Building an Interoperable Materials Infrastructure

May 2, 2016 Workshop

Co-Organizers

Jim Warren, Carrie Campbell, Ian Foster, Laura Bartolo

CHiMaD

Peter Voorhees, Juan de Pablo, Greg Olson

THE UNIVERSITY OF CHICAGO

NORTHWESTERN UNIVERSITY

Argonne NATIONAL LABORATORY

CHiMaD

Summary Agenda

- *7:30 – 8:10 Registration & Welcome*
- *8:10– 10:45 1st round project presentations*
- *11:00-12:50 2nd round project presentations*

- *12:50- 1:50 Lunch*

- *1:50 - 2:10 Exp & Comp data presentations*
- *2:10 - 3:10 Small group discussions*
- *3:10 - 3:55 Small group presentations*
- *3:55 - 4:30 Qs, Ds, & wrap up*
- *4:30 - 5:30 Informal demos*

All presentations will be made available afterwards

CHiMaD

Qs for Groups discussions

- Gp 1 Materials Research:
  What tools/services are needed?
- Gp 2 Tools & Services:
  T&S challenges - community, industry, other T&S
- Gp 3 Infrastructure:
  Federation infrastructure challenges – different platforms, diverse stakeholders
- Gp 4 Interoperability:
  Data reusability challenges – researchers, service providers

Groups 10 min Presentations

- Key points of Group discussion
- Proposed low barrier activity
- Requirements/needs/collaborations to accomplish activity

OCT 31-NOV 2, 2016
SAVE THE DATE!

MGI Global Summit: Materials Research, Advanced Manufacturing, & Data Infrastructure

- What? 2 ½ day international conference
- Where? Chicago, IL, USA
- For Whom? 200 invited international scientists, postdoctoral fellows, & graduate students in materials research, additive manufacturing, & data informatics

CHiMaD

Materials Data Facility - Data Services to Advance Materials Science Research

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

materialsdatafacility.org
globus.org

What is MDF?

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

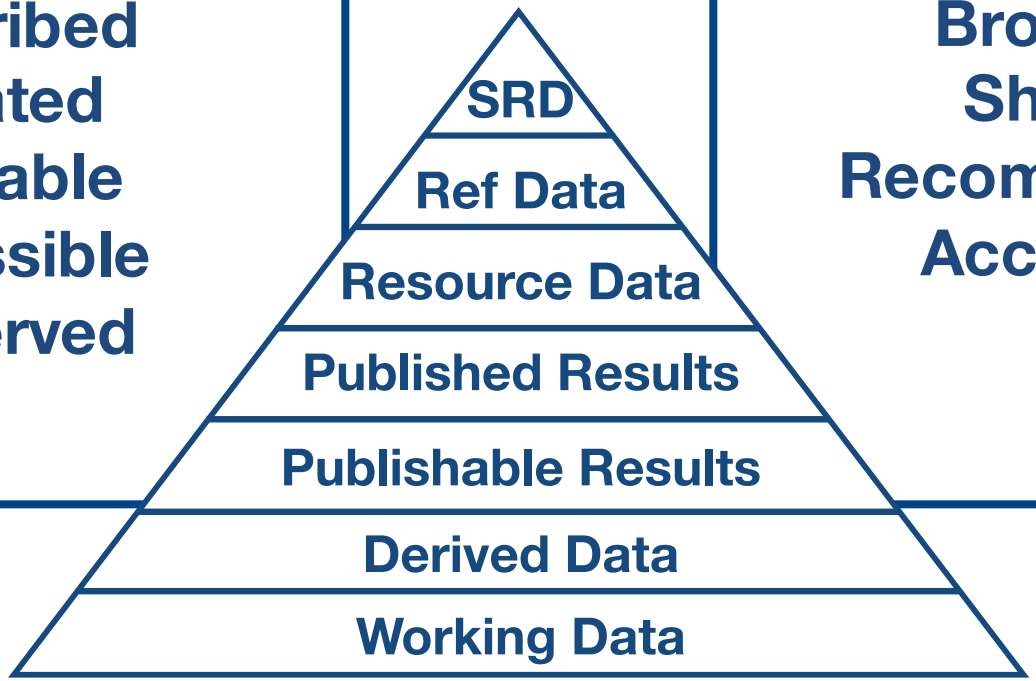
Published

Identified
Described
Curated
Verifiable
Accessible
Preserved

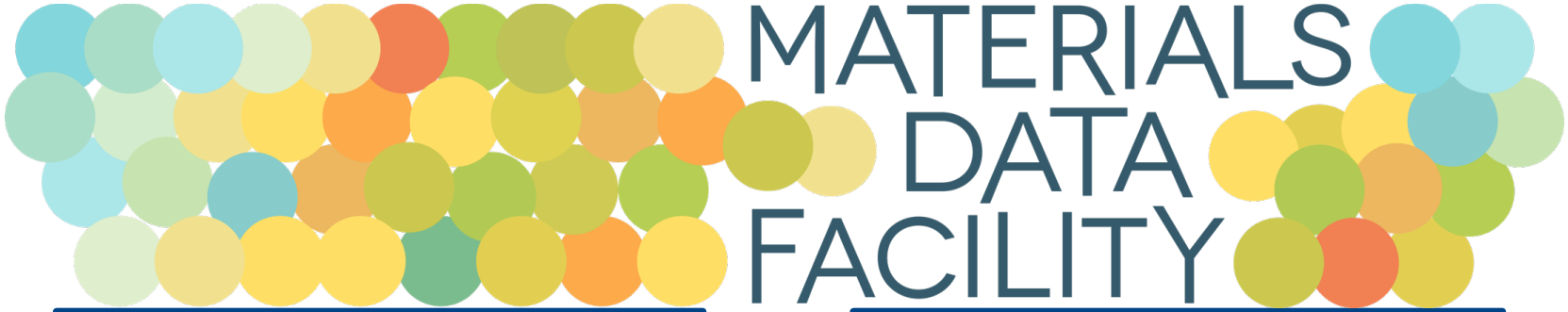
and

Discovered

Searched
Browsed
Shared
Recommended
Accessed



Data Service Infrastructure



Publication +

Discovery +

APIs

Data Interaction And Viz

Resource Registration

+ - Initial Foci

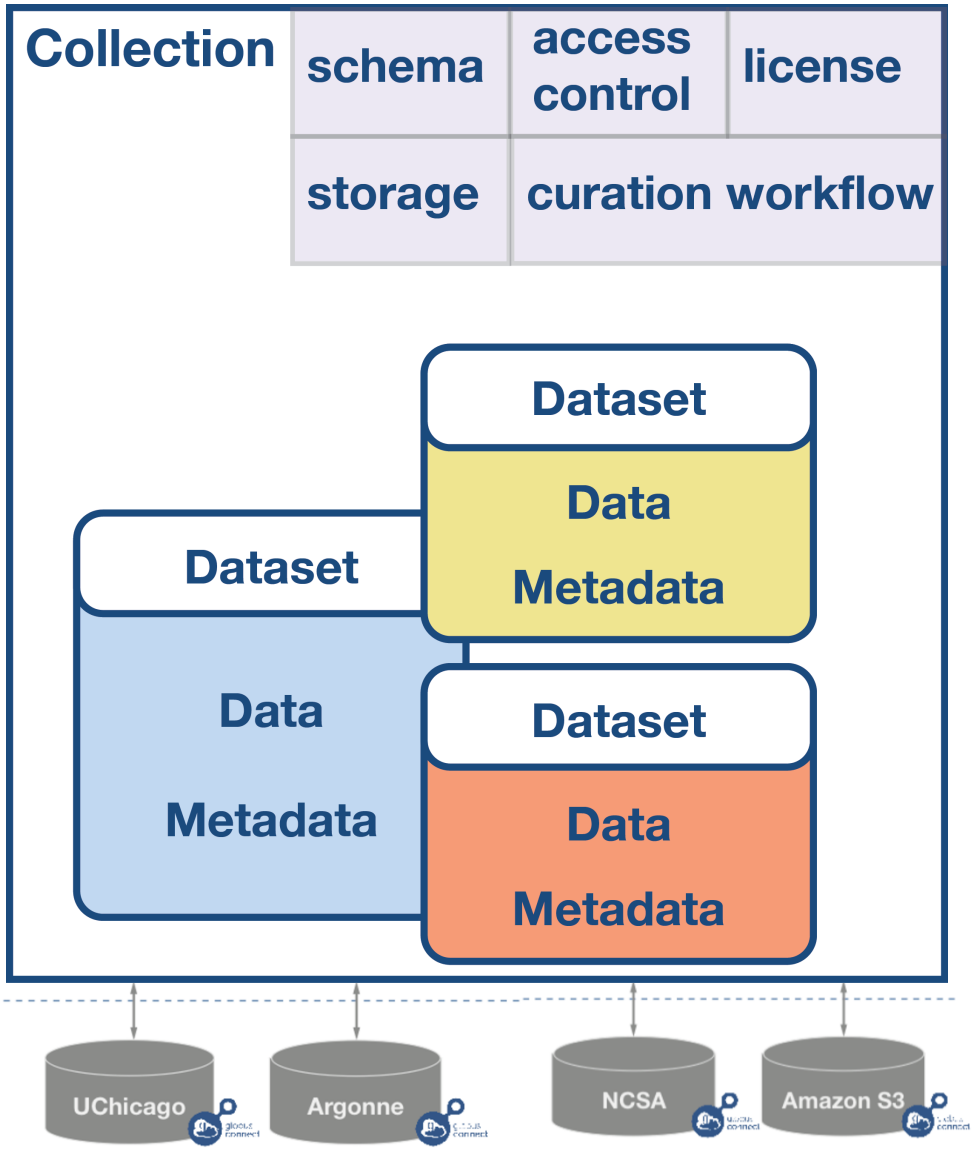
Publication

Opened to external users in mid Feb 2016

ca. 2.7 TB of data currently, >10 TB incoming (May/Jun)

- **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**

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

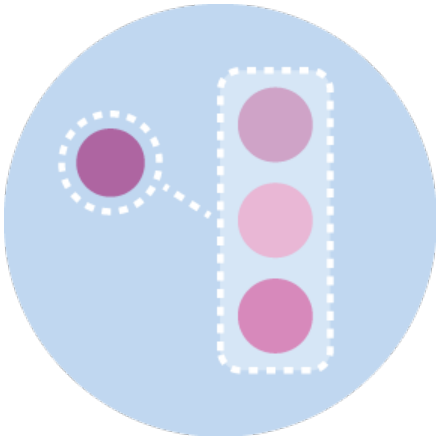
Publish Large Datasets

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

- Storage resources available now
  - 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)

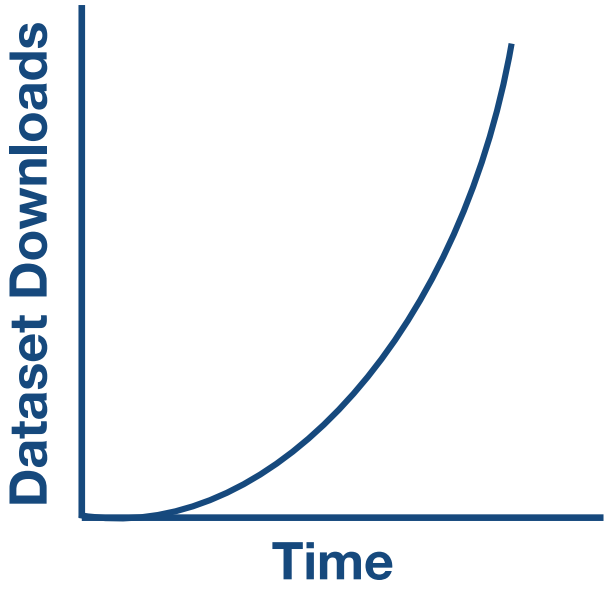
- Researchers can optionally use your own local or institutional storage

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**

MDF Submission Walkthrough

Example Use Case

Publishing Big, Remote Data

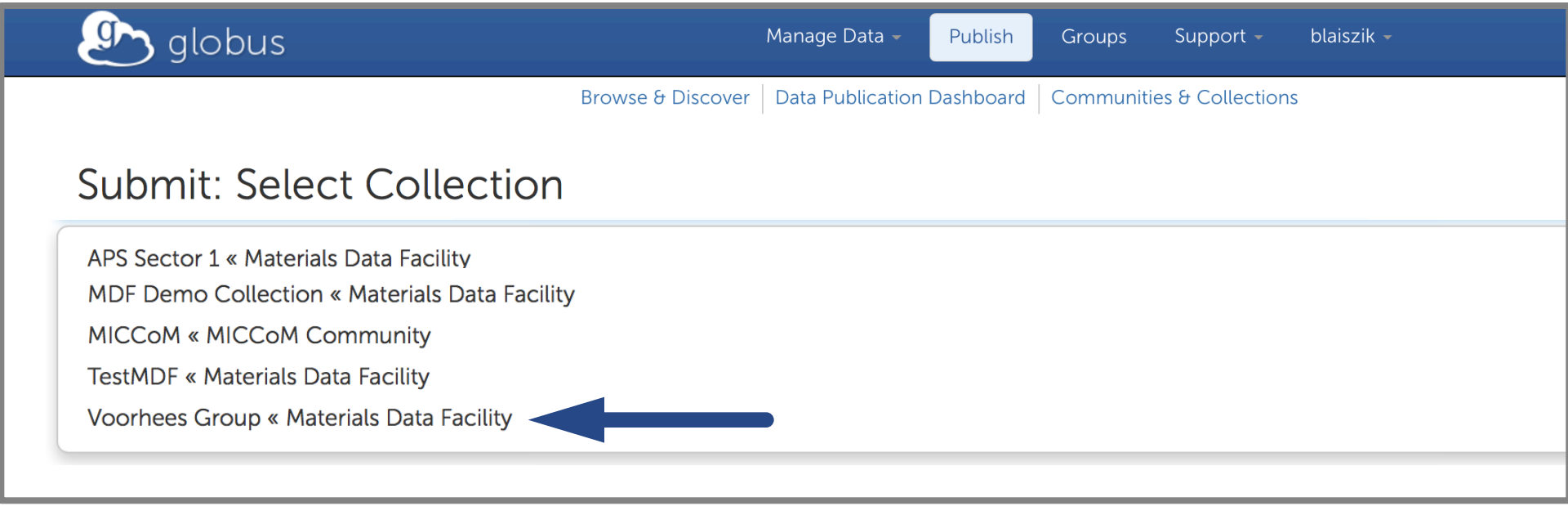
Collected multi TB of data at a light source

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 Collections

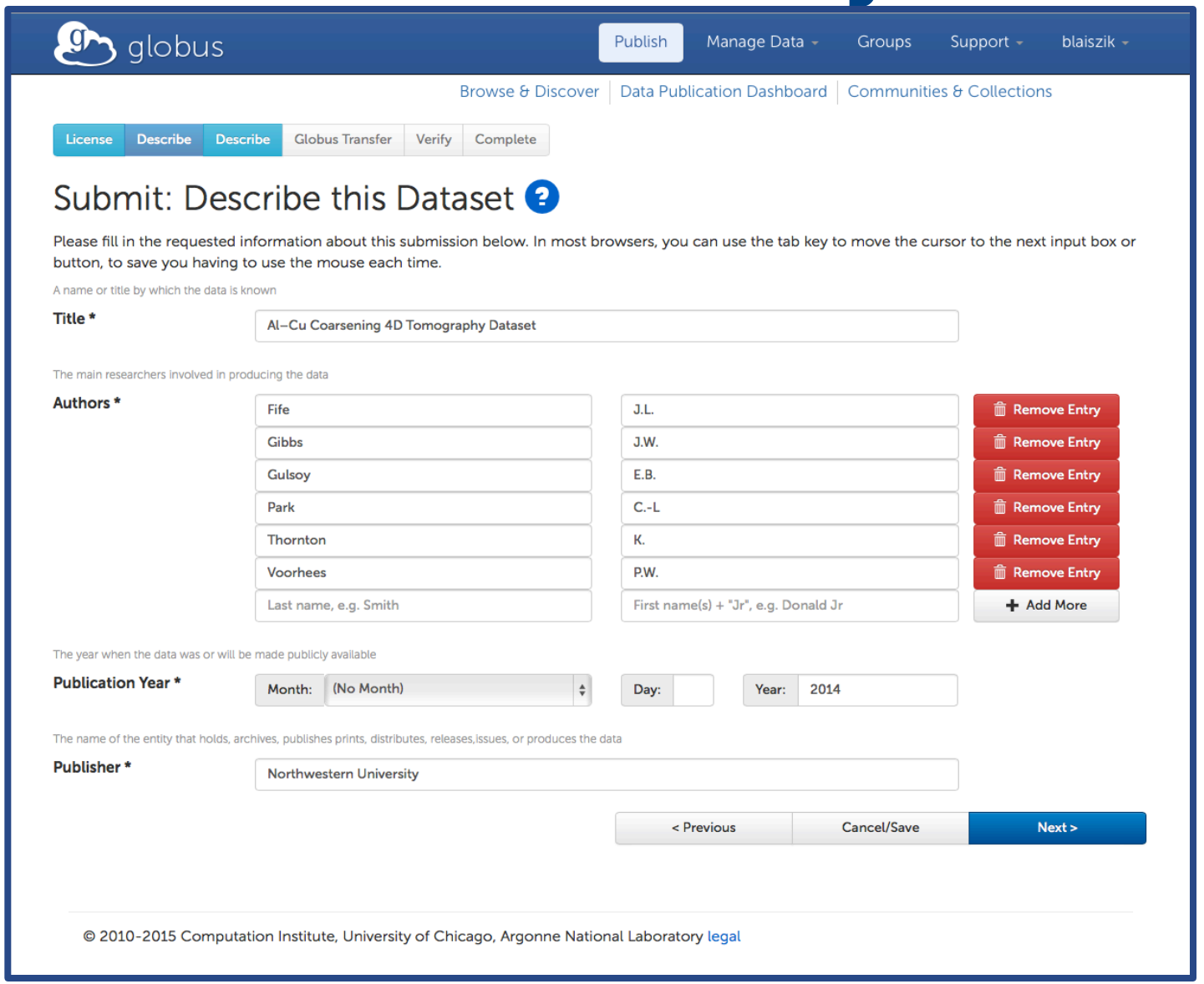


Recall: Policies Set at the Collection Level

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

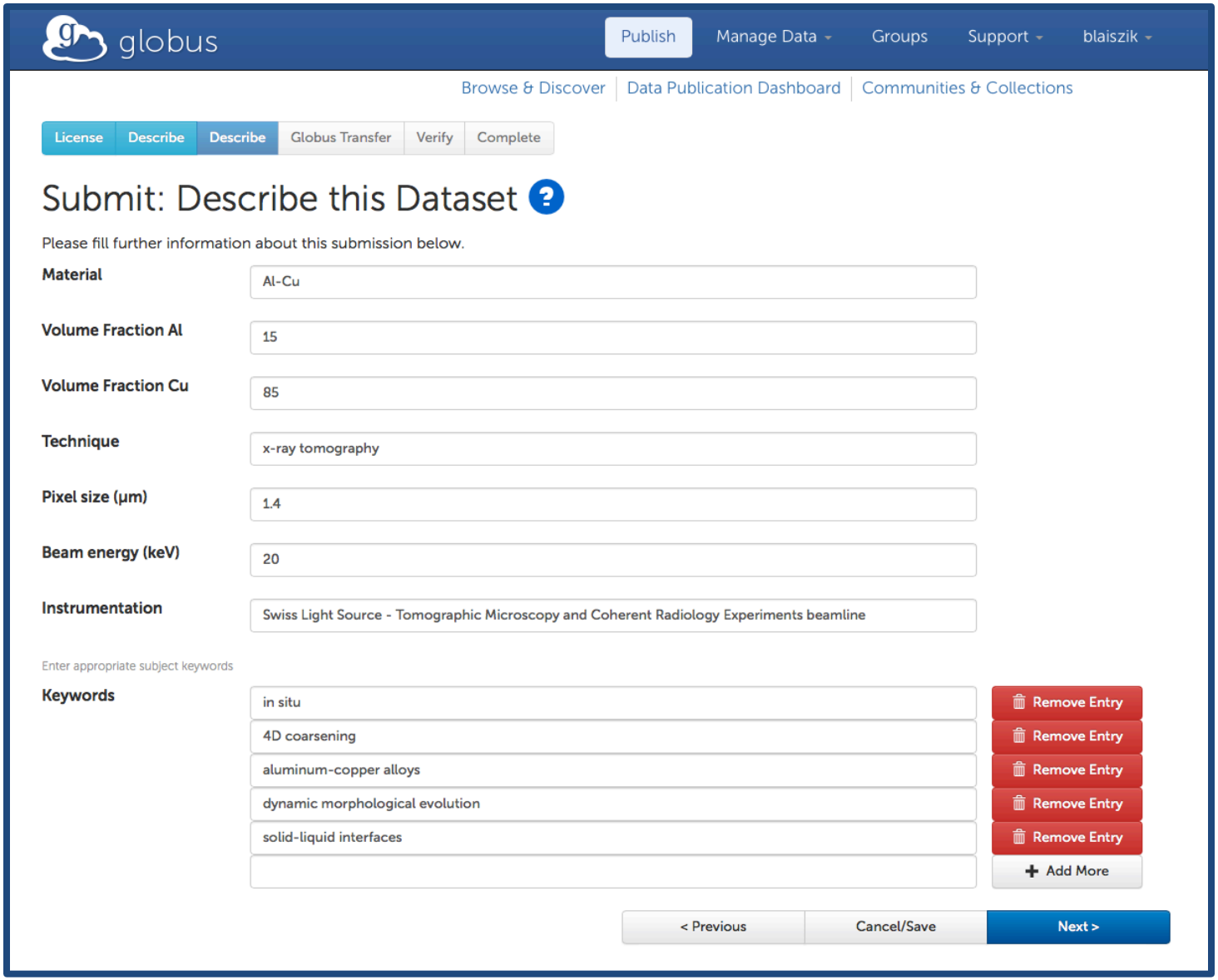
MDF Metadata Entry

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



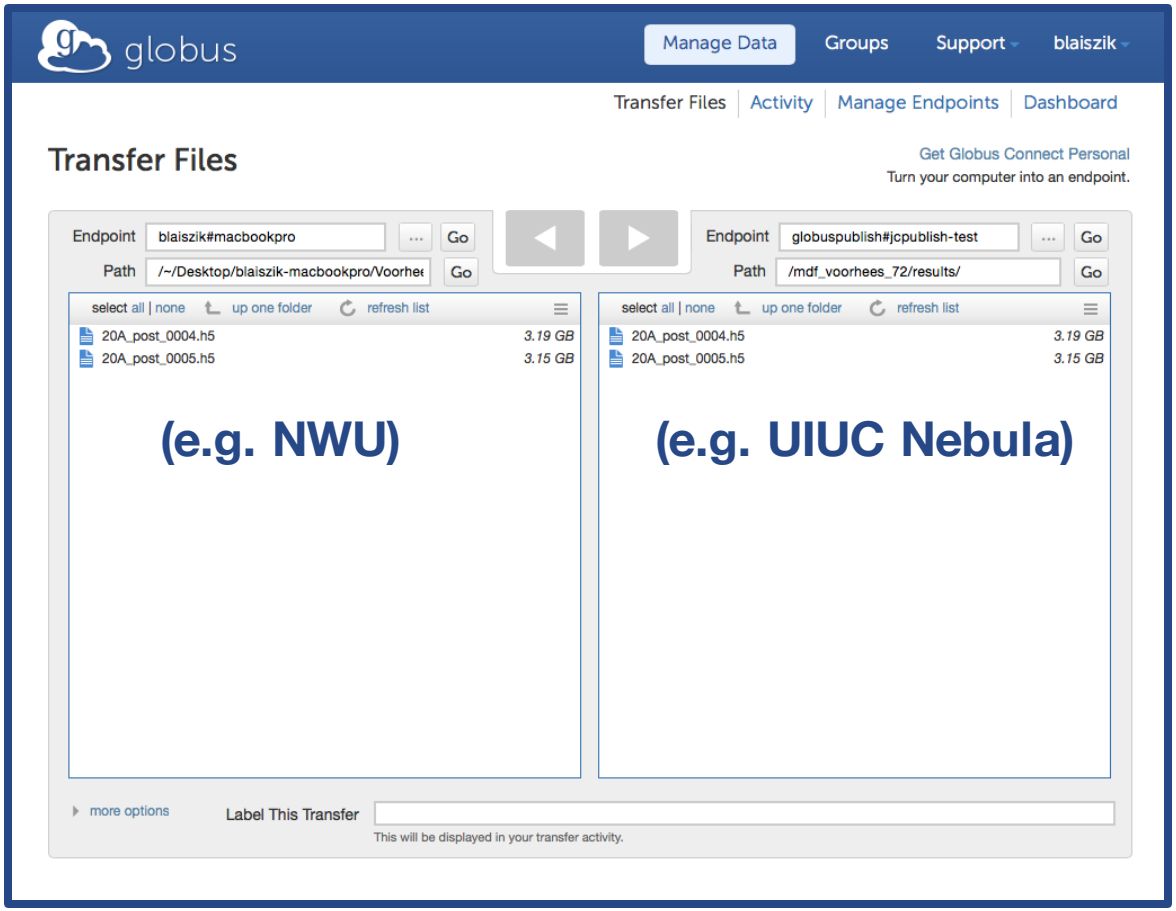
MDF Custom Metadata

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



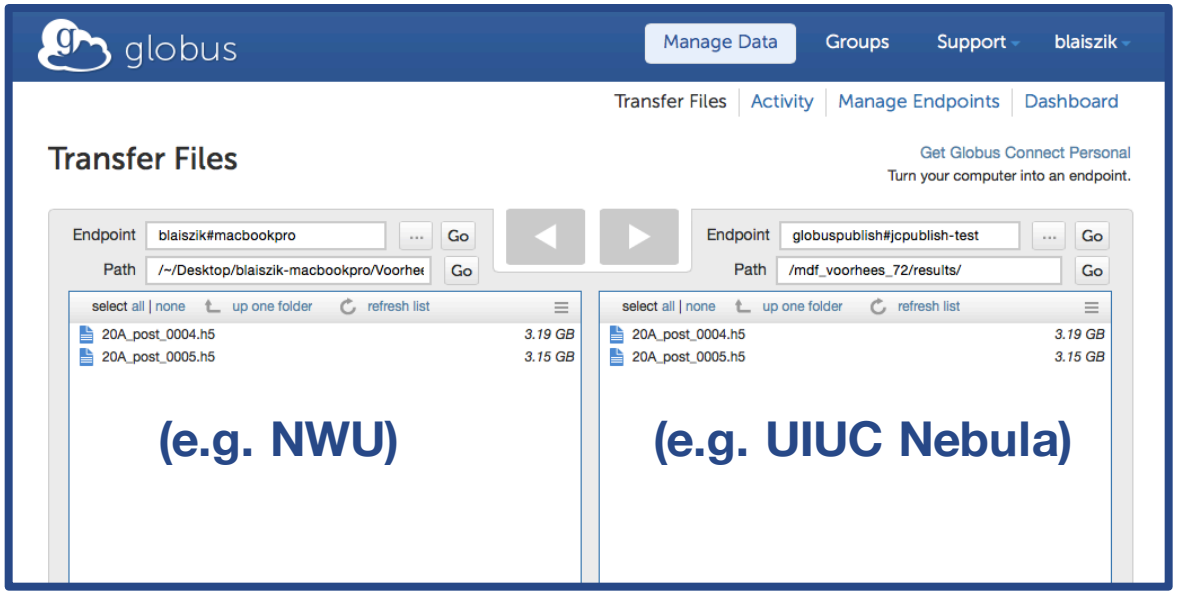
Dataset Assembly

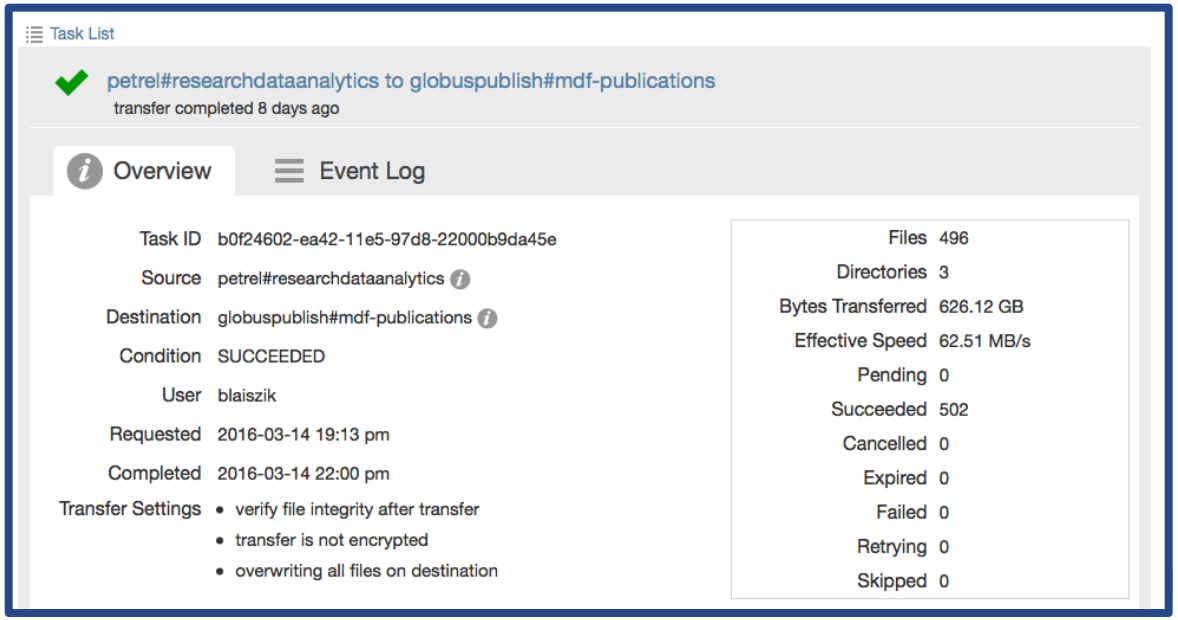
- Shared endpoint is auto-created on collection-specified data store
- Scientist transfers dataset files to a unique publish endpoint
- Dataset may be assembled over any period of time
- When submission is finished, dataset will be rendered immutable via checksum



Dataset Assembly

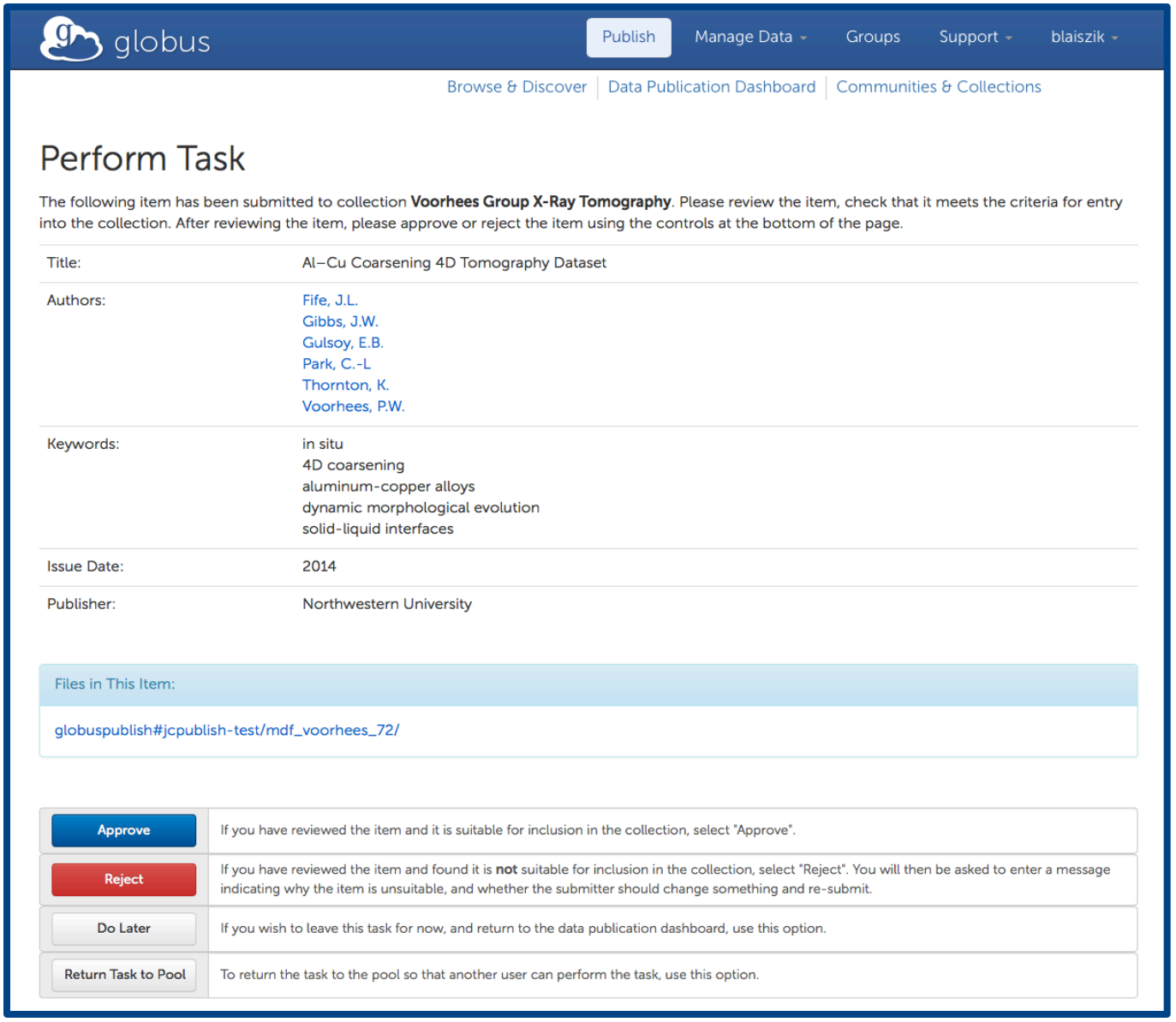
- Shared endpoint is auto-created on collection-specified data store
- Scientist transfers dataset files to a unique publish endpoint
- Dataset may be assembled over any period of time
- When submission is finished, dataset will be rendered immutable via checksum



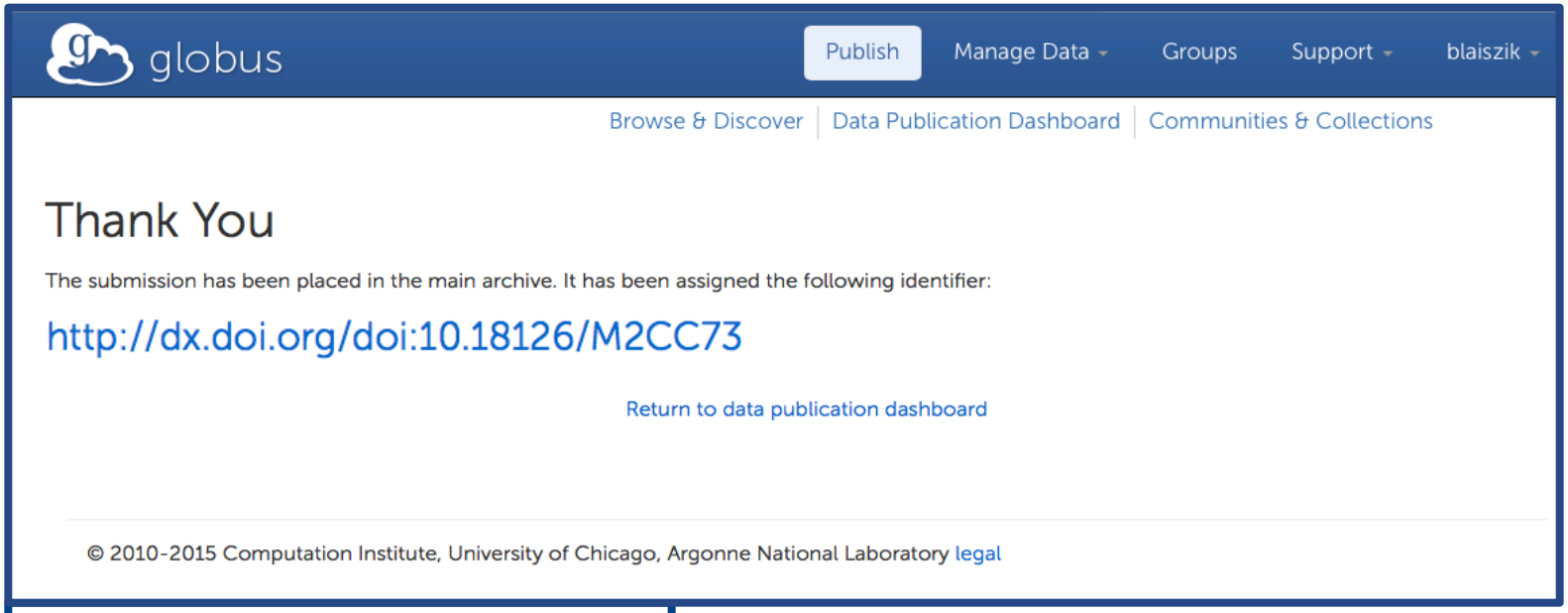


Dataset Curation

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

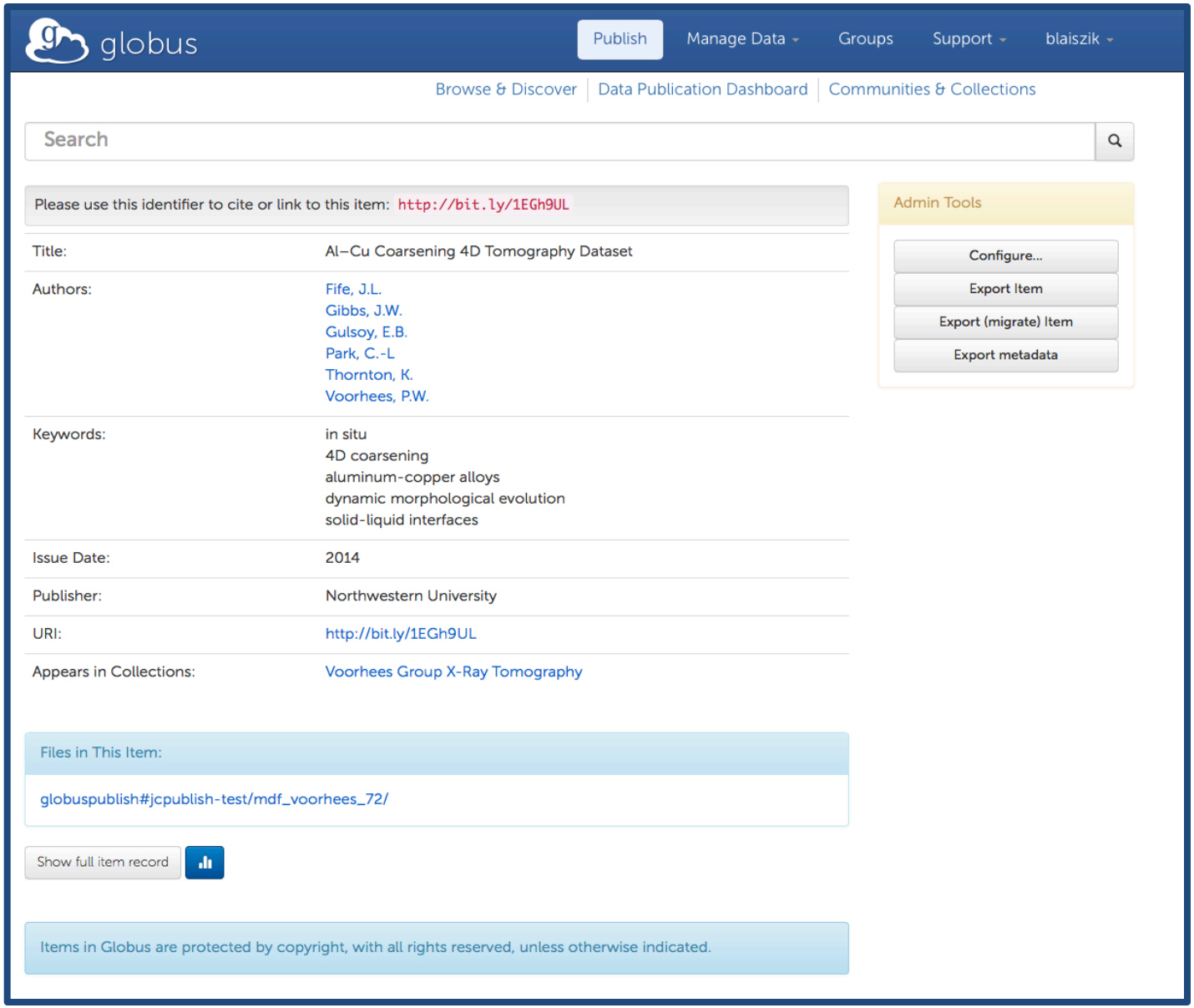


Mint a Permanent Identifier

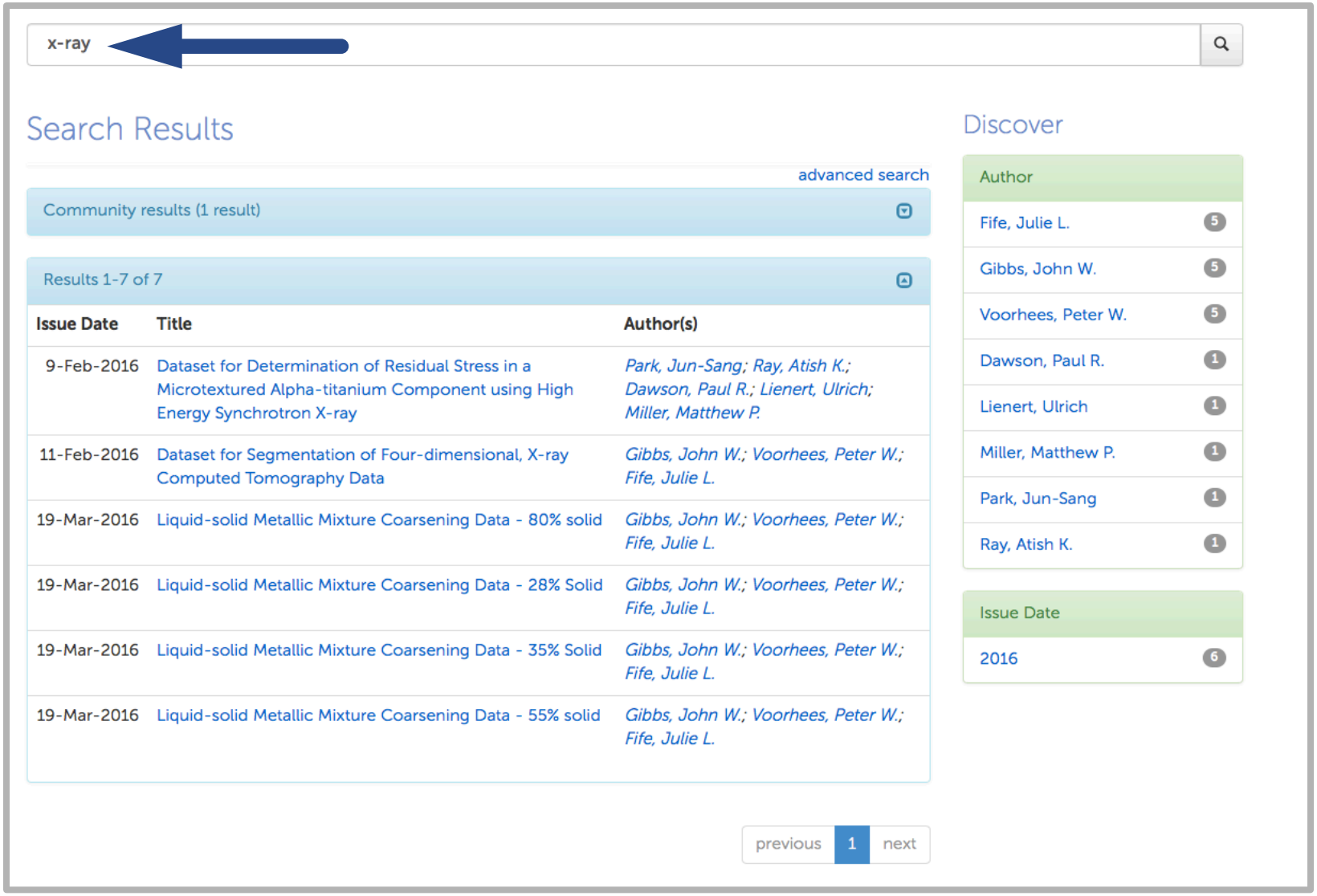


Can optionally be DOI or Handle

Dataset Record



Dataset Discovery

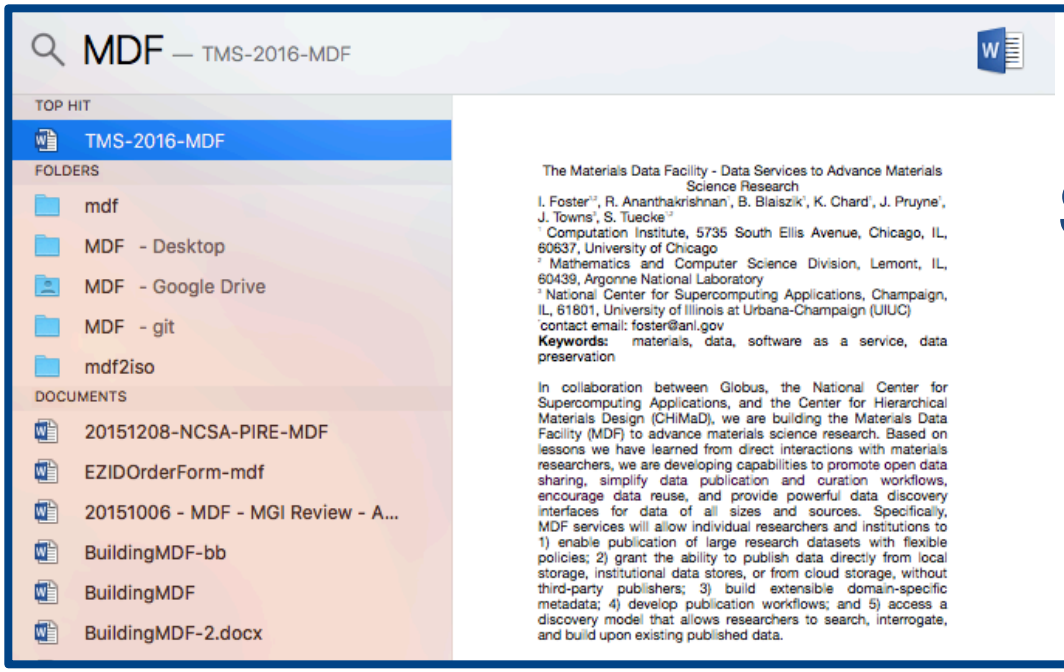


Discovery

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)

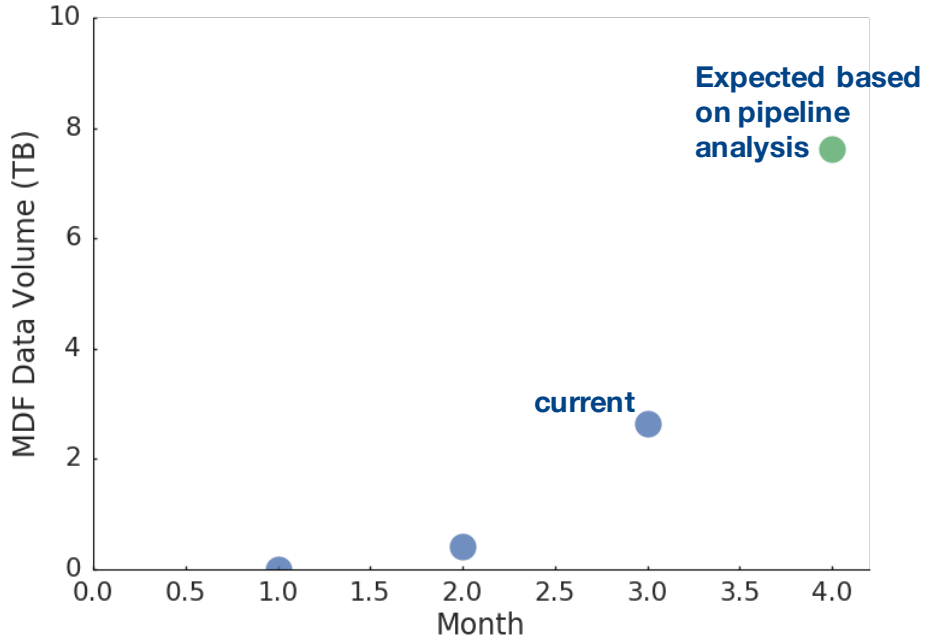


Future...

Spotlight for all data you have access to regardless of location

Summary

- Storage is allocated and available. Some early adopters are making use!
- Web UI is available, API under development
- Currently interacting with groups across multiple disciplines, institutions, and institution types

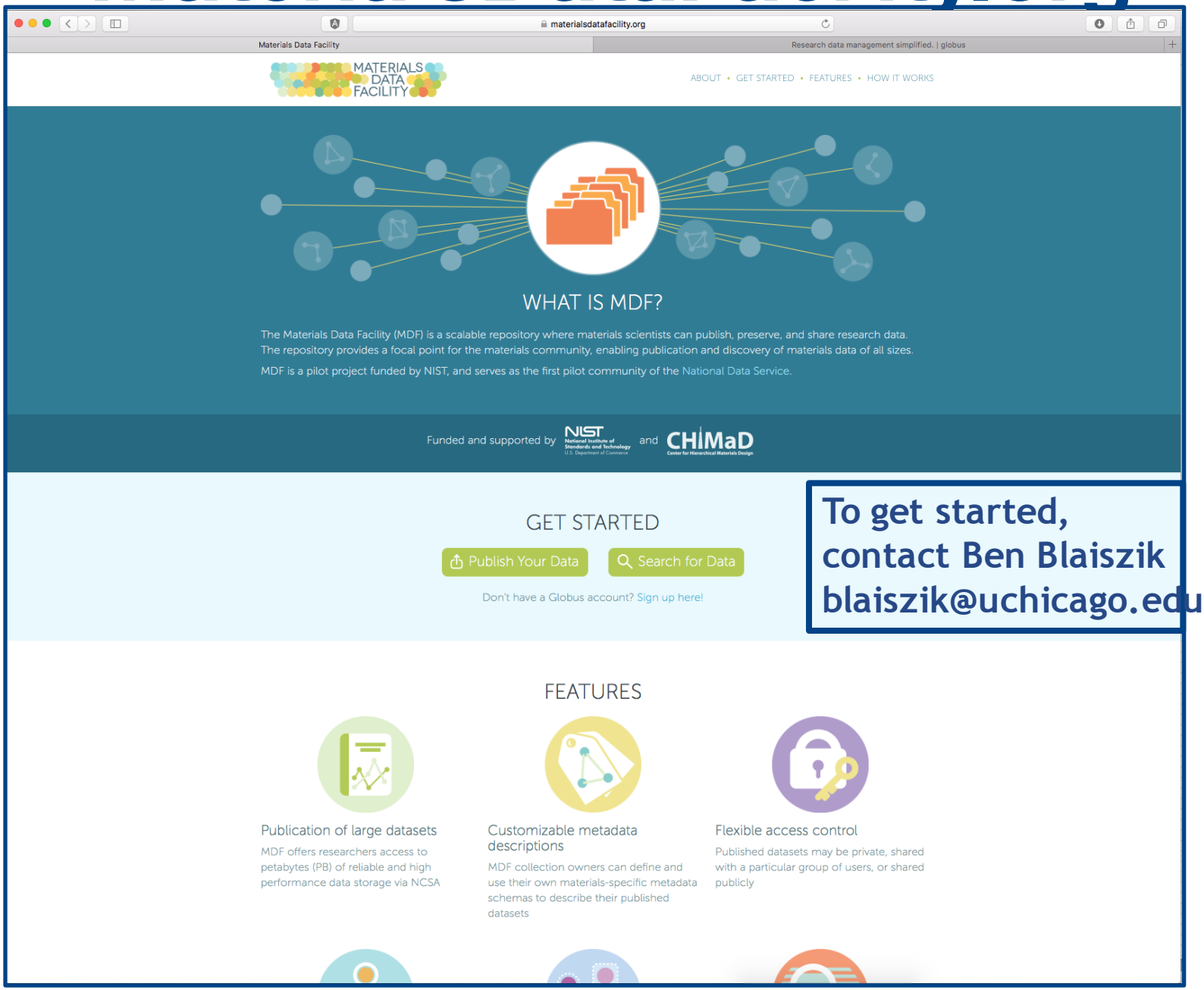


MDF Tutorial tomorrow!
https://github.com/blaiszik/materials-data-facility-training

Tentative Schedule

Time	Activity
9-9:30a	Overview and Discussion of the Materials Data Facility (MDF)
9:30-10a	Sign up for Globus and MDF, Set up an Endpoint
10-11a	Identification of Key Datasets and Metadata Formulation
11a-12p	Ingest datasets into MDF

MaterialsDataFacility.org



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

Thanks to Our Sponsors!

U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

THE UNIVERSITY OF CHICAGO

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

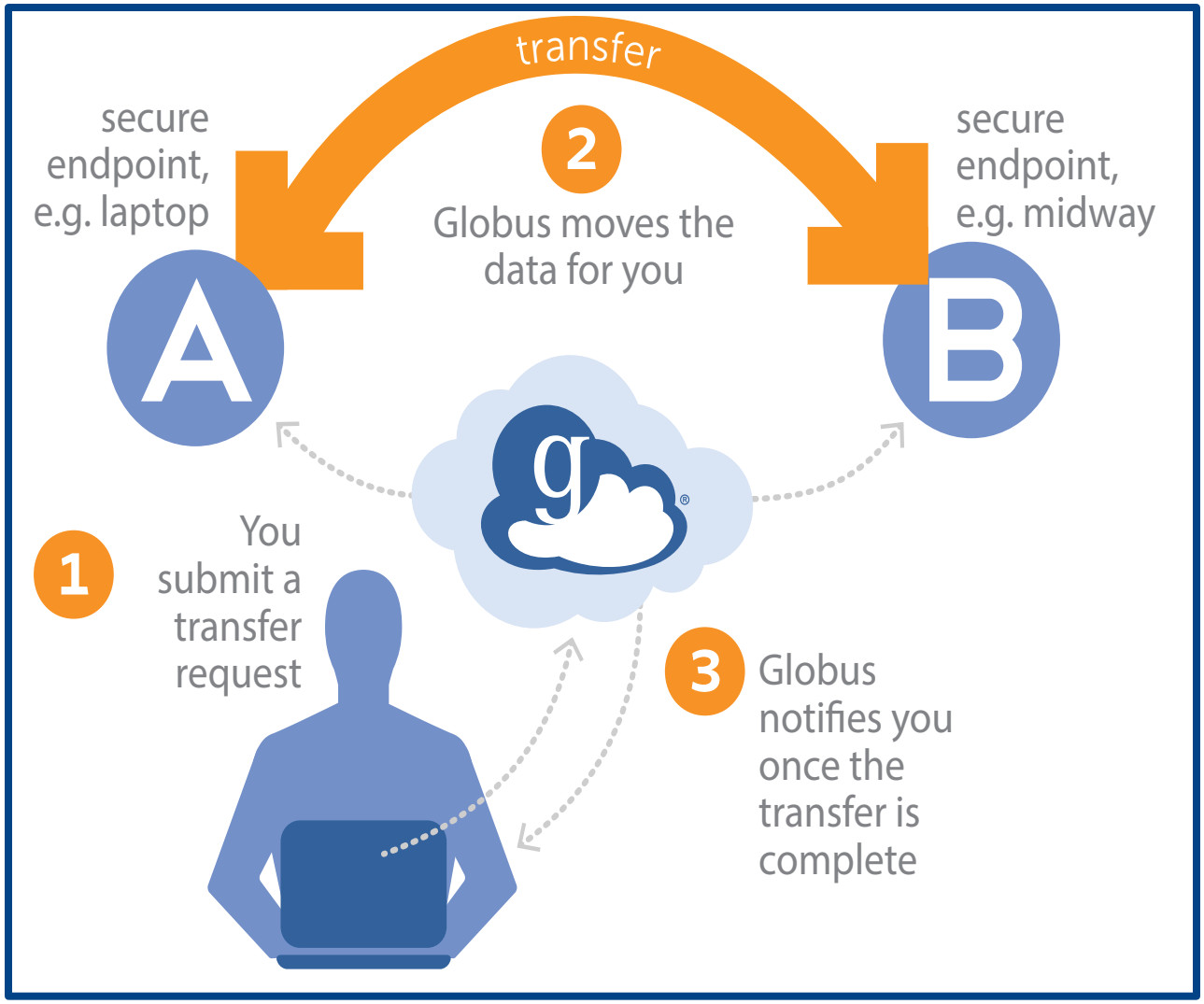
Globus Background

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



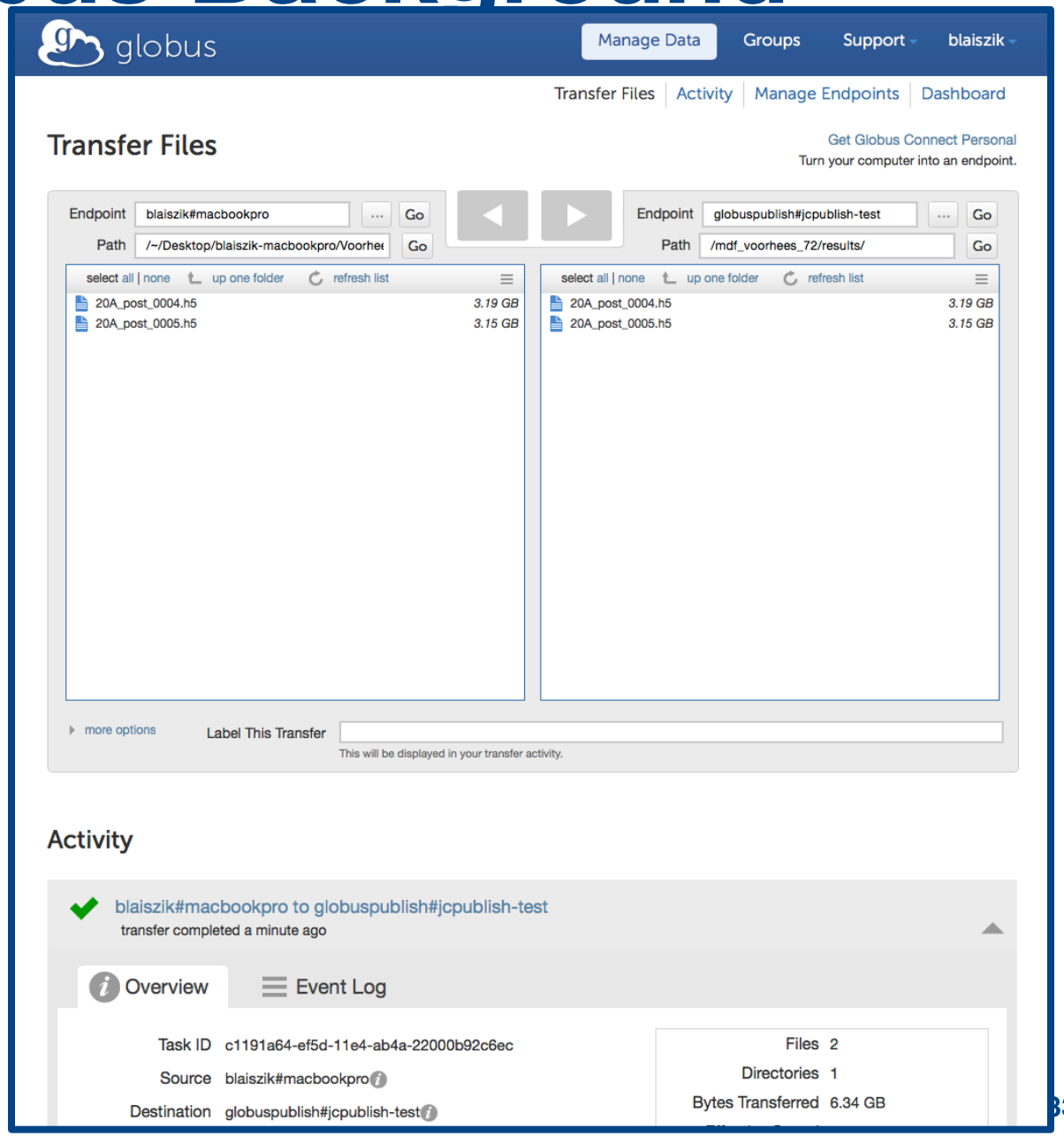
Globus Background

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

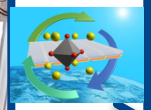
- REST API for automation and interoperability
- Web UI for convenience
- Optimizes and verifies transfers
- Handles auto-restarts
- Battle tested with big data



Where are we Now?

Data Publication

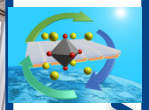
Materials Data Publication/Discovery is Often a Challenge



Materials Data Publication/Discovery is Often a Challenge

Want to Publish → ← **Want to Discover / Use**

Data Collection

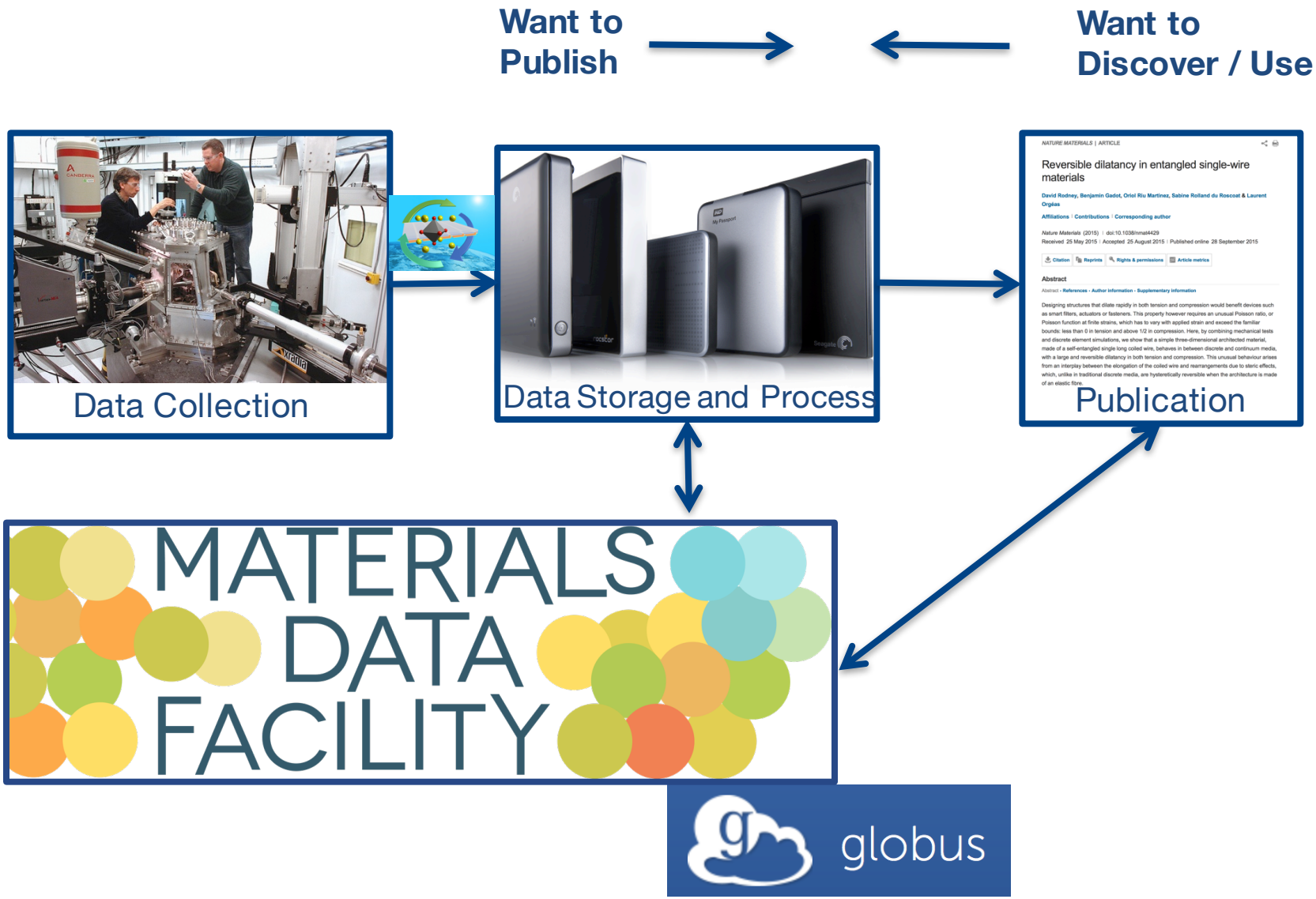


Data Storage and Process

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



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**

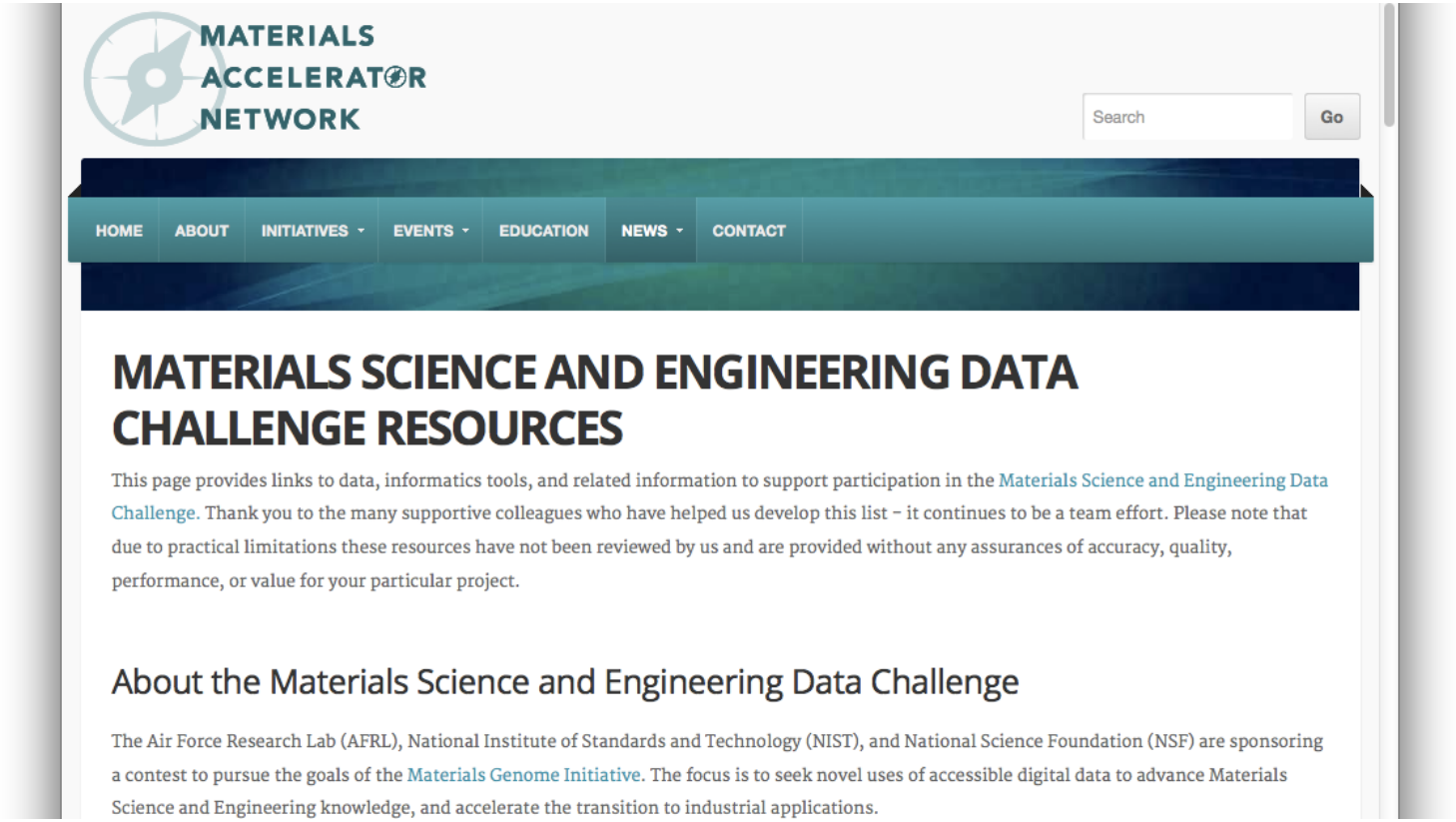
Future...

- **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

Registering Materials Resources

[w/ NIST – Youssef, Dima]

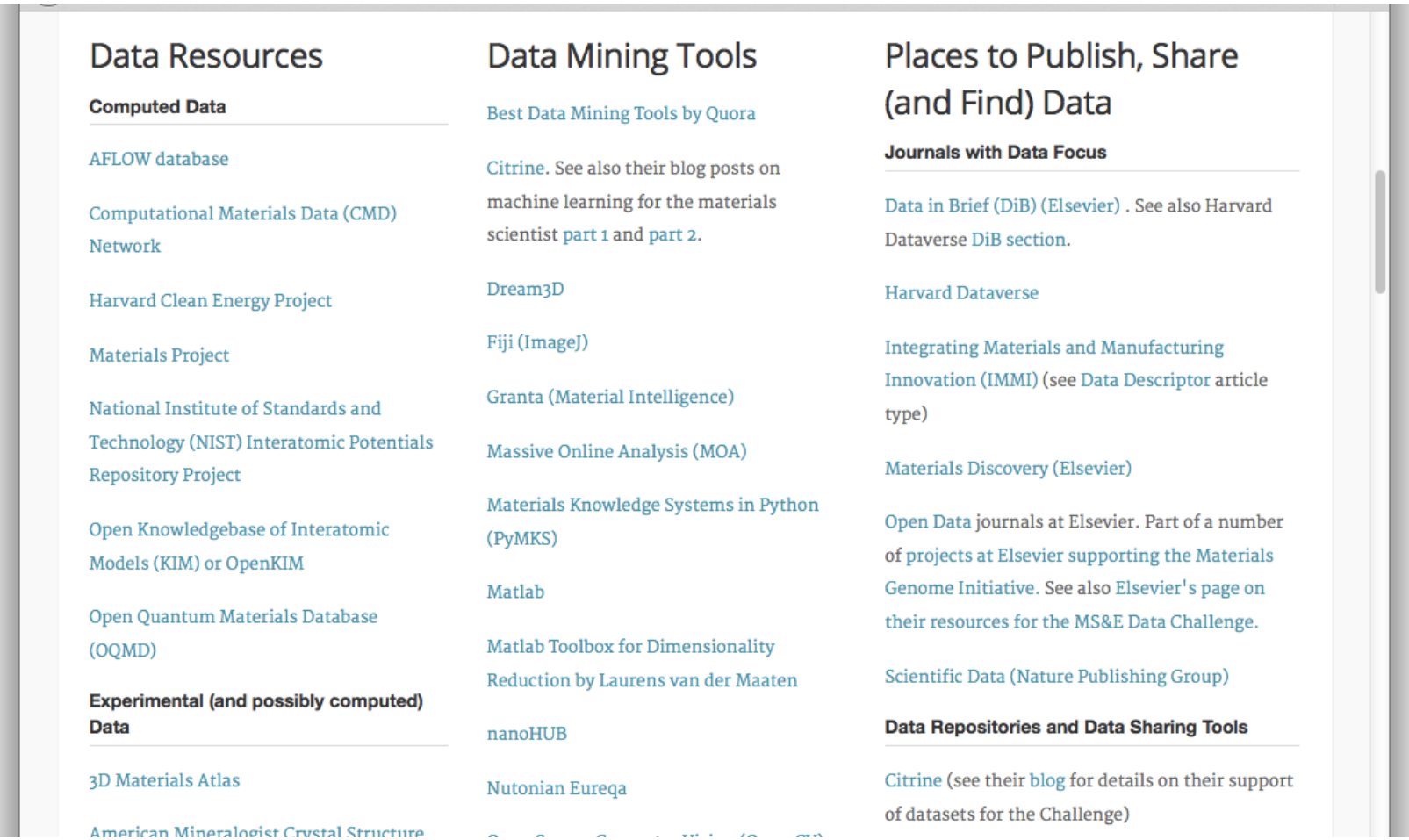
Materials Resource Registry



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.

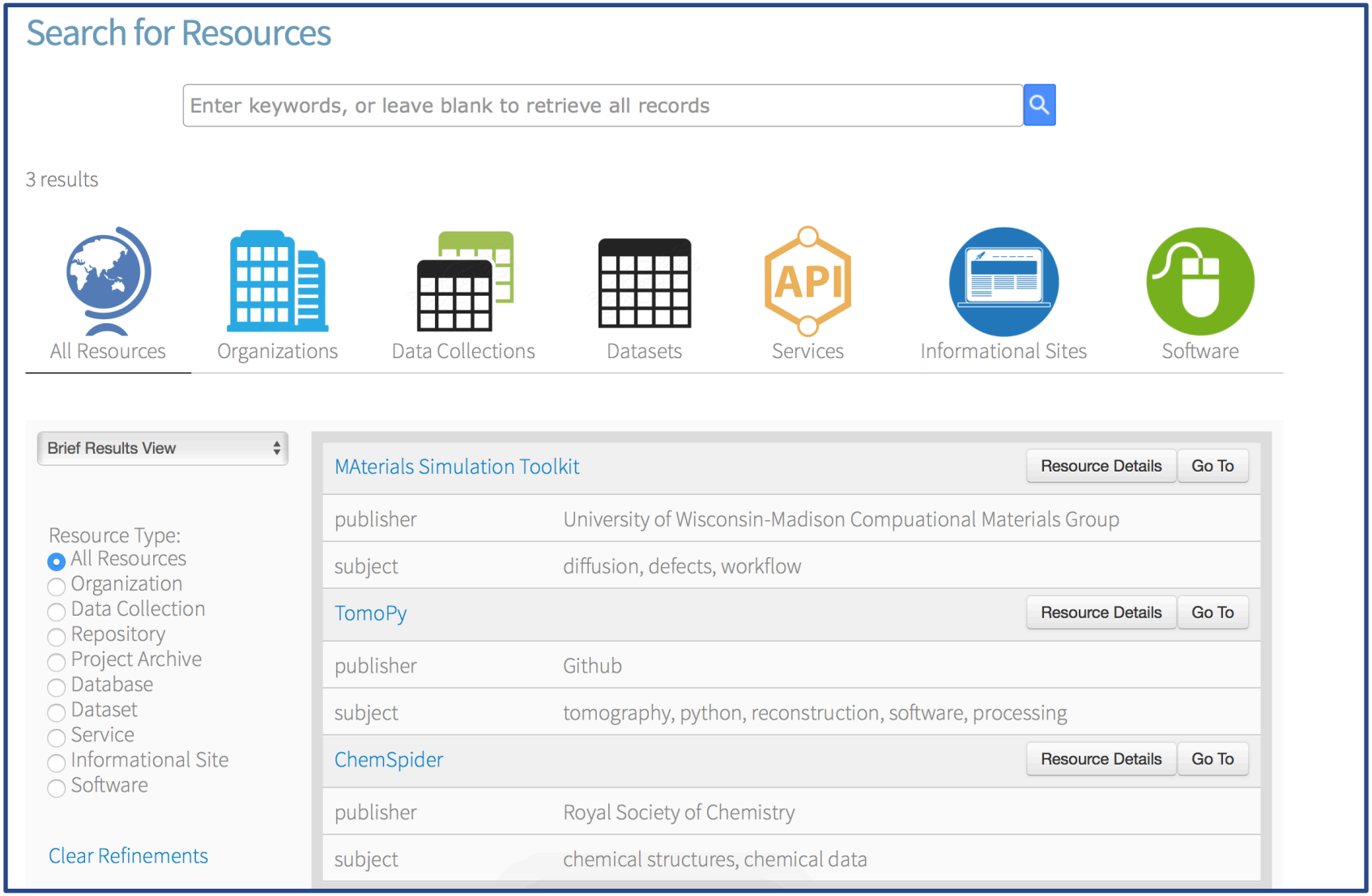
Scientific Data (Nature Publishing Group)

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]

What's Currently Available?

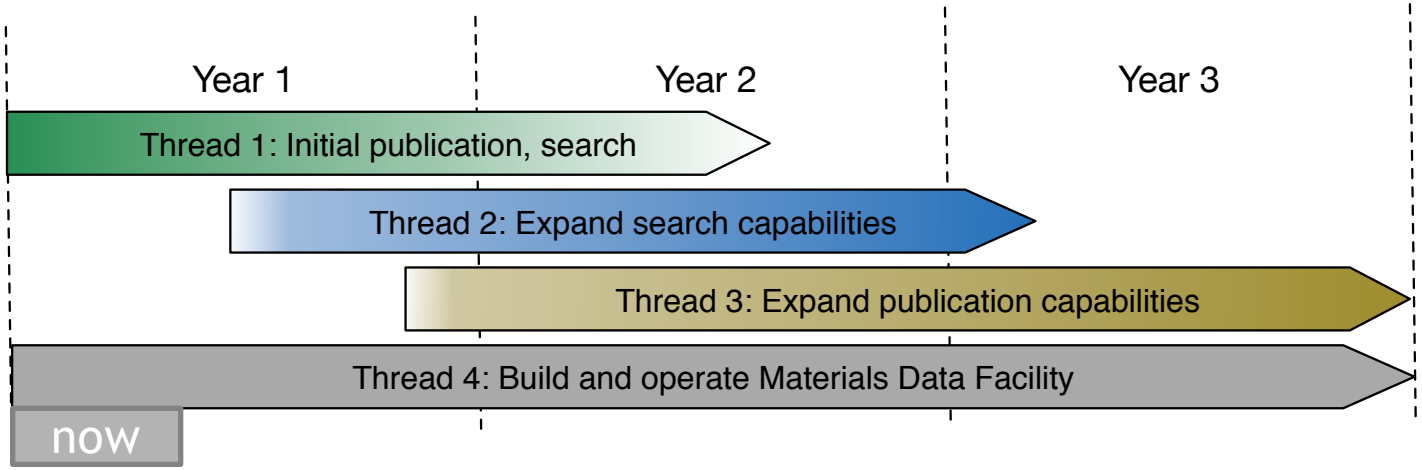
- Web interface to support data publication via Globus platform (identify management, user groups, optimized big data transfer)
- 100s of TB of storage at NCSA (scalable to many PB) more at Argonne (1.7 PB on Petrel)
- Help with developing metadata schemas to describe your research datasets
- Materials 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

Next Steps



- **Identify datasets to pilot publication pipelines and build schema repository**
- **Engage with researches working with materials data to understand use cases and learn friction points**
- **Please talk to us if you have data you want to share, publish, discover, …**

Presentations

- B. Blaiszik, K. Chard, R. Ananthakrishnan, J. Pruyne, J. Wozniak, M. Wilde, R. Osborn, S. Tuecke, I. Foster. "Globus Research Data Management Services and the Materials Data Facility", Sept. 2015, Center for PRedictive Integrated Structural Materials Science (PRISMS) Annual Meeting,
- B. Blaiszik, K. Chard, J. Pruyne, J. Towns, S. Tuecke, I. Foster. "The Materials Data Facility (MDF) – Data Services to Advance Materials Research", Oct. 2015, Materials Science and Technology Conference , Columbus, OH, USA.
- Ian Foster, B. Blaiszik, K. Chard. "The Materials Data Facility", Oct. 2015, 4th National Data Service Consortium Workshop, San Diego, CA, USA.
- B. Blaiszik, K. Chard, I. Foster. "The Materials Data Facility (MDF) – Data Services to Advance Materials Research", Dec. 2015, National Center for Supercomputing Applications Joint Materials Science Seminar, University of Illinois at Urbana-Champaign, Urbana, IL, USA.
- B. Blaiszik, K. Chard, I. Foster. "The Materials Data Facility (MDF) – Data Services to Advance Materials Research", Dec. 2015, Integrated Imaging Initiative Seminar, Argonne National Laboratory, Lemont, IL, USA.
- M. Ondrejcek, B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, J. Towns, K. McHenry, S. Tuecke, I. Foster. "Materials Data Facility - Data Services to Advance Materials Science Research", Feb. 2016, T2C2 DIBBS Meeting, University of Illinois at Urbana-Champaign, Urbana, IL, USA.
- B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, J. Towns, S. Tuecke, I. Foster. "Materials Data Facility - Data Services to Advance Materials Science Research", Feb. 2016, (invited) TMS 2016, Nashville, TN, USA.

Publications

- Kyle Chard, Jim Pruyne, Ben Blaiszik, Rachana Ananthakrishnan, Steven Tuecke, and Ian Foster. "Globus Data Publication as a Service: Lowering Barriers to Reproducible Science." In *e-Science (e-Science), 2015 IEEE 11th International Conference on*, pp. 401-410. IEEE, 2015.

- Ben Blaiszik, Kyle Chard, Jim Pruyne, Rachana Ananthakrishnan, John Towns, Steven Tuecke, and Ian Foster. "Building a Materials Data Facility- Data Services to Advance Materials Science Research", Materials Science and Technology 2015, 2015.

- Manuscript in prep for special issue of JOM on materials data services

Lessons Learned

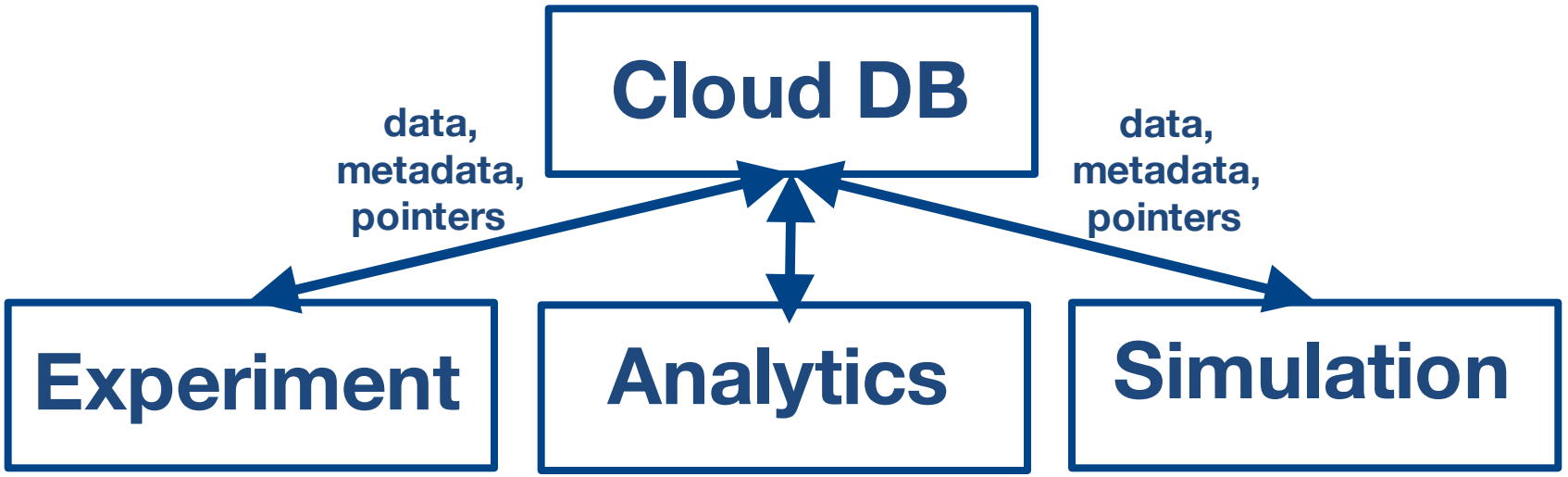
Lessons Learned

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

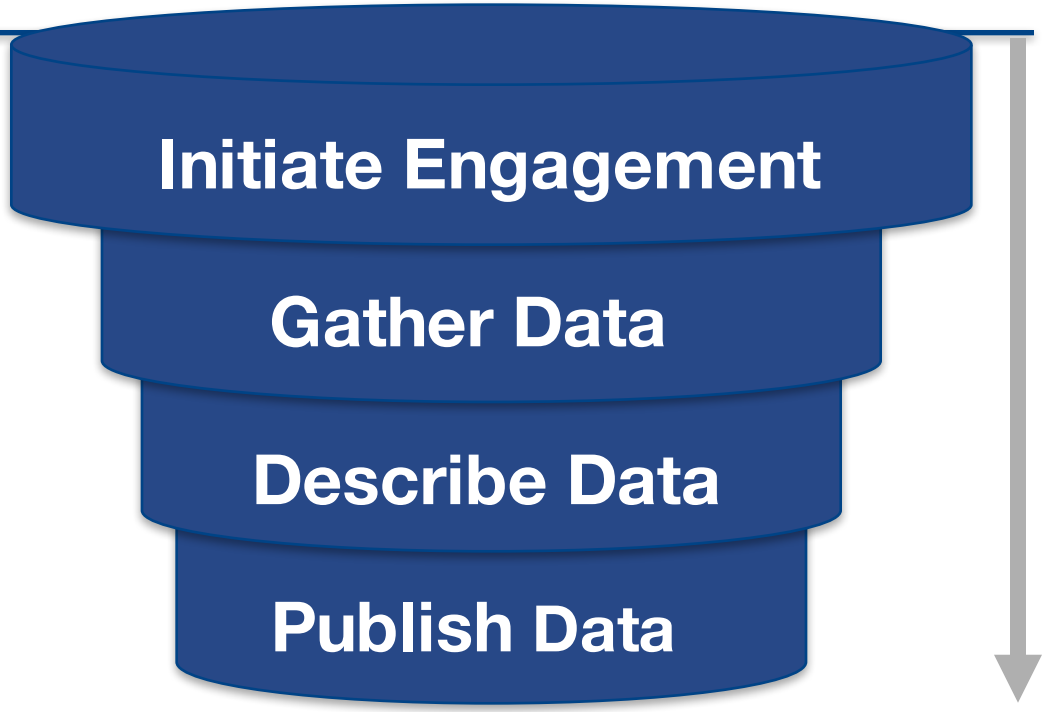
Data Interaction And Viz

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



Data Publication Pipeline Analysis

Enter	Proceed	Pending	x
23	15	7	1
14	6	8	0
6	3	3	0
3	2	1	0



- Numbers for late Nov. – early Feb.
  - A bit misleading since some of these are groups or centers rather than individuals
- No lossy stages detected yet
- Rate limiting step is data gathering+ data description
  - Building metadata schema
  - Populating schema with dataset values

The image part

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)

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