Data Publication and Discovery with Globus

GlobusWorld 2018
Kyle Chard
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
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.

Become a Contributor

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

https://frdr.ca/ https://materialsdatafacility.org
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
- 150 Authors
- 22 CHiMaD datasets
- >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

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

• **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)

- Auth
  - Get credentials

- Transfer
  - Create folder
  - Transfer data
  - Set ACL

- Describe
  - Get metadata

- Identifier
  - Mint DOI

- Search
  - Catalog
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
    o Location: Globus URL
    o Metadata: author, title, date
  – Lookup the identifier to find
    o Machine-accessible information
    o 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
    o Location & metadata
  – Set access permissions
    o Core metadata public
    o 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
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