Enhanced Research Data Management and Publication with Globus

Presenters

Steve Tuecke - *University of Chicago*
Vas Vasiliadis - *University of Chicago*

Tutorial Length

3 hours

Attendees

20-40

Abstract

The management and publication of large research datasets presents challenges for researchers, librarians, and campus data center providers alike. While existing campus data providers typically have large storage allocations they are limited in their reach and utility due, in part, to unreliable tools and a wide variety of storage systems with sub-optimal user interfaces. Further, they do not yet support the high-level workflows and submission and access policies required for data publication. In this session, participants will learn how to use the data management and publication capabilities provided by Globus to publish large research datasets.

Unlike other approaches, Globus provides research data management and publication capabilities as a hosted service, available to the broader research community. Globus can therefore be leveraged by researchers across the globe to seamlessly manage, transfer, synchronize, share and publish datasets without needing to manage the low-level technical challenges associated with such activities independent of data location, size or heterogeneity.

Attendees will be introduced to Globus and have the opportunity for hands-on interaction with these systems. We will describe how Globus services for public cloud storage integration and data publication can be used by institutions to meet an increasingly common need by campus libraries.

Tutorial Goals

During the past year, we have seen many institutions deploying Globus as a core component of a campus data service. As these deployments scale beyond initial proofs of concept in the research computing center, we are fielding a growing number of requests related to support for data publication capabilities directly from Globus-enabled storage. This tutorial will demonstrate how Globus' data management features combined with its newly released data publication capabilities can be leveraged to enable a robust and scalable campus data publication solution.

This tutorial focus not only on research data publication, it will more generally provide information about research data management. It will help participants answer these questions: What services can I offer to researchers for publishing and managing large datasets more efficiently? How can I integrate these services into existing campus computing infrastructure? What role can the public cloud play (and how does a service like Globus facilitate its integration)? How should such services be delivered to minimize the impact on my infrastructure? What issues should I expect to face (e.g. security) and how should I address them?

Target Audience

We intend to target two audiences: 1) librarians and researchers who want to manage and publish data, and 2) campus computing center providers who want to provide data management and publication services to their users. The material is appropriate for both management and technical staff; the second half of the tutorial will be somewhat more technical than the first, but anyone with basic knowledge of Linux should be able to follow the discussion and participate in the exercises.

Prerequisites and requirements

Attendees are required to bring laptops with a web browser. While attendees can follow the presentation without a computer, they will not be able to participate in the hands-on portions of the session. We also request that wired network access be available in the tutorial room to ensure the best possible user experience.

Tutorial Content

The tutorial will be split roughly equally between presentation, demonstration, and hands-on exercises so that participants can understand how and when to use the services and tools presented. Each exercise will build on prior material. Initially, participants will create an account for the Globus service, configure a data management endpoint on their laptops, and use it to move data between their laptop and a test endpoint. Participants will then experiment with Globus sharing and group management features that enable peer-to-peer data transfer and synchronization. Participants will experiment with Globus' data publication capabilities by publishing datasets on a shared tutorial collection, assembling a dataset composed of distributed data, and defining dataset metadata. Participants will then be guided through the curation process, enabling them to review and modify other participants' submissions. Finally participants will explore discovery capabilities to find and filter published datasets.

In the last part of the tutorial participants will learn how to create and configure a Globus endpoint on their campus server (using presenter-provided cloud servers), how to manage access for multiple users, and how to configure publication collections with policies, remote storage, metadata forms, and curation workflows. We will also provide time for group discussion on critical issues related to integration with campus identity systems such as InCommon, and techniques for troubleshooting common configuration problems.

In the past we have run similar tutorials that describe the Globus service for research data management and data publication at venues such as the annual Supercomputing conferences (e.g. https://www.globus.org/events/sc14/tutorial).

Agenda

- Research data publication challenges and emerging practices (20 min)
- Introduction to Globus and demonstration (10 min)
- *Demonstrations and Exercises*
    - *Account signup and configuration (10 min)*
    - *File sharing and group management (20 min)*
    - *Data publication and discovery (20 minutes)*
- Break (10 min)
- Administration and management of data publication collections overview (20 min)
- *Demonstrations and Exercises*
    - *Creating endpoints with Globus Connect Server (10 min)*
        - *Configuring Globus Connect Server (10 min)*
- *Configuring a data publication collection (20 min)*
- Wrap-up and Q&A (10 min)

Outcomes

We intend that participants will be able to:

- Create and manage data publication collections using the Globus data publication capabilities with remote storage, required metadata forms, customized submission and curation workflows, and configurable persistent identifier
- Publish and discover large research datasets including configurable metadata and utilizing local and remote storage
- Deploy services for data transfer, sharing, and publication on a high-performance computing system
- Analyze approaches for integrating SaaS data management and publication components into existing campus infrastructure.