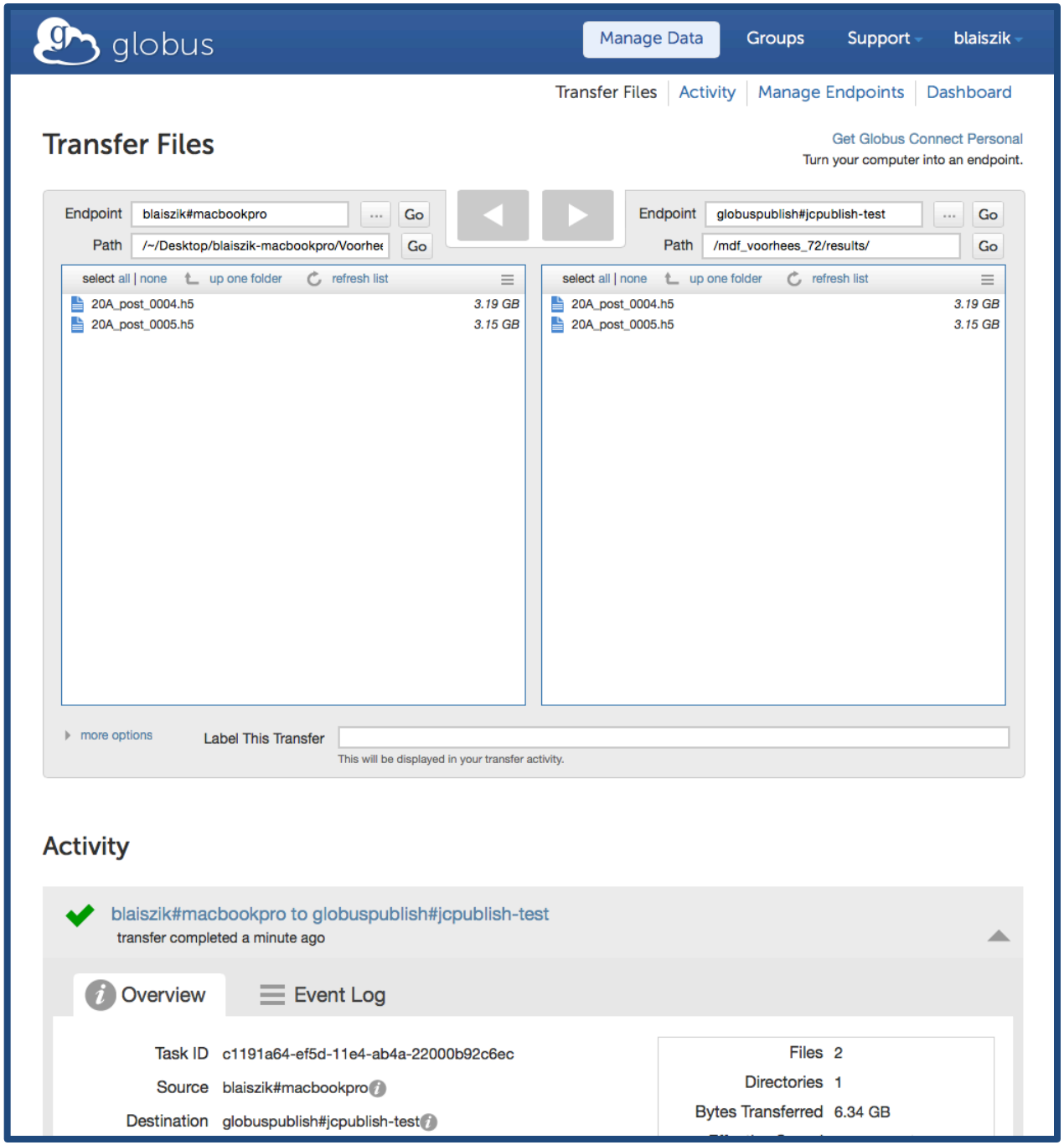
Globus Web UI

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



Globus APIs

- New version of core services released in Feb.
- New Python SDK available
  - https://github.com/globusonline/globus-sdk-python
- Jupyter Notebook Examples
  - https://github.com/globus/globus-jupyter-notebooks

Endpoint search

Globus has over 8000 registered endpoints. To find endpoints of interest you can access powerful search capabilities via the SDK. For example, to search for a given string across the descriptive fields of endpoints (names, description, keywords):

```
search_str = "Globus Tutorial Endpoint"
endpoints = tc.endpoint_search(search_str)
print("==== Displaying endpoint matches for search: '{}' ===".format(search_str))
for ep in endpoints:
    print("{} ({})".format(ep["display_name"] or ep["canonical_name"], ep["id"]))
```

Restricting search scope with filters

There are also a number of default filters to restrict the search for 'my-endpoints', 'my-gcp-endpoints', 'recently-used', 'in-use', 'shared-by-me','shared-with-me')

```
search_str = None
endpoints = tc.endpoint_search(
        filter_fulltext=search_str, filter_scope="recently-used")
for ep in endpoints:
    print("{} ({})".format(ep["display_name"] or ep["canonical_name"], ep["id"]))
```

Endpoint details

You can also retrieve complete information about an endpoint, including name, owner, location, and server configurations.

```
endpoint = tc.get_endpoint(tutorial_endpoint_1)
print("Display name:", endpoint["display_name"])
print("Owner:", endpoint["owner_string"])
print("ID:", endpoint["id"])
```

Transfer

Creating a transfer is a two stage process. First you must create a description of the data you want to transfer (which also creates a unique submission_id), and then you can submit the request to Globus to transfer that data.

If the submit_transfer fails, you can safely resubmit the same transfer_data again. The submission_id will ensure that this transfer request will be submitted once and only once.

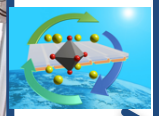
```
# help(tc.submit_transfer)
source_endpoint_id = tutorial_endpoint_1
source_path = "/share/godata/"

dest_endpoint_id = tutorial_endpoint_2
dest_path = "/~/"

label = "My tutorial transfer"

# TransferData() automatically gets a submission_id for once-and-only-once submission
tdata = globus_sdk.TransferData(tc, source_endpoint_id,
                                dest_endpoint_id,
                                label=label)

## Recursively transfer source path contents
tdata.add_item(source_path, dest_path, recursive=True)

## Alternatively, transfer a specific file
# tdata.add_item("/source/path/file.txt",
#                "/dest/path/file.txt"))

# Ensure endpoints are activated
tc.endpoint_autoactivate(source_endpoint_id)
tc.endpoint_autoactivate(dest_endpoint_id)

submit_result = tc.submit_transfer(tdata)
print("Task ID:", submit_result["task_id"])
```

Where are we Now?

Data 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

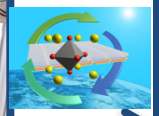
Designing structures that dilate rapidly in both tension and compression would benefit devices such as smart filters, actuators or fasteners. This property however requires an unusual Poisson ratio, or Poisson function at finite strains, which has to vary with applied strain and exceed the familiar bounds: 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

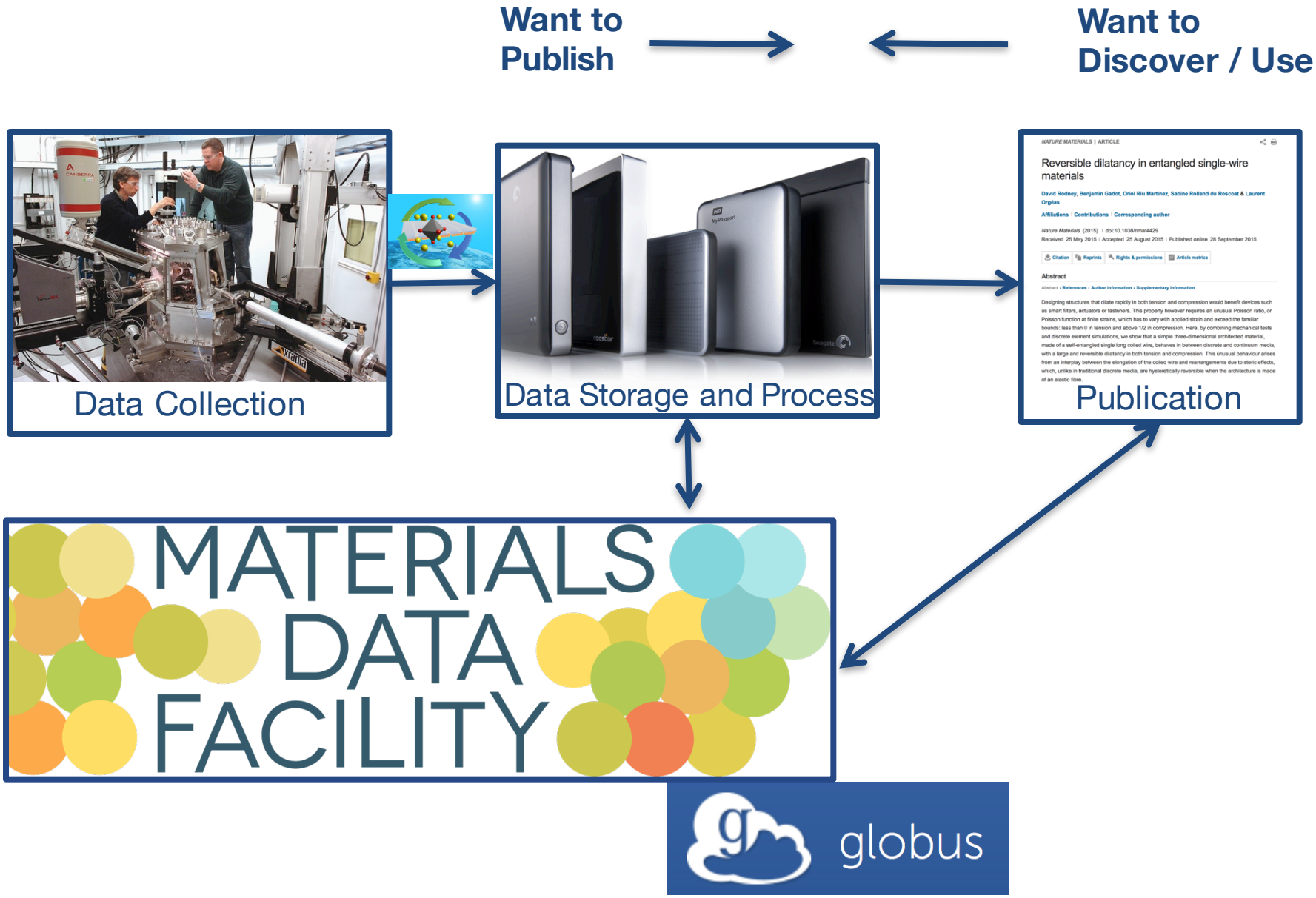


Data Storage and Process

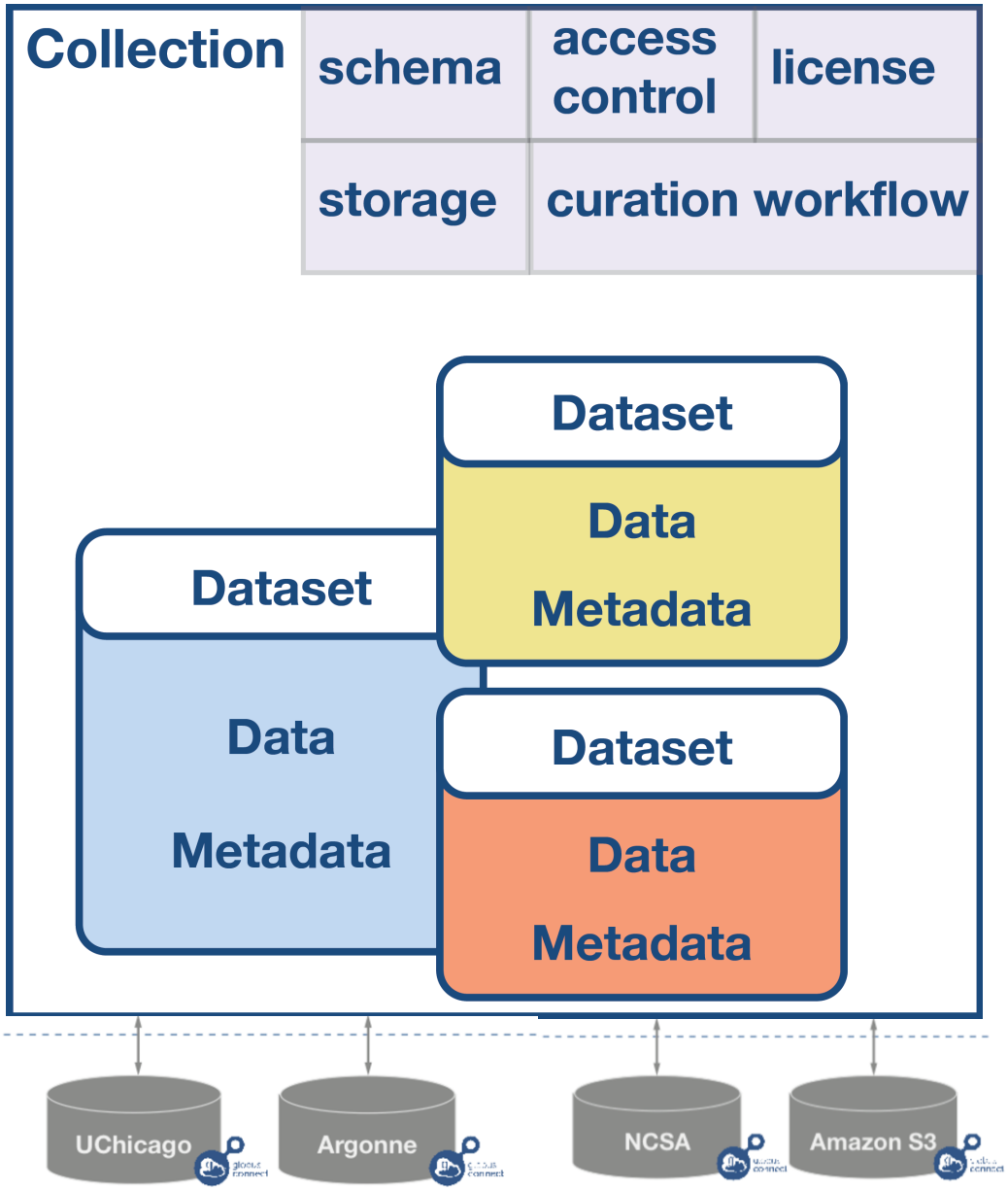
Publication

- 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



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

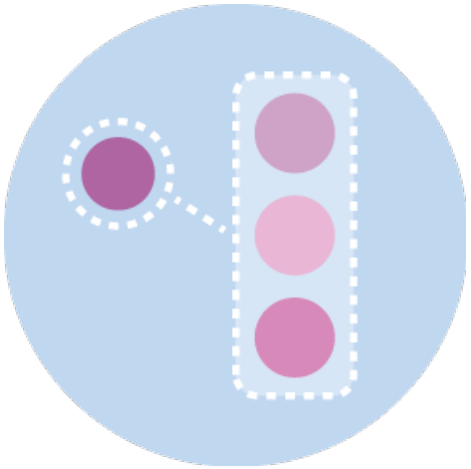
Publish Large Datasets

- Leverages Globus production capabilities for file transfer (i.e. dataset assembly), user authentication, and access control groups

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

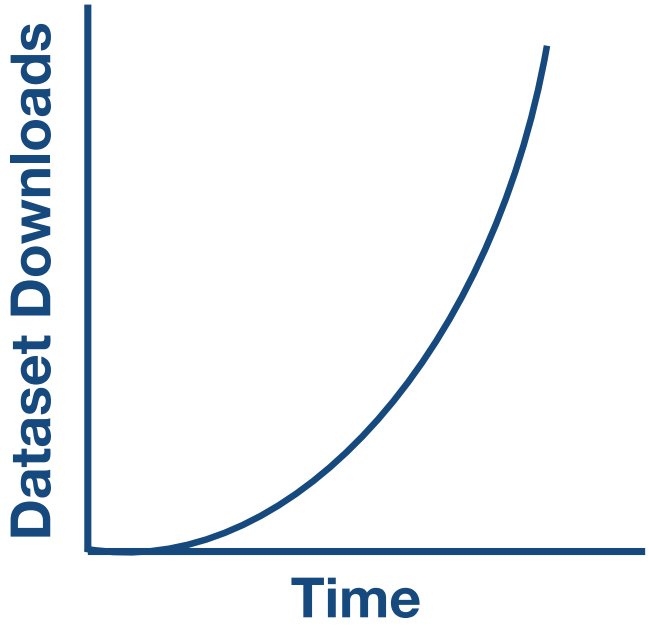
- Researchers can optionally use your own local or institutional storage

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**

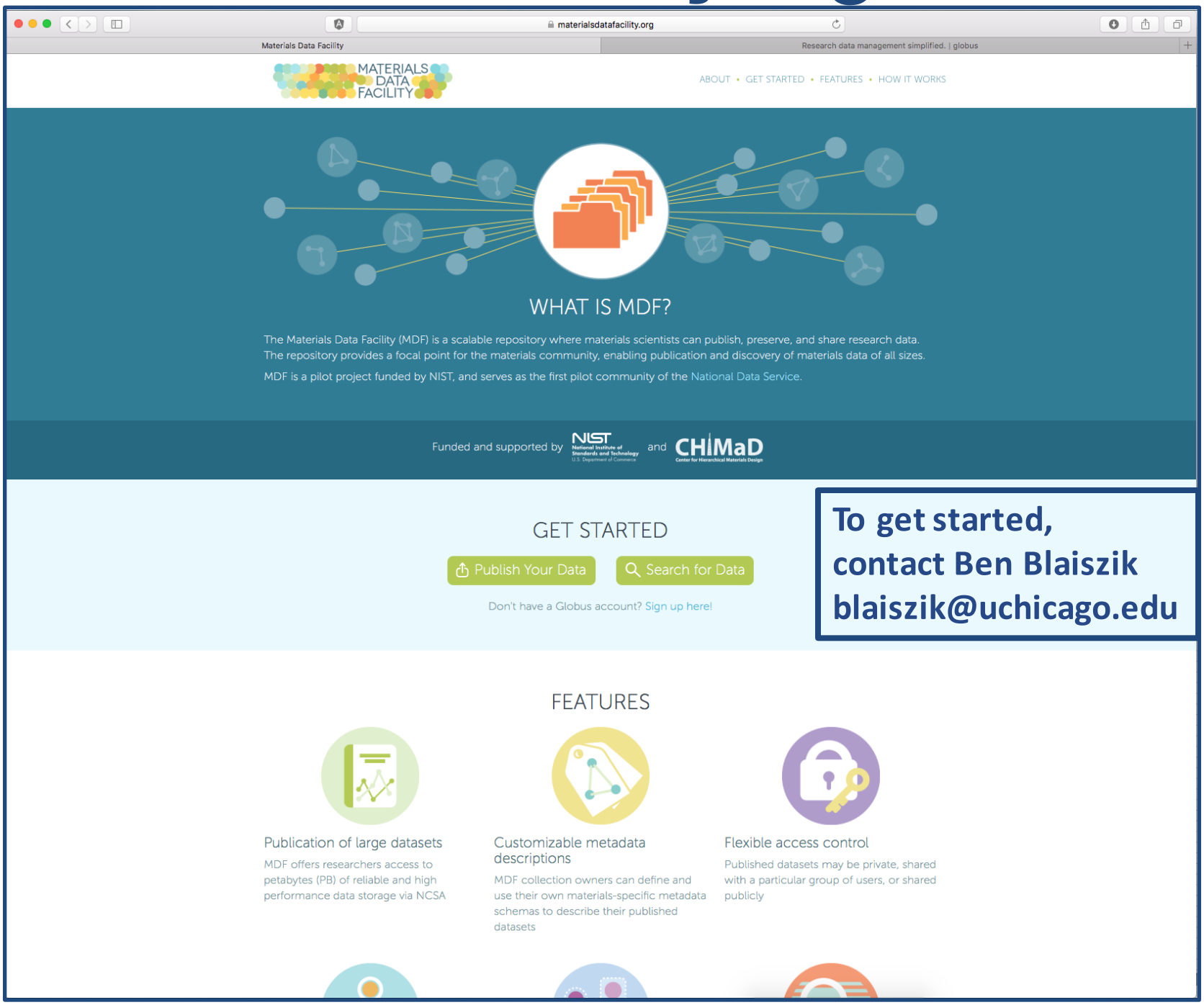
Future...

- **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



To get started,
contact Ben Blaiszik
blaiszik@uchicago.edu

MDF Submission Walkthrough

Example Use Case

Publishing Big, Remote Data

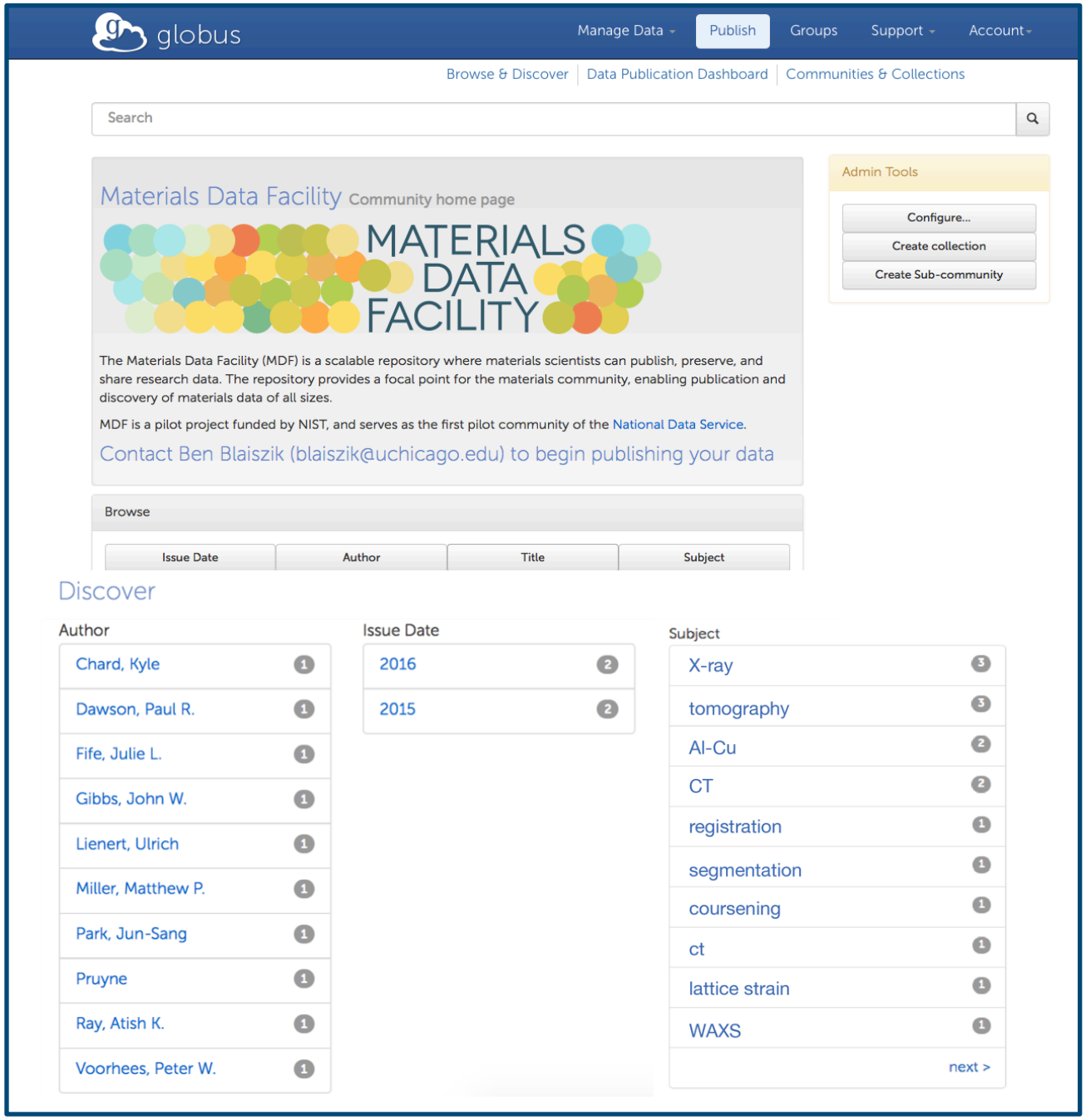
Collected multi TB of data at a light source

Bundle the data with metadata and provenance

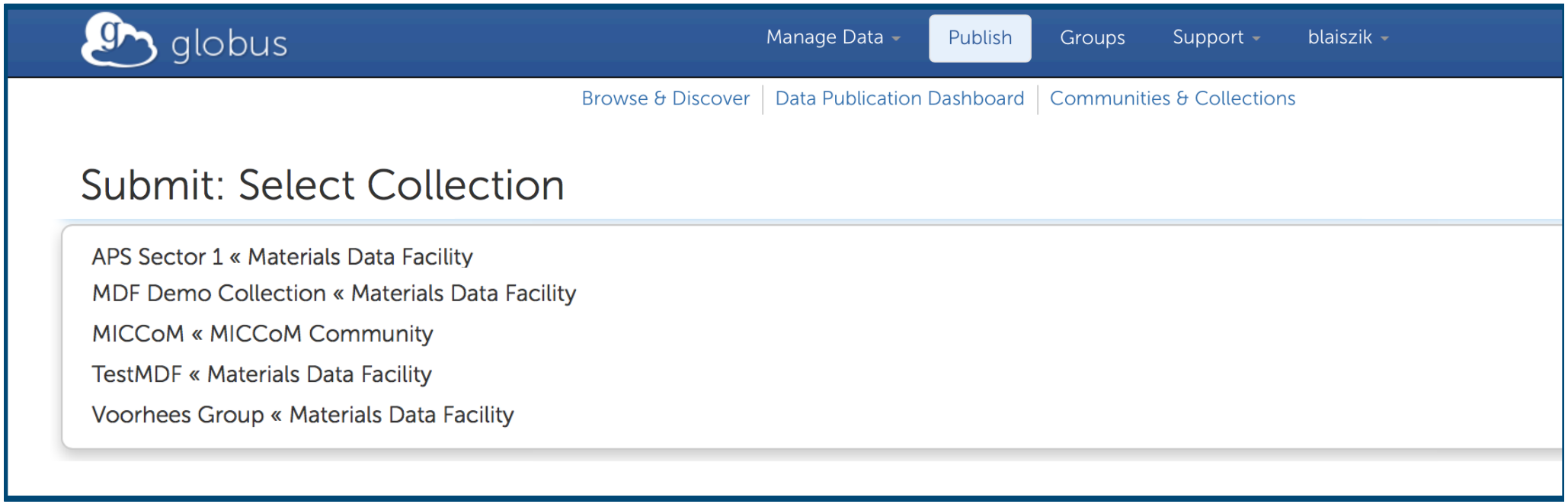
Want a citable DOI to share the raw and derived data with the community

Want their data to be discoverable by free text search and custom metadata

MDF Collection Home



MDF Collections

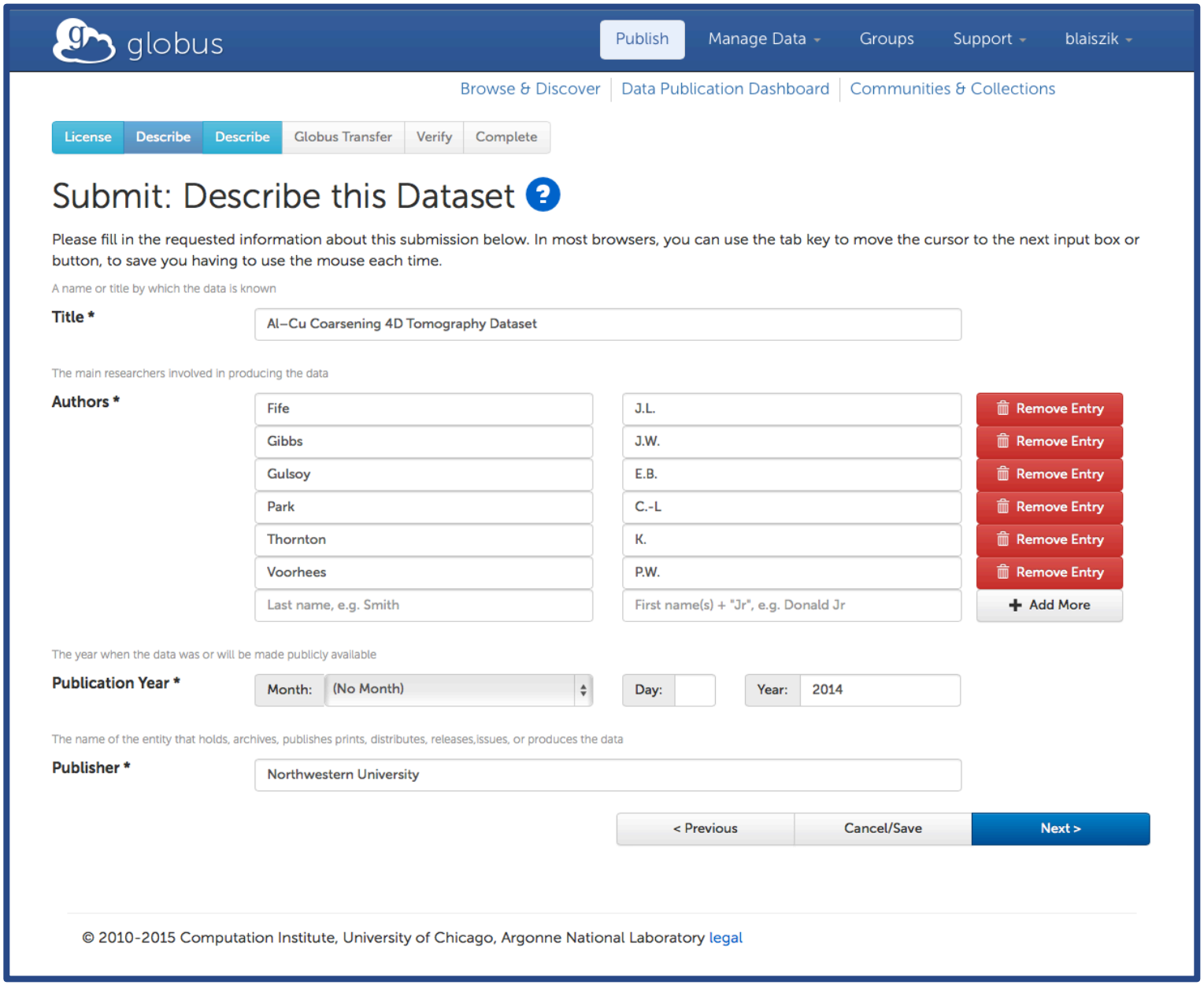


Recall: Policies Set at the Collection Level

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

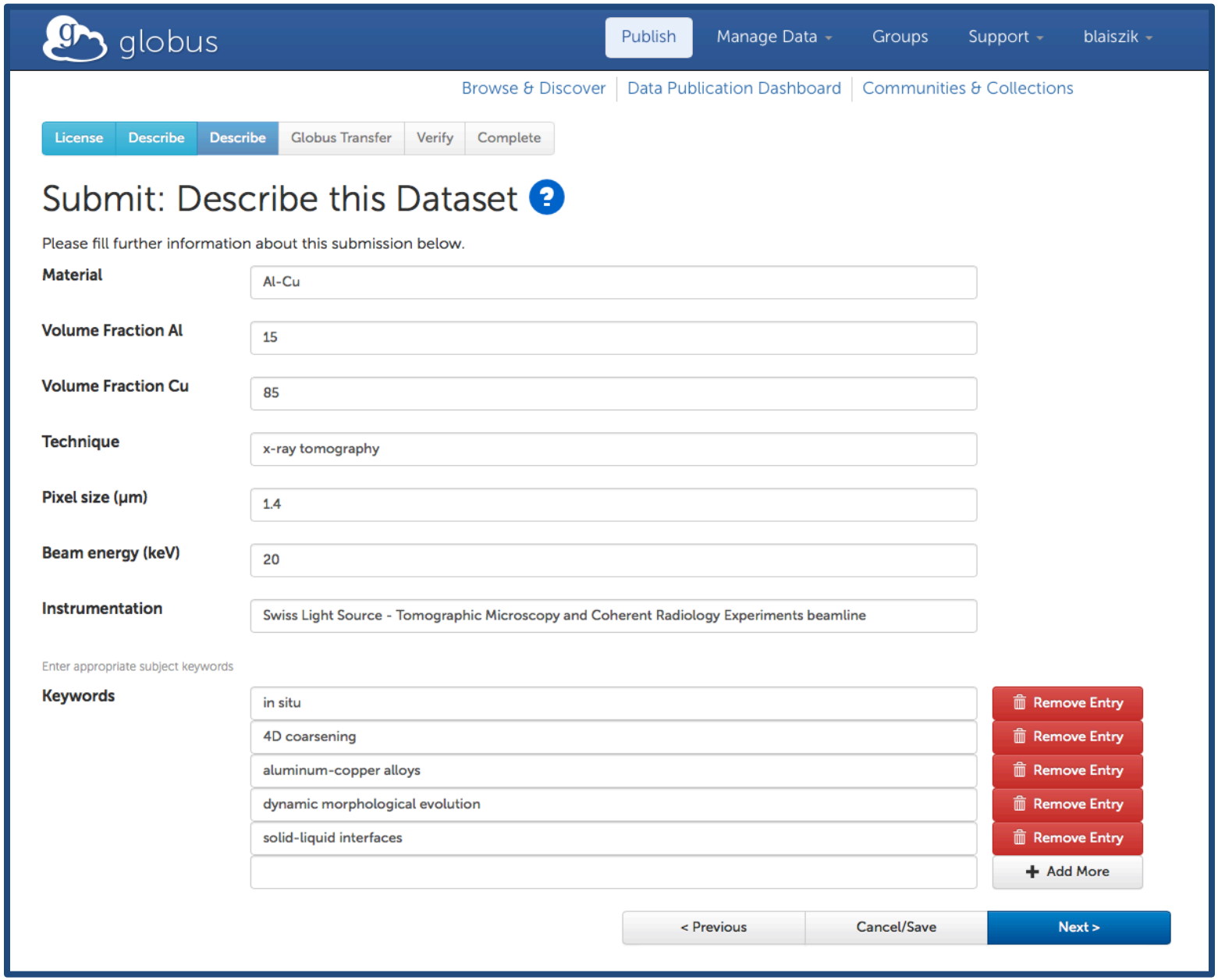
MDF Metadata Entry

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



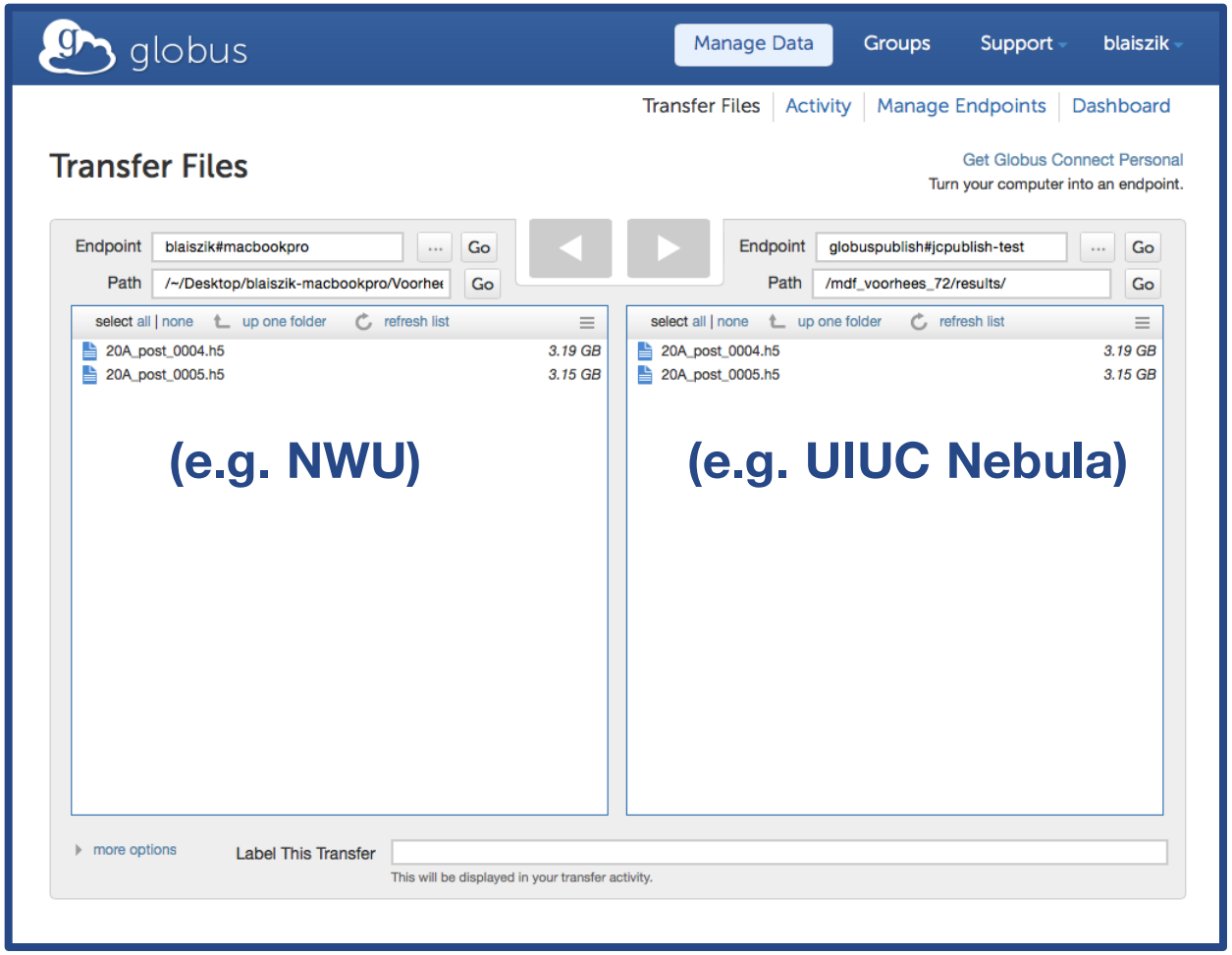
MDF Custom Metadata

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



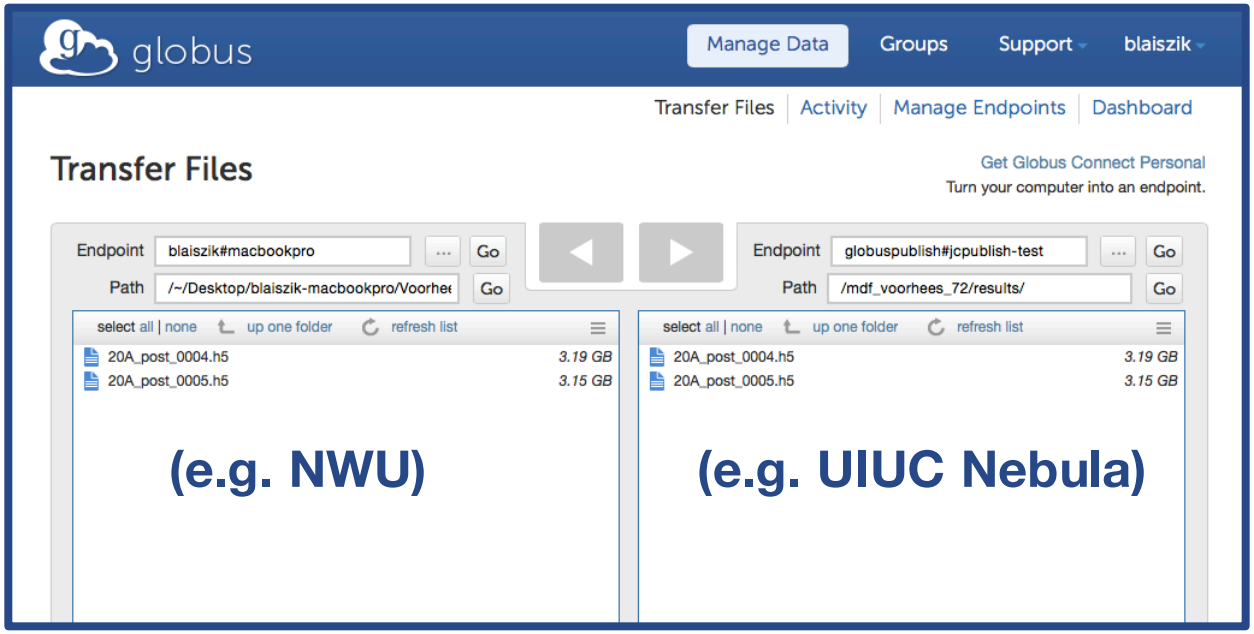
Dataset Assembly

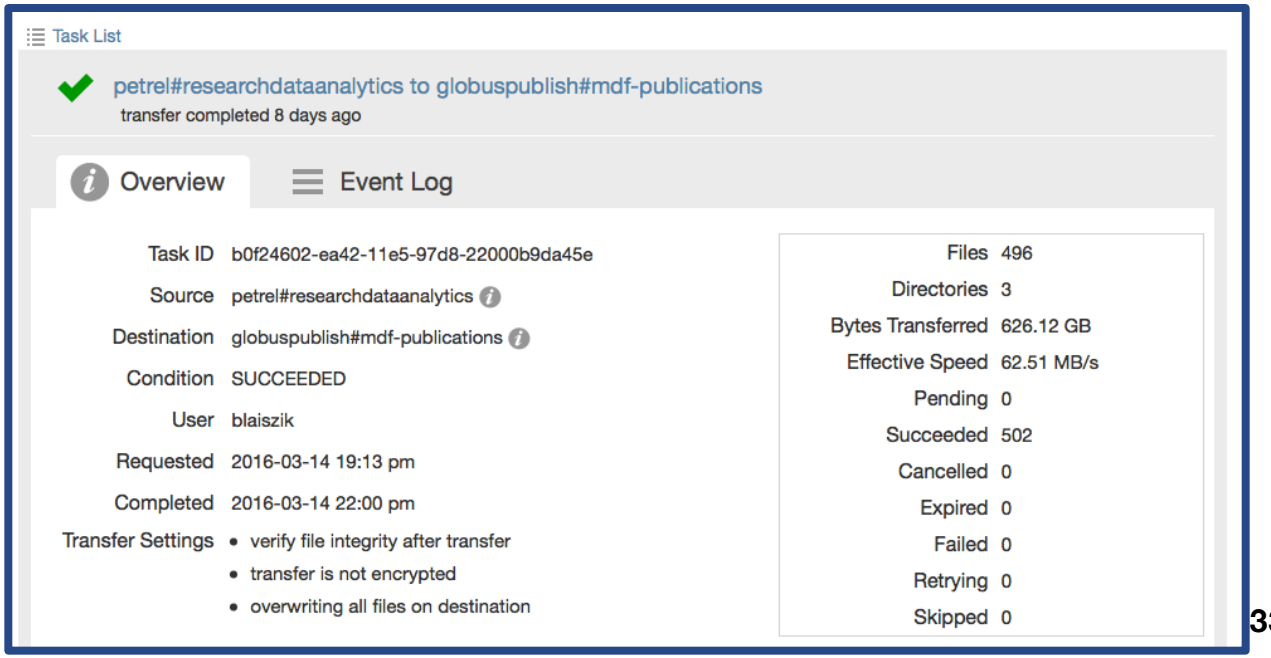
- Shared endpoint is auto-created on collection-specified data store
- Scientist transfers dataset files to a unique publish endpoint
- Dataset may be assembled over any period of time
- When submission is finished, dataset will be rendered immutable via checksum



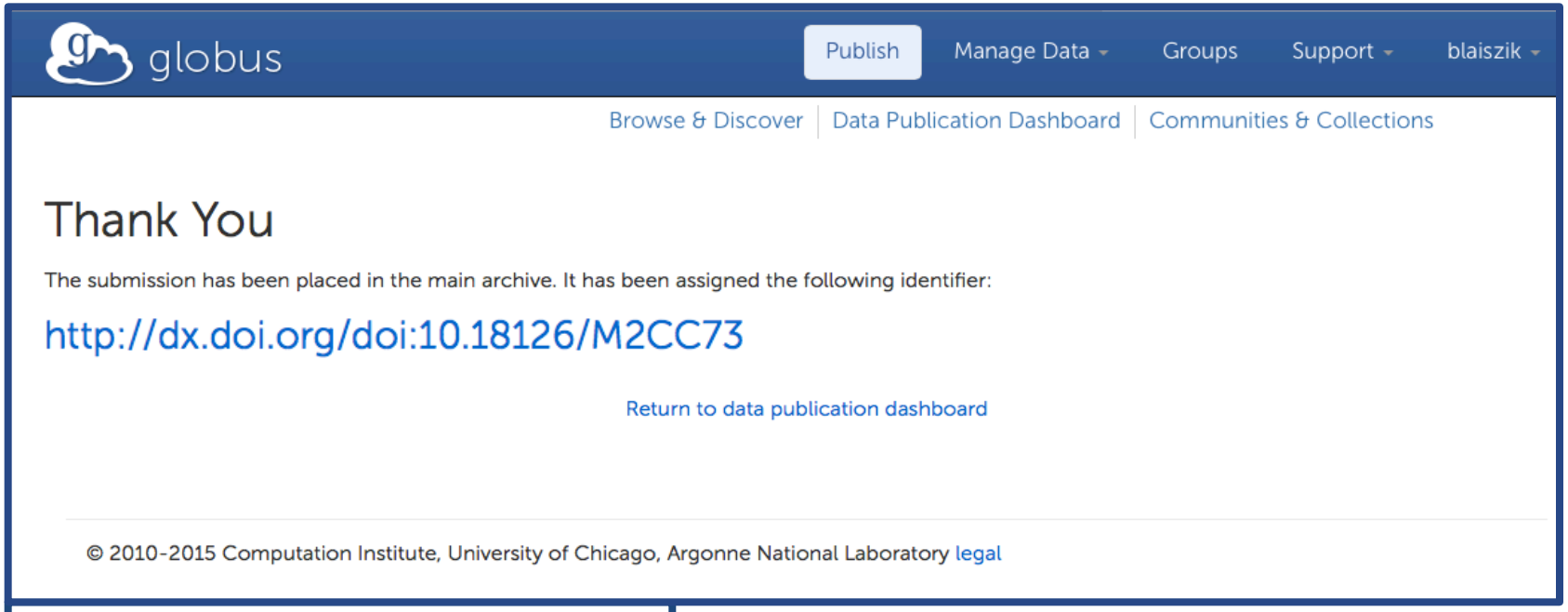
Dataset Assembly

- 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



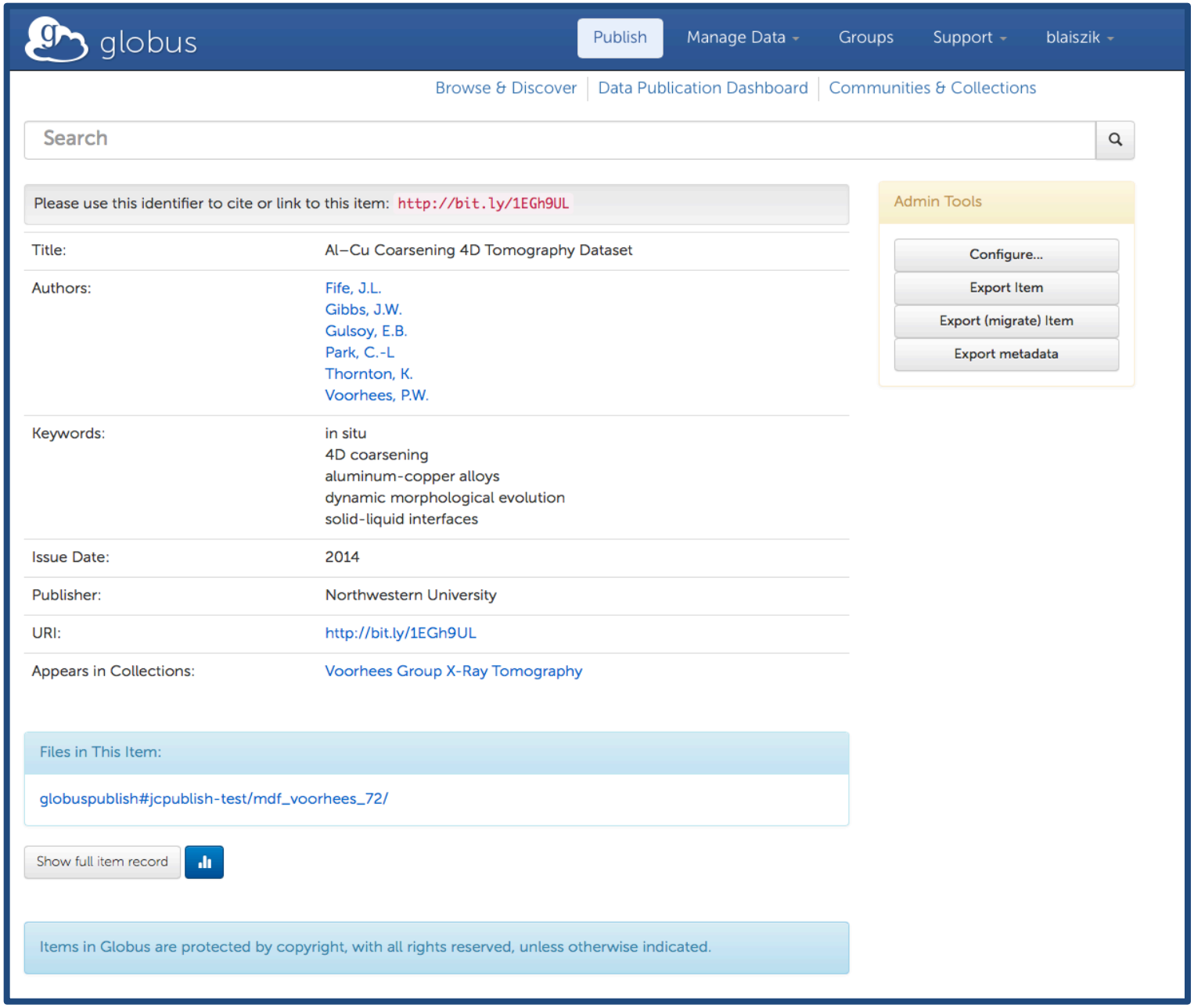


Mint a Permanent Identifier



Can optionally be DOI or Handle

Dataset Record



Dataset Discovery

x-ray

Search Results

advanced search

Community results (1 result)

Results 1-7 of 7

Issue Date	Title	Author(s)
9-Feb-2016	Dataset for Determination of Residual Stress in a Microtextured Alpha-titanium Component using High Energy Synchrotron X-ray	*Park, Jun-Sang; Ray, Atish K.; Dawson, Paul R.; Lienert, Ulrich; Miller, Matthew P.*
11-Feb-2016	Dataset for Segmentation of Four-dimensional, X-ray Computed Tomography Data	*Gibbs, John W.; Voorhees, Peter W.; Fife, Julie L.*
19-Mar-2016	Liquid-solid Metallic Mixture Coarsening Data - 80% solid	*Gibbs, John W.; Voorhees, Peter W.; Fife, Julie L.*
19-Mar-2016	Liquid-solid Metallic Mixture Coarsening Data - 28% Solid	*Gibbs, John W.; Voorhees, Peter W.; Fife, Julie L.*
19-Mar-2016	Liquid-solid Metallic Mixture Coarsening Data - 35% Solid	*Gibbs, John W.; Voorhees, Peter W.; Fife, Julie L.*
19-Mar-2016	Liquid-solid Metallic Mixture Coarsening Data - 55% solid	*Gibbs, John W.; Voorhees, Peter W.; Fife, Julie L.*

previous 1 next

Discover

Author

Fife, Julie L.	5
Gibbs, John W.	5
Voorhees, Peter W.	5
Dawson, Paul R.	1
Lienert, Ulrich	1
Miller, Matthew P.	1
Park, Jun-Sang	1
Ray, Atish K.	1

Issue Date

2016	6

Example Use Case

Publishing a Dataset and Evolving Analysis Code (or Env)

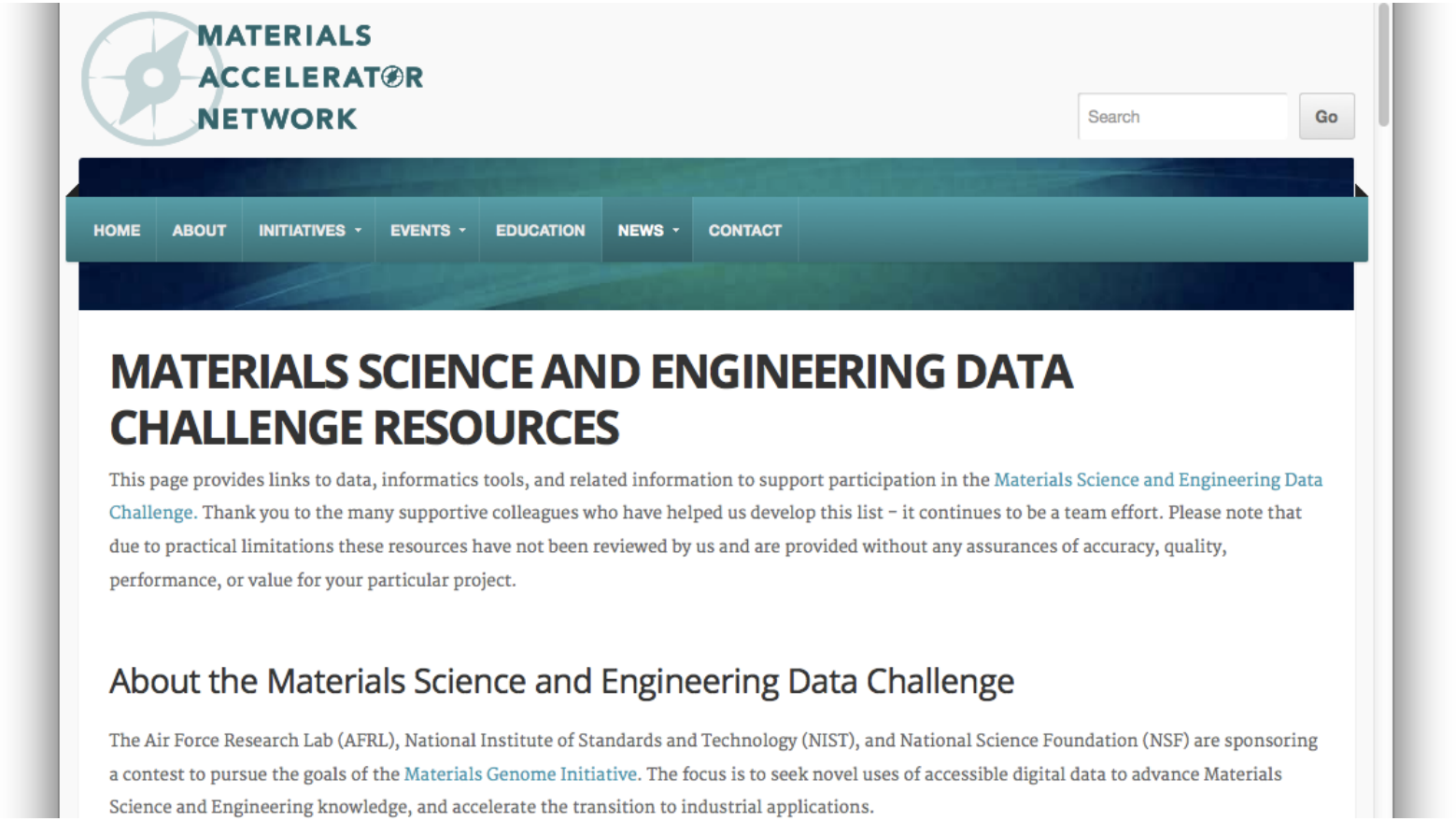
Code in a git repo or environment in DockerHub

Multi TB neutron scattering dataset

Registering Materials Resources

[w/ NIST – Youssef, Dima]

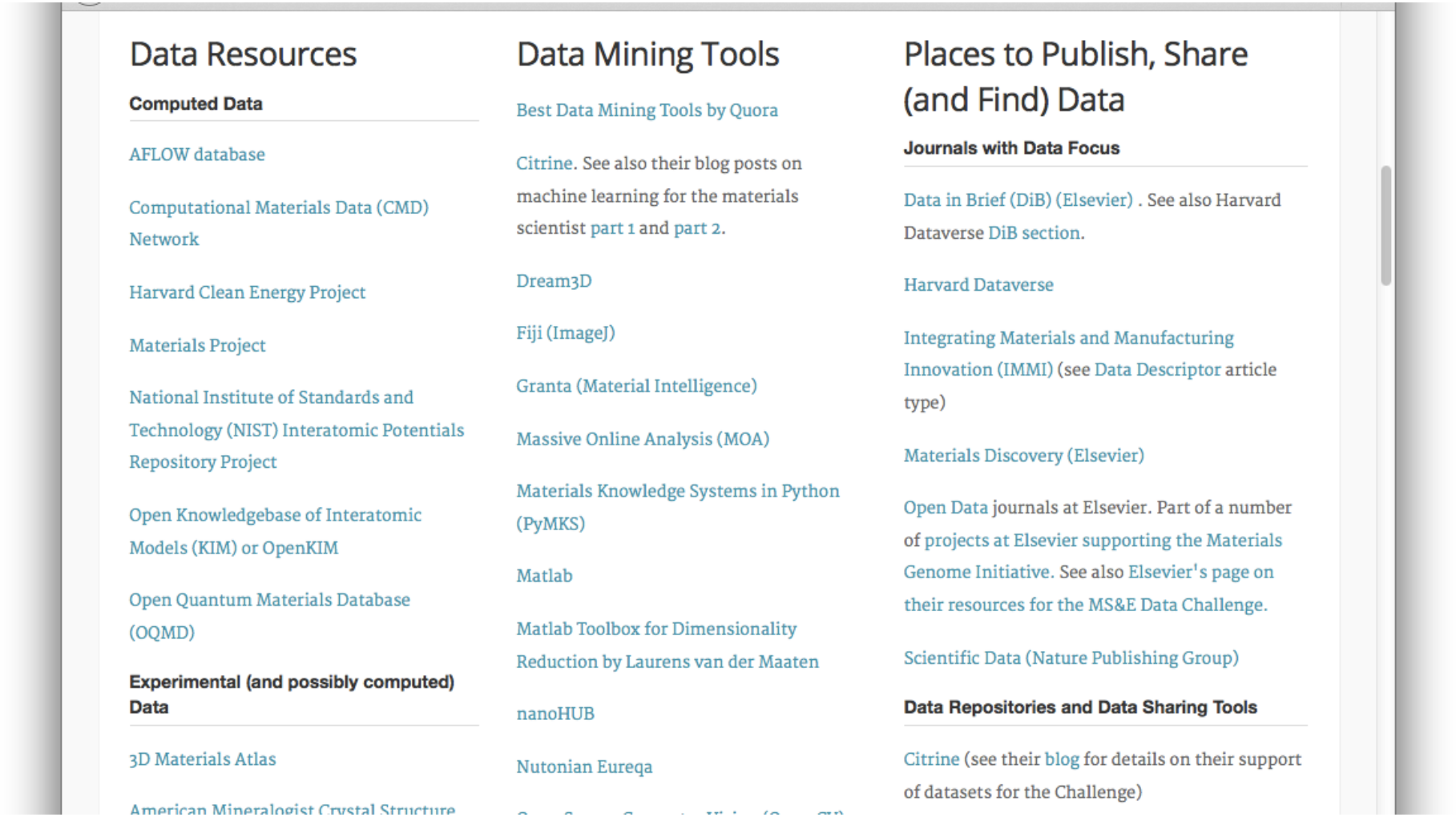
Materials Resource Registry



http://acceleratornetwork.org/mse-challenge/

Materials Science Data Challenge

Materials Resource Registry



Data Resources

Computed Data

AFLOW database

Computational Materials Data (CMD) Network

Harvard Clean Energy Project

Materials Project

National Institute of Standards and Technology (NIST) Interatomic Potentials Repository Project

Open Knowledgebase of Interatomic Models (KIM) or OpenKIM

Open Quantum Materials Database (OQMD)

Experimental (and possibly computed) Data

3D Materials Atlas

American Mineralogist Crystal Structure

Data Mining Tools

Best Data Mining Tools by Quora

Citrine. See also their blog posts on machine learning for the materials scientist part 1 and part 2.

Dream3D

Fiji (ImageJ)

Granta (Material Intelligence)

Massive Online Analysis (MOA)

Materials Knowledge Systems in Python (PyMKS)

Matlab

Matlab Toolbox for Dimensionality Reduction by Laurens van der Maaten

nanoHUB

Nutonian Eureqa

Places to Publish, Share (and Find) Data

Journals with Data Focus

Data in Brief (DiB) (Elsevier) . See also Harvard Dataverse DiB section.

Harvard Dataverse

Integrating Materials and Manufacturing Innovation (IMMI) (see Data Descriptor article type)

Materials Discovery (Elsevier)

Open Data journals at Elsevier. Part of a number of projects at Elsevier supporting the Materials Genome Initiative. See also Elsevier's page on their resources for the MS&E Data Challenge.

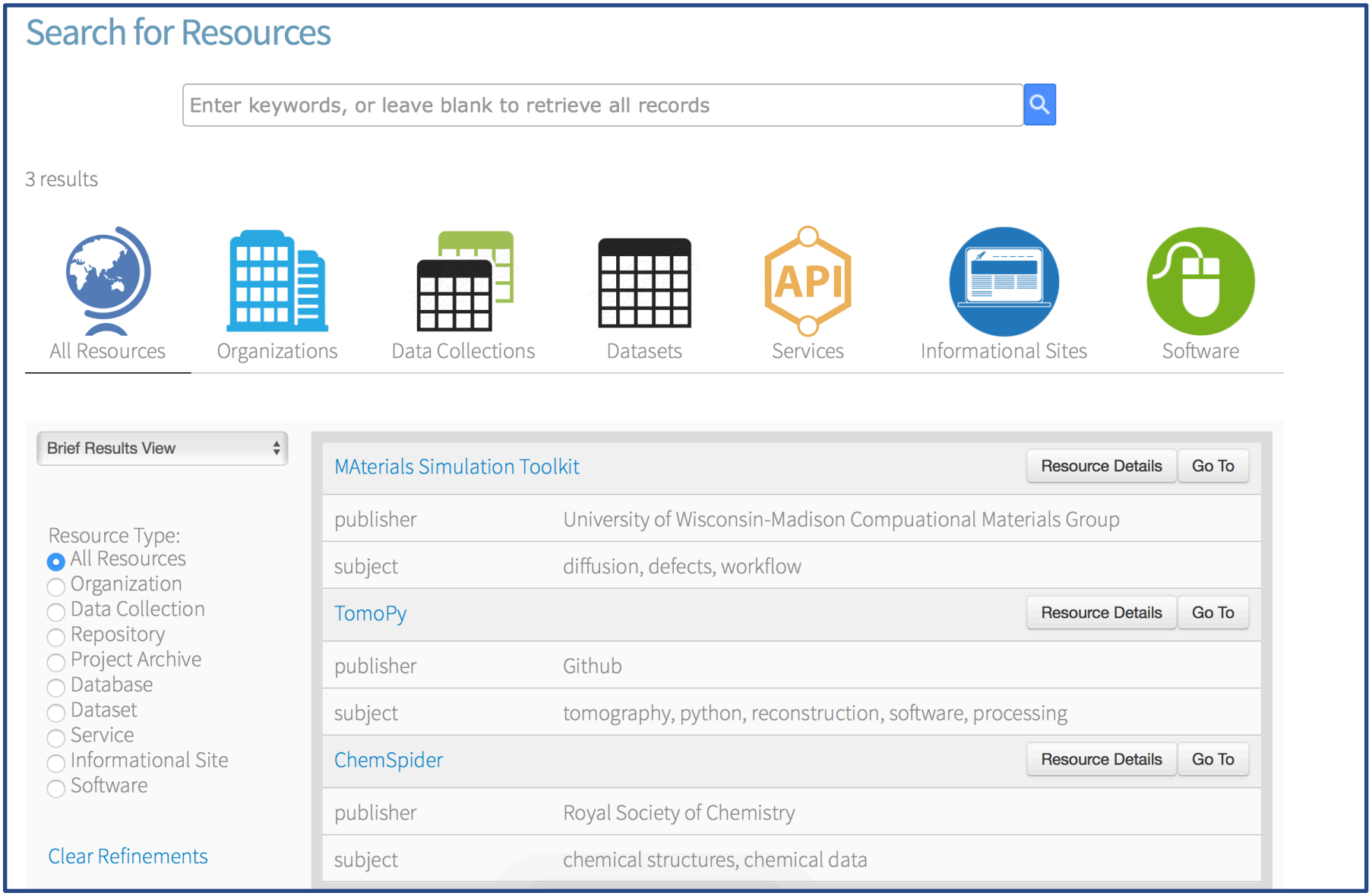
Scientific Data (Nature Publishing Group)

Data Repositories and Data Sharing Tools

Citrine (see their blog for details on their support of datasets for the Challenge)

Materials Accelerator Network

Materials Resource Registry



Browse Results

[w/ NIST - Youssef, Dima]

Lessons Learned

Understanding Incentives is Critical

Increasing Impact

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

Meeting Award Requirements

- Simplify DMP compliance

Smoothing Dislocations

- [Distance] Enable simple sharing among collaborators (near and far)
- [Personnel] Ease transitions between students
- [Format] 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]

Lessons Learned

- The demand is there from researchers and institutions
- Lots of cross-over with centers and projects
  - (NIST) CHiMaD
  - (DOE) MICCoM, JCESR, PRISMS, Argonne IT, I^3
  - (NSF) T2C2 [DIBBS], AMI-CFP (PIRE), HV/TMS (I/UCRC), BD Hubs, IMaD BD Spoke*

- Data Heterogeneity is a challenge
- Friction points
  - Data model (v 1.0)
    - Need data objects e.g. {"temperature":100, "unit":"K"}
    - Likely need finer grained metadata capabilities (i.e. file-dir level)
  - More data flavors (immutable alone is not enough)
  - Data gathering in retrospect
  - Schema generation and interoperability
    - Working with NIST, RDA, Citrine et al.
  - Differing approval processes
  - Lack of programmatic interface (planned). e.g. Integration with other institutional publication platforms
- Support for data interactivity and visualization
- Versioning

What's Currently Available?

- Web interface to support data publication (public-facing APIs coming soon)
- 100s of TB of storage at NCSA (scalable to many PB) more at Argonne (1.7 PB on Petrel)
- Help with developing metadata schemas to describe your research datasets
- Materials resource registry. Email me to get your resource added! (blaiszik@uchicago.edu)

What are we looking for?

- Early adopters, willing to get their hands dirty with the service and give honest feedback

- Key datasets and resources of all sizes, shapes, raw or derived, that might help us understand the process better

Thanks to Our Sponsors!

U.S. DEPARTMENT OF ENERGY

THE UNIVERSITY OF CHICAGO