Data Publication and Discovery with Globus

GlobusWorld 2018

Kyle Chard

globus®

Globus Data Publication V1

SaaS publication

BYO Storage & in-place publication

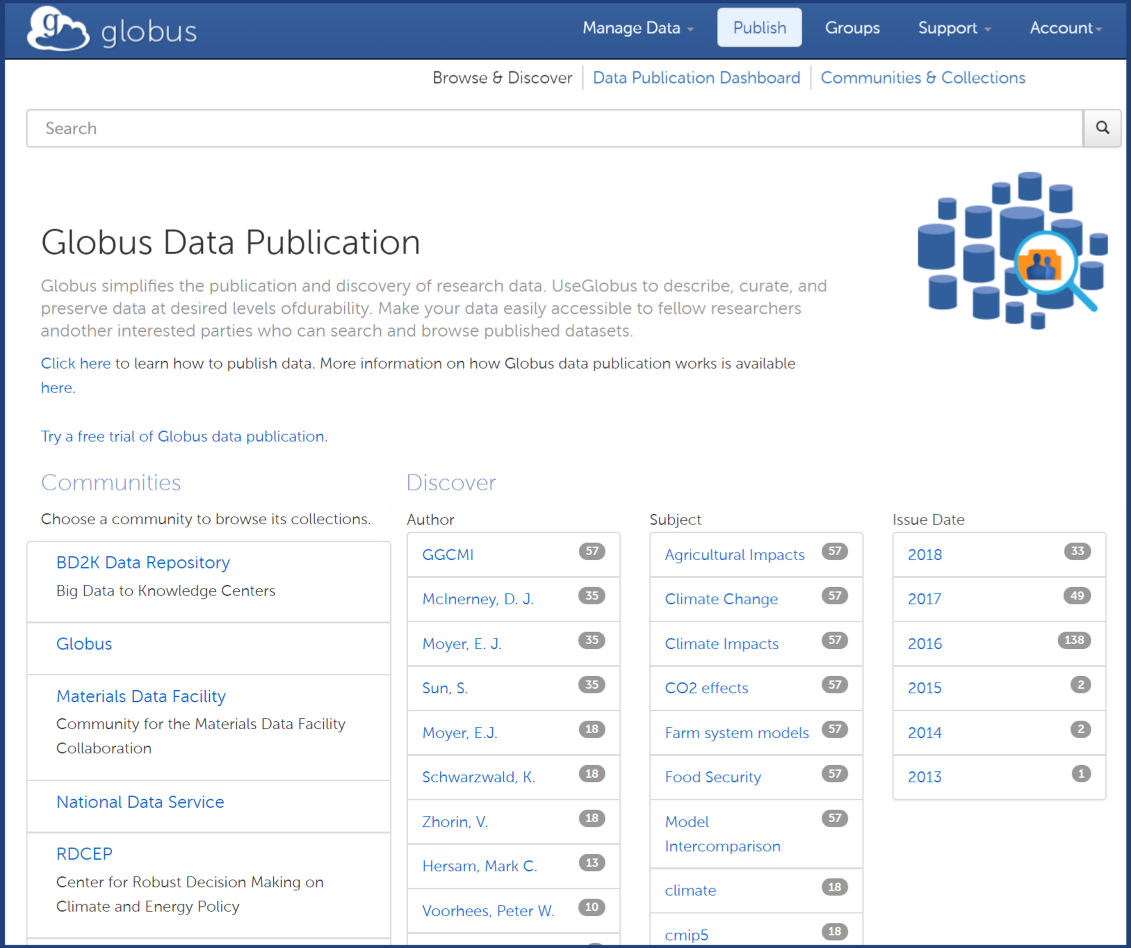
User-managed collections

Arbitrary metadata (with pre-defined schema)

Handle, DOI PIDs

Adoption since 2015:

>1800 users, >600 datasets



Globus Data Publication

Globus simplifies the publication and discovery of research data. UseGlobus to describe, curate, and preserve data at desired levels ofdurability. Make your data easily accessible to fellow researchers andother interested parties who can search and browse published datasets.

Click here to learn how to publish data. More information on how Globus data publication works is available here.

Try a free trial of Globus data publication.

Communities

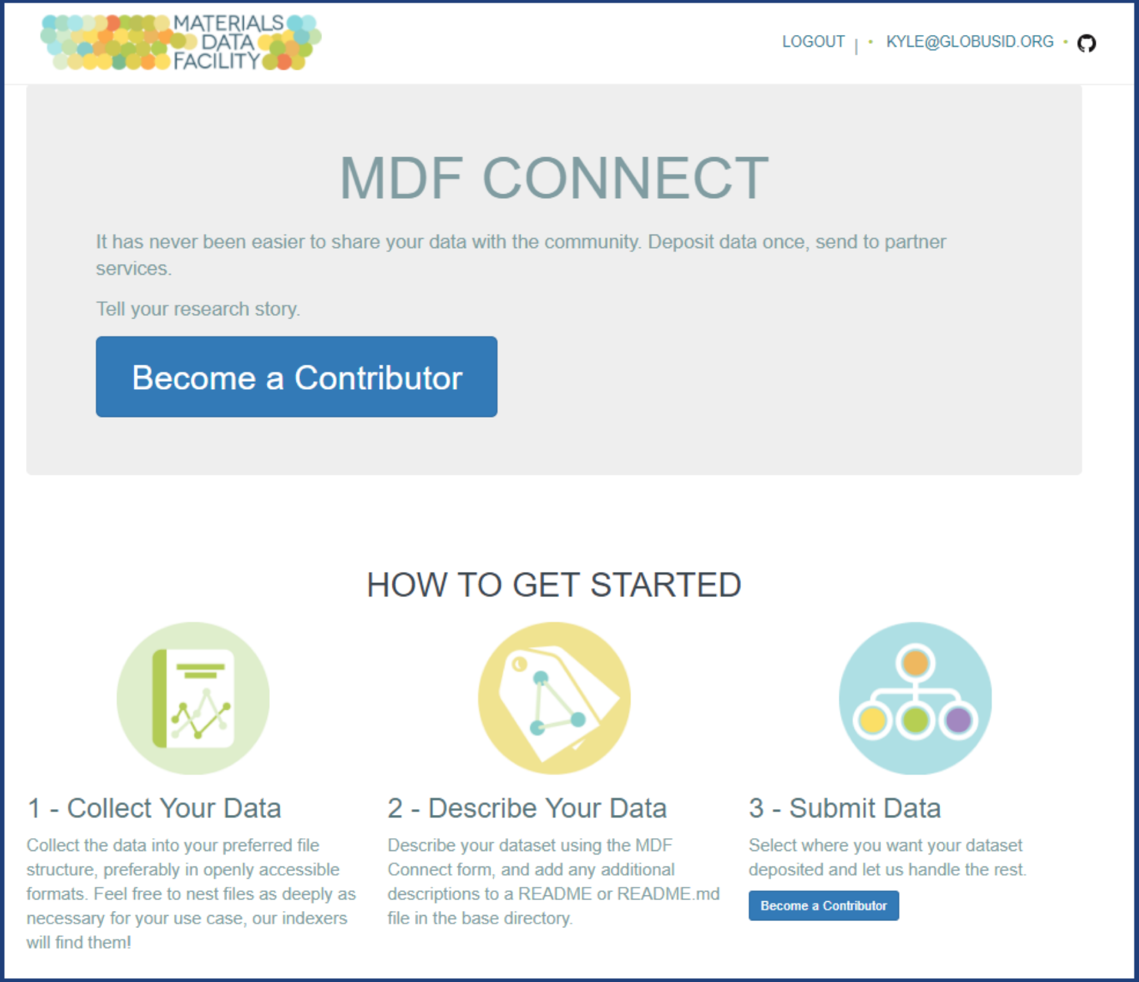
Choose a community to browse its collections.

- BD2K Data Repository
  Big Data to Knowledge Centers
- Globus
- Materials Data Facility
  Community for the Materials Data Facility Collaboration
- National Data Service
- RDCEP
  Center for Robust Decision Making on Climate and Energy Policy

Discover

Author		Subject		Issue Date	
GGCMI	57	Agricultural Impacts	57	2018	33
McInerney, D. J.	35	Climate Change	57	2017	49
Moyer, E. J.	35	Climate Impacts	57	2016	138
Sun, S.	35	CO2 effects	57	2015	2
Moyer, E.J.	18	Farm system models	57	2014	2
Schwarzwald, K.	18	Food Security	57	2013	1
Zhorin, V.	18	Model Intercomparison	57		
Hersam, Mark C.	13	climate	18		
Voorhees, Peter W.	10	cmip5	18		

Publication V1 success stories



MDF CONNECT

It has never been easier to share your data with the community. Deposit data once, send to partner services.

Tell your research story.

HOW TO GET STARTED

1 - Collect Your Data

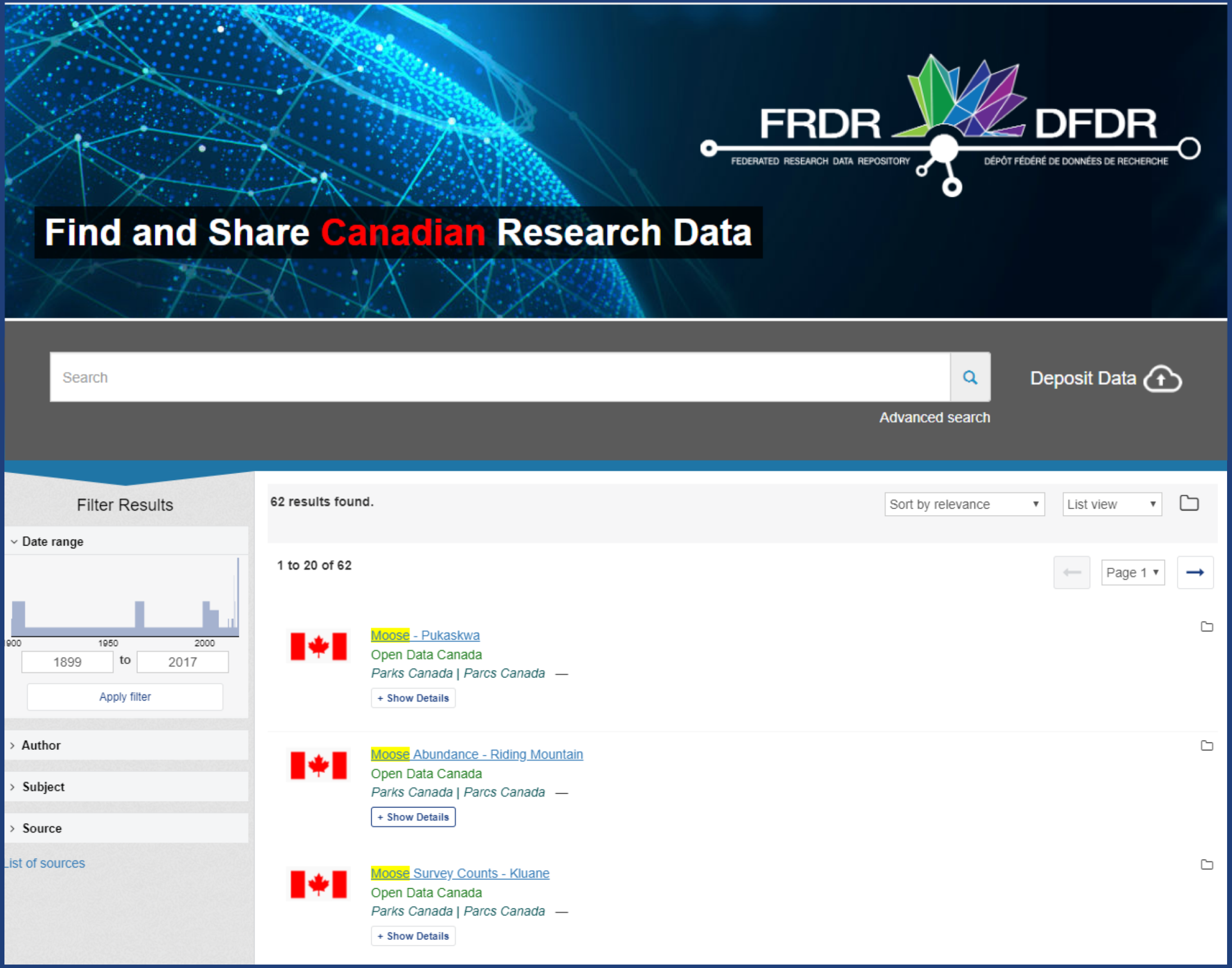
Collect the data into your preferred file structure, preferably in openly accessible formats. Feel free to nest files as deeply as necessary for your use case, our indexers will find them!

2 - Describe Your Data

Describe your dataset using the MDF Connect form, and add any additional descriptions to a README or README.md file in the base directory.

3 - Submit Data

Select where you want your dataset deposited and let us handle the rest.



Find and Share Canadian Research Data

62 results found.

1 to 20 of 62

Moose - Pukaskwa
Open Data Canada
Parks Canada | Parcs Canada —

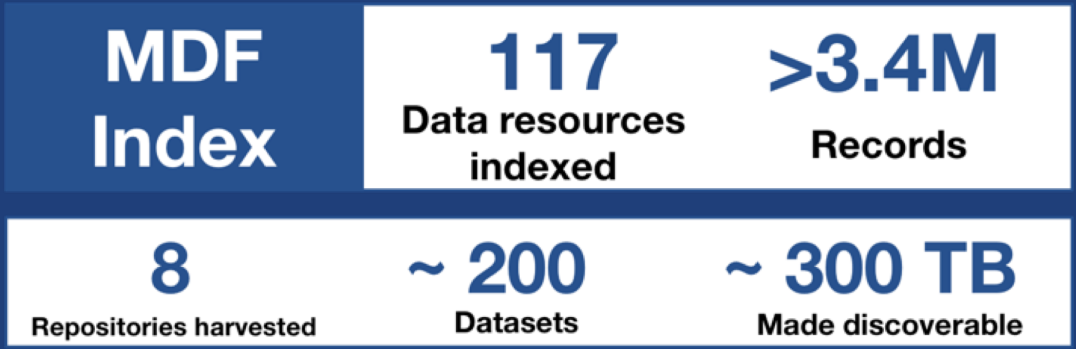
Moose Abundance - Riding Mountain
Open Data Canada
Parks Canada | Parcs Canada —

Moose Survey Counts - Kluane
Open Data Canada
Parks Canada | Parcs Canada —

https://materialsdatafacility.org

https://frdr.ca/

Publication V1 success stories



MDF Index

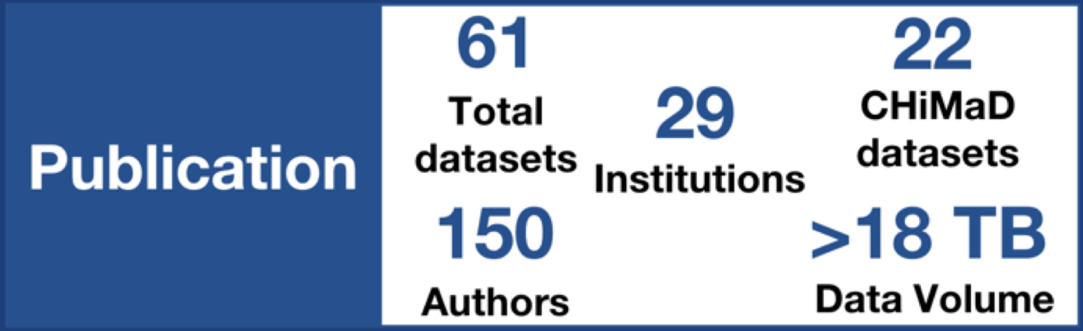
117 Data resources indexed

>3.4M Records

8 Repositories harvested

~ 200 Datasets

~ 300 TB Made discoverable



Publication

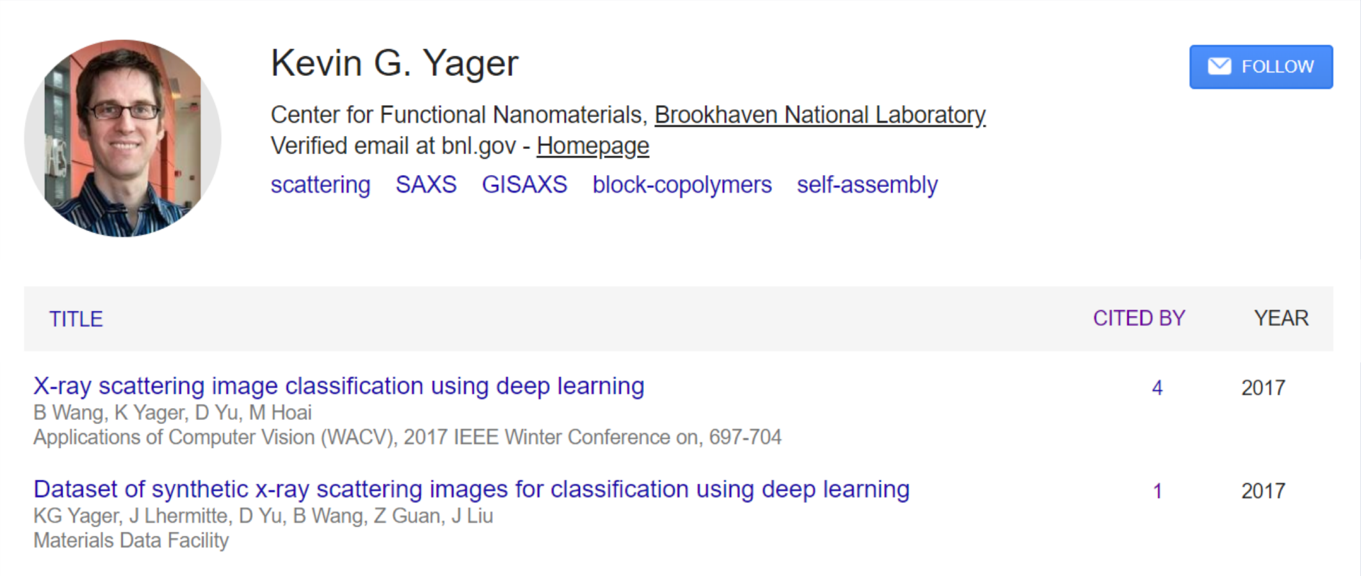
61 Total datasets

29 Institutions

22 CHiMaD datasets

150 Authors

>18 TB Data Volume



Kevin G. Yager

Center for Functional Nanomaterials, Brookhaven National Laboratory
Verified email at bnl.gov - Homepage

scattering SAXS GISAXS block-copolymers self-assembly

TITLE	CITED BY	YEAR
X-ray scattering image classification using deep learning B Wang, K Yager, D Yu, M Hoai Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, 697-704	4	2017
Dataset of synthetic x-ray scattering images for classification using deep learning KG Yager, J Lhermitte, D Yu, B Wang, Z Guan, J Liu Materials Data Facility	1	2017

Publication V1 lessons learned

Every domain, institution, researcher has

- Different definition of data publication
- Different publication requirements

Current systems are monoliths

- Little support for customization
- No way to combine the “good bits” of several services

Use cases demand flexibility, adaptability, and extensibility

Publication V2: Publication as a Platform

Publication as a Platform

- Decompose Globus Publish v1 into platform components
- Allow for flexible re-composition and adaptation by customers
- Enable extension and enhancement

Initial services

- Identifiers, search, (and data management)

Future services

- Description (metadata), automation (workflows)

Globus Search platform service

- **Search service:**
  - **Scalable**: to billions of entries
  - **Schema agnostic**: can use standard (e.g., DataCite) or custom metadata
  - **Fine grain access control**: only returns results that are visible to user
  - **Plain text search**: ranked results
  - **Faceted search**: for data discovery
  - **Rich query language**: ranges, expressions, regex, fuzzy, stemming, etc.
- **Limited production, generally available target year end**

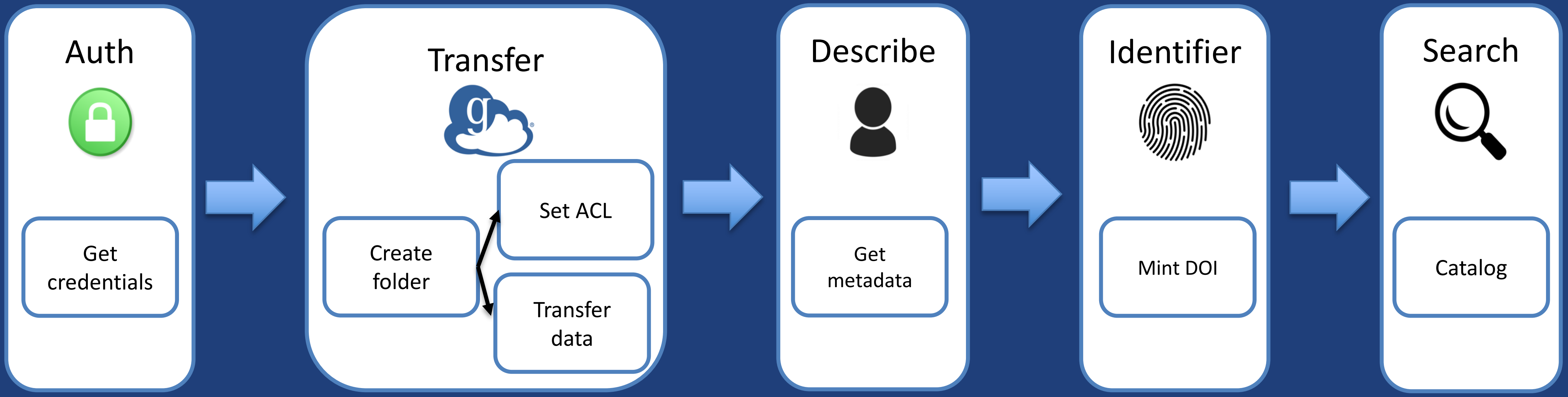
Globus Identifiers platform service

- **Issue persistent identifiers**
  - DOI, ARK, Handle, Globus
  - E.g., https://identifiers.globus.org/doi:10.1145/2076450.2076468
- **Within a namespace**
  - E.g., Your University's DataCite namespace
  - Control which identities and groups can create identifiers in your namespace
- **Each identifier has:**
  - **Link to data**: one or more https URLs, to file, folder or manifest
  - **Landing page**: provided by service, or by user
  - **Visibility**: which identities and groups can see identifier
  - **Checksum**: of the file or manifest
  - **Metadata**: as required by identifier (e.g., DataCite), extensible
  - **Replaces / Replaced-by**: for versioning
- **Limited beta available now, generally available year end**

Publication Platform Tutorial

What are we going to show?

– Creating a complete publication workflow composed of Globus publication platform services (in less than an hour)



1) Publish data

- **Goal:**
  - – Immutable, reliable, and accessible storage of files and directories
- **Steps:**
  - – Define a location for data storage
    - ○ On your endpoint, on a storage system, on the cloud, …
  - – Transfer data to that location
  - – Set access permissions to
    - ○ Make the data immutable (read-only)
    - ○ Make it accessible to appropriate users and groups

2) Associate an identifier

- **Goal:**
  - Persistent, unambiguous identifier for the dataset
- **Steps:**
  - Mint an ARK for the published data
    - Location: Globus URL
    - Metadata: author, title, date
  - Lookup the identifier to find
    - Machine-accessible information
    - Human-accessible landing page

3) Indexing metadata for discovery

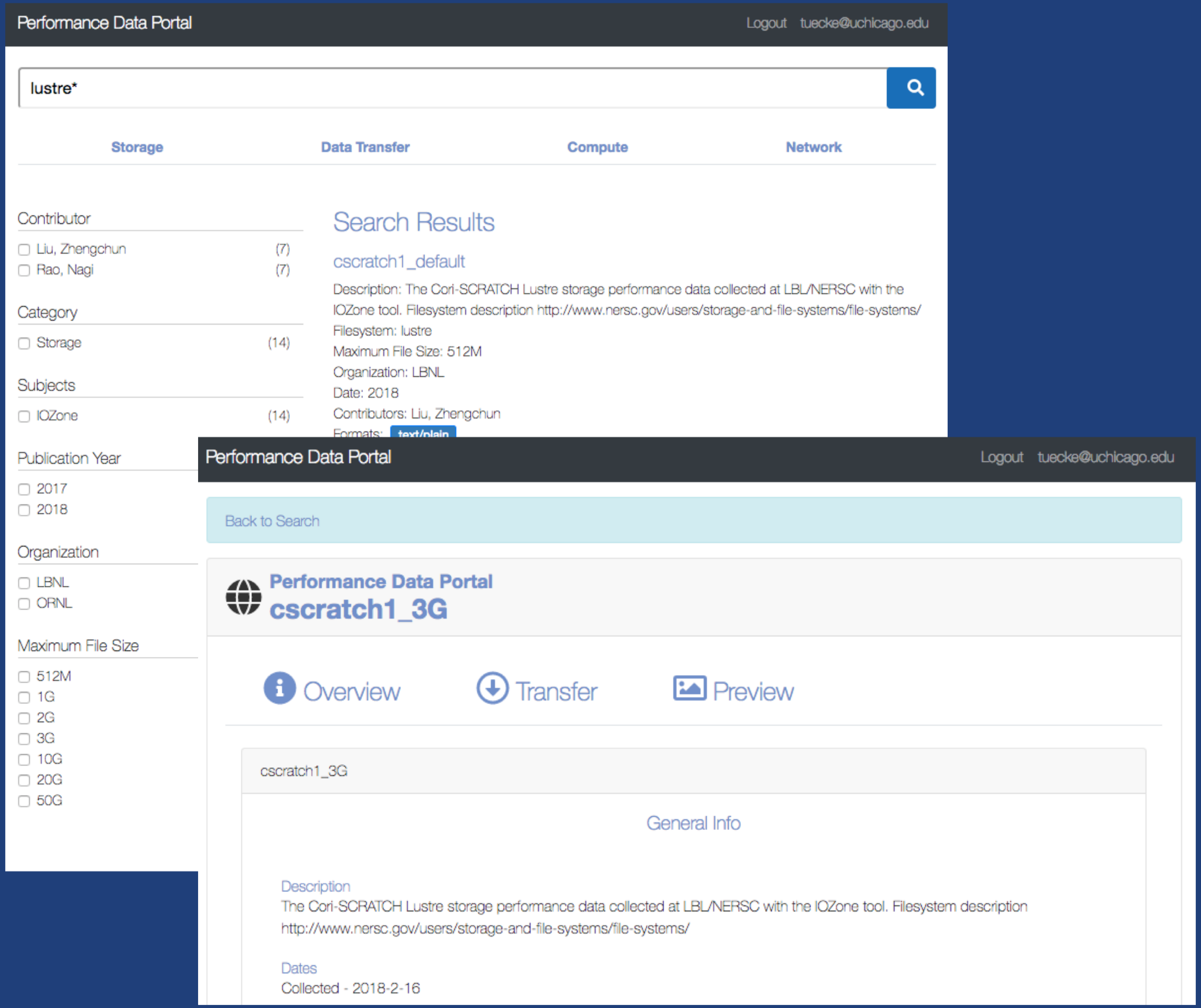
- **Goal:**
  - Index descriptive metadata, with access control, to allow others to discover the published dataset
- **Steps:**
  - Add the dataset to a search index
    - Location & metadata
  - Set access permissions
    - Core metadata public
    - Additional metadata restricted

3b) Indexing metadata for discovery

- **Goal:**
  - Search the index to discover published datasets
- **Steps:**
  - Explore query models and result formats
    - Free-text
    - Exact matches
    - Filtering and faceting

4) Creating a portal

- **Goal:**
  - Provide a GUI to discover, view, and download datasets
- **Steps:**
  - Use the example Django portal to find and download your datasets



lustre*

Contributor

☐ Liu, Zhengchun (7)
☐ Rao, Nagi (7)

Category

☐ Storage (14)

Subjects

☐ IOZone (14)

Publication Year

☐ 2017
☐ 2018

Organization

☐ LBNL
☐ ORNL

Maximum File Size

☐ 512M
☐ 1G
☐ 2G
☐ 3G
☐ 10G
☐ 20G
☐ 50G

Search Results

cscratch1_default

Description: The Cori-SCRATCH Lustre storage performance data collected at LBL/NERSC with the IOZone tool. Filesystem description http://www.nersc.gov/users/storage-and-file-systems/file-systems/
Filesystem: lustre
Maximum File Size: 512M
Organization: LBNL
Date: 2018
Contributors: Liu, Zhengchun
Formats: text/plain

Performance Data Portal
cscratch1_3G

cscratch1_3G

General Info

Description
The Cori-SCRATCH Lustre storage performance data collected at LBL/NERSC with the IOZone tool. Filesystem description http://www.nersc.gov/users/storage-and-file-systems/file-systems/

Dates
Collected - 2018-2-16

Summary

Monolithic publication systems are not sufficient for increasingly varied data publication scenarios and requirements

Globus data publication platform supports:

- Large datasets, *any* storage location, customizable metadata, flexible access control, user-oriented curation workflows, self service management, choice of persistent identifier, powerful search capabilities
- Users can build upon, extend, customize these services to develop publication pipelines for any scenario