

# 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



NORTHWESTERN  
UNIVERSITY



CHiMaD

# Summary Agenda

---

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

# 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

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

Ian Foster ([foster@uchicago.edu](mailto:foster@uchicago.edu))

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

Kyle Chard, Rachana Ananthakrishnan, Steven Tuecke

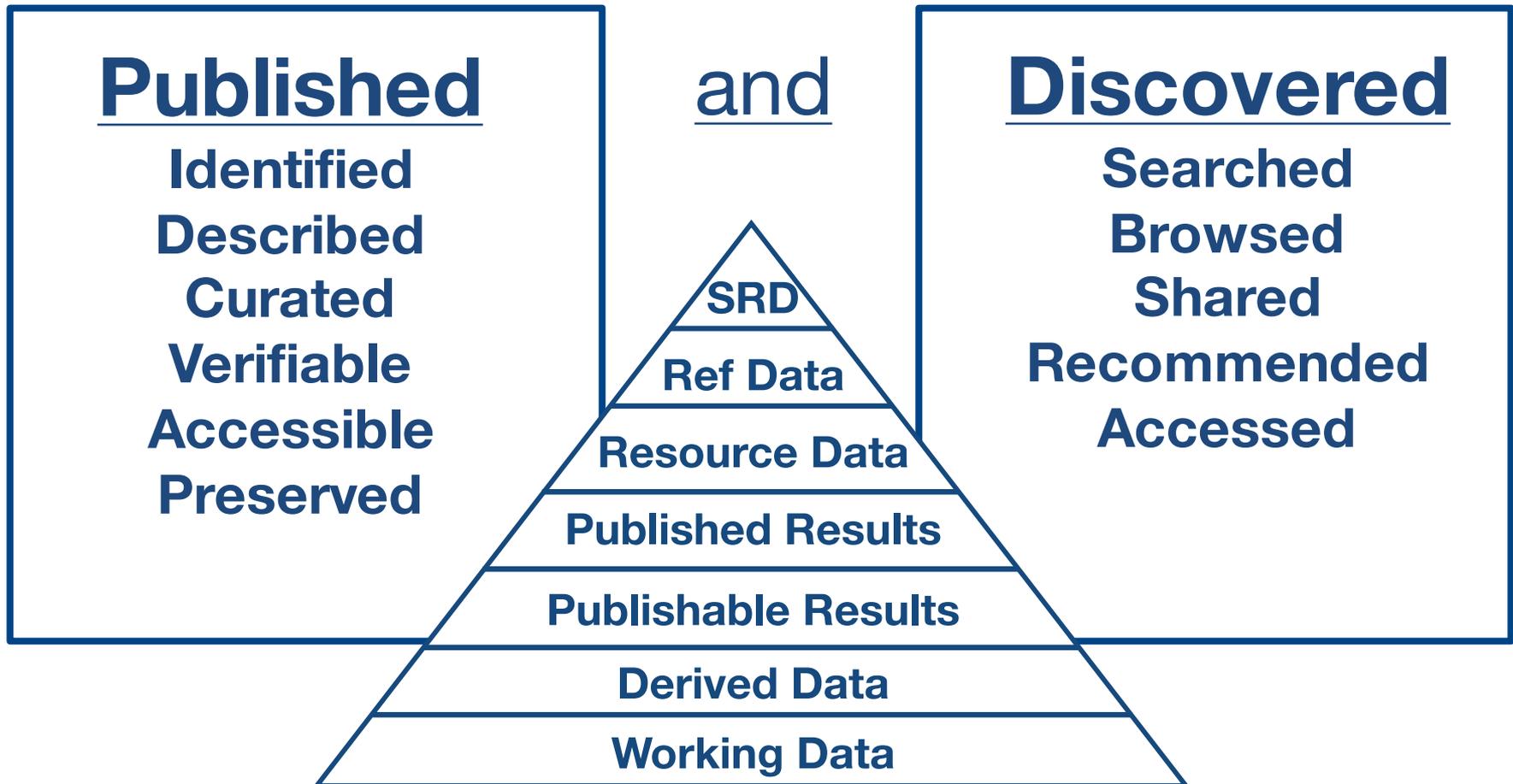
Michael Ondrejcek, Kenton McHenry, John Towns

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

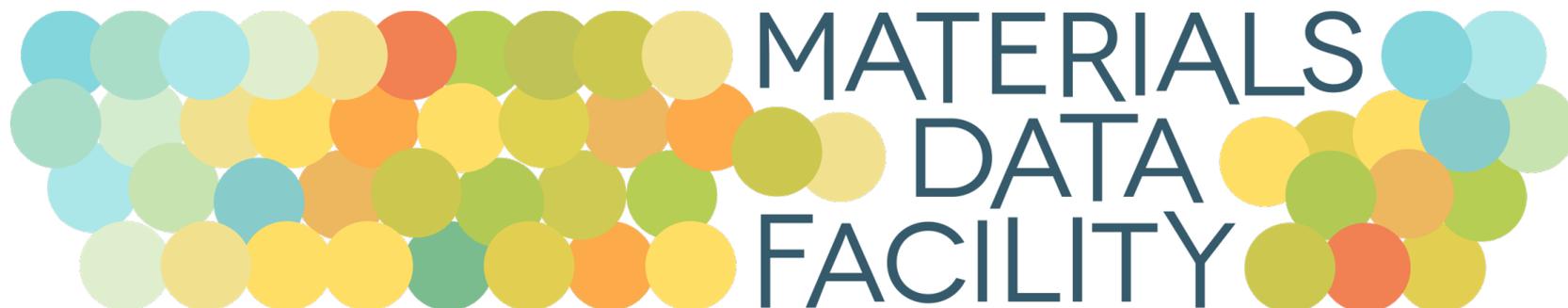
[globus.org](http://globus.org)

# What is MDF?

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



# Data Service Infrastructure



**Publication** +

**Discovery** +

**APIs**

**Data  
Interaction  
And Viz**

**Resource  
Registration**

# Publication

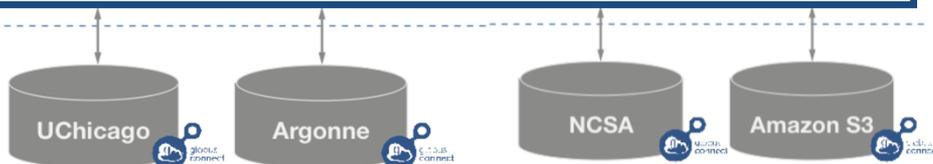
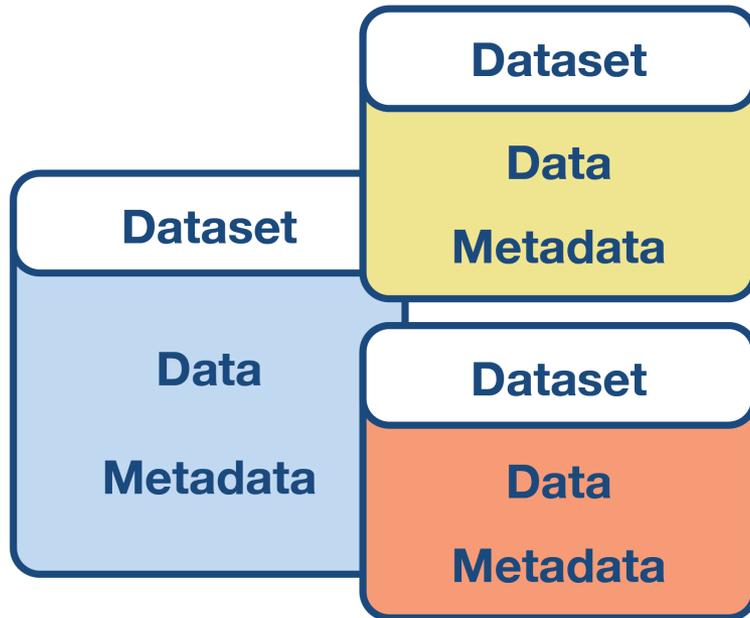
Opened to external users in  
mid Feb 2016

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

Collection	schema	access control	license
	storage	curation workflow	



- **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

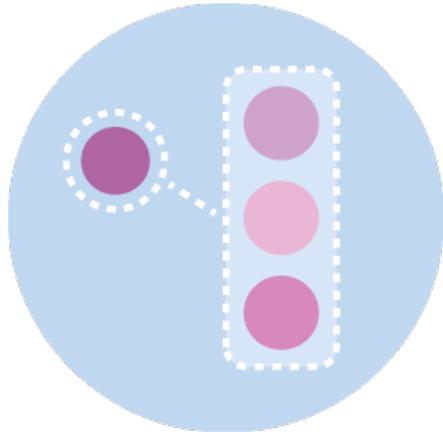


- Leverages Globus production capabilities for file transfer (i.e. dataset assembly), user authentication, and access control groups

**157,488,739,723** MB  
TRANSFERRED

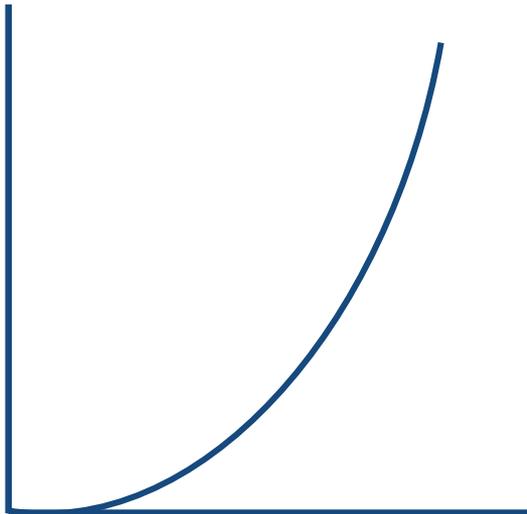
- **Storage resources available now**
  - 100s of TB of reliable storage @ NCSA, and more storage at Argonne
  - Globus endpoint at ncsa#mdf on Nebula
  - Expandable to many PBs as necessary
  - Automated tape backup for reliability (in progress)
- **Researchers can optionally use your own local or institutional 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

Dataset Downloads



Time

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

## 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 'Submit: Describe this Dataset' form in the Globus interface. The form is titled 'Submit: Describe this Dataset' with a help icon. It includes a progress bar with steps: License, Describe (selected), Describe, Globus Transfer, Verify, and Complete. The main form fields are:

- Title \***: A text input field containing 'Al-Cu Coarsening 4D Tomography Dataset'.
- Authors \***: A section for listing authors. It includes a list of input fields for names: Fife, Gibbs, Gulsoy, Park, Thornton, Voorhees, and a field for 'Last name, e.g. Smith'. To the right, there are two columns of input fields for initials: J.L., J.W., E.B., C.-L., K., and P.W., followed by a field for 'First name(s) + "Jr", e.g. Donald Jr'. Each author entry has a red 'Remove Entry' button, and there is an 'Add More' button at the bottom right of the authors section.
- Publication Year \***: A section for the year. It includes a 'Month' dropdown menu (set to '(No Month)'), a 'Day' input field, and a 'Year' input field (set to '2014').
- Publisher \***: A text input field containing 'Northwestern University'.

At the bottom of the form, there are three buttons: '< Previous', 'Cancel/Save', and 'Next >'. The footer of the page 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**

The screenshot shows the 'Submit: Describe this Dataset' form in the Globus Data Publication Dashboard. The form is titled 'Submit: Describe this Dataset' with a help icon. Below the title, it says 'Please fill further information about this submission below.' The form contains several input fields for metadata:

- 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

Below these fields, there is a section for 'Keywords' with the instruction 'Enter appropriate subject keywords'. The keywords entered are:

- in situ
- 4D coarsening
- aluminum-copper alloys
- dynamic morphological evolution
- solid-liquid interfaces

Each keyword has a 'Remove Entry' button to its right. There is also an '+ Add More' button at the bottom right of the keyword list. At the bottom of the form, there are three buttons: '< Previous', 'Cancel/Save', and '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

The screenshot displays the Globus 'Transfer Files' interface. At the top, the Globus logo and navigation links (Manage Data, Groups, Support, blaiszik) are visible. Below the navigation, there are tabs for 'Transfer Files', 'Activity', 'Manage Endpoints', and 'Dashboard'. The main content area is titled 'Transfer Files' and features two side-by-side file explorer panels. The left panel shows a local endpoint 'blaiszik#macbookpro' with a path of '/~/Desktop/blaiszik-macbookpro/Voorher' and contains two files: '20A\_post\_0004.h5' (3.19 GB) and '20A\_post\_0005.h5' (3.15 GB). The right panel shows a remote endpoint 'globuspublish#jcpublish-test' with a path of '/mdf\_voorhees\_72/results/' and contains the same two files. Both panels include navigation controls like 'select all', 'none', 'up one folder', and 'refresh list'. Below the panels, there is a 'Label This Transfer' field and a note: 'This will be displayed in your transfer activity.'

# 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

The screenshot shows the Globus 'Transfer Files' interface. At the top, there's a navigation bar with 'Manage Data', 'Groups', 'Support', and the user 'blaiszik'. Below that are links for 'Transfer Files', 'Activity', 'Manage Endpoints', and 'Dashboard'. The main area is titled 'Transfer Files' and includes a sub-header 'Get Globus Connect Personal Turn your computer into an endpoint.' Two file transfer windows are shown side-by-side. The left window has 'Endpoint: blaiszik#macbookpro' and 'Path: /~/Desktop/blaiszik-macbookpro/Voorhees'. It contains a file list with two items: '20A\_post\_0004.h5' (3.19 GB) and '20A\_post\_0005.h5' (3.15 GB). Below the list is the text '(e.g. NWU)'. The right window has 'Endpoint: globuspublish#jcpublish-test' and 'Path: /mdf\_voorhees\_72/results/'. It contains the same file list. Below the list is the text '(e.g. UIUC Nebula)'.

The screenshot shows the 'Task List' interface. At the top, there's a 'Task List' header. Below it, a green checkmark indicates a successful task: 'petrel#researchdataanalytics to globuspublish#mdf-publications' with the note 'transfer completed 8 days ago'. There are two tabs: 'Overview' (selected) and 'Event Log'. The 'Overview' tab displays task details in two columns. The left column shows: Task ID (b0f24602-ea42-11e5-97d8-22000b9da45e), Source (petrel#researchdataanalytics), Destination (globuspublish#mdf-publications), Condition (SUCCEEDED), User (blaiszik), Requested (2016-03-14 19:13 pm), and Completed (2016-03-14 22:00 pm). The right column shows: Files (496), Directories (3), Bytes Transferred (626.12 GB), Effective Speed (62.51 MB/s), Pending (0), Succeeded (502), Cancelled (0), Expired (0), Failed (0), Retrying (0), and Skipped (0). Below the task details, there are 'Transfer Settings' listed as: verify file integrity after transfer, transfer is not encrypted, and overwriting all files on destination.

# Dataset Curation

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

The screenshot shows the Globus interface for dataset curation. At the top, there is a navigation bar with the Globus logo, a 'Publish' button, and links for 'Manage Data', 'Groups', 'Support', and 'blaiszik'. Below the navigation bar, there are links for 'Browse & Discover', 'Data Publication Dashboard', and 'Communities & Collections'. The main heading is 'Perform Task'. Below this, a message states: 'The following item has been submitted to collection **Voorhees Group X-Ray Tomography**. Please review the item, check that it meets the criteria for entry into the collection. After reviewing the item, please approve or reject the item using the controls at the bottom of the page.'

The submission details are as follows:

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

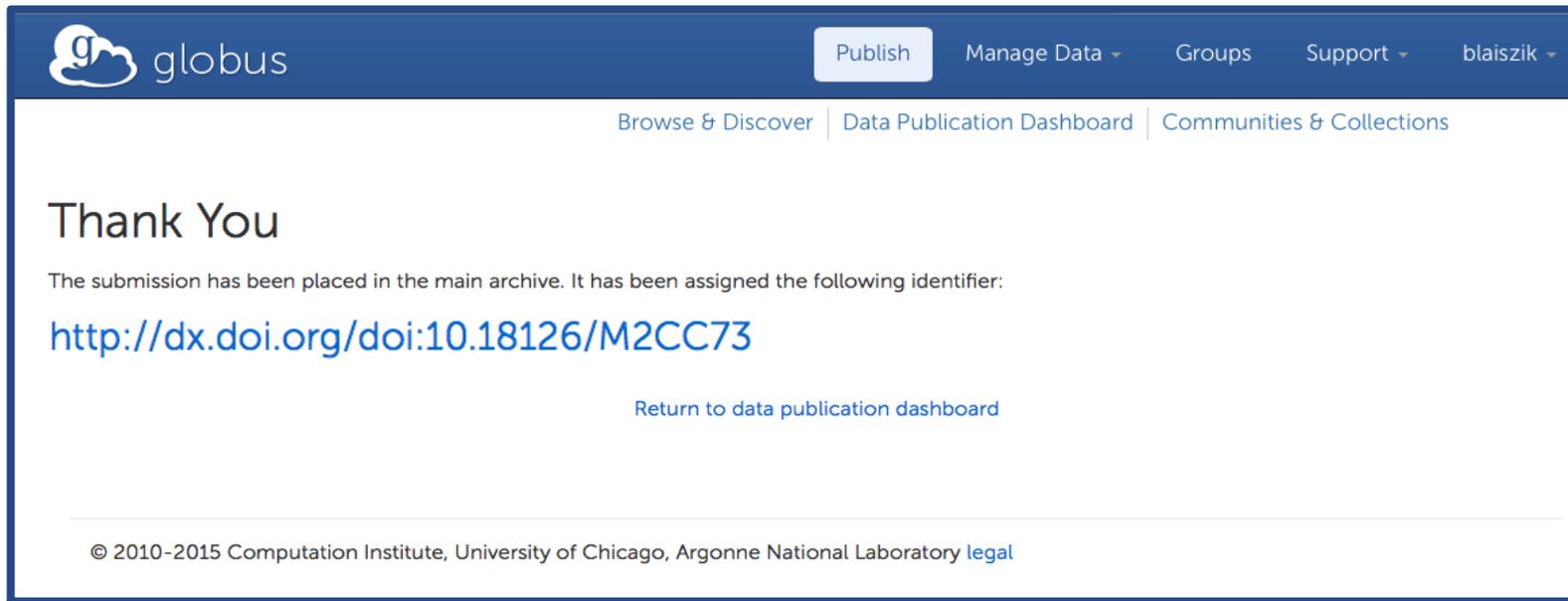
Files in This Item:

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

At the bottom, there are four action buttons with their respective descriptions:

<b>Approve</b>	If you have reviewed the item and it is suitable for inclusion in the collection, select "Approve".
<b>Reject</b>	If you have reviewed the item and found it is <b>not</b> suitable for inclusion in the collection, select "Reject". You will then be asked to enter a message indicating why the item is unsuitable, and whether the submitter should change something and re-submit.
<b>Do Later</b>	If you wish to leave this task for now, and return to the data publication dashboard, use this option.
<b>Return Task to Pool</b>	To return the task to the pool so that another user can perform the task, use this option.

# Mint a Permanent Identifier



The screenshot shows the Globus interface. At the top left is the Globus logo. To its right are navigation links: 'Publish', 'Manage Data', 'Groups', 'Support', and 'blaiszik'. Below the navigation bar are three main menu items: 'Browse & Discover', 'Data Publication Dashboard', and 'Communities & Collections'. The main content area displays a 'Thank You' message, stating that the submission has been placed in the main archive and assigned a DOI. The DOI link is <http://dx.doi.org/doi:10.18126/M2CC73>. Below the link is a button labeled 'Return to data publication dashboard'. At the bottom of the page, there is a copyright notice: '© 2010-2015 Computation Institute, University of Chicago, Argonne National Laboratory legal'.

Can optionally be DOI or Handle

# Dataset Record

Publish Manage Data ▾ Groups Support ▾ blaiszik ▾

[Browse & Discover](#) | [Data Publication Dashboard](#) | [Communities & Collections](#)

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

Title:	Al-Cu Coarsening 4D Tomography Dataset
Authors:	<a href="#">Fife, J.L.</a> <a href="#">Gibbs, J.W.</a> <a href="#">Gulsoy, E.B.</a> <a href="#">Park, C.-L</a> <a href="#">Thornton, K.</a> <a href="#">Voorhees, P.W.</a>
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

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	<i>Park, Jun-Sang; Ray, Atish K.; Dawson, Paul R.; Lienert, Ulrich; Miller, Matthew P.</i>
11-Feb-2016	Dataset for Segmentation of Four-dimensional, X-ray Computed Tomography Data	<i>Gibbs, John W.; Voorhees, Peter W.; Fife, Julie L.</i>
19-Mar-2016	Liquid-solid Metallic Mixture Coarsening Data - 80% solid	<i>Gibbs, John W.; Voorhees, Peter W.; Fife, Julie L.</i>
19-Mar-2016	Liquid-solid Metallic Mixture Coarsening Data - 28% Solid	<i>Gibbs, John W.; Voorhees, Peter W.; Fife, Julie L.</i>
19-Mar-2016	Liquid-solid Metallic Mixture Coarsening Data - 35% Solid	<i>Gibbs, John W.; Voorhees, Peter W.; Fife, Julie L.</i>
19-Mar-2016	Liquid-solid Metallic Mixture Coarsening Data - 55% solid	<i>Gibbs, John W.; Voorhees, Peter W.; Fife, Julie L.</i>

## 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
------	---

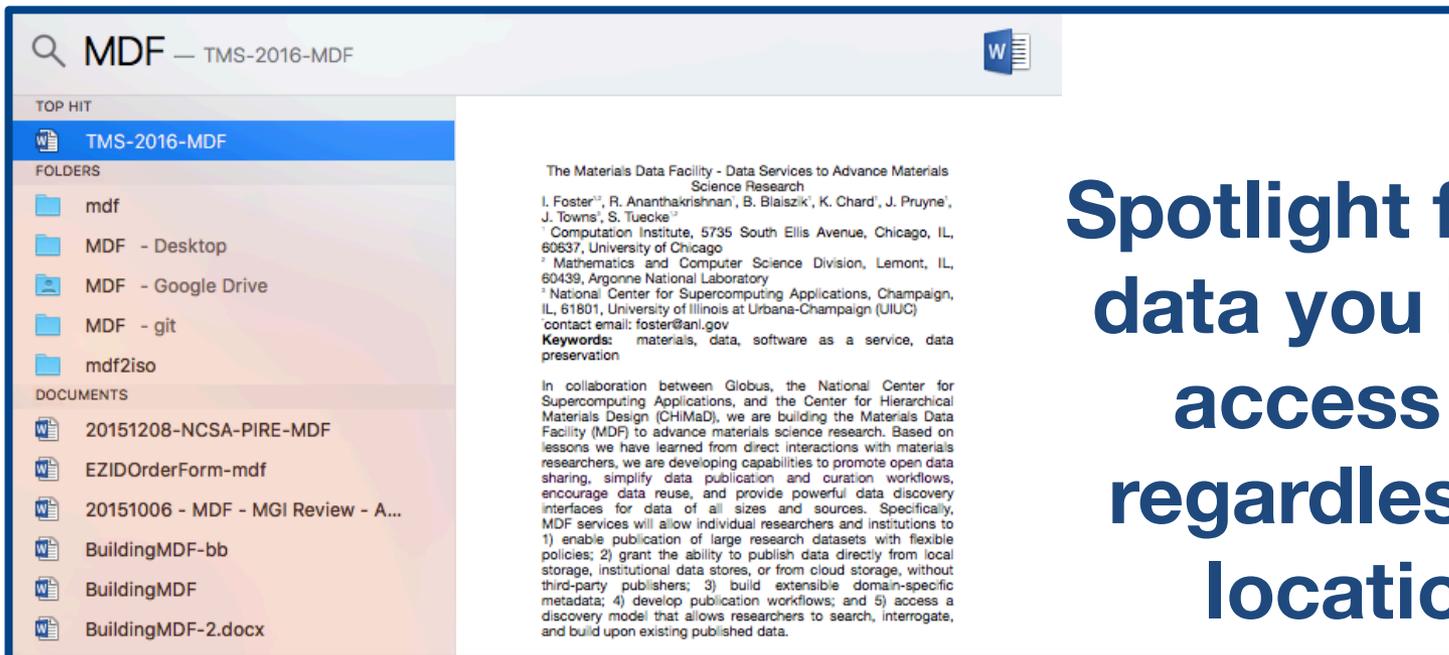
previous 1 next

# 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)



The screenshot shows a search interface for 'MDF' with the following structure:

- Search bar: **MDF** — TMS-2016-MDF
- TOP HIT: TMS-2016-MDF
- FOLDERS:
  - mdf
  - MDF - Desktop
  - MDF - Google Drive
  - MDF - git
  - mdf2iso
- DOCUMENTS:
  - 20151208-NCSA-PIRE-MDF
  - EZIDOrderForm-mdf
  - 20151006 - MDF - MGI Review - A...
  - BuildingMDF-bb
  - BuildingMDF
  - BuildingMDF-2.docx

The main content area displays a document titled 'The Materials Data Facility - Data Services to Advance Materials Science Research' by I. Foster<sup>1</sup>, R. Ananthakrishnan<sup>1</sup>, B. Blaiszik<sup>1</sup>, K. Chard<sup>1</sup>, J. Pruyne<sup>1</sup>, J. Towns<sup>1</sup>, S. Tuecke<sup>2</sup>. The document text includes:

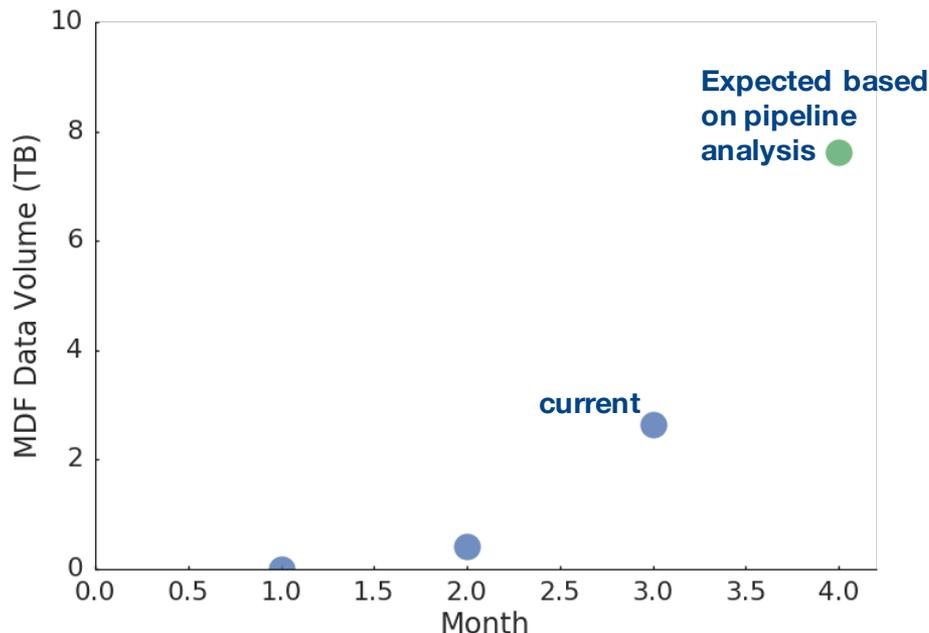
The Materials Data Facility (MDF) for advance materials science research. Based on lessons we have learned from direct interactions with materials researchers, we are developing capabilities to promote open data sharing, simplify data publication and curation workflows, encourage data reuse, and provide 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**

# 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

Materials Data Facility

Research data management simplified. | globus

MATERIALS DATA FACILITY

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  and 

## GET STARTED

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

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

## 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

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

# Thanks to Our Sponsors!

The logo for the National Institute of Standards and Technology (NIST), consisting of the letters "NIST" in a bold, black, sans-serif font.

U.S. DEPARTMENT OF  
**ENERGY**



**Argonne**  
NATIONAL LABORATORY

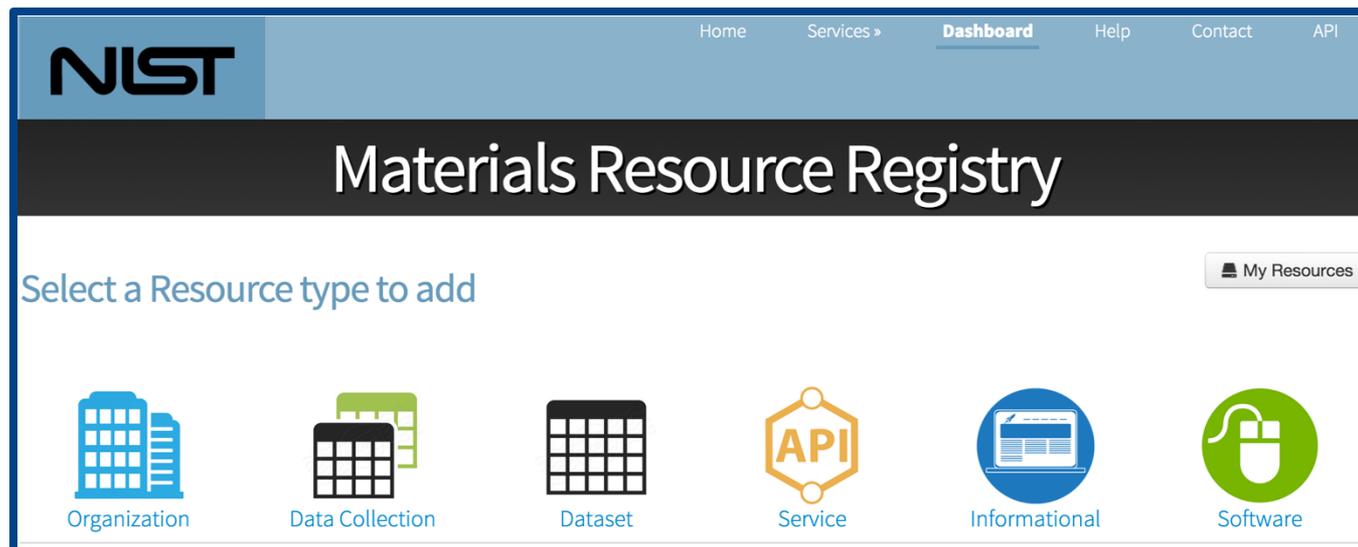


THE UNIVERSITY OF  
**CHICAGO**

# 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



# **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

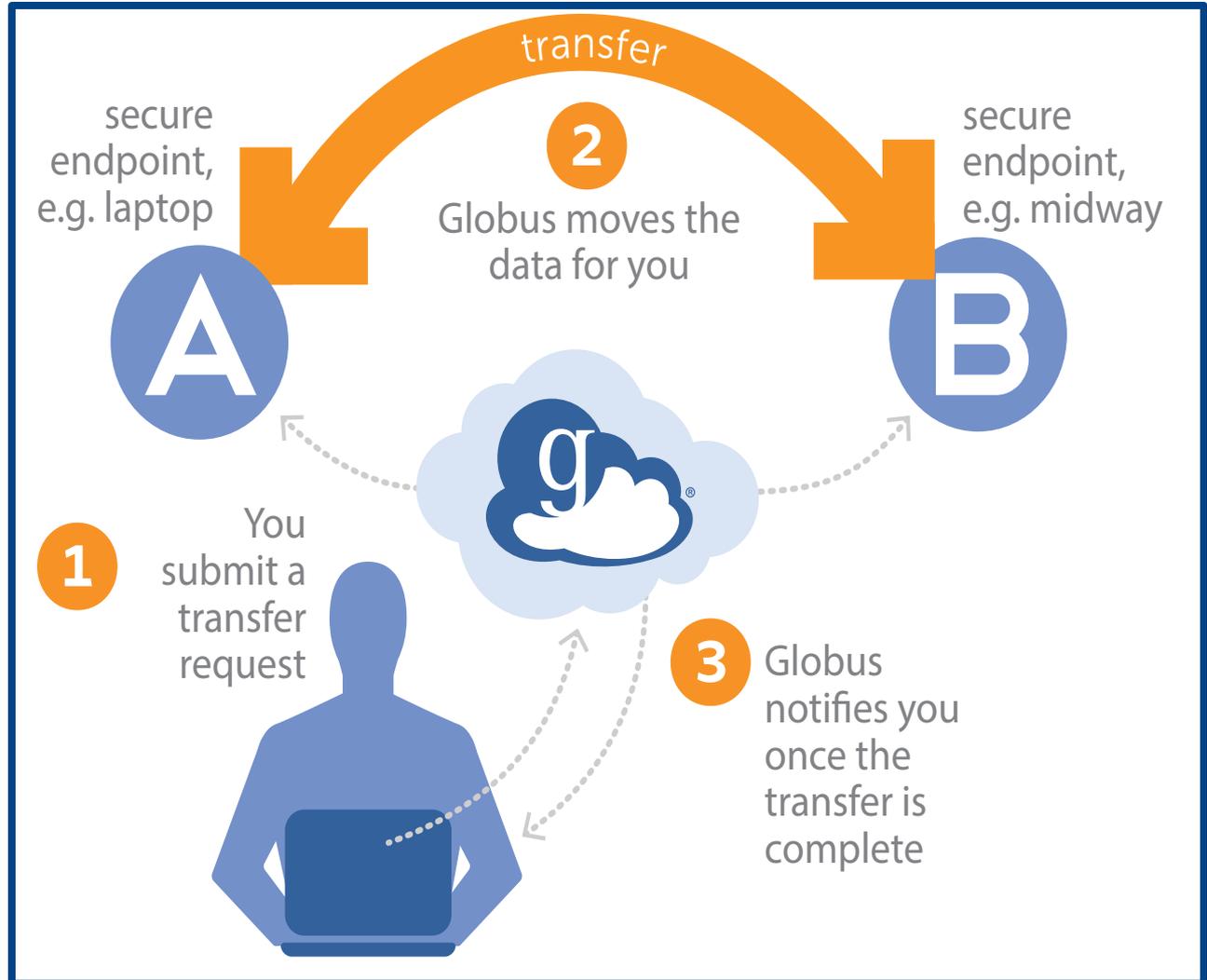
## Endpoint

- 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

The screenshot displays the Globus web interface. At the top, there's a navigation bar with 'Manage Data', 'Groups', 'Support', and 'blaiszik'. Below this, there are tabs for 'Transfer Files', 'Activity', 'Manage Endpoints', and 'Dashboard'. The main content area is titled 'Transfer Files' and includes a sub-header 'Get Globus Connect Personal Turn your computer into an endpoint.' The interface is split into two panels. The left panel shows a transfer from endpoint 'blaiszik#macbookpro' at path '/~/Desktop/blaiszik-macbookpro/Voorhees/' containing files '20A\_post\_0004.h5' (3.19 GB) and '20A\_post\_0005.h5' (3.15 GB). The right panel shows a transfer to endpoint 'globuspublish#jcpublish-test' at path '/mdf\_voorhees\_72/results/' containing the same two files. Below the panels, there's a 'Label This Transfer' field and a note 'This will be displayed in your transfer activity.' The bottom section is titled 'Activity' and shows a completed transfer from 'blaiszik#macbookpro' to 'globuspublish#jcpublish-test' a minute ago. It includes tabs for 'Overview' and 'Event Log'. The 'Overview' tab shows: Task ID c1191a64-ef5d-11e4-ab4a-22000b92c6ec, Source blaiszik#macbookpro, Destination globuspublish#jcpublish-test, Files 2, Directories 1, and Bytes Transferred 6.34 GB.

# Where are we Now?

**Data  
Publication**

# Materials Data Publication/Discovery is Often a Challenge

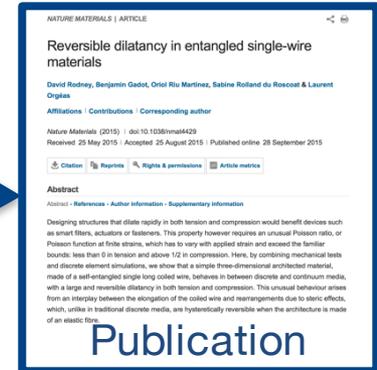
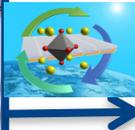


# Materials Data Publication/Discovery is Often a Challenge

Want to Publish



Want to Discover / Use



- 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

Want to Publish



Want to Discover / Use

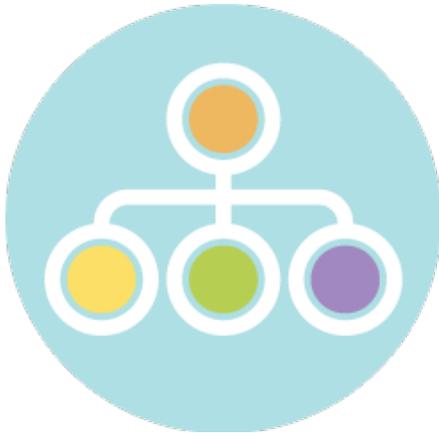


# Share Data with Flexible ACLs



- **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



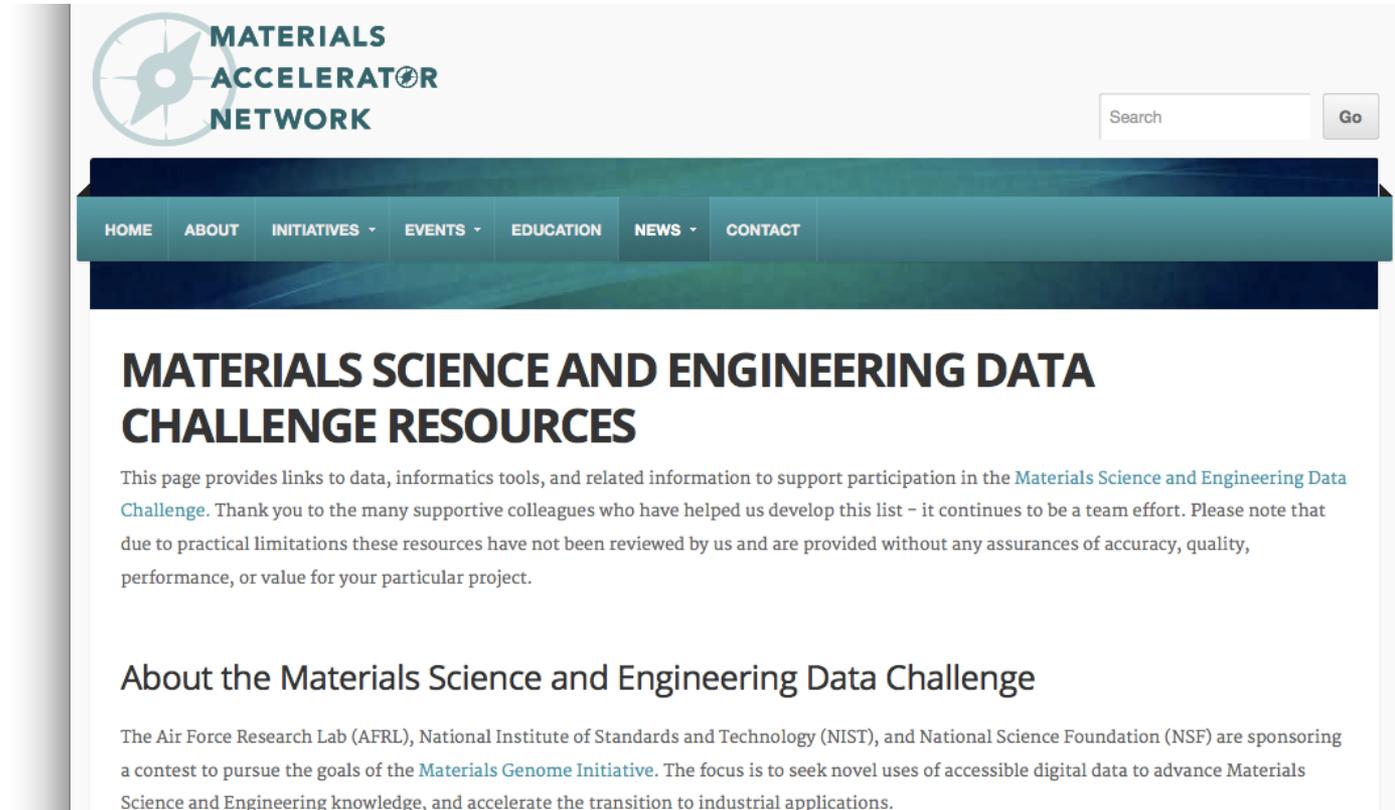
- **Build a custom metadata schema for your specific research data**
- **Re-use existing metadata schemas**
- **Working in conjunction with NIST researchers to define these schemas**

## 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

# Materials Resource Registry



The screenshot shows the Materials Accelerator Network website. The header includes the logo (a compass rose) and the text "MATERIALS ACCELERATOR NETWORK". A search bar with a "Go" button is located in the top right. The navigation menu contains links for HOME, ABOUT, INITIATIVES, EVENTS, EDUCATION, NEWS, and CONTACT. The main content area features the title "MATERIALS SCIENCE AND ENGINEERING DATA CHALLENGE RESOURCES" and a paragraph of introductory text. Below this is a section titled "About the Materials Science and Engineering Data Challenge" with a paragraph of descriptive text.

**MATERIALS ACCELERATOR NETWORK**

Search

HOME ABOUT INITIATIVES ▾ EVENTS ▾ EDUCATION NEWS ▾ CONTACT

## 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 Eureka](#)

## 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

Enter keywords, or leave blank to retrieve all records



3 results



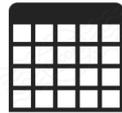
All Resources



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]<sup>43</sup>

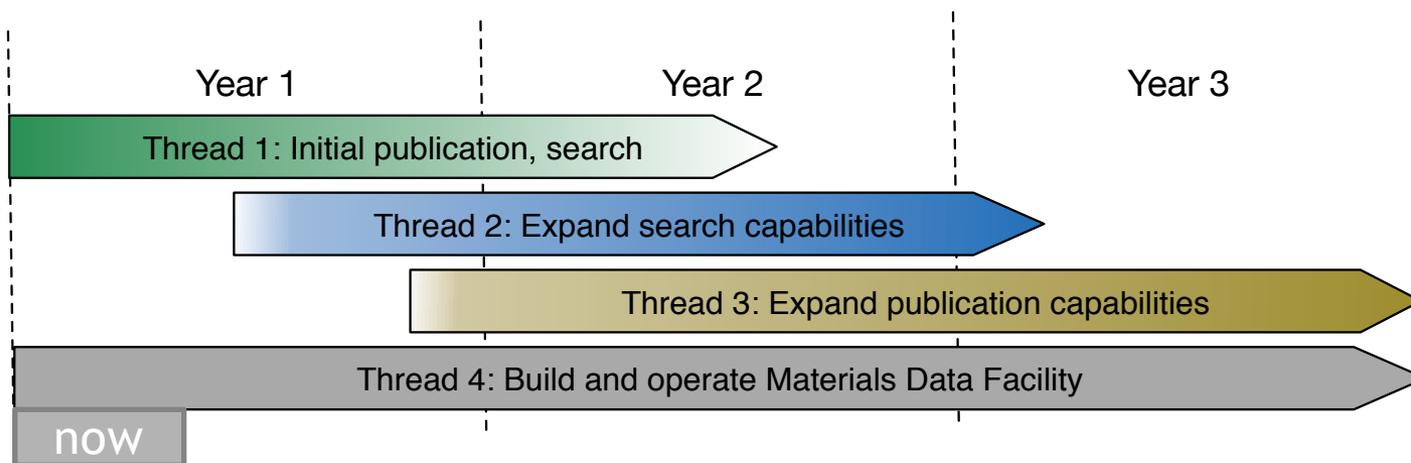
# What's Currently Available?

- **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

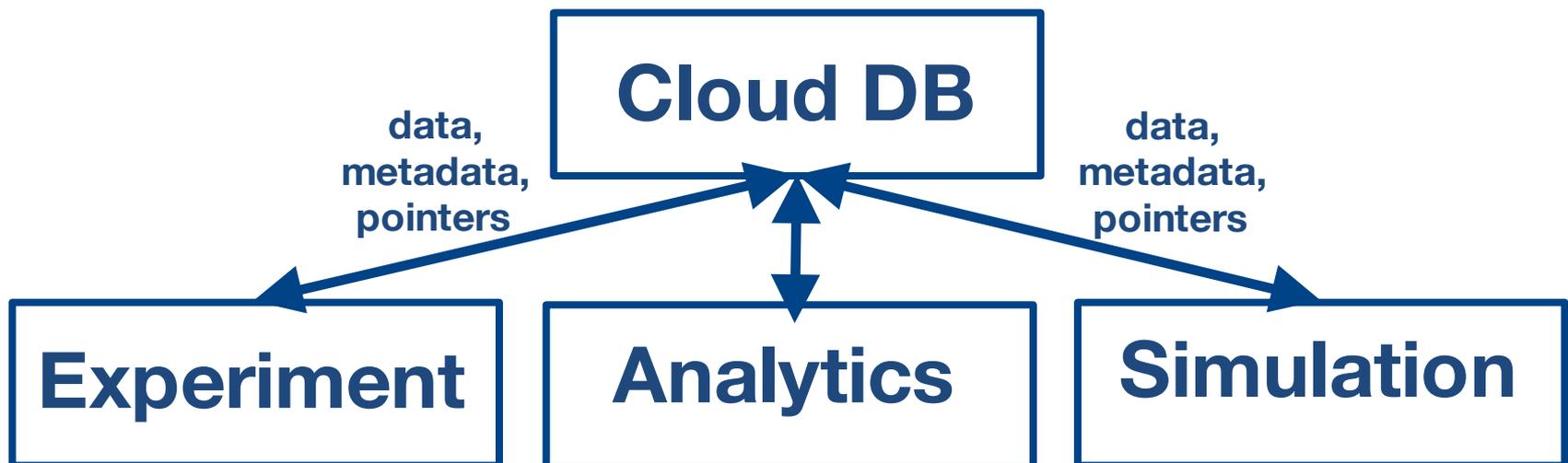
# Lessons Learned

- **The demand is there from researchers and institutions**
- **Lots of cross-over with centers and projects**
  - (NIST) CHiMaD
  - (DOE) MICCoM, JCESR, PRISMS, Argonne, I<sup>3</sup>
  - (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

Initiate Engagement

Gather Data

Describe Data

Publish Data

- 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

