Research Data Lifecycle Management as a Service

Ian Foster and the Globus Online team (www.globusonline.org)
Computation Institute
University of Chicago and Argonne National Laboratory

Introduction

Big increases in data generated within research laboratories and demands for more careful data management lead to increased pressure on investigators. Researchers need not data storage, but full-service data lifecycle management processes, encompassing data collection, storage, sharing, metadata, search, archiving, provenance, assignment of DOIs, security, etc. Establishing such processes would demand substantial time and resources that most researchers do not have, and cannot easily acquire.

We believe that the solution to this problem is not simply to define "best practices"—nor to provide researchers with software. Once defined, best practices must still be implemented. software still must be installed, operated, and maintained. Those implementation, installation, and operations steps are precisely where many investigators run into problems.

Instead, we should aim to outsource the entire lifecycle management process to a third party **Research Data Lifecycle Management service**. Ideally, this service will encompass discipline-specific practices and methods, so that the individual researcher can connect their lab and then have many of their problems taken care of—much as many outsource their email to Google today.

The Computation Institute is working to develop such a system. Our first attack on the problem, Globus Online, is operated as a hosted software-as-a-service (SaaS) product. Web 2.0 interfaces provide for convenient Web browser, REST, and command line interfaces. The ability to configure the details of personal, campus, and national resources makes interoperating between different resources straightforward. In its current instantiation, Globus Online provides just user profile and data movement services. Later this year, it will provide group management, data storage, and data sharing. We plan to build it out with progressively more functions, eventually encompassing the entire research data lifecycle.

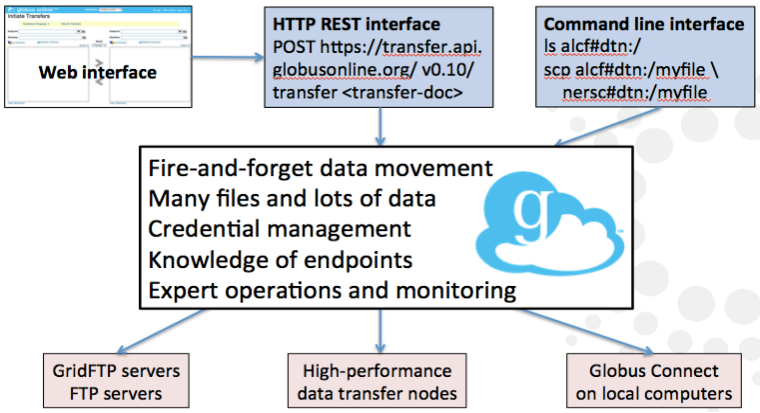
The Globus Online team is working to establish a campus deployment of what we intend to be a state-of-the-art Research Data Lifecycle Management service. We will use Globus Online to implement data management logic, both Amazon and local storage, campus credentials for authentication, and a set of UChicago and Argonne research laboratories (both small and large, and from a range of disciplines) to evaluate effectiveness.

We welcome input on what a Research Data Lifecycle Management service should look like, are looking for partners in this test deployment process.

Globus Online today

We started work on Globus Online in late 2009 with the goal of using SaaS capabilities to radically simplify research data transfer. The first publicly available Globus Online system, delivered in November 2010, implements methods for managing the transfer of single files, sets of files, and directories, as well as rsync-like directory synchronization. It manages security credentials, including for transfers across multiple security domains; selects transfer protocol parameters for high performance; monitors and retries transfers when

there are faults; and allows users to monitor status. REST, Web browser, and command line interfaces allow the casual user to initiate and monitor a transfer from a Web browser, while permitting the frequent user to integrate Globus Online calls into applications via command line scripting or REST messages.



The figure presents a user view of the system. Note the scp command used to illustrate the command line interface; this command has the same syntax as the commonly used but slow secure copy, but invokes high-performance, Globus Online-optimized GridFTP transfers that can move data 20 times faster or more that regular scp in many circumstances.

Response to Globus Online has been extremely positive from both individual users, who are delighted that previously time-consuming tasks are now automated, and from sites, who see user productivity and resource utilization increase, and support demands decrease.

Our planned campus deployment

We plan next to extend Globus Online capabilities to encompass a wider range of research data management tasks, and to make those extended capabilities within a campus setting that encompasses campus-supplied storage and computing facilities.

Our initial plan for this campus deployment envisions that Globus Online will be extended to allow not just data movement but data publication, discovery, analysis, and sharing. We envision a campus deployment enabling researchers to request that data be stored with certain quality of service guarantees (e.g., longevity), associate metadata with data, control how data is shared with colleagues, and request computation on data. We envision this extended service supporting data storage (and analysis) on both campus systems (e.g., archival storage systems operated by campus IT service organizations), specialized storage compute-storage clusters (sometimes termed, recently, “private clouds”), and commercial infrastructure as a service providers (e.g., Amazon Web Services).

We are developing requirements for this campus deployment, and plan to evaluate effectiveness, in partnership with a set of research partners. These partners include groups working in imaging, computational economics, geographic information systems, and large-scale simulation science.

Questions for workshop attendees

We are particularly interested in our fellow participants’ views on the following questions:

- What aspects of the research data management problem prove most time consuming for researchers today? How do you think the answer to this question will change over the next five years?

- What aspects of research data lifecycle management lend themselves to more general solutions, vs. requiring investigator/lab-specific customization?
- What are the concerns and trade-offs associated with hosting the data in the cloud (e.g., on Amazon) vs. on campus-provided storage?
- In order to campuses to adopt a SaaS solution for research data lifecycle management, what concerns must be addressed, technically and legally?
- What are investigators and campuses willing and able to pay for research data lifecycle management solutions, whether SaaS, locally deployed software, or custom solutions? Where are they likely to get this money?

Further reading (at www.globusonline.org)

Foster, I. Globus Online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Computing*(May/June):70-73, 2011.

Allen, B., Bresnahan, J., Childers, L., Foster, I., Kandaswamy, G., Kettimuthu, R., Kordas, J., Link, M., Martin, S., Pickett, K. and Tuecke, S. Globus Online: Radical Simplification of Data Movement via SaaS. Preprint CI-PP-05-0611, Computation Institute, 2011.