



Materials Data Facility - Data Services to Advance Materials Science Research

Ben Blaiszik (blaiszik@uchicago.edu),

Kyle Chard, Jim Pruyne, Rachana Ananthakrishnan

Michael Ondrejcek, Kenton McHenry

PIs: Ian Foster (foster@uchicago.edu), Steven Tuecke, John Towns

materialsdatafacility.org
globus.org



Materials Genome Initiative

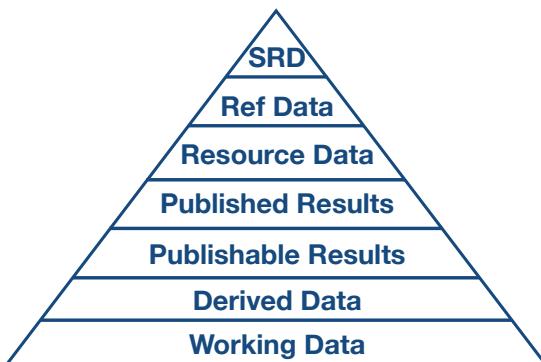


Publication

- Identify datasets with persistent identifiers (e.g., DOI or Handle)
- Describe datasets with appropriate metadata and provenance
- Verify dataset contents over time
- Handle big (and small) data: We have already ingested datasets with > 1.5M files and > 1TB in size

REST APIs

Materialsdatafacility
.org



- Search, query, and access datasets in modern ways
- Automatically index flexible metadata and harvest file contents
- Provide simple user interfaces (i.e., after Google and Amazon)

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

Publication

Discovery

Data transfer

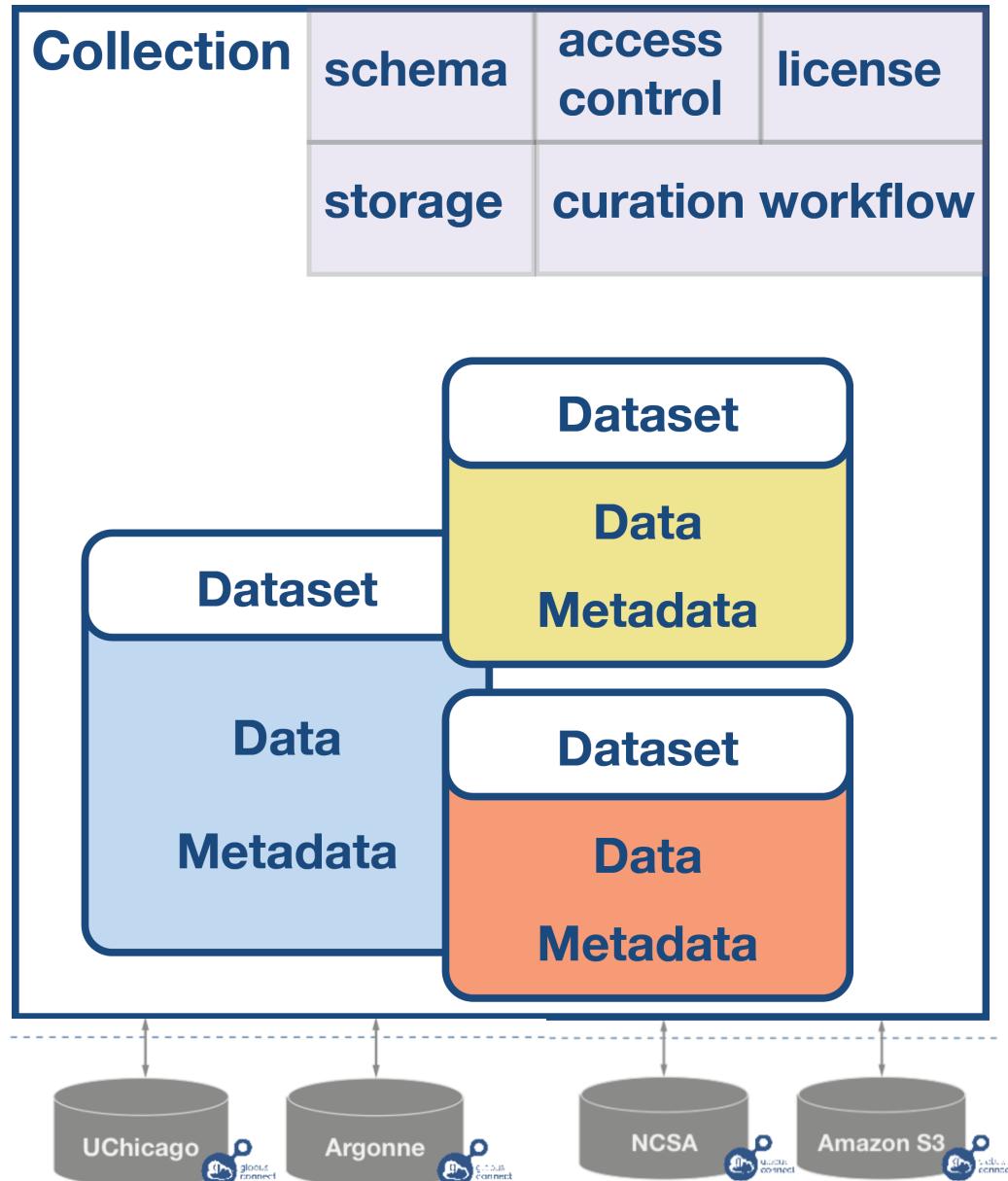
- High-performance data transfer from a web browser
- Optimize transfer settings and verify transfer integrity
- Add your laptop to the endpoint network with Globus Connect Personal

Data sharing

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

Data Publication Service

Publication Data Model



- **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

Create Dataset

1

Create dataset in web UI or via API

2

Assemble data files and metadata

3

Curate dataset (optional)

4

Mint dataset identifiers

- Customizable metadata can be added via auto-generated forms or REST API [coming soon] (schema is validated)
- Creates a directory on the collection-specified endpoint, handles ACLs

The screenshot shows the Globus Data Publication Dashboard. At the top, there are navigation links: Publish, Manage Data, Groups, Support, and a user dropdown for blaiszik. Below the navigation, there are tabs: License, Describe, Describe (which is selected), Globus Transfer, Verify, and Complete. The main content area has a heading "Submit: Describe this Dataset" with a question mark icon. It says "Please fill further information about this submission below." There are several input fields with placeholder text: Material (Al-Cu), Volume Fraction Al (15), Volume Fraction Cu (85), Technique (x-ray tomography), Pixel size (µm) (1.4), Beam energy (keV) (20), and Instrumentation (Swiss Light Source - Tomographic Microscopy and Coherent Radiology Experiments beamline). Below these, there is a section for "Enter appropriate subject keywords" with a list of terms: in situ, 4D coarsening, aluminum-copper alloys, dynamic morphological evolution, and solid-liquid interfaces. To the right of this list are red "Remove Entry" buttons and a "+ Add More" button. At the bottom of the form are buttons for < Previous, Cancel/Save, and Next >.

Assemble Dataset

1

Create dataset in
web UI or via API

2

Assemble data
files and metadata

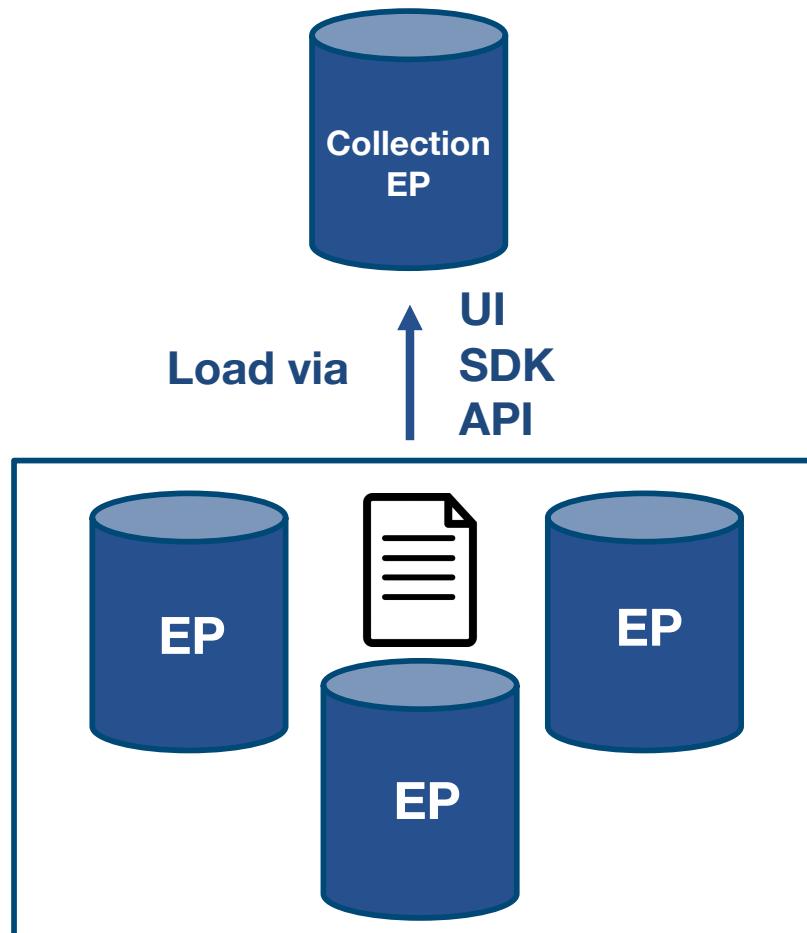
3

Curate dataset
(optional)

4

Mint dataset
identifiers

- Data can be assembled from multiple endpoints asynchronously



Curate Dataset (optional)

1

Create dataset in web UI or via API

2

Assemble data files and metadata

3

Curate dataset (optional)

4

Mint dataset identifiers

- If a collection specifies that curation is required, a notification is sent to the curator of the collection where the dataset is deposited
- Curator approves or sends the dataset back to the submitting user

Task notification

The controls at the bottom of the page to take

Dataset Title	Collection	Submitted By	Task
Gravure Printing of Graphene for Large-Area Flexible Electronics	MDF Open « Materials Data Facility	Ben Blaiszik	Check Submission

Title: Gravure Printing of Graphene for Large-Area Flexible Electronics
Authors: Hersam, Mark
Bergeron, Hadallia
Issue Date: 2016
Publisher: Materials Data Facility

Endpoint and path to dataset
82f1b5c6-6e9b-11e5-ba47-22000b92c6ec/unpublished/publication_383/ ← **Link to data**

Curation options

Approve	You have reviewed the dataset and it is suitable for publication in the collection.
Reject	You have reviewed the dataset and found that it is not suitable for publication in the collection. You will be asked to enter a message indicating why the dataset is unsuitable, and whether the submitter should change something and re-submit.
Edit Metadata	You have reviewed the dataset and found that you need to edit the dataset's metadata.
Do Later	Leave this task for now and return to the data publication dashboard.
Unclaim	Return this task to the pool so that another user can claim it.

Mint Dataset Identifiers

1

Create dataset in web UI or via API

2

Assemble data files and metadata

3

Curate dataset (optional)

4

Mint dataset identifiers

- Based on collection specifications, an identifier is minted (e.g., DOI, or Handle)
- A landing page with metadata summary and data link is also created for the DOI

Liu, Xiaolong; Balla, Itamar; Bergeron, Hadallia; Campbell, Gavin J.; Bedzyk, Michael J.; Hersam, Mark C., "Rotationally Commensurate Growth of MoS₂ on Epitaxial Graphene," 2016, <http://dx.doi.org/doi:10.18126/M2G59Q> Download Citation ▾

Title:	Rotationally Commensurate Growth of MoS ₂ on Epitaxial Graphene
Authors:	Liu, Xiaolong Balla, Itamar Bergeron, Hadallia Campbell, Gavin J. Bedzyk, Michael J. Hersam, Mark C.
Keywords:	molybdenum disulfide MoS ₂ silicon carbide SiC graphene chemical vapor deposition CVD van der Waals heterostructure scanning tunneling microscopy synchrotron X-ray scattering
Issue Date:	16-May-2016
Publisher:	Materials Data Facility
URI:	http://dx.doi.org/doi:10.18126/M2G59Q
Appears in Collections:	Hersam Group

Metadata summary

DOI

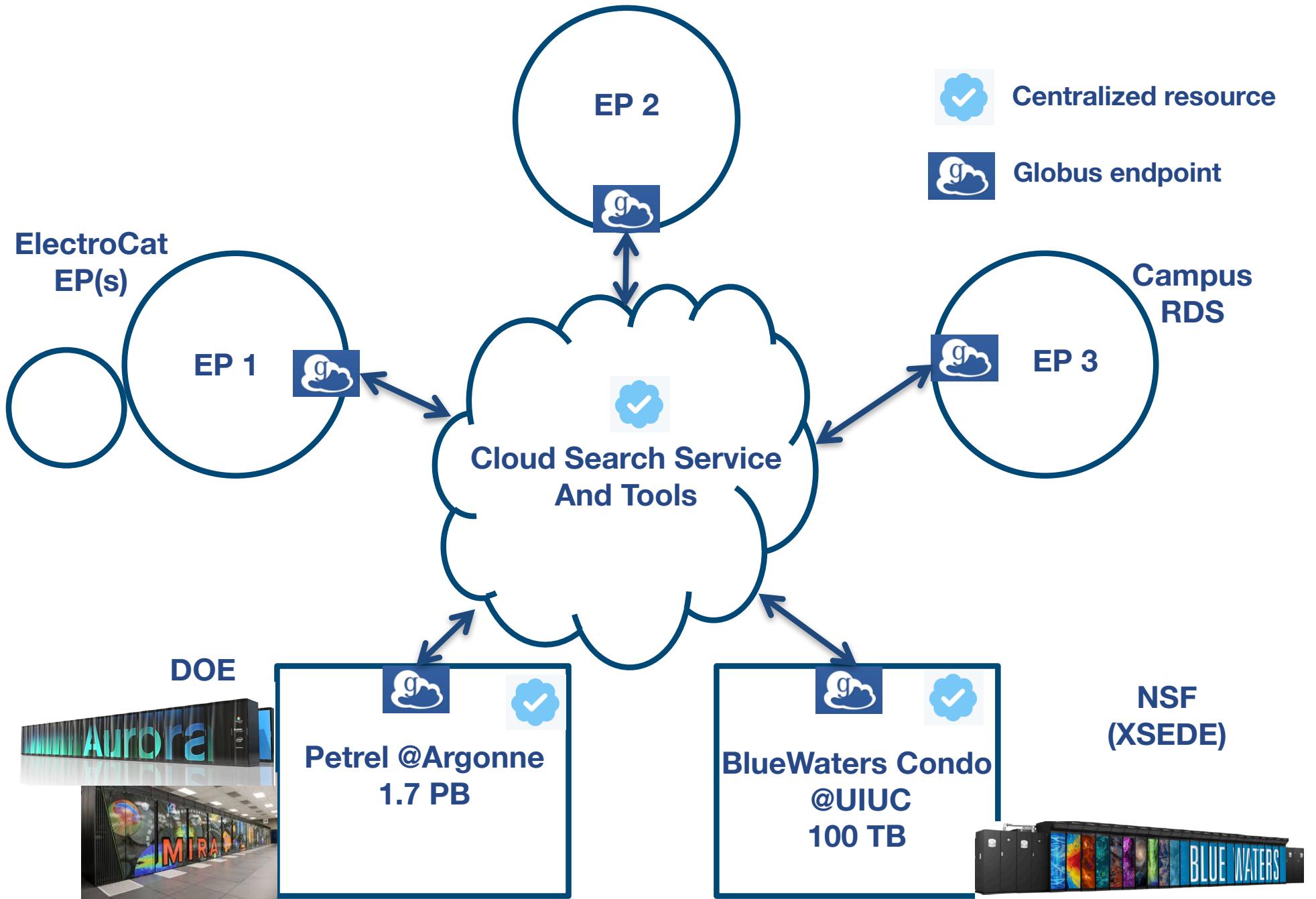
Link to data

Statistics

Endpoint and path to dataset
82f1b5c6-6e9b-11e5-ba47-22000b92c6ec/hersam/published/publication_102/

Show full record Return to data publication dashboard

Distributed Publication Model



Data Discovery Service

Data Search Service

Search Beta

The screenshot shows a search interface with a blue header bar. The search bar contains the query "aluminum^3 silicate". To the right of the search bar is a "Custom boosting" button. Below the search bar is a "Keywords" section containing a list of terms and their counts, such as "Form Descriptors F..." (34), "Nature and Environm..." (34), "Science and Technol..." (34), and "Alkaloids" (7). To the right of the keywords is a "Search Results" section. The first result is "Rotary forming of cast aluminum", with details: Collection: UBC Circle, Publication Date: 2013-08-13, Author: Roy, Matthew J. The second result is "Strengthening mechanisms in aluminum alloys", with details: Collection: UBC Circle, Publication Date: 2011-04-21, Author: Sahoo, Maheswar, Keywords: Aluminum alloys. The third result is "The anodic oxidation of aluminum", with details: Collection: UBC Circle, Publication Date: 2012-03-06, Author: Lye, Robert Glen. On the left side of the interface, there are two facets: "Keywords" and "Date". The "Keywords" facet lists terms like "Form Descriptors F...", "Nature and Environm...", "Science and Technol...", "Alkaloids", "Aluminum alloys", "Aluminum", "Indole", "Organic compounds -...", "Nuclear magnetic re...", and "Alloys -- Analysis". The "Date" facet lists years from 2011 to 2016, with counts: 2016 (11), 2015 (15), 2014 (22), 2013 (15), 2012 (44), and 2011 (114). A vertical arrow points upwards from the "Facets" label towards the facets.

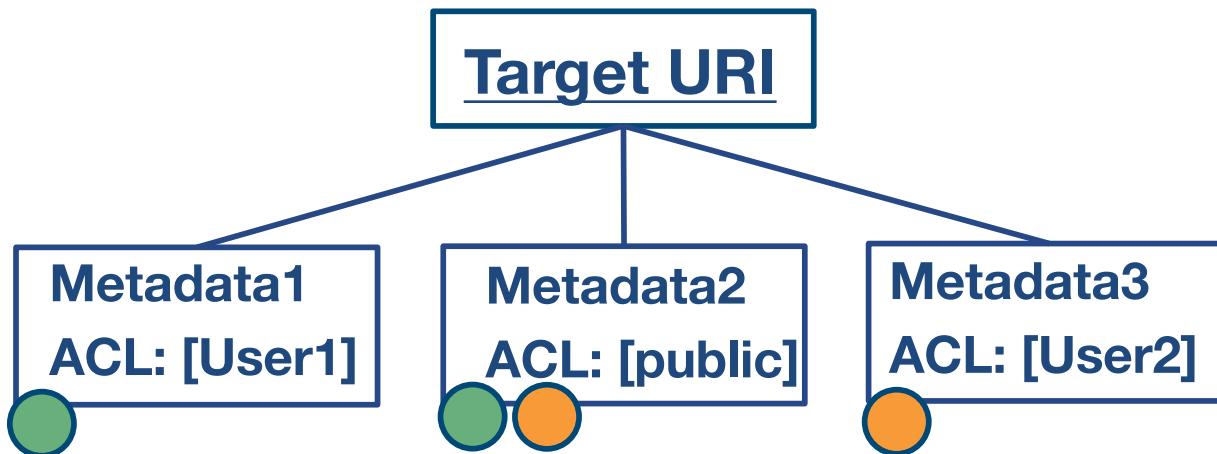
Facets

- **SaaS search service**
- **Indexing collections of data (many topics) and deep indexing specific datasets (e.g., atomistic modeling since there is a lot of data available)**
- **Will also be capable of indexing endpoint contents**

- **REST APIs and emerging python client**
- **Custom faceting and search boosting**
- **Partial matches, fuzzy matches, range queries**
- **Multi-scale search (current research topic)**

Data Search Model

Multiple sets of access-controlled metadata may be attached to a target



User1:
Sees union of
Metadata1, Metadata2

User2:
Sees union of
Metadata2, Metadata3

Data Search Document Format

```
{  
  "https://ir1.edu/pub" : {  
    "Target": can be a file, dir, or URI  
    "mimetype" : "application/json",  
    "visible_to": ["userid_1", "userid_2", "groupid_3"],  
    "content" : {  
      "@context" : {  
        "dc": "http://dublincore.org/documents/dcmi-terms",  
        "globus": "http://globus.org/publish-terms"  
      },  
      "dc:creator": "Researcher1",  
      "dc:title": "Data Publication Title",  
      "globus:shared_endpoint" : "<uuid>",  
      "globus:path": "/published/publication10"  
    }  
  }  
}
```

ACLs to be enforced on record during searches

Schema definitions and associated metadata in content field

- REST API
- Python
- Command line client



Data Search Example - OQMD

globus search Beta

OQMD

All Endpoints Files Publications

NOTE: Your search results are limited. Logging in can improve search results by displaying permissions associated with your account.

Search Results

- OQMD - N1P3
- OQMD - Se1Ti3
- OQMD - Cr1Yb3
- OQMD - Se1Sr3
- OQMD - Os1Rb3
- OQMD - Ge1Yb3
- OQMD - Re3Sm1
- OQMD - Ir1P3
- OQMD - S3Se1
- OQMD - Ga1N3

1 2 3 4

oqmd.org

OQMD: The Open Quantum Materials Database

Newsflash: OQMD v1.1 is out! (Download it [here](#).)

P₃N : ΔH_f = 0.858 eV/atom

Database Information

See also: [duplicates list](#)
Prototype: D0_22
Structure: 448415
Spacegroup: I4/mmm
of atoms: 4
Path: /home/oqmd/libraries/prototypes/binaries/D0_22/P_N
Associated keywords: prototype D0_22

Calculation History

Configuration	Total energy [eV/atom]	Band gap [eV]	Volume [Å ³ /atom]	# of ionic steps	Converged
initialize	-3.084	0	27.001	1	True
coarse_relax	-4.92	0	14.028	21	True
fine_relax	-5.037	0	14.102	5	True
standard	-5.041	0	14.102	1	True

Contact us by e-mail

Visualization

Crystal structure

IM: I4/mmm
a=7.000 Å
c=2.283 Å
c=6.513 Å
g=90.000°
p=160.000°
y=90.000°
z=90.000°

JSmol

Primitive Cell Conventional Cell

Download primitive or conventional cells (VASP format).

If you are using any results from this website, please reference this work as shown here

```
print(json.dumps(r.data['gmeta'][0], sort_keys=True, indent=4))
```

```
{
    "http://oqmd.org/materials/entry/302835": {
        "content": {
            "composition#comp": "N1P3",
            "http://dublincore.org/documents/dcmi-terms#title": "OQMD - N1P3",
            "http://www.oqmd.org#bandgap": 0.0,
            "http://www.oqmd.org#delta_e": 0.857929291944171,
            "http://www.oqmd.org#energy_pa": -5.0409296875,
            "http://www.oqmd.org#id": "302835",
            "http://www.oqmd.org#stability": 1.06730158537404,
            "http://www.oqmd.org#volume_pa": 14.1023
        },
        "mimetype": "application/json"
    }
}
```

Search Result Aggregation...

- REST API, command-line client, and python client
- Indexed ~10 materials-related sources (OQMD, HOPV, JANAF-NIST, Khazana, Nanomine, Ab Initio Solute-Solvent Diffusion Database, etc.) ~1M records

Authenticate Search Client

Uses a valid authentication token if it exists, redirects to web if not

```
client = globus_auth.login("https://datasearch.api.demo.globus.org/")
```

Search for Data and Aggregate into a DataFrame

```
params = {  
    'q': "OQMD",  
    'highlight': True,  
    'resource_type': None,  
    'count': 20,  
    'from': 0,  
    'stats': True,  
    'facets': None,  
    'filters': None,  
}  
  
r = client.search(**params)  
print("Found %d results"%(r.data['gstats']['total']))  
  
Found 264251 results
```

Aggregate a Larger Query

```
n_records = 1000  
offset = 0  
vectors = []  
  
for i in range(0,5):  
    params = { 'q': "OQMD",  
               'count': n_records,  
               'from': i*n_records}  
  
    r = client.search(**params)  
    result_iterator = r.data['gmeta']  
    for result in result_iterator:  
        vectors.append(get_vectors(result))  
  
df = pd.DataFrame(vectors, columns=["composition", "delta_e",  
                                     "energy_pa", "volume_pa",  
                                     "magmom_pa", "stability"])
```

Search Result Aggregation...

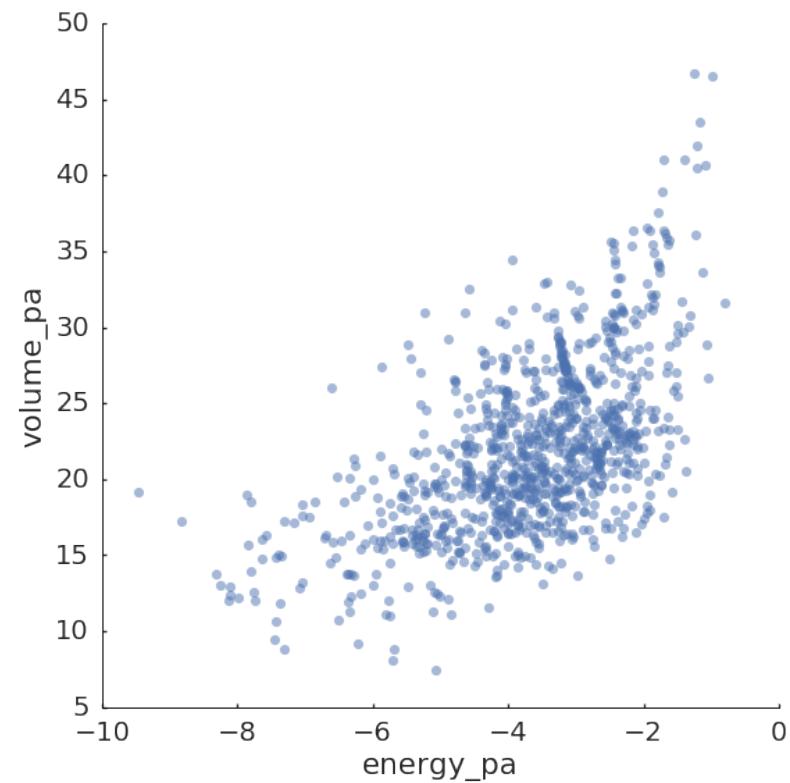
- REST API, command-line client, and python client
- Indexed ~10 materials-related sources (OQMD, HOPV, JANAF-NIST, Khazana, Nanomine, Ab Initio Solute-Solvent Diffusion Database, etc.) ~1M records

Visualize the Results

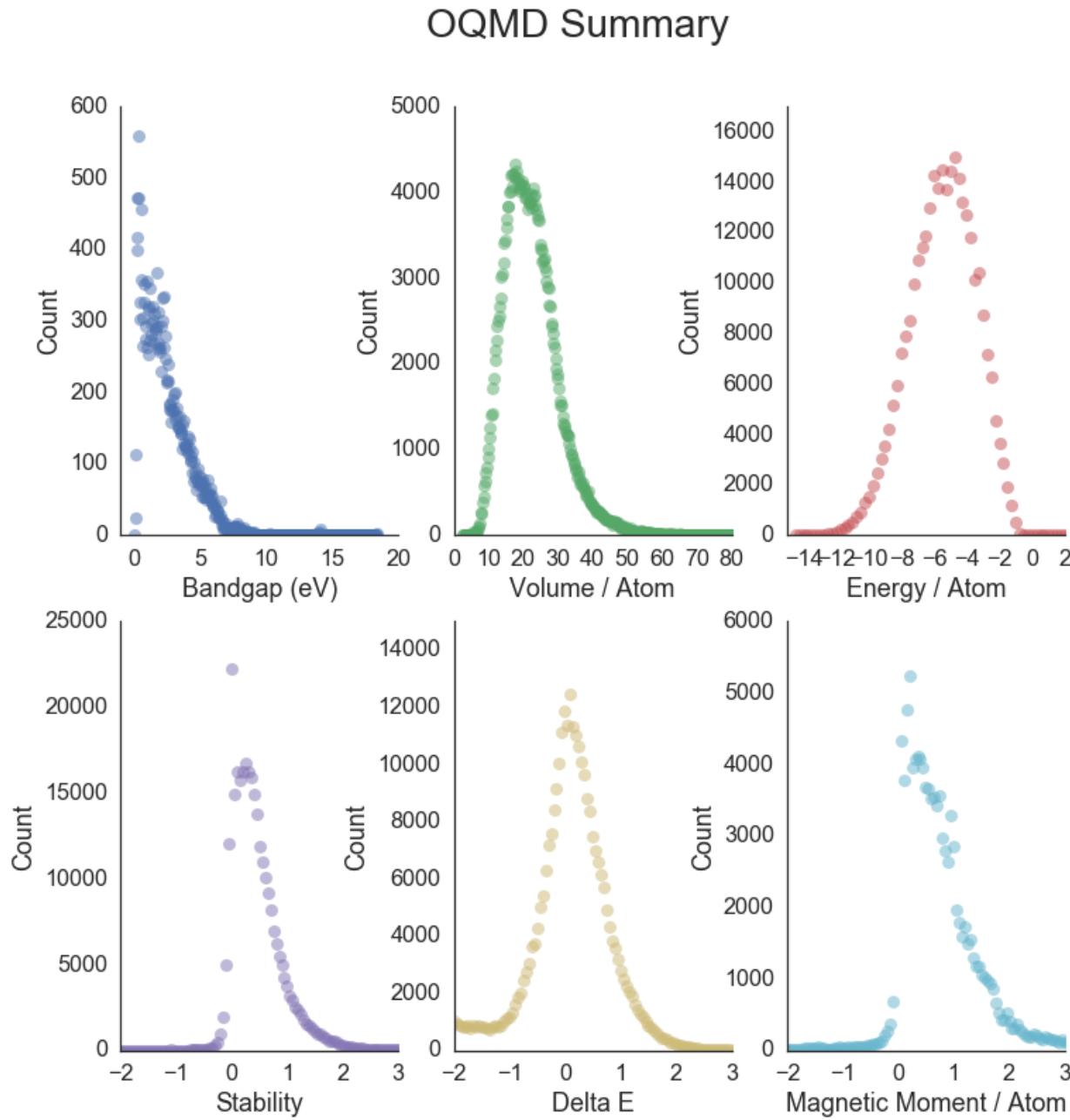
```
pal = sns.color_palette("muted")

x_col = "energy_pa"
y_col = "volume_pa"

# Make plot
g = sns.lmplot(x=x_col, y=y_col, data=df, fit_reg=False,
                 palette=pal, size=8, scatter_kws={"s": 50,
`"alpha": 0.5})
```



Dataset Summary Aggregation



CHiMaD Nanomine Example

Perform Nanomine Query

```
params = {
    'q': "nanomine",
    'count': 10,
    'from': 0
}

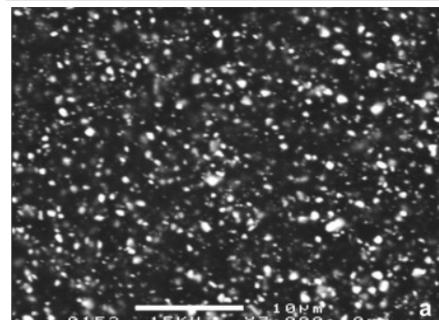
r = client.search(**params)
res = pop_top_gmeta(r)
head = res[0]['content'][0][ '#content'][ '#PolymerNanocomposite']

head.keys()

dict_keys(['#PROCESSING', '#ID', '#PROPERTIES', '#MATERIALS', '#CHARACTERIZATION', '#DATA_SOURCE', '#MICROSTRUCTURE'])
```

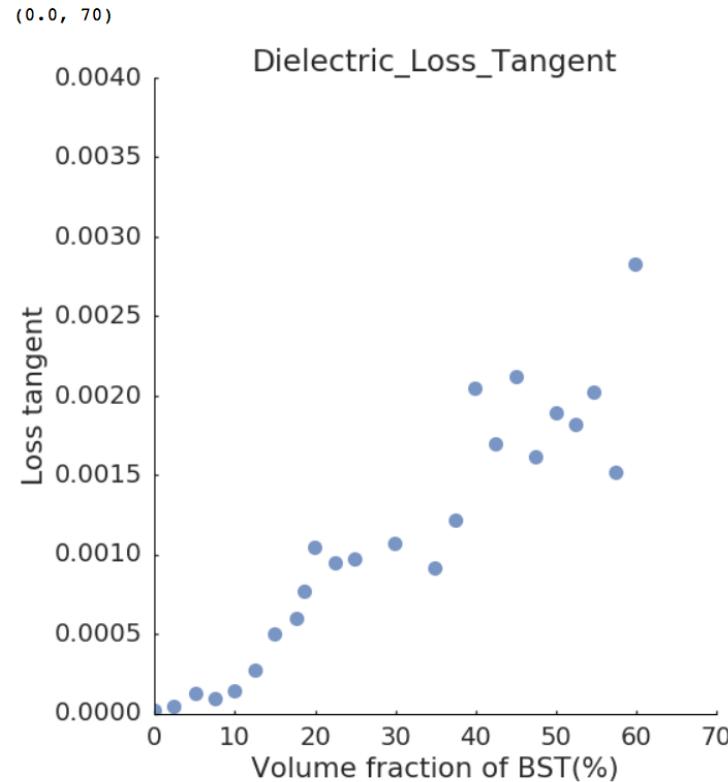
Get Microstructure Image

```
microstructure = head['#MICROSTRUCTURE']
Image(url=microstructure['#ImageFile'][ '#File'])
```



Example Plot of Dielectric Loss Tangent

```
pal = sns.color_palette("muted")
g = sns.lmplot(x=headers[0], y=headers[1], data=df_nm,
                fit_reg=False, palette=palette, size=8,
                scatter_kws={"s": 125, "alpha": 0.75})
g.ax.set_title('${}_{{}}({}_{})')
g.ax.set_xlim(0.0, 0.004)
g.ax.set_ylim(0.0, 70)
```



Where Can I get More Information?

JOM
August 2016, Volume 68, Issue 8, pp 2045–2052

The Materials Data Facility: Data Services to Advance Materials Science Research

B. Blaiszik [✉](#), K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, I. Foster [✉](#)

Article
First Online: [06 July 2016](#)
DOI: [10.1007/s11837-016-2001-3](#)

Cite this article as:
Blaiszik, B., Chard, K., Pruyne, J. et al. JOM (2016) 68: 2045. doi:10.1007/s11837-016-2001-3

2 Citations 3 Shares 147 Views

<http://dx.doi.org/10.1007/s11837-016-2001-3>

Links:

materialsdatafacility.org

globus.org

Contact:

blaiszik@uchicago.edu

foster@uchicago.edu

MATERIALS DATA FACILITY

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](#) and [CHIMaD](#)

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.

Contributors (unordered)

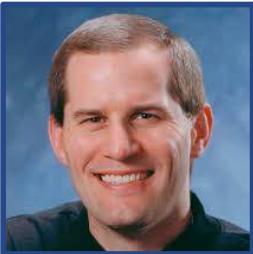
UC/Argonne 



Ian Foster (PI)



Ben Blaiszik



Steve Tuecke (PI)



Jim Pruyne



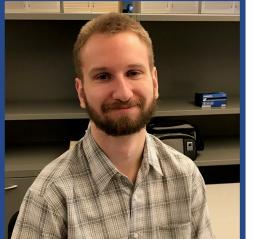
Rachana



Kyle Chard



Logan Ward



Jonathon Gaff



Stephen Rosen

Illinois (Urbana-Champaign)



John Towns (PI)



Kenton McHenry



Michal Ondrejcek

Thanks to Our Sponsors!



CHIMaD
Center for Hierarchical Materials Design



U.S. DEPARTMENT OF
ENERGY



Argonne
NATIONAL LABORATORY



THE UNIVERSITY OF
CHICAGO

REST APIs, Clients, and Docs

- **Globus Docs**
 - <https://docs.globus.org>
- **New Python SDK available**
 - <https://github.com/globusonline/globus-sdk-python>
- **Jupyter Notebook Examples**
 - <https://github.com/globus/globus-jupyter-notebooks>
- **Sample Data Portal**
 - <https://github.com/globus/globus-sample-data-portal>
- **(alpha) MDF Data Publication Service API**

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"], 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"], 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"])
```

Integration with the Community is Key

