

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

Ben Blaiszik (blaiszik@uchicago.edu)

Ian Foster (foster@uchicago.edu)

Steve Tuecke, Kyle Chard,

Rachana Ananthakrishnan, Jim Pruyne (UC)

Michal Ondrejcek, Kenton McHenry,

John Towns (UIUC – NCSA)

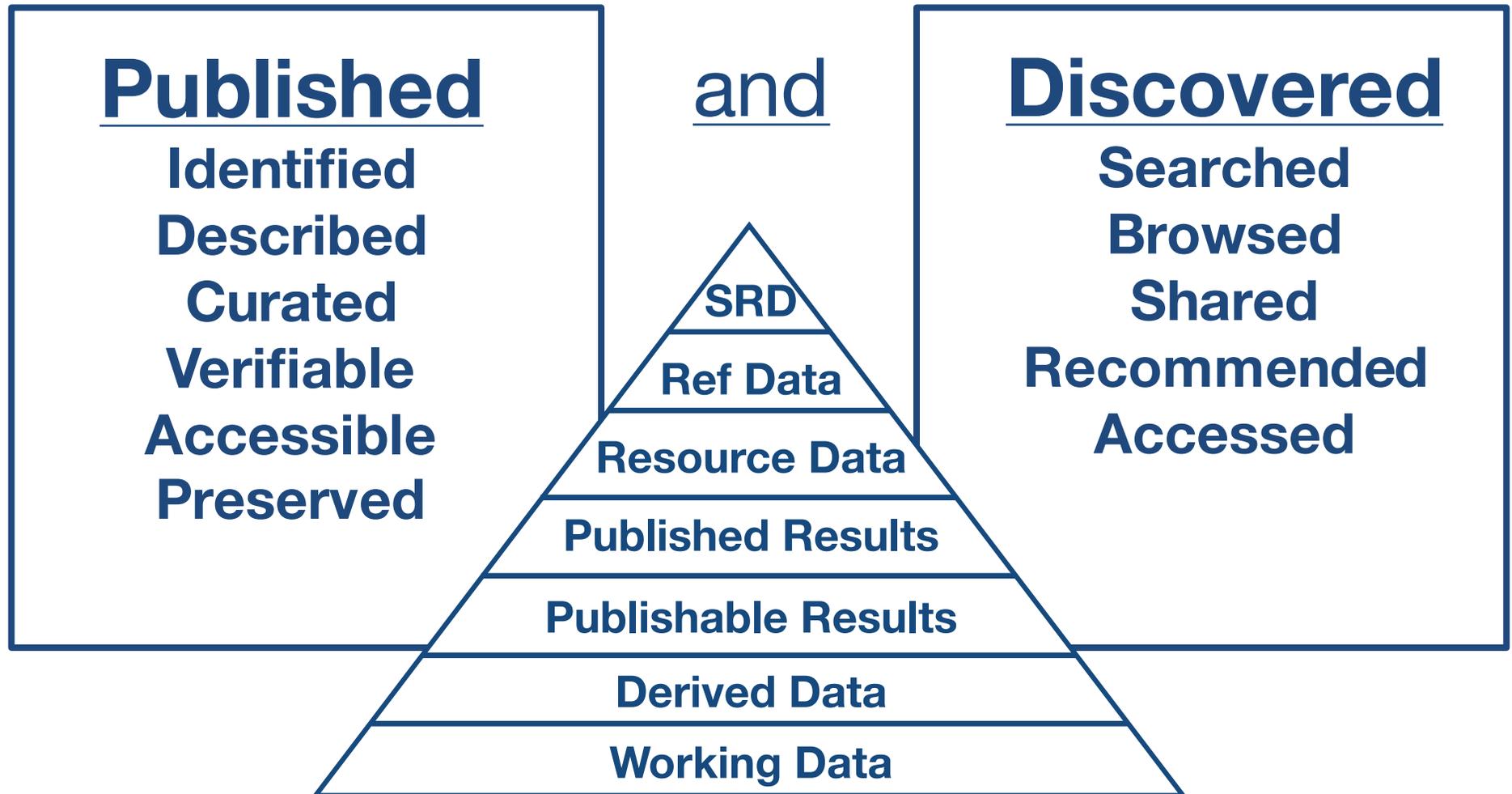
materialsdatafacility.org

Outline

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

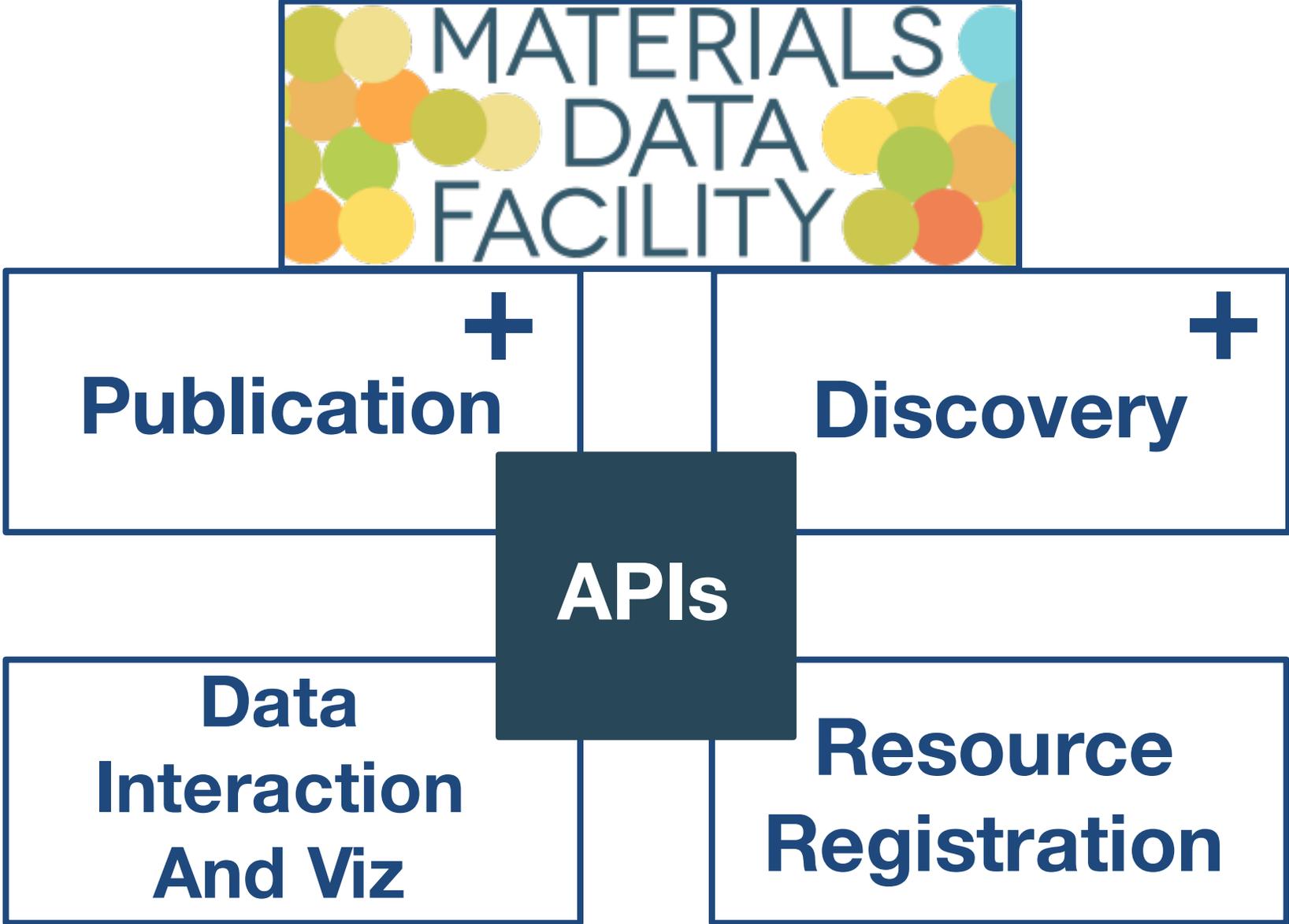
What is MDF?

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



**What infrastructure do
we need to effectively
support materials
researchers?**

Service Infrastructure



+ - Initial Foci

Publication

Some key features

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

Discovery

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

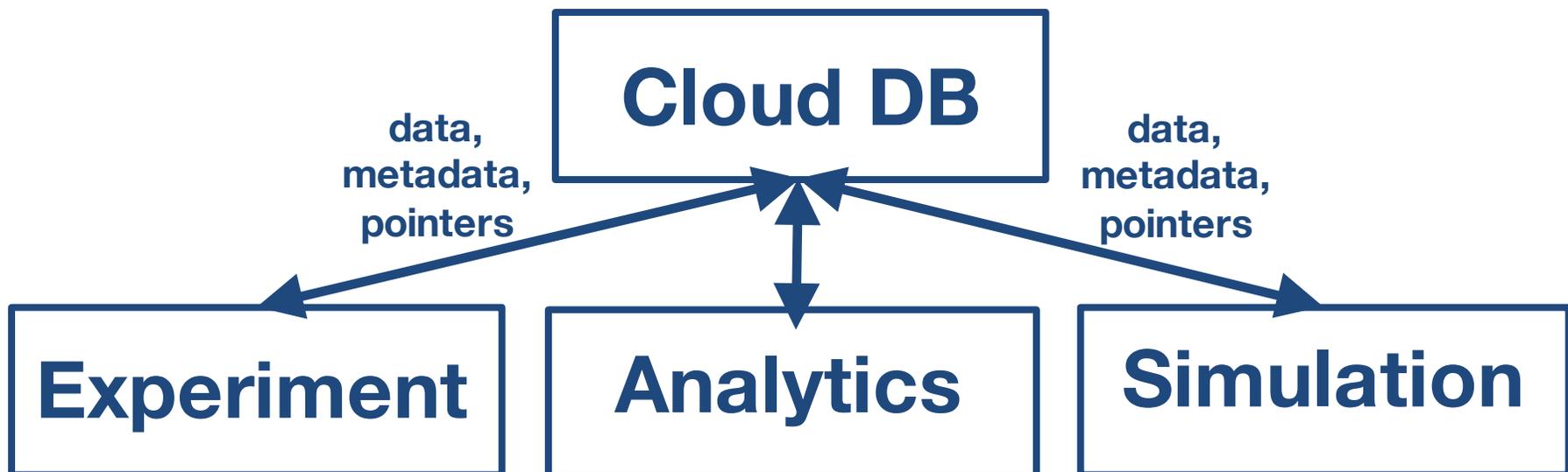
Resource Registration

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

Data Interaction And Viz

**See me for
discussion**

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



Understanding Incentives is Critical

Increasing Impact

- Increase paper citations¹
- Add dataset citation capabilities

Meeting Award Requirements

- Simplify DMP compliance

Smoothing Dislocations

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

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

**So where are
we now?**

Publication

Materials Data Publication/Discovery is Often a Challenge



Materials Data Publication/Discovery is Often a Challenge

Want to Publish



Want to Discover / Use



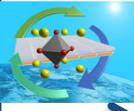
- Need networked storage, sometimes many TB ?
- Need to uniquely identify data for search/cite ?
- Need custom metadata descriptions ?
- Need a data curation workflow
- Need automation capabilities

Materials Data Publication/Discovery is Often a Challenge

Want to Publish



Want to Discover / Use



Globus Platform-as-a-Service

Identity management

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

User groups

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

Data publication

Data transfer

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

Data sharing

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

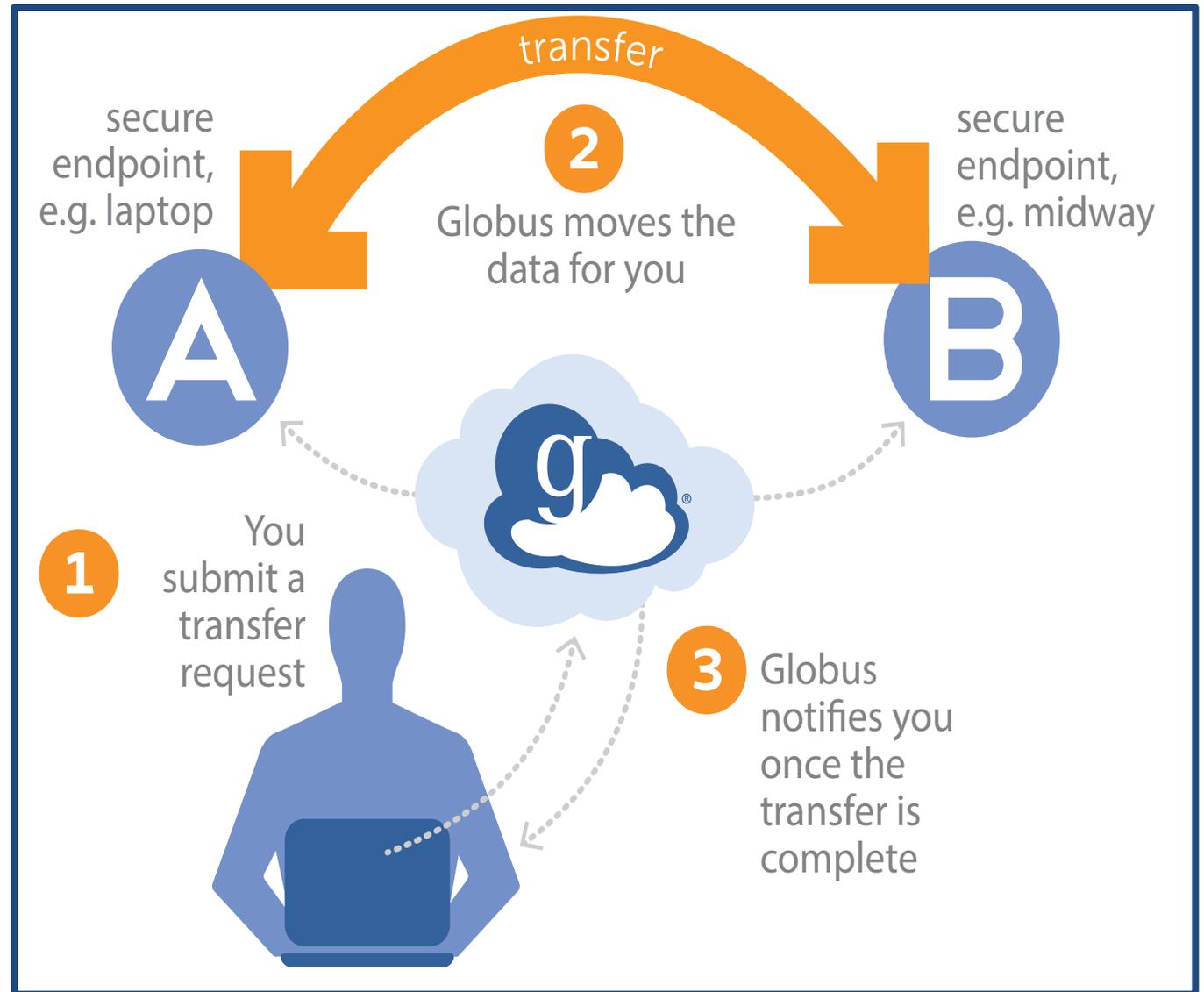
Globus Transfer Background

Endpoint

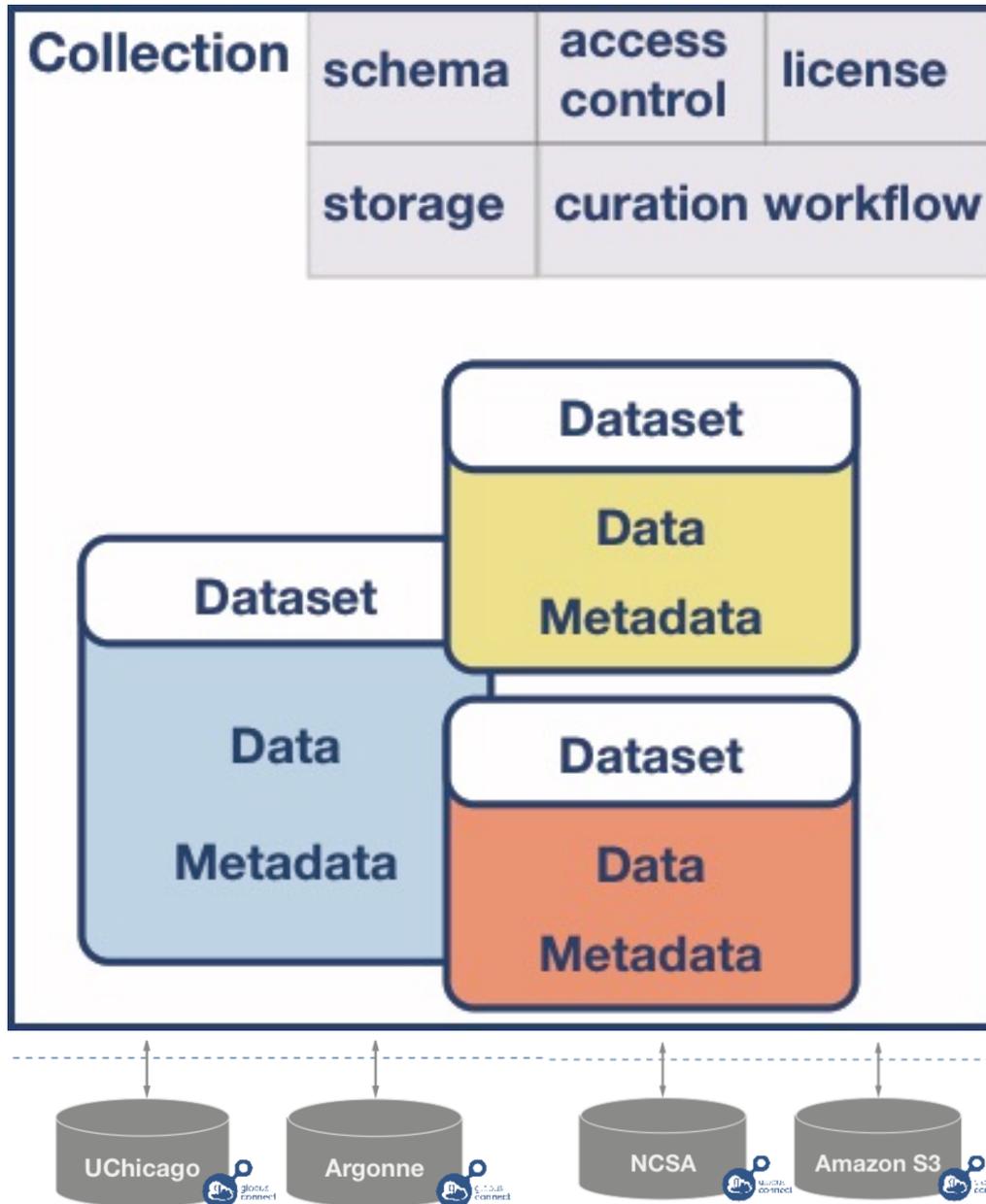
- E.g. laptop or server running a Globus client (e.g. Dropbox client)
- Enables advanced file transfer and sharing
- Currently GridFTP, future GridFTP + HTTP

Some Key Features

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



Collection Model



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

Publish Large Datasets

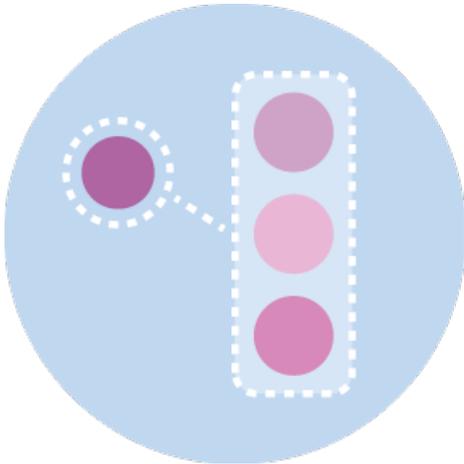


- Leverages Globus production capabilities for file transfer, user authentication, and groups

126,000,848,075 MB
TRANSFERRED

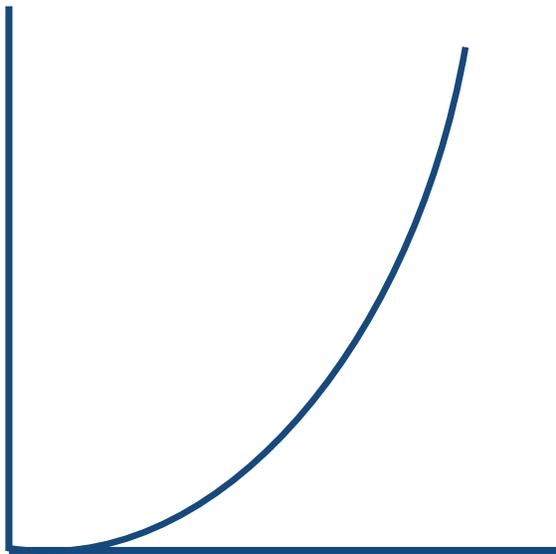
- **100 TB of reliable storage @ NCSA, and more storage at Argonne**
 - Globus endpoint at ncsa#mdf
 - Expandable to PBs as necessary
 - Automated tape backup for reliability (in progress)
- **Optionally use your own local or institutional storage**

Uniquely Identify Datasets



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

Dataset Downloads



Time

Future...

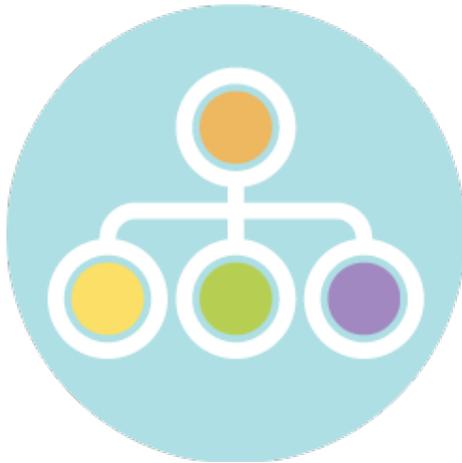
- Your work has been cited 153 times in the last year
- Researchers from 30 institutions have downloaded your datasets

Share Data with Flexible ACLs



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

Leverage Curation Workflows



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

Customize Metadata



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

Future...

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

Discover Research Datasets



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

MaterialsDataFacility.org

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

MDF

Submission

Walkthrough

Example Use Case

Publishing Big, Remote Data

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

MDF Collection Home



Publish

Log In

Sign Up

Search



Materials Data Facility Community home page

Browse

Issue Date

Author

Title

Subject

Discover

Author

Cahill, D.G.	12
Plante, Raymond	6
Felarca, Mario	5
Lee, S-M	5
Pohl, R.O.	3
Pruyne, Jim	3
Venkatasubramanian, R.	3
Selinder, T.I.	2
Watson, S.K.	2
Ash, B.J.	1

next >

Subject

Amorphous solids, thermal conduct...	3
thermal conductivity, superlattic...	3
ozone, thermal conductivity, chem...	2
thermal conductivity, single crys...	2
another test	1
example test	1
Film	1
pyrex temperature	1
test demo	1
thermal conductivity, ac techniqu...	1

next >

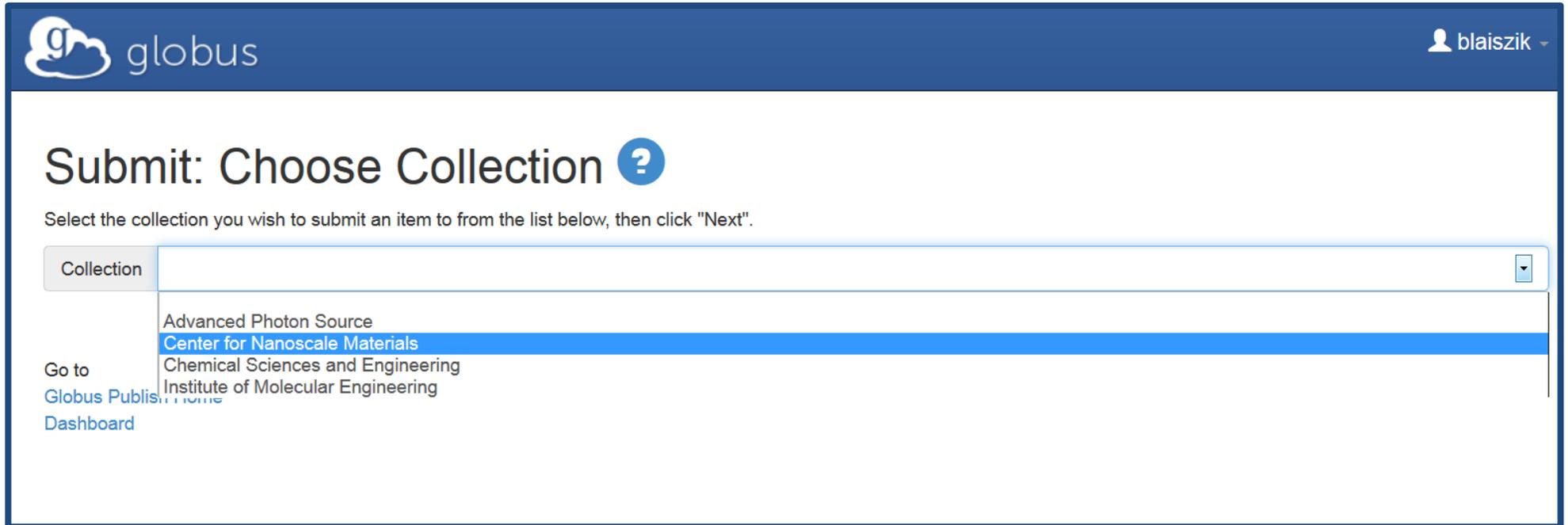
Date issued

2010 - 2015	14
2000 - 2009	8
1990 - 1999	13
1987 - 1989	1

Sub-communities within this community

Computational Materials Science

MDF Collections



The screenshot shows the Globus interface for submitting an item to a collection. The header includes the Globus logo and the user name 'blaiszik'. The main heading is 'Submit: Choose Collection' with a help icon. Below the heading is a instruction: 'Select the collection you wish to submit an item to from the list below, then click "Next".' A dropdown menu is open, showing a list of collections: 'Advanced Photon Source', 'Center for Nanoscale Materials' (highlighted), 'Chemical Sciences and Engineering', and 'Institute of Molecular Engineering'. On the left side, there are links for 'Go to Globus Publisher Home' and 'Dashboard'.

globus blaiszik

Submit: Choose Collection ?

Select the collection you wish to submit an item to from the list below, then click "Next".

Collection

- Advanced Photon Source
- Center for Nanoscale Materials
- Chemical Sciences and Engineering
- Institute of Molecular Engineering

Go to
[Globus Publisher Home](#)
[Dashboard](#)

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 Metadata Entry interface. 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 content area is titled 'Submit: Describe this Dataset' with a help icon. Below the title, there is a progress bar with steps: License, Describe (selected), Describe, Globus Transfer, Verify, and Complete. The form itself is titled 'Submit: Describe this Dataset' and contains the following fields:

- Title ***: A text input field containing 'Al-Cu Coarsening 4D Tomography Dataset'.
- Authors ***: A section for entering the main researchers. It consists of two columns of input fields. The left column contains: Fife, Gibbs, Gulsoy, Park, Thornton, Voorhees, and a label 'Last name, e.g. Smith'. The right column contains: J.L., J.W., E.B., C.-L., K., P.W., and a label 'First name(s) + "Jr", e.g. Donald Jr'. To the right of each input field is a red 'Remove Entry' button. At the bottom right of the authors section is a '+ Add More' button.
- Publication Year ***: A section for entering the year. It includes a 'Month' dropdown menu (currently set to '(No Month)'), a 'Day' input field, and a 'Year' input field (currently 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 contains the copyright information: '© 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 'Describe this Dataset' form in the Globus Data Publication Dashboard. The form is titled 'Submit: Describe this Dataset' and includes a progress bar with steps: License, Describe (selected), Describe, Globus Transfer, Verify, and Complete. The form fields are as follows:

Field	Value
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 the form 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 of the keyword list. At the bottom of the form, there are three navigation 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 interface for a file transfer. At the top, the Globus logo and navigation links are visible. The main section is titled 'Transfer Files' and shows two panels for source and destination endpoints. The source endpoint is 'blaiszik#macbookpro' and the destination is 'globuspublish#jcpublish-test'. Both panels show a list of files: '20A_post_0004.h5' (3.19 GB) and '20A_post_0005.h5' (3.15 GB). The source panel is labeled '(e.g. Switzerland)' and the destination panel is labeled '(e.g. Northwestern)'. Below the panels is a 'Label This Transfer' field. The 'Activity' section below shows a green checkmark and the message 'blaiszik#macbookpro to globuspublish#jcpublish-test transfer completed a minute ago'. It also shows an 'Overview' tab with details: Task ID c1191a64-ef5d-11e4-ab4a-22000b92c6ec, Source blaiszik#macbookpro, Destination globuspublish#jcpublish-test, Files 2, Directories 1, and Bytes Transferred 6.34 GB.

Dataset Curation

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

The screenshot shows the Globus Data Publication Dashboard interface. 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 content area is titled 'Perform Task' and contains a message: '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 item details are as follows:

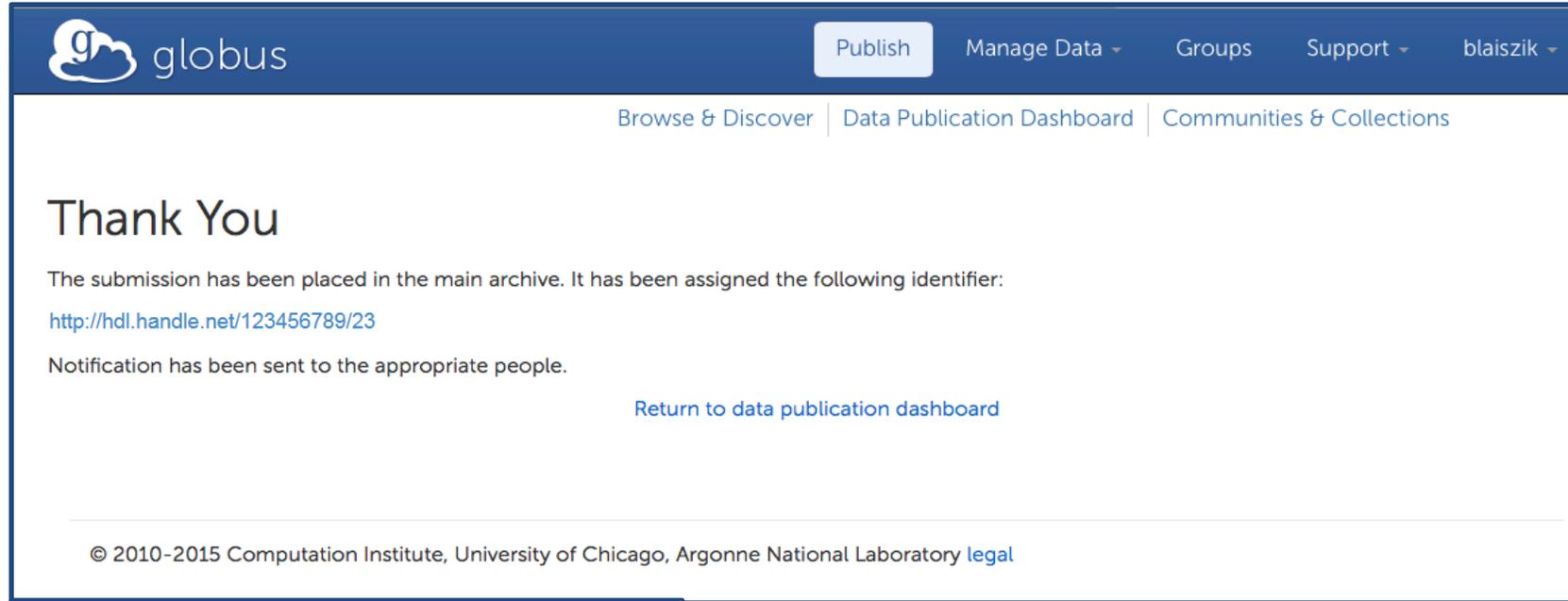
- Title:** Al-Cu Coarsening 4D Tomography Dataset
- Authors:** Fife, J.L., Gibbs, J.W., Gulsoy, E.B., Park, C.-L., Thornton, K., Voorhees, P.W.
- Keywords:** in situ, 4D coarsening, aluminum-copper alloys, dynamic morphological evolution, solid-liquid interfaces
- Issue Date:** 2014
- Publisher:** Northwestern University

Below the item details, there is a section titled 'Files in This Item:' with a link: [globuspublish#jcpublish-test/mdf_voorhees_72/](#)

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

Approve	If you have reviewed the item and it is suitable for inclusion in the collection, select "Approve".
Reject	If you have reviewed the item and found it is not 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.
Do Later	If you wish to leave this task for now, and return to the data publication dashboard, use this option.
Return Task to Pool	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 user interface. At the top left is the Globus logo. The top navigation bar includes a 'Publish' button and links for 'Manage Data', 'Groups', 'Support', and 'blaisik'. Below the navigation bar are three main menu items: 'Browse & Discover', 'Data Publication Dashboard', and 'Communities & Collections'. The main content area features a large 'Thank You' heading, followed by a message stating the submission has been placed in the main archive and assigned an identifier. The identifier is shown as a blue hyperlink: <http://hdl.handle.net/123456789/23>. Below this, it says 'Notification has been sent to the appropriate people.' and provides a blue link to '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:	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:	http://bit.ly/1EGh9UL
Appears in Collections:	Voorhees Group X-Ray Tomography

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

The screenshot displays the Globus dataset discovery interface. At the top, the Globus logo and navigation links (Publish, Manage Data, Groups, Support, blaiszik) are visible. Below the navigation bar, there are tabs for 'Browse & Discover', 'Data Publication Dashboard', and 'Communities & Collections'. A search bar contains the text 'Voorhees'. The search results are displayed in a table format. The table has columns for 'Issue Date', 'Title', and 'Author(s)'. The first result is from 2014, titled 'Al-Cu Coarsening 4D Tomography Dataset', with authors 'Fife, J.L.; Gibbs, J.W.; Gulsoy, E.B.; Park, C.-L.; Thornton, K.; Voorhees, P.W.'. To the right of the search results, there are two panels: 'Discover' and 'Subject'. The 'Discover' panel lists authors: Fife, J.L., Gibbs, J.W., Gulsoy, E.B., Park, C.-L., Thornton, K., and Voorhees, P.W., each with a count of 1. The 'Subject' panel lists subjects: 4D coarsening, aluminum-copper alloys, dynamic morphological evolution, in situ, and solid-liquid interfaces, each with a count of 1. At the bottom of the search results, there are navigation buttons for 'previous', '1', and 'next'.

globus Publish Manage Data Groups Support blaiszik

Browse & Discover | Data Publication Dashboard | Communities & Collections

Voorhees

Search Results

Collection results (1 result) advanced search

Results 1-2 of 2

Issue Date	Title	Author(s)
2014	Al-Cu Coarsening 4D Tomography Dataset	Fife, J.L.; Gibbs, J.W.; Gulsoy, E.B.; Park, C.-L.; Thornton, K.; Voorhees, P.W.

Discover

Author

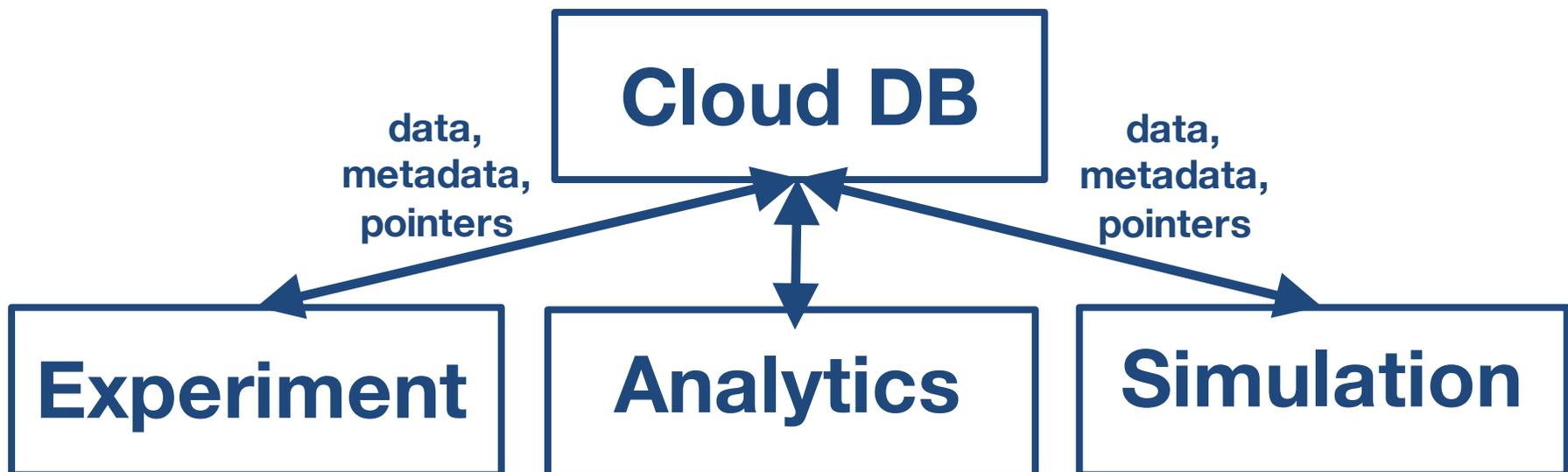
- Fife, J.L. 1
- Gibbs, J.W. 1
- Gulsoy, E.B. 1
- Park, C.-L. 1
- Thornton, K. 1
- Voorhees, P.W. 1

Subject

- 4D coarsening 1
- aluminum-copper alloys 1
- dynamic morphological evolution 1
- in situ 1
- solid-liquid interfaces 1

previous 1 next

Data Interaction



Catalog

alpha software

Catalog → Datasets → Members
Datasets/Members have typed tags

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



iPhoto, ca. 2014

vs

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

Standard file/directory structure

Catalog

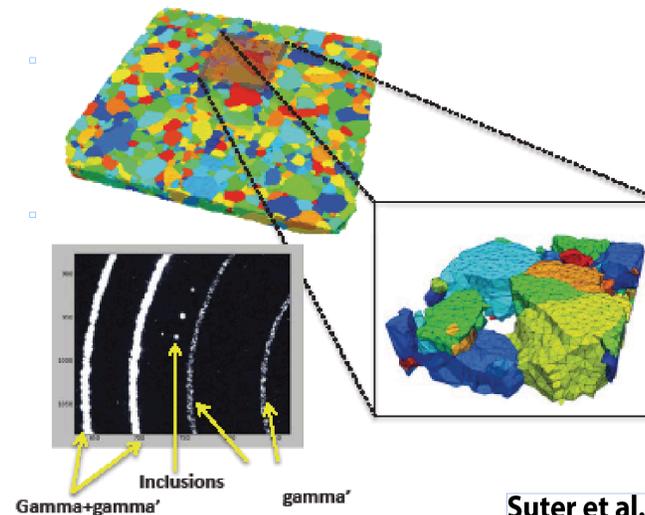
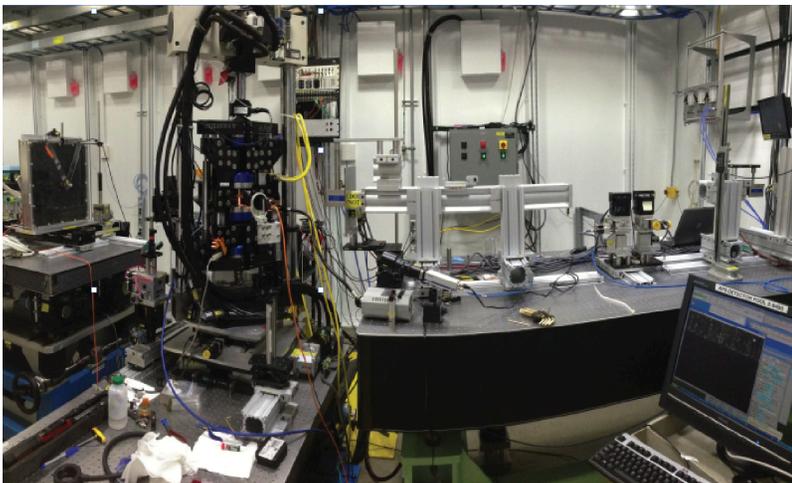
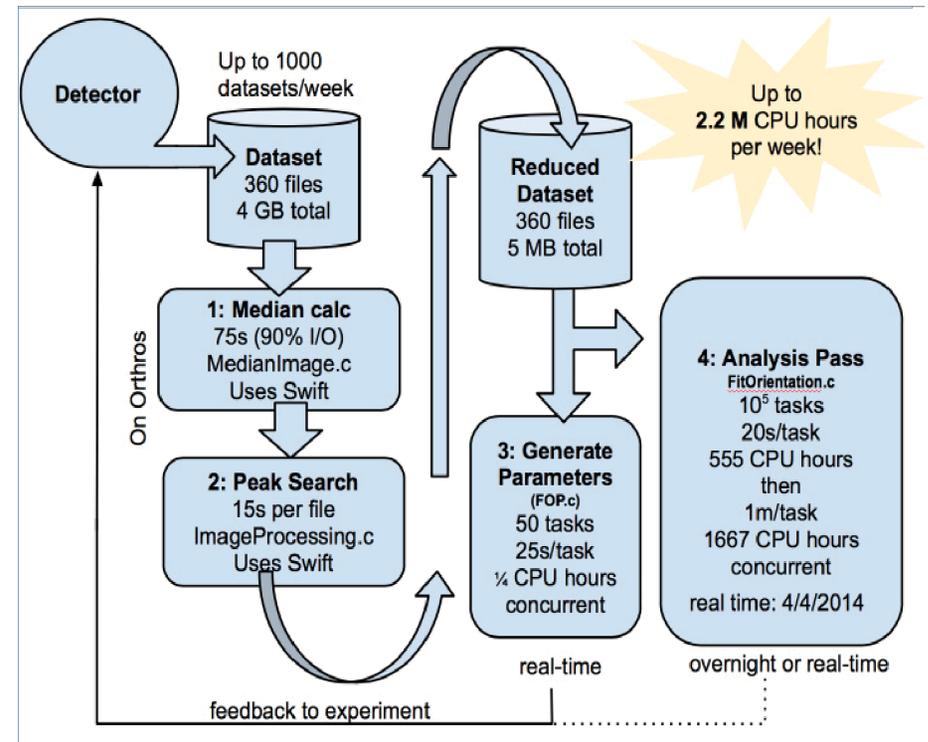
alpha software

Catalog → Datasets → Members
Datasets/Members have typed tags

The screenshot shows the Globus NeXus demo 2 interface. At the top left is the Globus logo and the text "globus". Below it is the title "NeXus demo 2" and a search bar labeled "Search catalog". A toolbar contains icons for refresh, share, and delete, along with a "Sort" button. A sharing section is active, showing "Share 5 selected datasets in NeXus demo 2" with a text input field containing "ian" and checkboxes for "Read" and "Write" permissions. The main content area displays a dataset entry for "lsmo52_400k". This entry includes a list of tags: "PI: Ray Osborn", "beamline: ANL APS Sector 6", "date: 2014-03-28 14:33:09", "host: mlra", "path: /gpfs/mlra-fs0/projects/ExM/DE/bigdata/lsmo52_400k", "sample: LSMO", "size: 637G", and "temperature: 400". Below the tags are sections for "Members" and "Share Settings". To the right of the dataset entry, there are icons for link, user, and date, with the user "u:wozniak" and date "2014-08-07" displayed. A second dataset entry for "lsmo52_350k" is partially visible at the bottom.

Workflows and Integrations

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



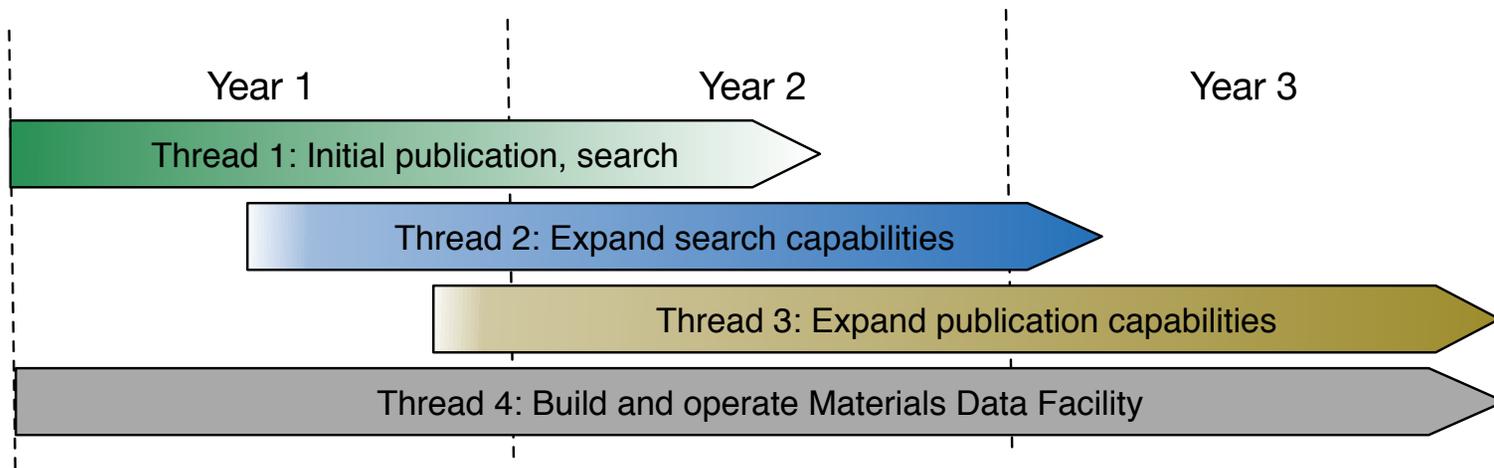
What's Available?

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

What are we looking for?

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

First Steps



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

Thanks to Our Sponsors!

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



THE UNIVERSITY OF
CHICAGO