



# Materials Data Facility - Data Services to Advance Materials Science Research

Ben Blaiszik ([blaiszik@uchicago.edu](mailto:blaiszik@uchicago.edu)),  
Ian Foster ([foster@uchicago.edu](mailto:foster@uchicago.edu))

Kyle Chard, Rachana Ananthakrishnan, Steven Tuecke  
Michael Ondrejcek, Kenton McHenry, John Towns

[materialsdatafacility.org](http://materialsdatafacility.org)  
[globus.org](http://globus.org)

# Outline

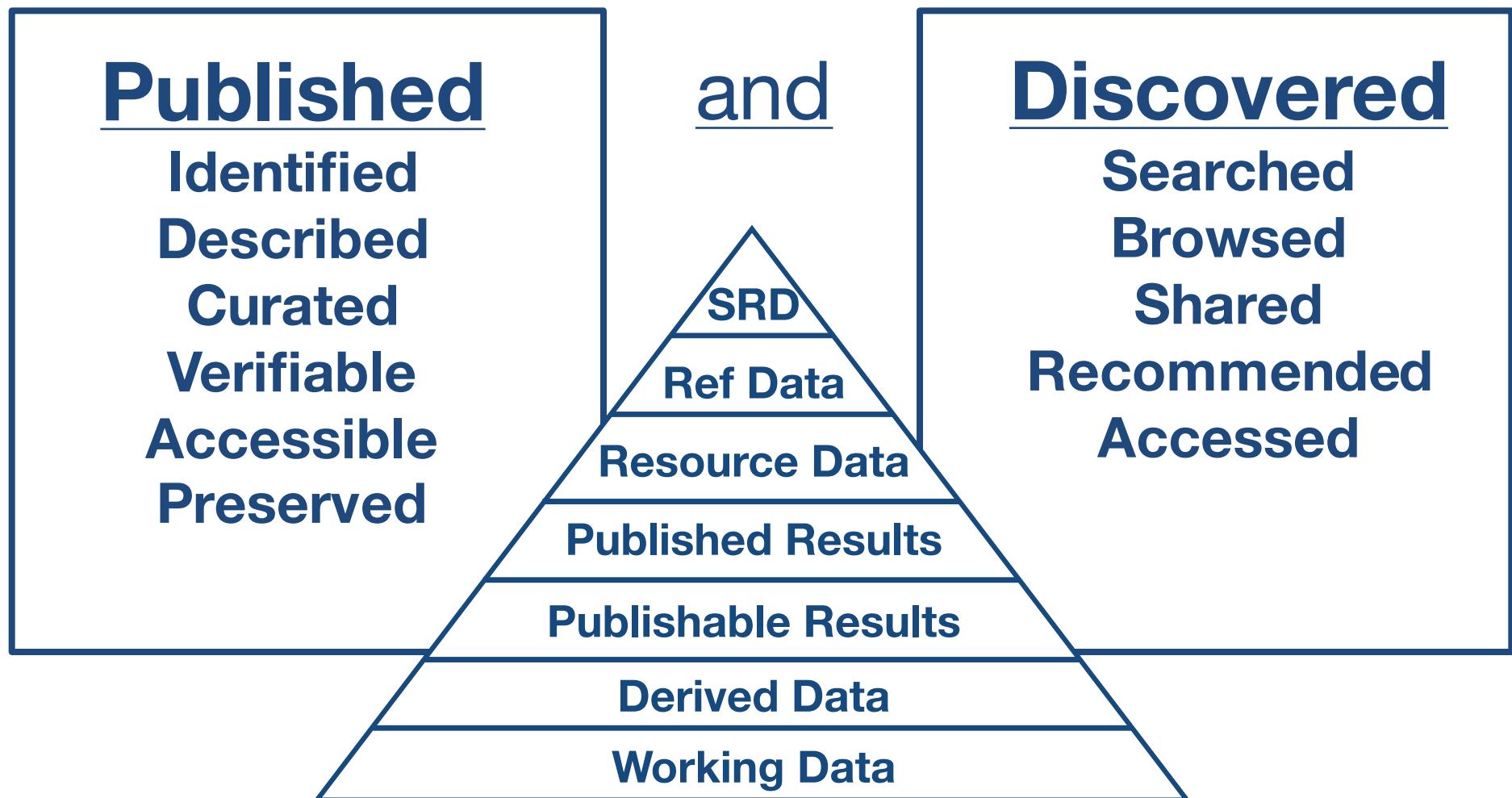
- Overview
  - MDF
  - Globus quick introduction
- Data Publication
  - Key MDF features
  - Publication walk-through
  - Early lessons learned
- Materials Resource Registration
- Next Steps

# Lessons Learned this Week

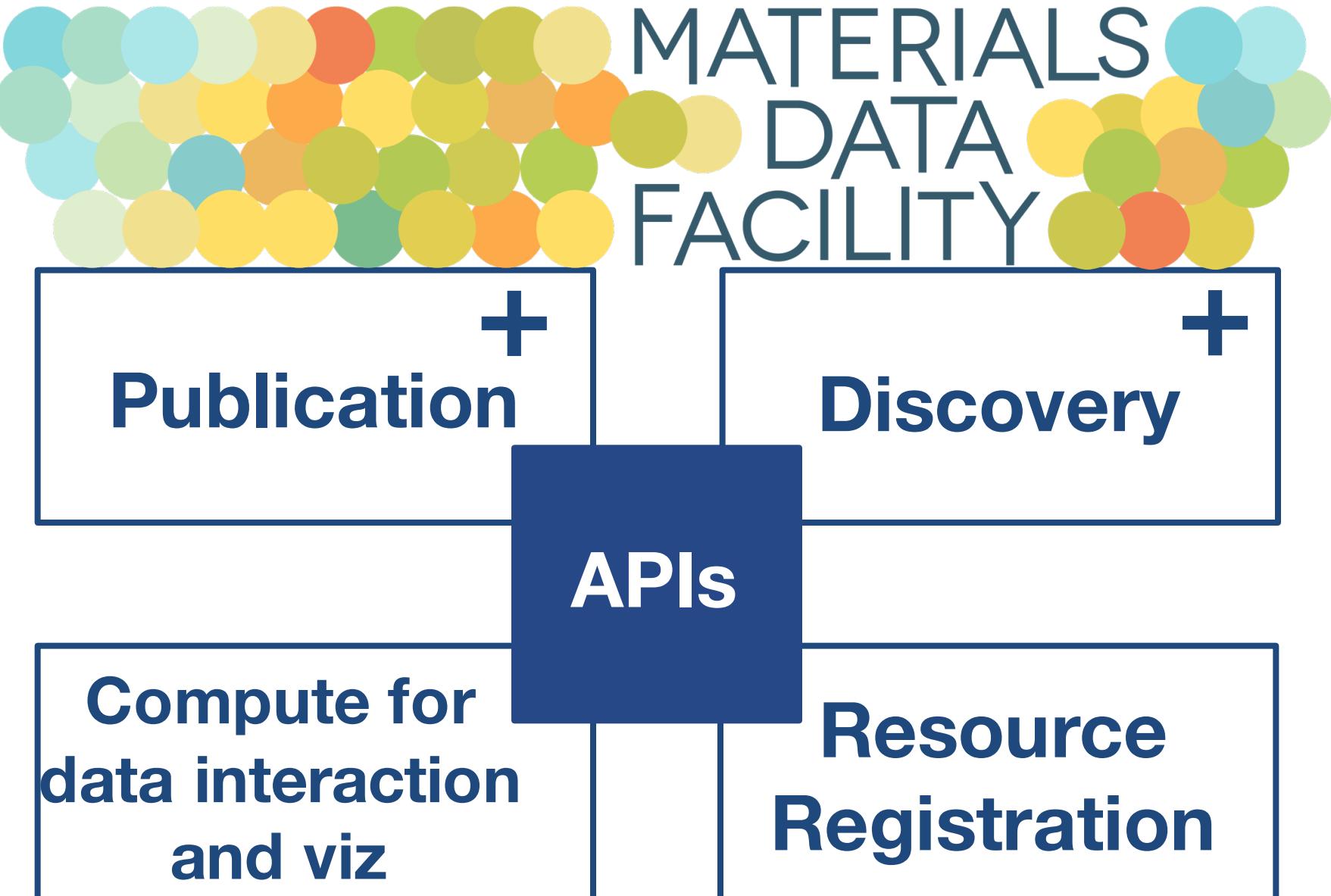
- **Finding good materials datasets from individual groups is hard**
- **Finding well-described materials datasets is hard**
- **Combining materials datasets across groups is harder**
- **Securing rights and permissions to those datasets for mining and sharing is hard**

# What is MDF?

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



# Data Service Infrastructure



+ - 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)**

The screenshot shows a search interface for the Materials Data Facility (MDF). The search term "MDF — TMS-2016-MDF" is entered in the search bar. The results are displayed under the heading "TOP HIT". The first result is "TMS-2016-MDF", which is highlighted with a blue bar. Below this, there are sections for "FOLDERS" and "DOCUMENTS". The "FOLDERS" section lists "mdf", "MDF - Desktop", "MDF - Google Drive", "MDF - git", and "mdf2iso". The "DOCUMENTS" section lists "20151208-NCSA-PIRE-MDF", "EZIDOrderForm-mdf", "20151006 - MDF - MGI Review - A...", "BuildingMDF-bb", "BuildingMDF", and "BuildingMDF-2.docx". To the right of the search results, there is a detailed description of the "TMS-2016-MDF" entry. It includes author information (I. Foster<sup>1,2</sup>, R. Ananthakrishnan, B. Blaiszik<sup>1</sup>, K. Chard<sup>1</sup>, J. Pruyne<sup>1</sup>, J. Towns<sup>1</sup>, S. Tuecke<sup>1,2</sup>), the Computation Institute, 5735 South Ellis Avenue, Chicago, IL, 60637, University of Chicago; Mathematics and Computer Science Division, Lemont, IL, 60439, Argonne National Laboratory; National Center for Supercomputing Applications, Champaign, IL, 61801, University of Illinois at Urbana-Champaign (UIUC); contact email: foster@anl.gov. It also lists keywords: materials, data, software as a service, data preservation. Below this, a paragraph describes the collaboration between Globus, the National Center for Supercomputing Applications, and the Center for Hierarchical Materials Design (CHIMaD), stating they are building the Materials Data Facility (MDF) to advance materials science research. Based on lessons learned from direct interactions with materials researchers, they are developing capabilities to promote open data sharing, simplify data publication and curation workflows, encourage data reuse, and provide powerful data discovery interfaces for data of all sizes and sources. Specifically, MDF services will allow individual researchers and institutions to 1) enable publication of large research datasets with flexible policies; 2) grant the ability to publish data directly from local storage, institutional data stores, or from cloud storage, without third-party publishers; 3) build extensible domain-specific metadata; 4) develop publication workflows; and 5) access a discovery model that allows researchers to search, interrogate, and build upon existing published data.

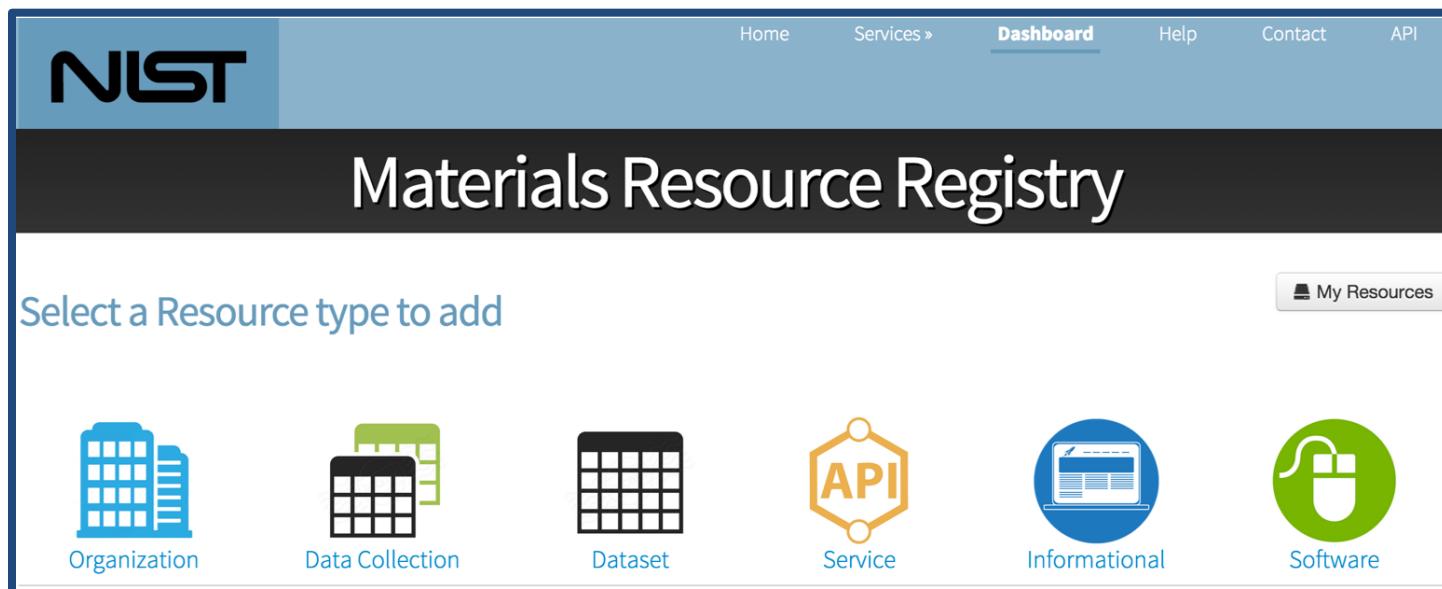
**Future...**

**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

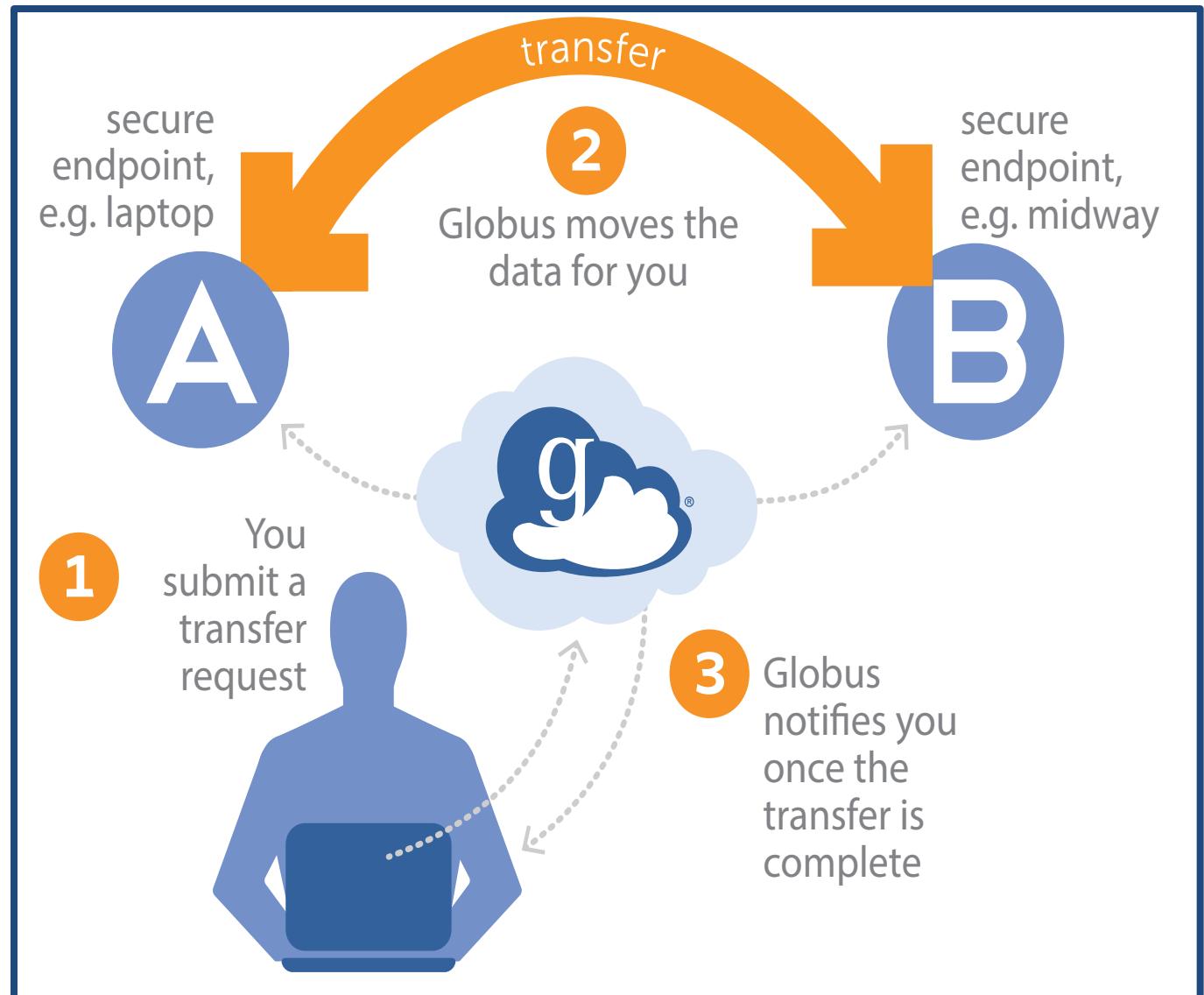
# Globus Background

## 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

The screenshot shows the Globus Web UI interface. At the top, there's a navigation bar with the Globus logo, 'Manage Data', 'Groups', 'Support', and a user dropdown for 'blaiszik'. Below the navigation is a banner for 'Get Globus Connect Personal'.

**Transfer Files**: This section shows two file selection panes. The left pane has an endpoint of 'blaiszik#macbookpro' and a path of '/~/Desktop/blaiszik-macbookpro/Voorhees'. It contains files '20A\_post\_0004.h5' (3.19 GB) and '20A\_post\_0005.h5' (3.15 GB). The right pane has an endpoint of 'globuspublish#jcpublish-test' and a path of '/mdf\_voorhees\_72/results/'. It also contains files '20A\_post\_0004.h5' (3.19 GB) and '20A\_post\_0005.h5' (3.15 GB). There are 'select all' checkboxes at the top of each pane.

**Activity**: This section shows a completed transfer. A green checkmark icon indicates success. The transfer details are: 'blaiszik#macbookpro to globuspublish#jcpublish-test', 'transfer completed a minute ago'. Below this, there are tabs for 'Overview' (selected) and 'Event Log'. The 'Overview' tab displays task ID 'c1191a64-ef5d-11e4-ab4a-22000b92c6ec', source 'blaiszik#macbookpro', destination 'globuspublish#jcpublish-test', and bytes transferred '6.34 GB'. The 'Event Log' tab is currently empty.

# 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

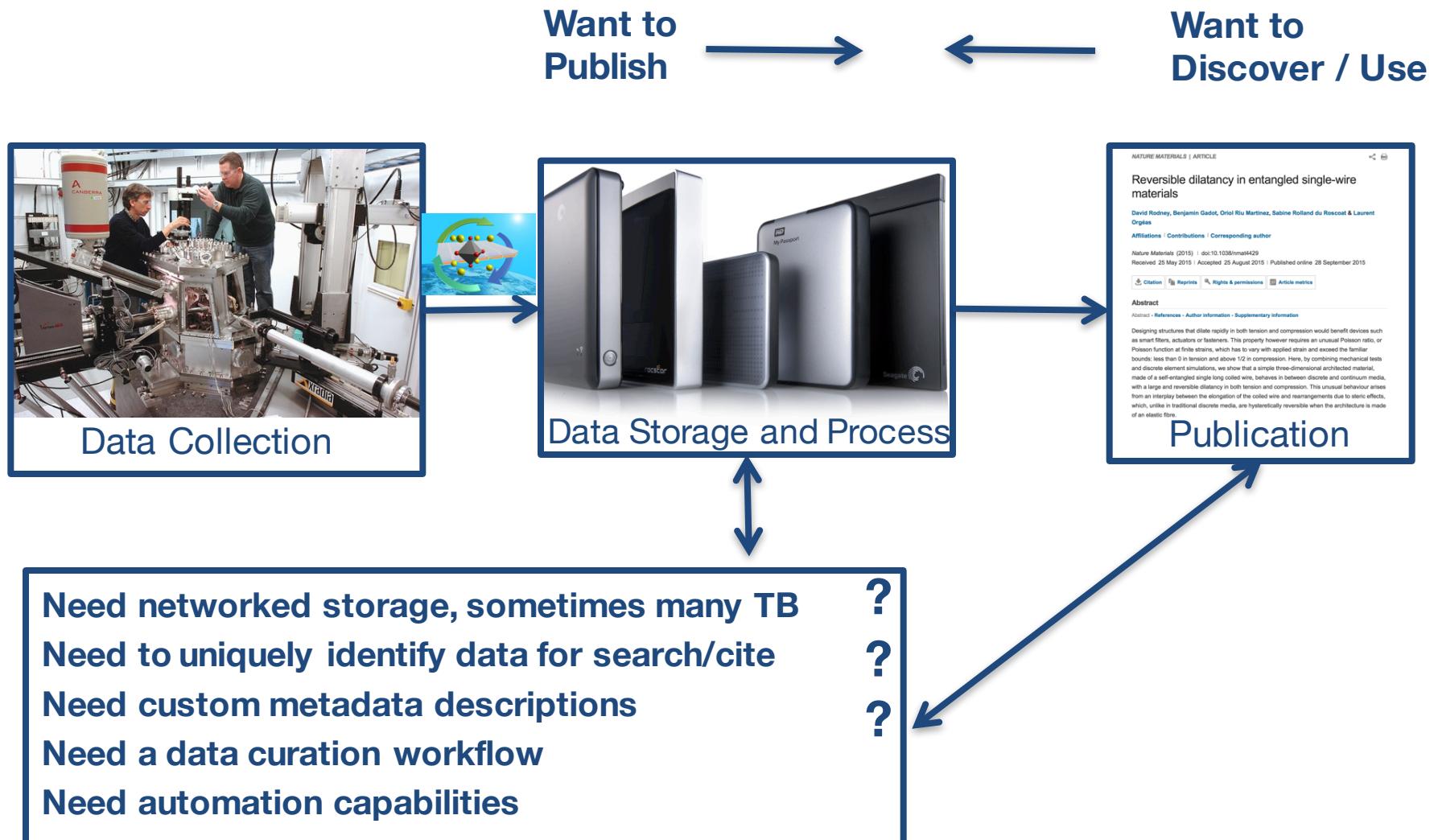


Data Storage and Process

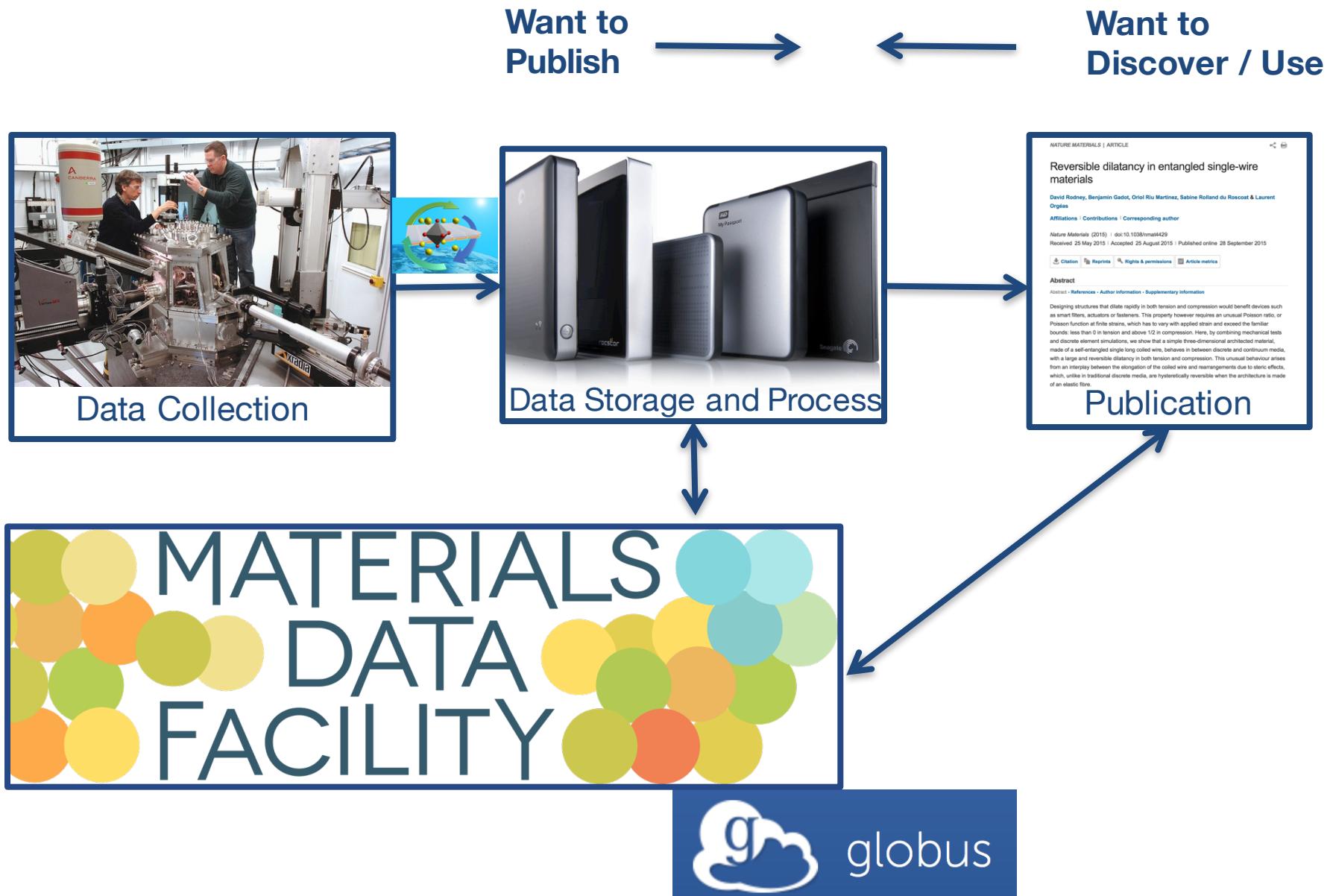
NATURE MATERIALS | ARTICLE  
Reversible dilatancy in entangled single-wire materials  
David Rodney, Benjamin Gadot, Oriol Riu Martinez, Sabine Rolland du Roscoat & Laurent Orgeas  
Affiliations | Contributions | Corresponding author  
Nature Materials (2015) | doi:10.1038/nmat4429  
Received: 28 May 2015 | Accepted: 25 August 2015 | Published online: 28 September 2015  
Citation | Reproductive rights & permissions | Article metrics  
Abstract  
Abstract · References · Author information · Supplementary information  
Describing 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 extensive element simulations, we show that a simple three-dimensional architected material, made of a wire entangled in a long chain of loops, exhibits this behaviour between tension and compression, with a large and reversible dilation 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 hysteresitically reversible when the architecture is made of an elastic fibre.

Publication

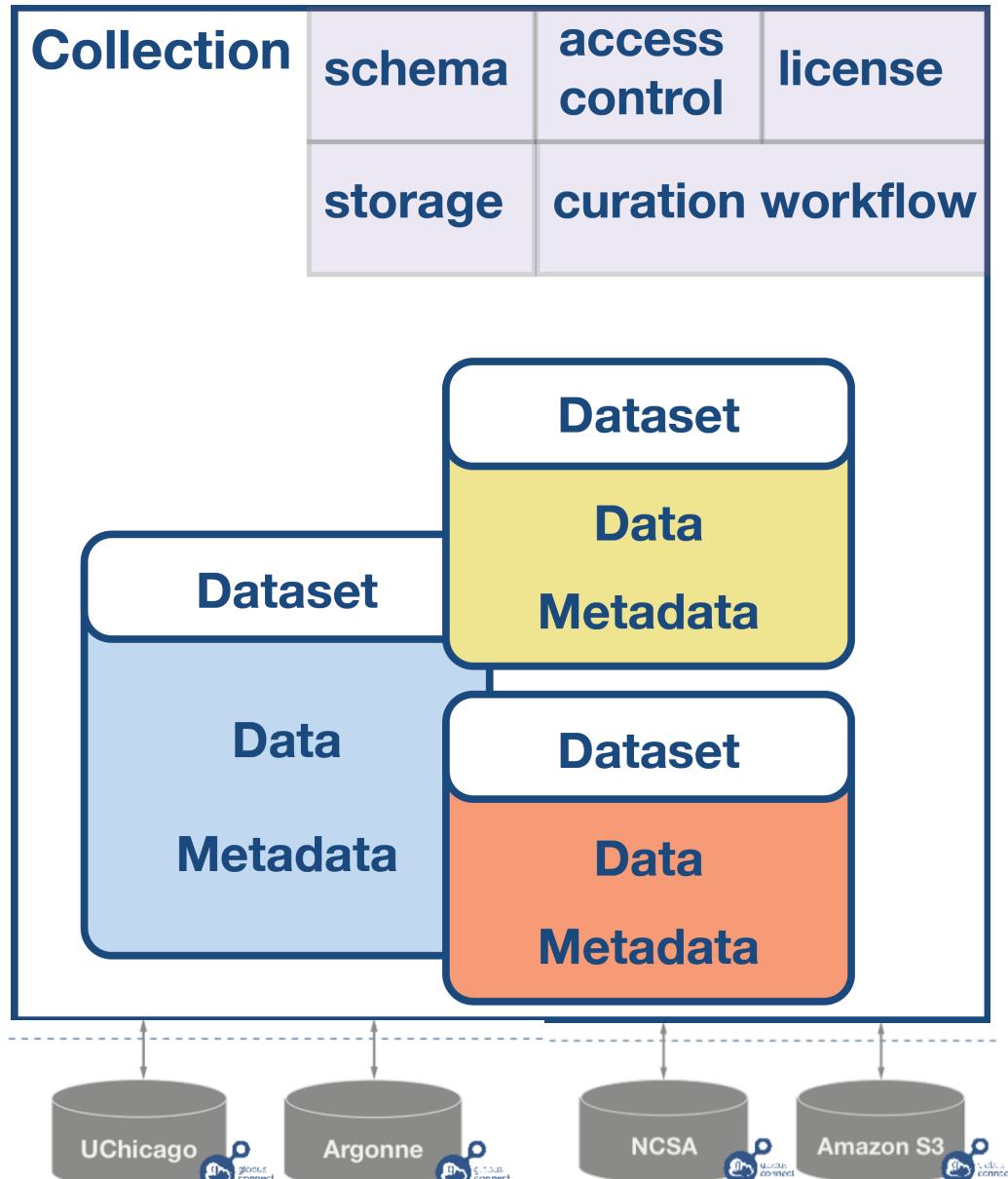
# Materials Data Publication/Discovery is Often a Challenge



# Materials Data Publication/Discovery is Often a Challenge



# Collection Model



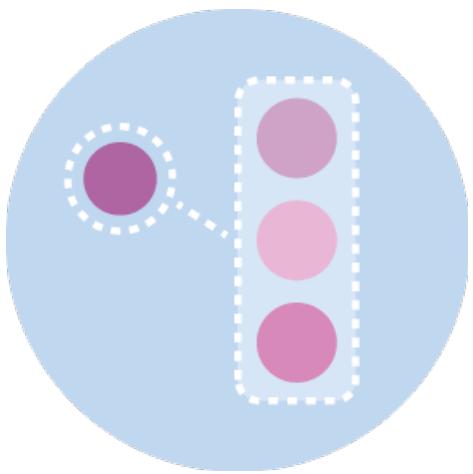
- Collections might be a research group or a research topic...
- Collections have specified
  - Mapping to storage endpoint
    - Currently handled as automatically created shared endpoints
  - Metadata schemas
  - Access control policies
  - Licenses
  - Curation workflows
- Collections contain
  - 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
- **156,835,245,562 MB TRANSFERRED**
- **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



# Share Data with Flexible ACLs



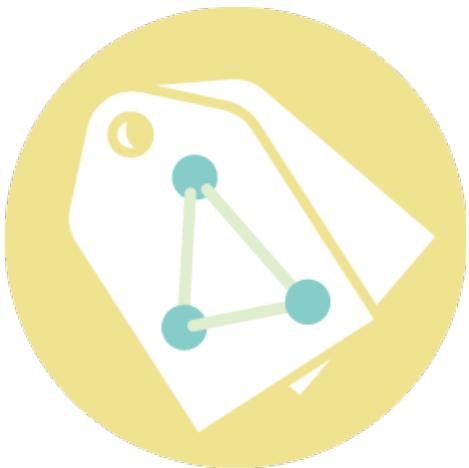
- Share data publicly, with a set of users, or keep data private

# 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 facetting (e.g. Amazon-style)**

# MaterialsDataFacility.org

The screenshot shows the homepage of MaterialsDataFacility.org. At the top, there's a navigation bar with links for 'ABOUT', 'GET STARTED', 'FEATURES', and 'HOW IT WORKS'. Below the navigation is a large central graphic featuring a white circle containing an orange icon of stacked files, with several smaller blue circles connected by lines around it, symbolizing a network or repository. The background is a dark teal color. Below this graphic, the text 'WHAT IS MDF?' is centered. A detailed description follows: 'The Materials Data Facility (MDF) is a scalable repository where materials scientists can publish, preserve, and share research data. The repository provides a focal point for the materials community, enabling publication and discovery of materials data of all sizes. MDF is a pilot project funded by NIST, and serves as the first pilot community of the National Data Service.' At the bottom of the main section, it says 'Funded and supported by' with logos for 'NIST National Institute of Standards and Technology U.S. Department of Commerce' and 'CHiMaD Center for Hierarchical Materials Design'. On the left side of the main content area, there's a 'GET STARTED' section with 'Publish Your Data' and 'Search for Data' buttons, and a note about signing up for a Globus account. To the right, a call-to-action box contains the text 'To get started, contact Ben Blaiszik blaiszik@uchicago.edu'. The bottom section, titled 'FEATURES', includes three items: 'Publication of large datasets' (with a green circular icon), 'Customizable metadata descriptions' (with a yellow circular icon), and 'Flexible access control' (with a purple circular icon). Each feature has a brief description below it.

ABOUT • GET STARTED • FEATURES • HOW IT WORKS

WHAT IS MDF?

The Materials Data Facility (MDF) is a scalable repository where materials scientists can publish, preserve, and share research data. The repository provides a focal point for the materials community, enabling publication and discovery of materials data of all sizes. MDF is a pilot project funded by NIST, and serves as the first pilot community of the National Data Service.

Funded and supported by **NIST**  
National Institute of Standards and Technology  
U.S. Department of Commerce and **CHiMaD**  
Center for Hierarchical Materials Design

GET STARTED

[Publish Your Data](#) [Search for Data](#)

Don't have a Globus account? [Sign up here!](#)

To get started,  
contact Ben Blaiszik  
[blaiszik@uchicago.edu](mailto:blaiszik@uchicago.edu)

FEATURES

Publication of large datasets

MDF offers researchers access to petabytes (PB) of reliable and high performance data storage via NCSA

Customizable metadata descriptions

MDF collection owners can define and use their own materials-specific metadata schemas to describe their published datasets

Flexible access control

Published datasets may be private, shared with a particular group of users, or shared publicly

25

# 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

The screenshot shows the MDF Collection Home page within the Globus interface. The top navigation bar includes links for Manage Data, Publish, Groups, Support, and Account. Below the navigation is a search bar and a menu for Browse & Discover, Data Publication Dashboard, and Communities & Collections.

The main content area features the MDF logo, which consists of a grid of colored circles (green, yellow, blue, red) followed by the text "MATERIALS DATA FACILITY". A brief description of the MDF is provided, stating it is a scalable repository for materials scientists to publish, preserve, and share research data. It is mentioned as a pilot project funded by NIST and part of the National Data Service. Contact information for Ben Blaiszik is also listed.

On the right side, there is an "Admin Tools" sidebar with options for Configure..., Create collection, and Create Sub-community.

The "Discover" section allows users to browse data by Author, Issue Date, Title, or Subject. The "Author" list includes Chard, Kyle; Dawson, Paul R.; Fife, Julie L.; Gibbs, John W.; Lienert, Ulrich; Miller, Matthew P.; Park, Jun-Sang; Pruyne; Ray, Atish K.; and Voorhees, Peter W. The "Issue Date" section shows 2016 and 2015. The "Subject" section lists X-ray (3), tomography (3), Al-Cu (2), CT (2), registration (1), segmentation (1), coursening (1), ct (1), lattice strain (1), and WAXS (1). A "next >" link is visible at the bottom of the subject list.

# MDF Collections

The screenshot shows the Globus MDF Collections interface. At the top, there's a blue header bar with the Globus logo and the word "globus". On the right side of the header are links for "Manage Data", "Publish", "Groups", "Support", and a user account named "blaiszik". Below the header, there are three navigation links: "Browse & Discover", "Data Publication Dashboard", and "Communities & Collections". The main content area has a light gray background and features a heading "Submit: Select Collection". Below this, there's a list of collection names, each preceded by a small icon: "APS Sector 1 « Materials Data Facility", "MDF Demo Collection « Materials Data Facility", "MICCoM « MICCoM Community", "TestMDF « Materials Data Facility", and "Voorhees Group « Materials Data Facility".

Submit: Select Collection

- APS Sector 1 « Materials Data Facility
- MDF Demo Collection « Materials Data Facility
- MICCoM « MICCoM Community
- TestMDF « Materials Data Facility
- Voorhees Group « Materials Data Facility

## 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**

The screenshot shows the Globus Data Publication Dashboard with the 'Describe' tab selected. The main heading is 'Submit: Describe this Dataset'. Below it, a note says: 'Please fill in the requested information about this submission below. In most browsers, you can use the tab key to move the cursor to the next input box or button, to save you having to use the mouse each time.' A placeholder text 'A name or title by which the data is known' is in the 'Title \*' field, which contains the value 'Al-Cu Coarsening 4D Tomography Dataset'. The 'Authors \*' section lists several names in separate input fields: Fife, Gibbs, Gulsoy, Park, Thornton, Voorhees, and a placeholder 'Last name, e.g. Smith'. To the right of these fields is a vertical column of red 'Remove Entry' buttons, each with a trash icon. A '+ Add More' button is at the bottom. The 'Publication Year \*' section includes dropdowns for Month ('(No Month)'), Day (''), and Year ('2014'). The 'Publisher \*' field contains 'Northwestern University'. At the bottom are navigation buttons: '< Previous', 'Cancel/Save', and a large blue 'Next >' button. A copyright notice at the bottom reads: '© 2010-2015 Computation Institute, University of Chicago, Argonne National Laboratory [legal](#)'.

# MDF Custom Metadata

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

globe globus

Publish Manage Data Groups Support blaiszik

Browse & Discover | Data Publication Dashboard | Communities & Collections

License Describe Describe Globus Transfer Verify Complete

## Submit: Describe this Dataset ?

Please fill further information about this submission below.

**Material** Al-Cu

**Volume Fraction Al** 15

**Volume Fraction Cu** 85

**Technique** x-ray tomography

**Pixel size (μm)** 1.4

**Beam energy (keV)** 20

**Instrumentation** Swiss Light Source - Tomographic Microscopy and Coherent Radiology Experiments beamline

Enter appropriate subject keywords

**Keywords**

in situ

4D coarsening

aluminum-copper alloys

dynamic morphological evolution

solid-liquid interfaces

Remove Entry

Remove Entry

Remove Entry

Remove Entry

Remove Entry

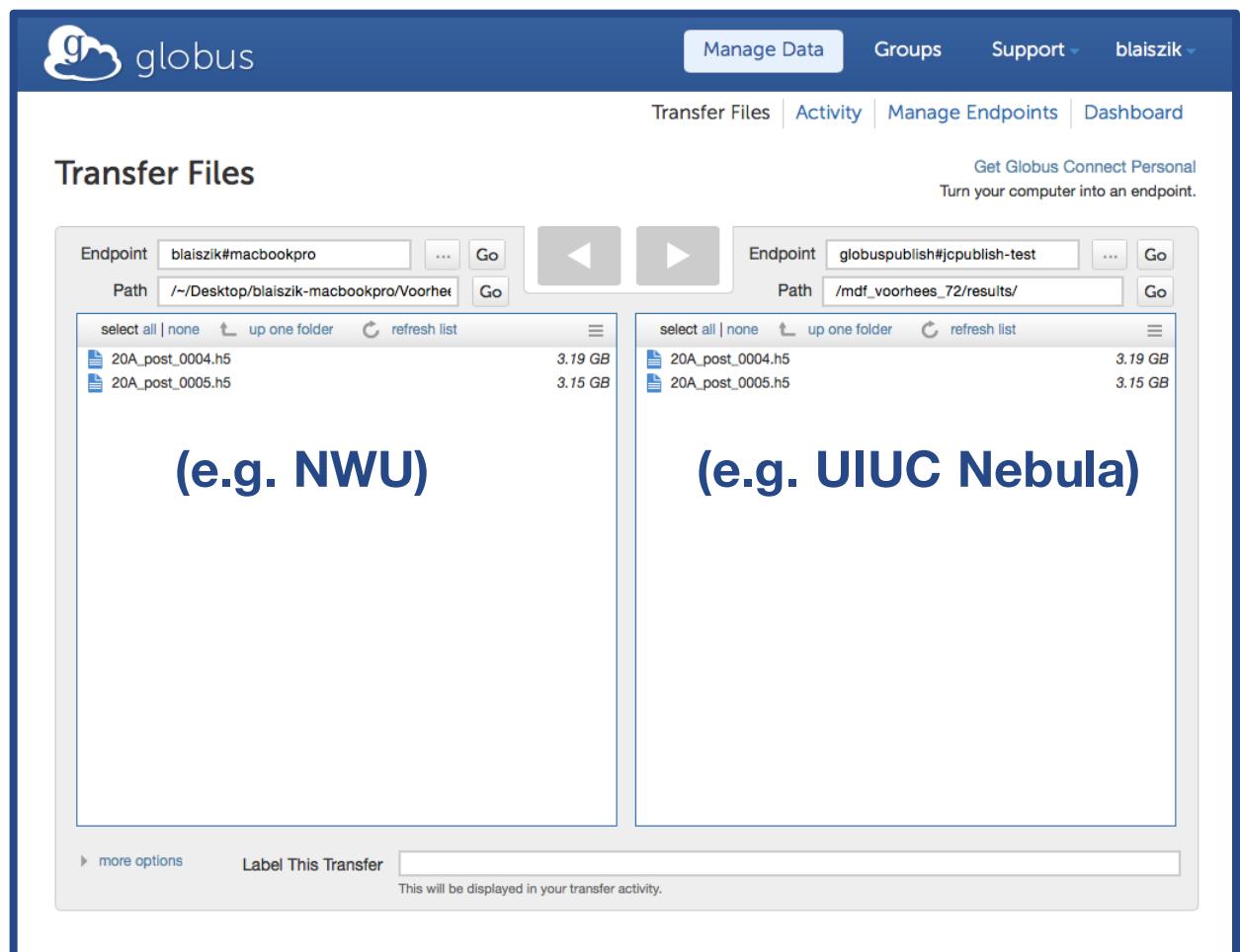
Remove Entry

+ Add More

< Previous Cancel/Save Next >

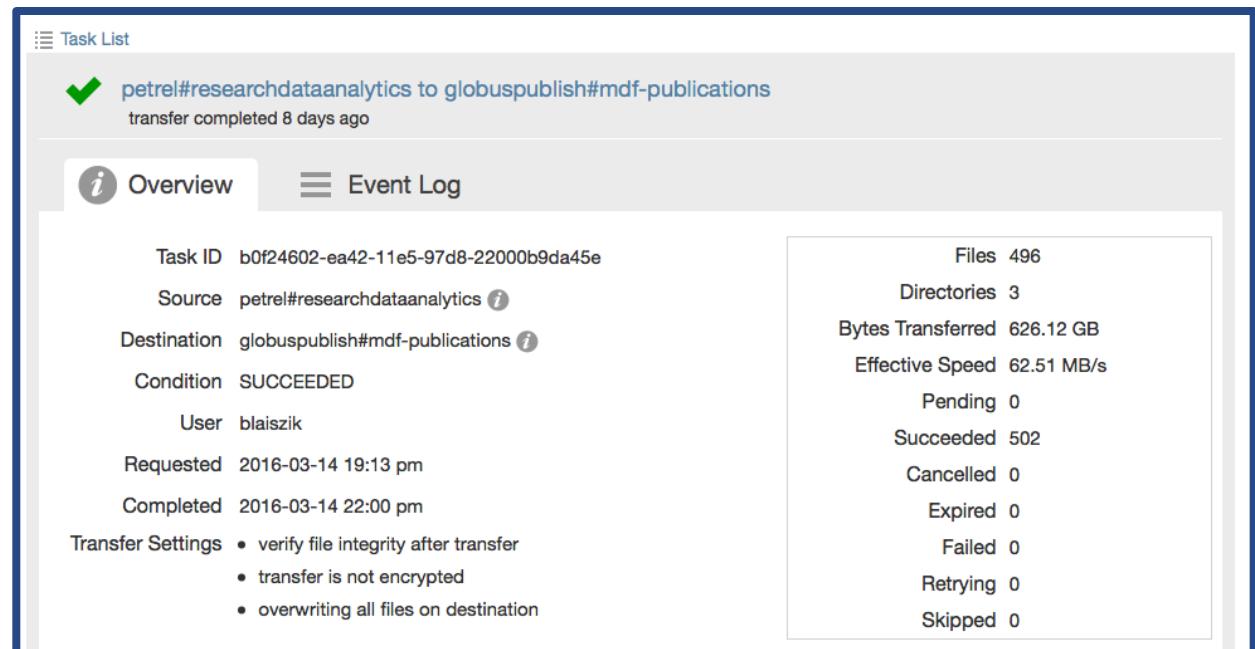
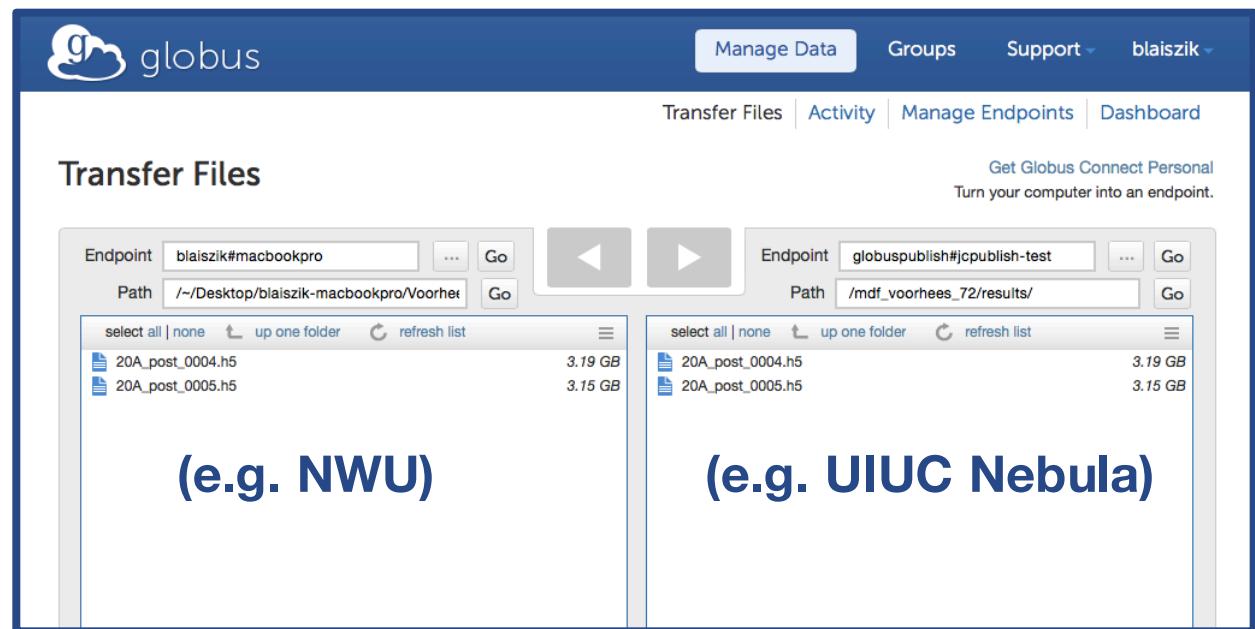
# Dataset Assembly

- 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

The screenshot shows a screenshot of the Globus Data Publication Dashboard. At the top, there's a blue header bar with the Globus logo on the left and navigation links: "Publish", "Manage Data", "Groups", "Support", and a user account dropdown for "blaiszik". Below the header, there are three links: "Browse & Discover", "Data Publication Dashboard", and "Communities & Collections". The main content area has a white background. It starts with a large "Thank You" heading. Below it, a message states: "The submission has been placed in the main archive. It has been assigned the following identifier:". Underneath that is a blue link: "<http://dx.doi.org/doi:10.18126/M2CC73>". At the bottom of the content area, there's a "Return to data publication dashboard" link. A thin horizontal line separates the content from the footer. The footer contains copyright information: "© 2010-2015 Computation Institute, University of Chicago, Argonne National Laboratory legal".

Thank You

The submission has been placed in the main archive. It has been assigned the following identifier:

<http://dx.doi.org/doi:10.18126/M2CC73>

[Return to data publication dashboard](#)

© 2010-2015 Computation Institute, University of Chicago, Argonne National Laboratory legal

Can optionally be DOI or Handle

# Dataset Record

globus

Publish Manage Data Groups Support blaiszik

Browse & Discover | Data Publication Dashboard | Communities & Collections

Search 

Please use this identifier to cite or link to this item: <http://bit.ly/1EGh9UL>

Title:	Al-Cu Coarsening 4D Tomography Dataset
Authors:	Fife, J.L. Gibbs, J.W. Gulsoy, E.B. Park, C.-L Thornton, K. Voorhees, P.W.
Keywords:	in situ 4D coarsening aluminum-copper alloys dynamic morphological evolution solid-liquid interfaces
Issue Date:	2014
Publisher:	Northwestern University
URI:	<a href="http://bit.ly/1EGh9UL">http://bit.ly/1EGh9UL</a>
Appears in Collections:	<a href="#">Voorhees Group X-Ray Tomography</a>

Admin Tools

Configure...  
Export Item  
Export (migrate) Item  
Export metadata

Files in This Item:

[globuspublish#jcpublish-test/mdf\\_voorhees\\_72/](#)

Show full item record 

Items in Globus are protected by copyright, with all rights reserved, unless otherwise indicated.

# Dataset Discovery

## Search Results

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.

[advanced search](#)

## Discover

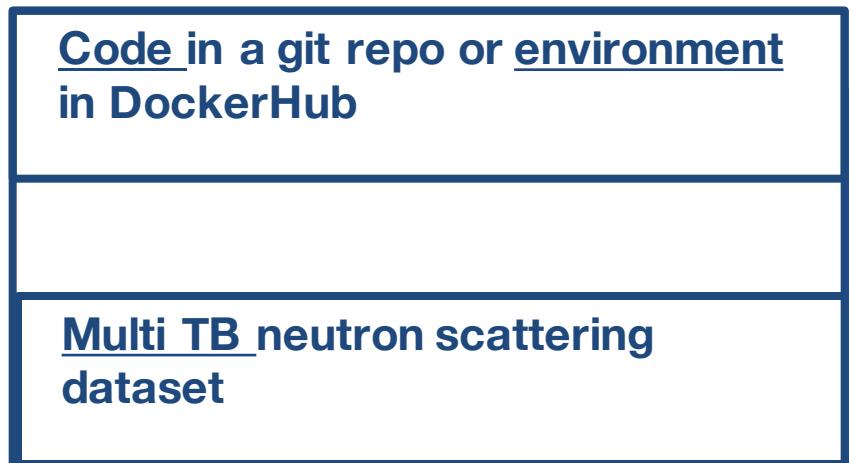
Author	Count
Fife, Julie L.	5
Gibbs, John W.	5
Voorhees, Peter W.	5
Dawson, Paul R.	1
Lienert, Ulrich	1
Miller, Matthew P.	1
Park, Jun-Sang	1
Ray, Atish K.	1

Issue Date	Count
2016	6

[previous](#) [1](#) [next](#)

# Example Use Case

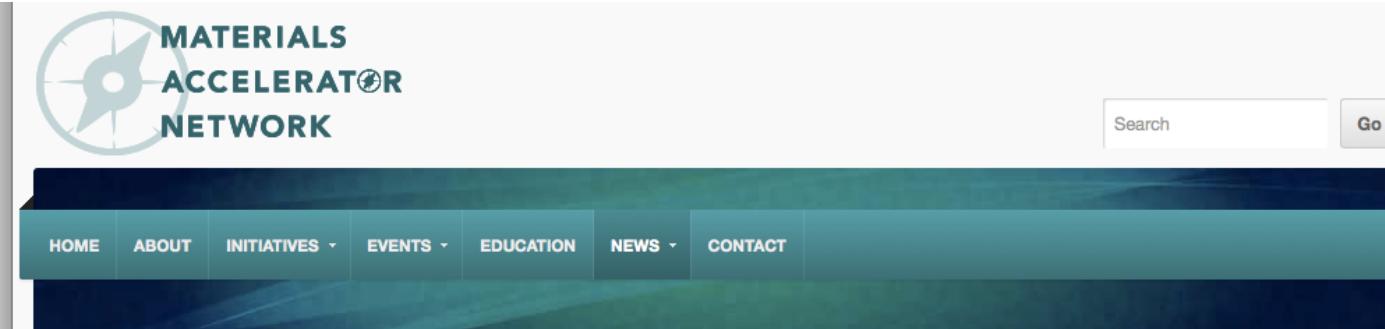
## Publishing a Dataset and Evolving Analysis Code (or Env)



# Registering Materials Resources

[w/ NIST – Youssef, Dima] 39

# Materials Resource Registry



## MATERIALS SCIENCE AND ENGINEERING DATA CHALLENGE RESOURCES

This page provides links to data, informatics tools, and related information to support participation in the [Materials Science and Engineering Data Challenge](#). Thank you to the many supportive colleagues who have helped us develop this list – it continues to be a team effort. Please note that due to practical limitations these resources have not been reviewed by us and are provided without any assurances of accuracy, quality, performance, or value for your particular project.

### About the Materials Science and Engineering Data Challenge

The Air Force Research Lab (AFRL), National Institute of Standards and Technology (NIST), and National Science Foundation (NSF) are sponsoring a contest to pursue the goals of the [Materials Genome Initiative](#). The focus is to seek novel uses of accessible digital data to advance Materials Science and Engineering knowledge, and accelerate the transition to industrial applications.

<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

## Search for Resources

3 results



All Resources



Organizations



Data Collections



Datasets



Services



Informational Sites



Software

Brief Results View

- Resource Type:
- All Resources
  - Organization
  - Data Collection
  - Repository
  - Project Archive
  - Database
  - Dataset
  - Service
  - Informational Site
  - Software

[Clear Refinements](#)

### MAterials Simulation Toolkit

[Resource Details](#) [Go To](#)

publisher University of Wisconsin-Madison Computational Materials Group

subject diffusion, defects, workflow

### TomoPy

[Resource Details](#) [Go To](#)

publisher Github

subject tomography, python, reconstruction, software, processing

### ChemSpider

[Resource Details](#) [Go To](#)

publisher Royal Society of Chemistry

subject chemical structures, chemical data

## Browse Results

[w/ NIST - Youssef, Dima] 42

# Lessons Learned

# Understanding Incentives is Critical

## Increasing Impact

- Increase paper citations<sup>1</sup>
- 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)

<sup>1</sup> 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<sup>3</sup>
  - (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!

NIST



U.S. DEPARTMENT OF  
**ENERGY**



THE UNIVERSITY OF  
**CHICAGO**