OR2015 | 10th International Conference on Open Repositories
June 8-11, 2015, Indianapolis, Indiana, USA

Globus Software as a Service data publication and discovery

Kyle Chard, University of Chicago – Computation Institute, chard@uchicago.edu
Jim Pruyne, University of Chicago – Computation Institute, pruyne@uchicago.edu
Rachana Ananthakrishnan, University of Chicago – Computation Institute, ranantha@uchicago.edu
Steve Tuecke, University of Chicago – Computation Institute, tuecke@uchicago.edu
Ian Foster, University of Chicago – Computation Institute, foster@anl.gov

Session Type (select one)

- ☐ Panel
- ☒ Presentation

Abstract

Globus is software-as-a-service for research data management, used at dozens of institutions and national facilities for moving, sharing, and publishing big data. Recent additions to Globus include services for data publication and discovery that enable: publication of large research datasets with appropriate policies for all types of institutions and researchers; the ability to publish data directly from locally owned storage or from cloud storage; extensible metadata that can describe the specific attributes of any field of research; flexible publication and curation workflows that can be easily tailored to meet institutional requirements; public and restricted collections that give complete control over who may access published data; and a rich discovery model that allows others to search and use published data. This presentation will give an overview of these services.

Conference Themes

Select the conference theme(s) your proposal best addresses:

- ☐ Supporting Open Scholarship, Open Science, and Cultural Heritage
- ☒ Managing Research (and Open) Data
- ☐ Integrating with External Systems
- ☐ Re-using Repository Content
- ☐ Exploring Metrics and Assessment
- ☐ Managing Rights
- ☐ Developing and Training Staff
- ☐ Building the Perfect Repository

OR2015 | 10th International Conference on Open Repositories
June 8-11, 2015, Indianapolis, Indiana, USA

Keywords

Globus, data publication, software-as-a-service repositories, research data management

Audience

This submission will be of interest to institutional repository managers, research computing centers, campus data centers, data producers, data publishers, librarians and others who face challenges managing and publishing large amounts of research data.

Background

Our submission presents Globus' data publication capabilities, an effort to support the creation of open, accessible and scalable repositories for research data building upon Globus' research data management capabilities and extending the commonly used DSpace system. Important features include the ability to reliably and efficiently publish huge amounts of data via a "bring-your-own" storage model in which repository providers may use their own storage resources. This approach ensures complete autonomy and control over all published data. Globus' data publication capabilities are offered through a professionally hosted service that is available to the wider research community without requiring installation or management of any software locally. This approach provides a flexible solution for institutional data repositories, project or group data repositories, national data repositories, and publisher data repositories.

Presentation content

We are at an important stage in the development, deployment and adoption of open data repositories. The demands on repositories are growing in multiple dimensions: the size of the data, the number and type of users who must interact with the data throughout the publication life-cycle and not least of all the rate of demand for repositories as funding agencies, publishers and researchers require data be made openly accessible. While demand is clear, suitable systems that support the deposit and controlled access to large research data are few. The challenges associated with publishing data, in comparison with publishing documents, makes repurposing existing repositories non-trivial and the increasing need for publication of "Big Data" has necessitated the use of distributed and cloud-based approaches.

Many domains, particularly library sciences, have invested significant effort developing data publication solutions. Examples include Dataverse [1], figshare [2], PURR [3], and ScholarSphere [4]. Each enables researchers to create and publish documents and datasets for wider dissemination. However, such systems are typically operated at the level of an individual institution, so they support only institutional users and rely on institutional storage, thus negating

the opportunity for easy publication of data in multi-institutional collaborations while also forming isolated silos that are difficult to search across. Service-based approaches such as figshare, provide global collections through which datasets can be found, however many researchers and institutions require creation and management of their own collections through which they can control every aspect of the publication process. In most cases, due to their generic focus, existing repositories support only common metadata schemas rather than extensible domain-specific schemas. In addition, they have often evolved from digital content repositories and therefore do not support the deposit and storage of increasingly large datasets. For example, Dataverse limits dataset uploads to 2 GB, and Figshare to 1 GB, insufficient for large experimental datasets.

Research data is growing rapidly, it is often diverse and less well-structured than the digital content traditionally stored by repositories. New approaches based on automation and outsourcing are required to meet the unique challenges associated with research data publication. Just as Flickr transformed photo preservation and sharing by making it easier to store data online than not, so too can appropriate technology transform data preservation and sharing practices for research data. We believe that the key is to leverage modern cloud computing to enable the automation and outsourcing of currently manual practices for collecting, organizing, sharing, searching, and accessing large quantities of research data.

In this presentation we describe Globus' recently released data publication capabilities. Over the past four years, Globus has become a foundational data management service that scientists rely upon for everyday research activities. Globus has been used to transfer over 77 PB of data, across more than 3 billion files and has grown to support more than 20,000 users. Globus' primary offerings support high-performance data transfer, secure storage management, in-place data sharing, and distributed user authentication and group management. Each of these capabilities is provided via a professionally hosted Software as a Service (SaaS) offering which any researcher can access using their browser without requiring installation of client software.

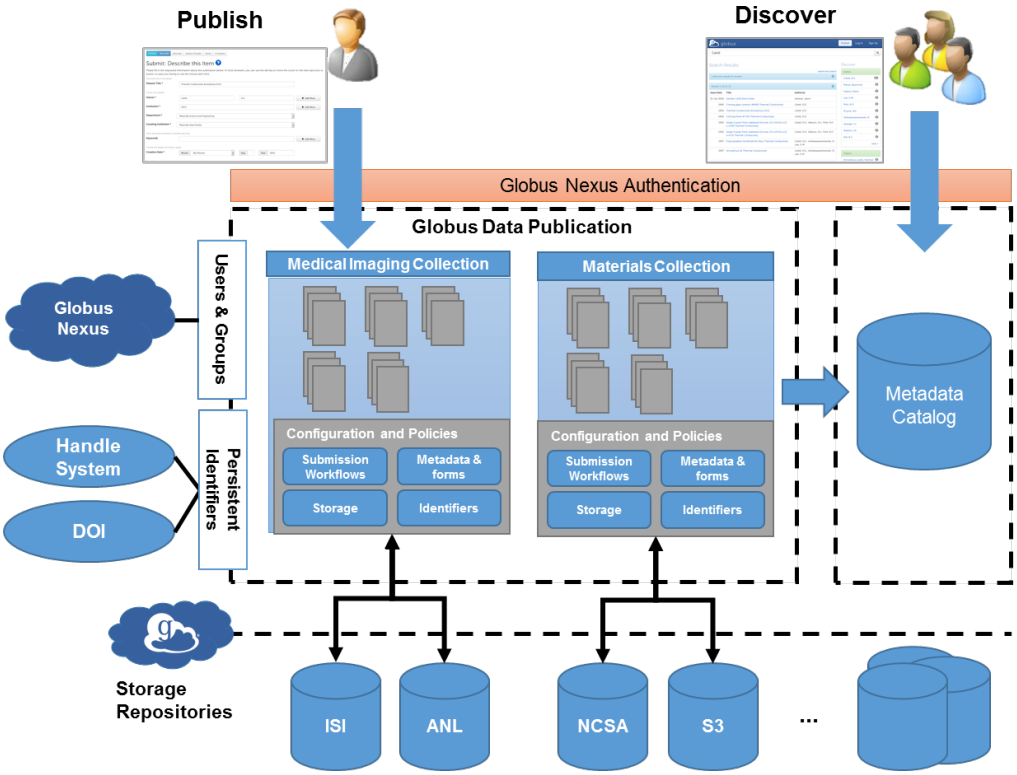
Globus Data Publication builds upon Globus' data management capabilities by providing a self-service system enabling users to create and manage their own data publication "collections." A collection is a repository for related published data such as may be maintained by a single research project, a department or a group within an institution or organization. Within a collection other users can publish datasets comprised of: a set of set of one or more files and directories, a set of metadata describing the characteristics, provenance, etc., of the dataset and a unique persistent identifier, which can be resolved to locate the dataset even if it's physical or network location changes. Collections are governed by user-defined policies established when they are created. These policies dictate which users can submit, view and discover entries in the collection, what metadata will describe the entries in the collection, what curation steps must be performed before a submission is entered in to the collection, and how and where the data is stored. Data storage policies allow users to provide their own, distributed storage avoiding the bottleneck of a central storage location and thus enabling transfer and storage of Big Data.

Like other Globus capabilities, Globus Data Publication is offered via a SaaS model through which any researcher, librarian, campus data storage manager or preservationist can create and manage their own collection and then allow others to publish and discover data interactively via a browser or programmatically via REST APIs. Collectively, these services provide intuitive and accessible interfaces for repository creation and depositing, describing, curating, sharing, searching, retrieving, and analyzing data sets of arbitrary size.

A schematic of the operational structure of the Globus Data Publication system is shown in the figure. The core capability is built upon DSpace [5] -- an institutional repository system designed for building open digital repositories. We chose to build upon DSpace as it provides an intuitive publication model, customizable publication workflows, granular access control, easy to use and extensible web interfaces, and it has been widely adopted. Also critical to our adoption of DSpace is its Open Source development model and licensing which has permitted us to augment it and integrate it with the other components of Globus and to operate it in a hosted, SaaS model. We have enhanced the DSpace system, leveraging rich identity and group support provided by Globus, scalable data upload, storage and access using Globus' transfer capabilities, and a self-service management model that allows collection owners to customize their collection in its entirety. We describe the unique components of our system below.

Globus Nexus provides user identity and group management across Globus services. A unique feature of Globus Nexus is its ability to manage "linked identities" a capability that allows a user to associate a variety of external identities such as Google, InCommon and other systems common in research environments with their Globus identity. This allows users to begin using and collaborating using Globus Data Publication without creation of a new identity or set of credentials. Globus Nexus also provides robust group management capabilities. We use groups to define roles and capabilities for users such as which collections a user can submit to and which users are responsible for curating these collections. Importantly, changes to group memberships are reflected instantly across the entire Globus system.

As a SaaS offering, it is critical that users are able to manage their use of the system in a self-service manner without intervention by an administrator. We provide users a direct interface for creating and managing collections. When defining a collection, the user selects policies that customize the collection as they desire. One policy is how persistent identifiers will be created. We allow users to select between issuing handles and digital object identifiers (DOIs). Users may provide credentials allowing Globus Data Publication to issue identifiers on their behalf or they may have identifiers issued using Globus' credentials with these systems.



Other policies govern workflows, curation and metadata. Users select what steps will be performed during the submission and curation workflows as well as what user group (managed by Globus Nexus) will be responsible for curation for the collection. The metadata used to describe items within the collection can also be selected from among a pre-defined set of schemas matching common publication scenarios such as the Datacite terminology [6].

The final class of policies governs storage. Users define the storage system where datasets will be stored. They also determine whether published datasets will be visible to only members of the collection, other groups of users or to the general public. Finally, users may request that the metadata values associated with each published dataset also be stored with the data. This insures users full openness since both the data and the metadata are present on storage systems they control.

Our remote data storage model is built upon Globus' data management capabilities, enabling reliable high-speed transfer of large datasets when depositing and downloading submissions, as well as supporting flexible data access. We leverage these capabilities to allow users to assemble datasets for publication, storing only references to dataset contents in the cloud and therefore enabling direct data transfer between the source data location and the designated collection storage. This model also supports authenticated access to published data and direct transfer to a desired location. Globus' support for in-place sharing of data directly from existing

storage repositories underlies the federated and distributed storage model. Through this approach users who wish to leverage remote data storage can simply create a shared Globus endpoint hosted on their storage resources. Globus Data Publication is then able to control access to this data remotely. Collection owners and data submitters can manage access to published data so that, for example, after submission access permissions on the shared endpoint are shared (read-able) with curators, and at the conclusion of the curation process access permissions on the shared endpoint are set to share with groups that have permission to view data in that collection.

A further benefit of the SaaS nature of our repository is the consolidation of metadata values and search capability. All collections have their metadata stored in a single index allowing searches to cross the collections and the organizational or discipline barriers that often separate them. Thus, users wishing to discover relevant research data are much more likely to find their results quickly and reliably with a single search rather than searching multiple, siloed repositories as is necessary when each institution operates a separate repository.

Conclusion

Globus' data publication capabilities provide a SaaS solution that is available to researchers worldwide irrespective of their domain. The model focuses on flexibility and self-service, allowing researchers to rapidly create and manage their own publication collection customized to their particular requirements. The unique "bring-your-own" storage model allows arbitrarily large data sizes while maintaining complete control of data and metadata on collection owners' infrastructure. Other capabilities enable complete control of submission and curation workflows, persistent identifiers, metadata schema and input forms. Globus Data Publication acts as a mediator, enforcing collection policies and providing a rich global discovery model across collections for controlled and public discovery of published research data.

References

[1] Dataverse. http://thedata.org. Web site. Accessed: January 1, 2015.
[2] figshare. http://figshare.com. Web site. Accessed: January 1, 2015.
[3] PURR: Purdue University Research Repository. http://purr.purdue.edu. Web site. Accessed: January 1, 2015.
[4] ScholarSphere. https://scholarsphere.psu.edu/. Web site. Accessed: January 1, 2015.
[5] DSpace. http://dspace.org. Web site. Accessed: January 1, 2015.
[6] DataCite. https://www.datacite.org. Web site. Accessed: January 1, 2015.