DURASPACE™

P.O. Box 7144
Ithaca, NY 14851-7144
(607) 216-4548 • http://www.duraspace.org

Josh Greenberg
Alfred P. Sloan Foundation
630 Fifth Avenue, Suite 2550
New York, NY, 10111

March 9, 2011

Dear Josh,

Thank you very much for your initial interest in our grant proposal ideas for the DuraSpace not-for-profit organization. As part of our mission, we are interested in providing new technologies and services to assist researchers and scientists in confronting the actual realities of the metaphorical "data deluge." The increased production and collection of scientific data presents research communities with significant data management challenges with both social and technical implications. In our work at DuraSpace, we have observed that researchers, scientists, and data curators have begun to respond to these challenges by identifying and evaluating new approaches to managing their research data.

With this contextual backdrop, we are writing this letter of inquiry to explore the prospect of the Sloan Foundation funding a new pilot program focused on the use of our new *DuraCloud* platform for research data. Currently, there is notable interest in cloud-based storage services for research data. However, researchers and data managers need direct evidence that cloud services can serve as alternatives to institutional systems, elaborate data grid solutions, servers in their laboratories, or personal disk drives in their offices. In developing DuraCloud, we understood that researchers and scientists require an easy, reliable way to store, manage, monitor, and preserve their data. Furthermore, they need a way to easily share data and run computational services upon the data in an easy-to-use infrastructure.

We envision the potential of a "direct-to-researcher" cloud platform that provides researchers an easy and reliable way to store and manage their data. A first step is ensuring that data is stored safely and securely. From there, data specialists within institutions such as libraries, research laboratories, and universities can benefit from the same platform to facilitate robust data curation activities. Finally, entire research communities can benefit from a platform that can open up avenues for data access and sharing. Ultimately, we envision a cloud-based data platform that can enable all of these things, as well as supporting a "commons" where users can contribute and run services for data mining and cross-disciplinary analysis.

We propose a two-phase project to investigate the potential of DuraCloud for these purposes. Phase 1 will be the planning stage of the project. During this stage, we will conduct interviews with candidate participants from within data-intensive organizations that DuraSpace is currently affiliated (e.g., Data Conservancy, University of Virginia, MIT, NCAR). Next, we propose strategic workshops to define goals and plan for the DuraCloud pilot program. Outputs of the workshops will include reports on user and system requirements, key questions that must be answered during the

pilots, and metrics for evaluating DuraCloud as a solution. During Phase 1 we will also develop a set of detailed use cases and a project plan for running pilots. We estimate Phase 1 to have a duration of approximately six months and a cost of approximately $150,000.

Once planning activities are complete, we propose to launch Phase 2 to actually run the pilot program. During this hands-on phase we will work with selected partners to load data into to the DuraCloud platform and test core services. We will continuously capture new requirements for the software and engage with software developers to evolve the platform. Each pilot partner will focus on a set of use cases that are designed to test the viability of DuraCloud in these areas: (1) utility and performance of data loading and integrity checking, (2) usability and fit within researcher and scientific workflows, (3) ease of data discovery and access, and (4) suitability for the integration of data analysis services. We estimate Phase 2 to have a duration of 18 months at a cost of $400,000 with a breakdown of approximately $150,000 to run the pilots and $250,000 towards work on technical architecture and software development. We envision three six-month cycles that will result in major updates to the DuraCloud open source software.

At a time when both institutional systems and national cyberinfrastructure are in early stages of emergence, DuraCloud is positioned to provide an easy, practical solution for today's needs. Our goal is to demonstrate and evaluate DuraCloud as a cloud-based data management solution offering both preservation and access services. While offering immediate benefit on its own, DuraCloud can also be a building block for other emerging solutions since it is architected to naturally integrate with other systems.

DuraCloud is currently available as open source software (beta) and will be launched as a hosted cloud service in Q3 2011. Our interests align with researchers, scholars, librarians, and data curators who depend upon open technologies for ensuring durable access to their digital data. Our status as a not-for-profit organization is significant since it positions us as a trusted partner with key communities working to meet the challenges of data management.

We look forward to new opportunities and the prospect of making contributions in the important area of data management. We would be delighted to engage with the Sloan Foundation in pursuing this new work.

Sincerely,

Sandra Payette

Digitally signed by Sandy Payette
DN: cn=Sandy Payette, o=DuraSpace, ou, email=spayette@duraspace.org, c=US
Date: 2011.03.09 16:08:10 -05'00'

Sandy Payette
Executive Advisor

*signing on behalf of:*

| | | |
|---|---|---|
| Michele Kimpton | Brad McLean | Jonathan Markow |
| Chief Executive Officer | Chief Technology Officer | Chief Strategy Officer |