DURASPACE™

DuraCloud DTR ("Direct To Researcher")

Data Management in the Cloud for Researchers and Scientists

Summary

DuraSpace, with grant funding from the Alfred P. Sloan Foundation, is undertaking a project to provide enhanced cloud-based storage for research data. A software platform optimized for the needs of researchers and scientists will deliver durable storage and flexible management of data. The application, building on the DuraCloud project, will provide preservation services in a secure environment that safeguards privacy. In addition, researchers will be able to provide secure access to data curators and institutional data management staff to ensure that project data may be preserved for future use.

Background

The significant increase in the production and collection of scientific data has appropriately been referred to as a "data deluge." Researchers and scholars confronting this deluge and are faced with new data management challenges that have both social and technical dimensions. From the social perspective there are fundamental open questions pertaining to roles, responsibilities, and processes for responsible data stewardship and curation. From the technical perspective, there are currently many types of systems in use for data management including an array of non-standard systems, grid-based storage networks and, to a lesser degree, institutional repositories such as DSpace, Fedora, and enterprise solutions from various vendors--not to mention idiosyncratic "data under the desk" approaches, where researchers and scientists go it alone by storing and managing valuable data in their personal computing environment with commodity computers and local storage devices.

Currently, there is notable interest in cloud-based storage services for research data. While concerns exist (e.g., trust, security, location of data, guarantees against data loss), there is increasing interest in taking advantage of the scalability and low cost of utility cloud providers.

DuraSpace is an established 501(c)(3) not-for-profit organization committed to providing leadership and innovation in the development of open technologies that promote durable, persistent access to digital data for the research, scientific, and higher education communities. We collaborate with our communities in creating practical solutions to help ensure that current and future generations have access to our collective digital heritage. The organization is the home of Fedora and DSpace, two well-established open source digital repository solutions. In addition to repositories, DuraSpace develops new innovative technologies, most recently *DuraCloud*, a cloud-based platform providing digital preservation and access services over a network of cloud providers.

We aspire to make DuraCloud a compelling alternative to utility cloud services that are intended to serve a generic user. With a grant from the Alfred P. Sloan Foundation, we plan to build DuraCloud DTR, a cloud platform that provides a secure, user-friendly, and reliable way to store and manage research data, as well as a suite of services to enable preservation, long-term archiving, and specialized forms of data access. In developing the Direct-To-Researcher variation of DuraCloud, the first step is to ensure that the researcher can be in control and that data is stored safely and securely. Once data is under safe cover, DuraCloud DTR will enable a researcher to delegate responsibility for data curation, preservation, and archiving to specialists within supporting organizations such as research libraries, data management organizations, or university technology departments.

Existing cloud solutions targeted at end users, such as DropBox, allow an individual with a user account to store copies of local digital content in a single cloud provider. In contrast, the DuraCloud service will provide many additional benefits:

- store data in a community-driven cloud service run by a not-for-profit organization
- permit easy, intuitive access to project data
- provide authentication and access control; ensure privacy
- enable delegation of data curation tasks to other users after data has been deposited
- replicate data to *multiple* cloud providers using one unified user console
- perform data integrity checking at every stage of the deposit and ingest process
- run preservation-oriented services that monitor and "health check" data, over time
- run specialized compute services upon the data after it has been stored

In our collaborations with researchers and scientists we have repeatedly heard concerns about the time and expertise required to effectively manage data. Interruption of the research process is a very significant concern. From a user perspective, our goal is to offer technologies ensuring that essential capabilities for data management and preservation fit naturally into research and scientific workflows and to add significant value to these processes on behalf of researchers.

Consistent with our prior work, we are committed to deploying DuraCloud as open source software to provide transparency and the ability for communities to participate in the software development process of a new cloud solution. Unlike most proprietary cloud offerings, it is part of our *mission* as a not-for-profit organization to ensure that our technology is transparent, standards-based, and sustainable.

The Approach

In consultation with an advisory council, we plan to engage a set of researchers and data specialists through multiple cycles of design, development, and testing - with the goal of evolving the DuraCloud platform with researcher sensibilities at the forefront. Throughout the course of the project, there are several key questions that we will examine as we build DuraCloud DTR in response to researcher reactions and insights:

- *How can we empower the researcher to have autonomy in using DuraCloud, and also enable an institutional "back door" for curators, data mangers, and archivists?*
- *How can we demonstrate the added value to researchers that will motivate a sustained partnership with data curation specialists?*
- *How should we support researcher and institutional requirements in the areas of security, data privacy, and other legal aspects of archiving?*

- *What are the most practical economics of a cloud solution for managing research data?*

During the project, we will run two invitational workshops to validate target issues, define scenarios, develop use cases, and identify technical requirements. We will address a number of “hot issues” that will be relevant to the use of cloud technologies for research data, specifically (1) user identity, (2) data security, and (3) privacy. In addition we will attempt to reconcile issues around how to make it easy and natural for individual researchers to store and access their data, but also ensure that stored data is amenable to curation and long-term management.

DuraCloud DTR may be viewed as an end-user service that will not require immediate mediation by a library or data management institution. Longer term, however, these supporting institutions play an important role in the overall process of managing and preserving data. Thus, in developing new features, we anticipate a mechanism by which researchers can delegate authority to other specialists to perform data curation and grooming tasks upon the data to improve its ability to be managed. Moreover, we believe that the most effective curation outcomes will result from an early partnership between researchers and data managers.

Our work will be structured so that we can iterate through several pilot cycles, each involving software development, system testing, and evaluation. We will continuously capture new requirements for the software and engage with researchers and data specialists to validate and test them.

Finally, we plan to focus on economic models that permit practical, widespread implementation of these practices.

Conclusion

As providers of software technologies, we measure our broader impact in terms of the contributions we make to the greater cause of enabling scientists, researchers, and data curators in managing, accessing, and analyzing data for the purpose of bringing forth new insights. In the short run, our direct impact can be measured by how widely our DuraCloud DTR service is adopted, the quality and integrity of the software, the health of our community-driven processes, and the overall quality of the user experience (i.e., natural fit into the workflows of users). In the longer run, our impact will be measured by the availability of the high quality data that has been successfully curated and archived. Long-term impact for data management systems can be achieved with sustainable software, flexible systems, open standards, and community-driven processes. As part of its not-for-profit mission, the DuraSpace organization is explicitly committed to these practices and will make them key components of the DuraCloud DTR project.