Data publication and discovery with Globus

Questions and comments to *outreach@globus.org*

The Globus data publication and discovery services make it easy for institutions and projects to establish “collections”, which are dataset repositories with customized curation policies, metadata schema, access policies, etc.; for researchers to create and populate collections with datasets comprising (big) data and metadata; and for researchers to search, browse, and access datasets contained within any collections they are authorized to access. A software-as-as-service (SaaS) approach means that the costs associated with establishing, accessing, and operating repositories and collections are low.

Globus data publication and discovery services were first demonstrated in prototype form at the GlobusWorld conference in April 2015. A video of the demonstration is available at www.globus.org/data-publication. The initial production version of the services was released in October 2015.

Globus data publication features

- SaaS for publishing large data
- Bring your own storage
- Extensible metadata
- Publication and curation workflows
- Public and restricted collections
- Rich discovery model

Globus background

Globus leverages cloud-based software-as-a-service (SaaS) methods to deliver powerful research data management capabilities over the network. For researchers, Globus accelerates discovery by providing easy-to-use, self-service capabilities for rapid, reliable, and secure data management. For campuses and other resource providers, Globus enhances the value, usability, and manageability of valuable network and storage system resources. For tool developers, Globus can serve as a platform, allowing applications to outsource time-consuming tasks such as federated user and group management, and data management.

Globus currently supports file transfer, sharing, and publication. As of April 2017, it has 56,000 registered users, more than 10,000 active endpoints (i.e., storage systems connected to Globus), and has moved more than 256 petabytes in 36 billion files. It is recommended by major research facilities such NSF’s XSEDE and DOE’s NERSC.

Globus development and operations are supported by the DOE, NIH, NSF, Amazon, Sloan Foundation, and University of Chicago. A subscription model for premium features provides for long-term sustainability (details on Globus Subscriptions are at www.globus.org/subscriptions).

Data publication and discovery

A combination of federal mandates and good research practice is driving increased interest in data publication capabilities. There are limited tools currently available to librarians, digital media managers, and others on campus organizations tasked with managing data publication. Typical approaches involve developing, installing, and configuring various software components, and integrating these with existing campus identity and storage systems. This is a costly and time-consuming activity that few can afford.

In response to these challenges, Globus publication offers features that make it easy to identify, describe, curate, verify, access, and preserve data at appropriate levels of durability. Additional planned features will enable rich discovery by making it possible to search, browse, and access large published data sets, irrespective of where they may be stored.

Globus data publication services benefit many stakeholders including researchers, librarians, and campus IT managers. Researchers have a straightforward way to publish their data; they can define specific and relevant metadata; they can easily annotate and describe large data sets with the help of automation; and they can discover relevant data sets published by others, both within and beyond their field of study. Librarians can leverage robust, scalable infrastructure to curate and preserve diverse data collections at modest cost; they have complete control over the curation process; and they do not need to invest in custom solutions or acquire unnecessary technical skills. Campus IT managers can leverage existing resources as storage repositories and campus identity systems; and they can provide advanced services to students, faculty, and staff with minimal new investment.

System Overview

Globus publication capabilities are delivered through a hosted service, meaning that no software need be installed or operated to run the service. Published data and associated metadata is stored on campus, institutional, or group resources that are managed and operated by their respective administrators. A copy of metadata is also stored in the cloud and indexed for facilitating rich discovery models, including cross collection and community searches. To enable a storage resource as a repository for data publication, administrators configure a Globus endpoint for sharing and then associate the endpoint with the data collection through Globus. Creating a Globus endpoint is a straightforward process requiring just a few commands to install and configure Globus Connect Server (www.globus.org/globus-connect-server).

Datasets are published into “collections”, which in turn can be organized into “communities” (see Figure 1). For example, an Argonne National Laboratory community has several member collections: Advanced Photon Source; Center for Nanoscale Materials; and Computing, Environment and Life Sciences, to name a few. Often, collections will map to a department or group within an institution, but this is not required. Globus users can create and manage their own communities and collections through the data publication service. A collection enables the submission of datasets with policies regarding access.

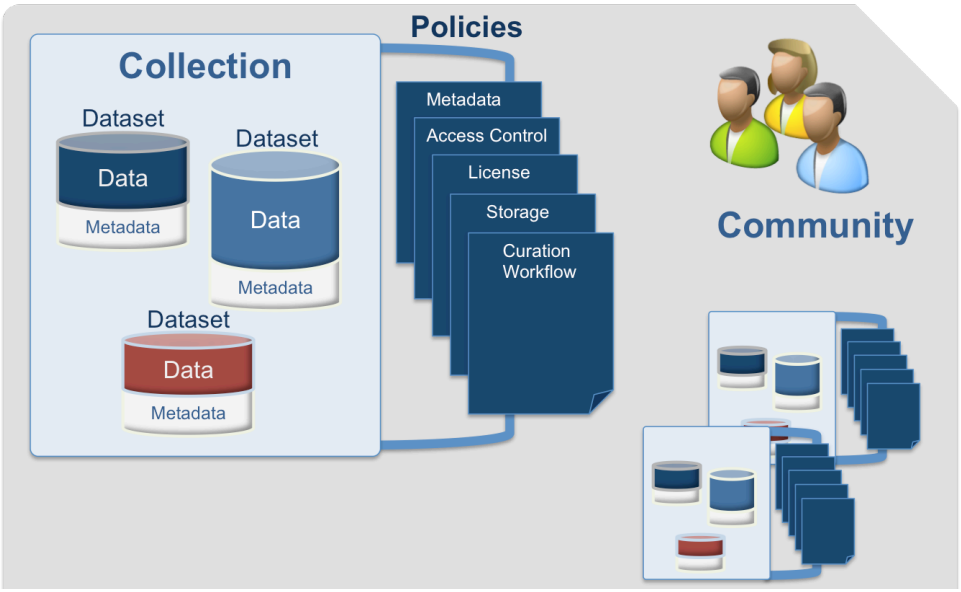

Figure 1: Globus Publishing Model

A dataset comprises data and metadata. Policies can be set on entire communities and/or individual collections to manage: metadata (schema, requirements); access control (user and group based); curation workflow; submission and distribution license; and storage (e.g., which determines durability).

Data Publication

Datasets undergo curation based on a workflow defined by the community that will publish the data. Each community may customize the workflow to capture their specific metadata, and to reflect the community's review process.

In a typical workflow, a scientist who has just published a paper wishes to publish the data associated with his publication. Using Globus:

(1) the scientist selects the community that will publish his data, and the collection that will house his datasets;
(2) the scientist describes the submission using both publication (Dublin core) and collection-specific scientific metadata (e.g., see Figure 2), as defined by the collection administrator;
(3) a unique Globus endpoint is created and only the scientist is granted permission to write to the endpoint for this submission;
(4) the scientist assembles a dataset on this endpoint by transferring files from one or more systems, e.g., from the campus computing cluster where the final analysis was run, and from the high-resolution microscope used to capture the raw experiment data (the scientist can assemble this dataset over a long period of time and can return to the submission workflow when all the desired data are in place);
(5) the scientist agrees to the submission license specified by the community for that collection and submits the data for publication (the license allows the submitting user to grant rights to the collection and the Globus system to manage and disseminate the dataset based on the agreed upon policies);

(6) the curators/reviewers for the community are notified that a submission awaits their approval, and are granted permission to access the submission;
(7) the curators view (and, if desired, edit) the metadata and assembled files of the dataset, and approve the submission;
(8) the submission is published in the collection and assigned a DOI.

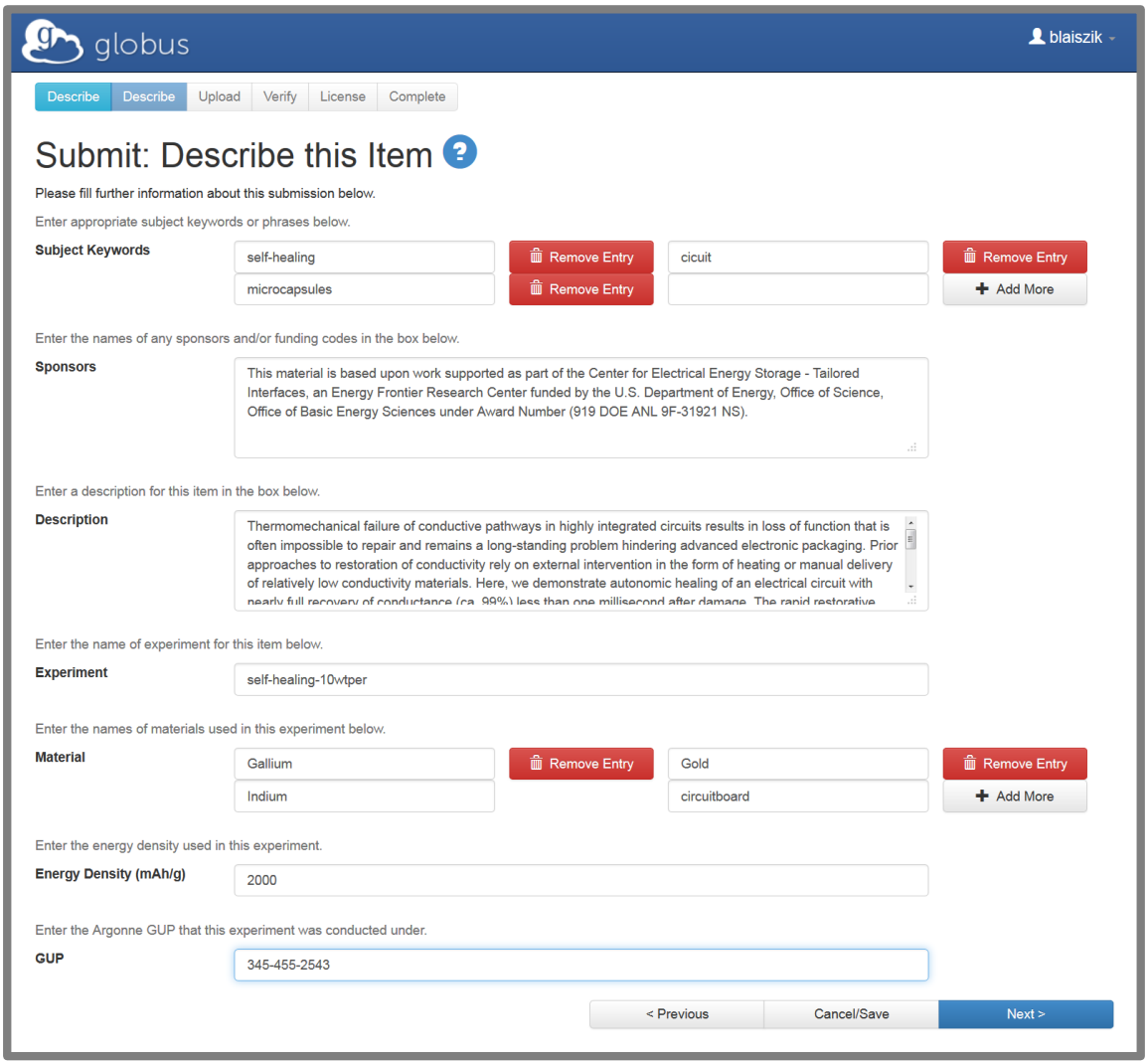

Figure 2: The user supplies domain-specific metadata

Most steps in this process may be customized to meet the requirements of the community. Globus supports a wide range of “data publication” workflows, ranging from formal archival publication to informal sharing within a group of collaborators.

Data discovery

Globus discovery capabilities make it easy to efficiently search and browse collections. After a dataset is published, it is discoverable using both free text and faceted search that allows the researcher to progressively filter results and rapidly focus in on the data of interest.

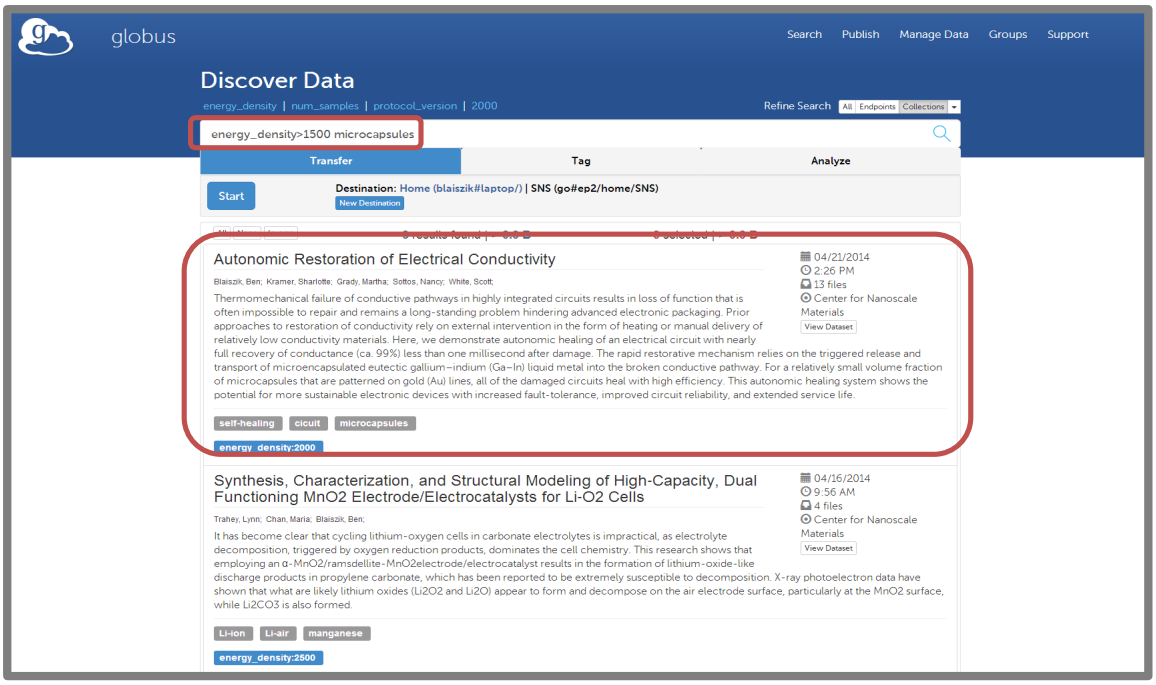

Figure 3: Users can search within and across collections

The interface for discovery of datasets is not unlike Amazon's interface for discovery of products. The first step of discovery is to define the context in which the user wants to search. The user may use free text search terms, key-value terms, or even range queries. A user may limit the search context to a single collection, to collections owned by a community, or to all collections to which the user has access, including collections published by other communities. Future plans call for the search context to include both collections (i.e., published datasets) and endpoints (i.e., uncurated, but accessible, datasets).

Search results show a brief summary of each published dataset, including information about the publication time, collection, summary information about the data files, name, authors, description, and a set of keyword tags as well as key-value tags. Each field can be used to search for a particular dataset. Results are ranked according to their relevance to the search criteria. The researcher can iteratively refine the search by expanding the number of search terms and/or by adding more specific values to the query (see Figure 3).

Once data of interest are found, the researcher may transfer them to another Globus endpoint for further inspection and processing.

DOIs and ORCIDs

We expect that many data repository managers will want to assign Digital Object Identifiers (DOIs: www.doi.org) to datasets and associate Open Researcher and

Contributor Identifiers (ORCIDs: www.orcid.org) for data authors with datasets. The Globus data publication and discovery service supports both DOIs and ORCIDs.

Linking to publishers

An increasing number of academic publishers allow researchers to associate datasets with publications, typically by supplying DOIs. As part of our work with the National Data Service (www.nationaldataservice.org), we are working with publishers to determine how to incorporate such linkages into Globus data publication workflows.

Implementation

The Globus data publication and discovery prototype leverages and extends elements of the Globus service for user and file management, and DSpace (www.dspace.org) for publication and curation workflows, all hosted in the Amazon Web Services cloud.

For more information

See www.globus.org/data-publication for the latest on Globus data publication and discovery. That page also contains pointers to a talk and slides showing a demonstration of an early version of data publication and discovery functionality. Please give the service a try at https://trial.publish.globus.org/.