DuraCloud "Direct To Researcher"
Easy Data Management in the Cloud for Researchers and Scientists

1. What is the main issue, problem or subject and why is it important?

The significant increase in the production and collection of scientific data is appropriately referred to as a "data deluge" [1]. 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. A recent survey of interdisciplinary science leaders found that over 50% archive their data in their laboratory and 38% on university servers. Respondents expressed concern that even within a single institution, there are no standards for storing data, resulting in ad hoc approaches and variability across departments and individuals [2]. Unfortunately, there are also 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. 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. Our recent discussions with scientists have revealed an interest in looking towards utility cloud-based storage solutions with the

hopes of (1) attaining more autonomy by subscribing to a cloud-based service, (2) mitigating complexities of having to negotiate with intermediaries such as libraries and other data management units, and (3) gaining relief from the burdens of backup and disaster recovery when data is stored in local systems. This is consistent with an overall trend towards cloud services for obtaining benefits such as easily provisioned storage services and web-based service points for upload and access. While concerns exist (e.g., trust, control of data location, guarantees against data loss) and are well documented [30], there is increasing interest in taking advantage of the scalability and low cost of utility cloud providers [3]. Research commissioned by DuraSpace found that technology decision makers concerns about risk were mitigated by the prospect of trusted organizations providing oversight to data being stored with cloud providers.

We write this proposal from our position as an organization whose mission is aligned with the needs of research and academic communities. 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 research, scientific, and higher education communities [4]. 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 [5] and DSpace [6], two well-established open source digital repository solutions. In addition to repositories, DuraSpace is developing new innovative technologies, most recently *DuraCloud*, a cloud-based platform providing digital preservation and access services over a network of cloud providers [7].

We aspire to make DuraCloud a compelling alternative to utility cloud services that are intended to serve a generic user. Consistent with our mission to serve the research community, we envision a "Direct-To-Researcher" cloud platform that provides an easy 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 can enable a researcher to delegate responsibility for data curation, preservation, and archiving to specialists within supporting institutions such as research libraries, data management organizations, or university technology departments.

Existing cloud solutions targeted at end users, such as DropBox,[1] simply allow an individual with a user account to store copies of local digital content in a single cloud provider [8]. In contrast, the DuraCloud service can provide the following additional benefits and capabilities:

1. store data with ease in a cloud service run by a not-for-profit organization
2. replicate data to *multiple* cloud providers using one unified user console
3. perform data integrity checking at every stage of the deposit and ingest process
4. enable delegation of data curation tasks to other users after data has been deposited
5. run preservation-oriented services that monitor and "health check" data, over time
6. permit easy access to the data, at the discretion of researcher
7. facilitate authentication and access control to the data

[1] DropBox is a Web-based service targeted at non-technical users that makes it easy to store local files in the DropBox cloud. Local files are continuously synchronized with cloud storage. Files can be accessed from the user's computer, smart phone, or directly from the DropBox website.

8. run specialized compute services upon the data after it has been stored
9. provide transparency in underlying system design to permit complete auditing

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 that ensure that essential capabilities for data management and preservation fit naturally into research and scientific workflows. Essentially, such capabilities should be seamless to the end user. Taking inspiration from Jonathan Ive, Senior VP of Design at Apple, a design is done well when "…it feels almost inevitable, almost un-designed and it feels, almost, like of course it is that way. Why would it be any other way?" [9]

2. What is the major related work in this field?

There are several well-known cloud providers offering storage and compute services. Notable players include Amazon [10,11] and RackSpace [12] who both offer general-purpose "utility clouds" accessible over the Web via APIs. In contrast, DropBox presents an appealing paradigm for supporting individual users with an easily installed application that synchronizes local files on a desktop computer with the DropBox storage cloud. While all of these can provide an effective storage solution, they are generic services with limitations in their ability to address the particular needs of researchers and substantive issues related to research data. An exception may be Microsoft's Azure [13] cloud platform since initiatives of Microsoft Research and a partnership with NSF have resulted in a focus on researchers. However, the primary focus in Azure appears to be

around data sharing and cloud computing for algorithmic data analysis. It is important to note that none of these cloud vendors provide a mechanism to easily transfer data to another cloud vendor, as there are no standard interfaces between providers.

Our own prior work in this area is exemplified in the beta version of DuraCloud and the completion of two phases of pilot testing with funding from the Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP) [14]. Using agile and iterative process, we developed the core DuraCloud technology and tested with fourteen NDIIPP pilot partners. Each partner ingested a minimum of 2TB and maximum of 10TB of digital content and tested DuraCloud replication, resulting in multiple copies of each file across different cloud providers. The DuraCloud integrity checking services were invoked at every transfer point in the ingest process. Also, the partners extensively tested synchronization of content from Fedora and DSpace repositories. DuraCloud was designed with a service plug-in architecture to enable easy integration of new capabilities in the future. Pilot partners tested an initial set of services, including file format transformation, and advanced image viewing, and video streaming.

There are other projects related to research data that are targeted at institutions, and less at the individual researcher. The NSF-funded Data Conservancy [15] is building a data curation and archiving system to contribute to emerging cyberinfrastructure [16, 17]. In the open source repository community, a number of Fedora repository projects are emerging to support data management. Notable is Islandora [18], a Drupal and Fedora open source stack that has generated interest in government organizations including NASA/Goddard and the Smithsonian for "virtual research environments." In the data grid

community, the iRODS open source software [19] has been deployed by San Diego Super Computer Center in the context of the Chronopolis service to support data archiving [20].

3. Why is the proposer(s) qualified to address the issue or subject for which funds are being sought?

The DuraSpace organization has a long track record working with universities in the areas of digital preservation, archiving, digital libraries, and open source repositories. We have active partnerships with the NSF Data Conservancy and research institutions that are tackling the challenges of digital preservation of all types of content, including data sets. We are the developers and stewards of widely-deployed open source repository technologies to serve global communities of research institutions, universities, libraries, and others. Our expertise positions us to confidently assert that data management is more than just storage and backup.

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 many proprietary cloud offerings, it is part of our *mission* as a not-for-profit organization to ensure that our technology is transparent and sustainable. This is especially relevant to data management and digital preservation, which present the paradoxical challenge of sustaining both the data and the software that manages it.

DuraCloud offers cloud storage with preservation-oriented services including basic replication across multiple cloud storage providers, data integrity checking, format conversion, and rich media viewing. Mediating multiple cloud providers is a strategy to mitigate risks and overcome obstacles of storing data at any one provider. It also avoids a

single point of failure for data storage and frees the researcher from data lock-in to a single storage provider [21]. Currently, there are DuraCloud connectors to three commercial storage clouds: Amazon, Rackspace, and Microsoft's Azure. There is also a DuraCloud connector to the Chronopolis grid-based service hosted by the San Diego Supercomputer Center. We anticipate future work with DuraCloud connecting to university storage clouds such as the Open Cloud Consortium [22], led by the University of Chicago.

The DuraSpace leadership team brings years of experience working with academic, library, and scholarly/scientific communities. Together we have a strong record of demonstrated results in research, open source software, and entrepreneurship which positions them well to achieve the goals of this proposal. DuraSpace CEO, Michele Kimpton, brings extensive background in technology-oriented not-for-profit organizations, with entrepreneurial and business experience. In 2010, the Library of Congress recognized Kimpton as a "Digital Preservation Pioneer." Previously she was Chief Business Officer of DuraSpace, Executive Director of the DSpace Foundation, and founder of Archive-It, a subscription service of the Internet Archive for preserving digital content. DuraSpace Strategic Advisor, Sandy Payette, brings vision and leadership with a research background in digital preservation and technical architectures for digital content management. Previously, Payette was CEO DuraSpace, Executive Director of Fedora Commons, Researcher in Information Science at Cornell University, and the creator of the original Flexible Extensible Digital Object Repository Architecture (Fedora). DuraSpace Chief Technology Officer, Brad McLean brings extensive prior leadership experience in open source software and technology startups related to information management. Previously McLean was Technical Director of the DSpace Foundation and Chief Systems Architect at

Constant Contact, Inc. DuraSpace Community Strategy Officer, Jonathan Markow, is the newest member of the DuraSpace leadership team and brings valuable expertise in building open source communities for Higher Education. Previously Markow was Executive Director of JASIG, a not-for-profit organization focused on open source software, and also a director of information technology initiatives at Columbia University.

4. What is the approach being taken?

Our prior experience with user-centered design, agile development, and collaborative open source process has been the hallmark of our prior work in deploying successful open source technologies. By continuing with this approach in the Sloan project, 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. Short turn-around times in our open source software process enables rapid response to user feedback. Throughout the course of the Sloan project, there are several key questions that we will examine as we evolve DuraCloud in response researcher reactions and insights:

a. *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?* We will examine the means of "opening the back the door" of a researcher's DuraCloud space to enable curation and archiving of data. We will work with researchers, curators, and data managers to define requirements for DuraCloud features that enable the delegation of responsibility by the researcher to supporting specialists. We anticipate that an authorized person can be permitted by the researcher to work

on the data, either in situ or by transferring data to another location (another DuraCloud space or an external location). If it is desirable to have curatorial and archival work done directly on the data in its original DuraCloud location (i.e., the researcher's space) then the researcher should be empowered to delegate by initiating a state change and permission event indicating that the data is ready for further processing.

b. *How should we support institutions in their need for security, data privacy, and other legal aspects of archiving?* Researchers and institutions must trust DuraCloud as a service provider. We want our technologies to support, not impede, institutions in fulfilling their legal and administrative responsibilities. While we do not plan to build overly complex security features into DuraCloud, we will address security, encryption, and access control *appropriately* to be a good partner to researchers and institutions with regulatory concerns.

c. *What are the economics of a cloud solution for research data? What business models are viable while being compelling to both individual researchers and university subscribers?* We will examine several possible business models to see which ones resonate with best with researchers, while also contributing revenue to run and support the DuraCloud platform.

4.1 Enlisting Pilot Partners from the Research Community

The DuraSpace organization has a well-developed community network that affords the opportunity to engage with researchers, scientists, librarians, and data managers on the range of issues related to both short term and long term data management. As a partner in

the Data Conservancy, we are already working with Johns Hopkins University and a team of multi-disciplinary researchers on building infrastructure to address a range of issues along the spectrum of preservation and access. Researchers and scientists are eager to have data storage solutions that do not disrupt their research workflows. They also want easy access data and the ability to integrate formerly disparate data sets to unleash new insights.

Several DuraSpace "Gold Sponsor" institutions, especially the University of Virginia and MIT, are highly motivated to advance the integration of DuraCloud with Fedora or DSpace repositories to manage research data. Also, existing DuraCloud pilot partners such as ICPSR and Columbia University have expressed a high interest in moving forward with their DuraCloud instances to explore new approaches for managing research data.

With these pre-established connections, we will be able to jump-start our work in the Sloan project. Since the project will be stronger if we attract a diverse set of participants beyond only large universities, we plan to build out our researcher network to include smaller institutions, government institutions, and research centers with a goal of 50% of the participants from universities and 50% from other settings.

4.2 Engaging Researchers on Data Management Scenarios

During the project, we will run two invitational workshops to define scenarios, validate use cases, and identify technical requirements. We hypothesize that there are 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) data privacy.

Of particular interest in the Sloan project are scenarios where researchers are directly collecting data "in the field." The DuraSpace organization has had opportunities to learn about these data collection practices from our collaborators, especially the

Biodiversity Heritage Library, the Cornell Lab of Ornithology, the Smithsonian, the National Center for Atmospheric Research (NCAR), Cornell University Library, and the Woods Hole Marine Biology Laboratory. The basic idea in these cases is that data is collected by the researcher in many forms such as photographs, observations recorded in lab notebooks, and measurements from devices in the field. These types of data are found in a variety of digital formats including digital images, spreadsheets, text files with delimiters (e.g. comma separated), text files with markup (e.g., XML), and many other formats. These types of data sets are often referred to as "small data" and are more closely integrated into the researcher's direct workflow. Researchers will often keep this type of data on their own personal computers or commodity storage disks. We see significant opportunities for DuraCloud to assist researchers in easily getting this type of data under safe cover, and subsequently into a more robust data curation and preservation workflow.

Another opportunity area involves scenarios that begin with automated front-end data collection systems that store large quantities of "raw" data. These systems suit their purpose at early stages of data lifecycles and enable researchers to deploy workflows for tasks such as analysis, transformation and summarization of raw data. However, these same systems are not necessarily designed to support long-term preservation and archiving. Raw data that has been down-sampled, summarized, cleaned, or normalized are often referred to as "higher level" data products. These types of data are of great interest to researchers for re-use, sharing, and publication. They have also been identified as *objects of interest* for data curation, preservation, and archiving.

One interesting aspect of our proposed Sloan work will be reconciling 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. In this context, we hypothesize several key themes are in play when the researcher is empowered:

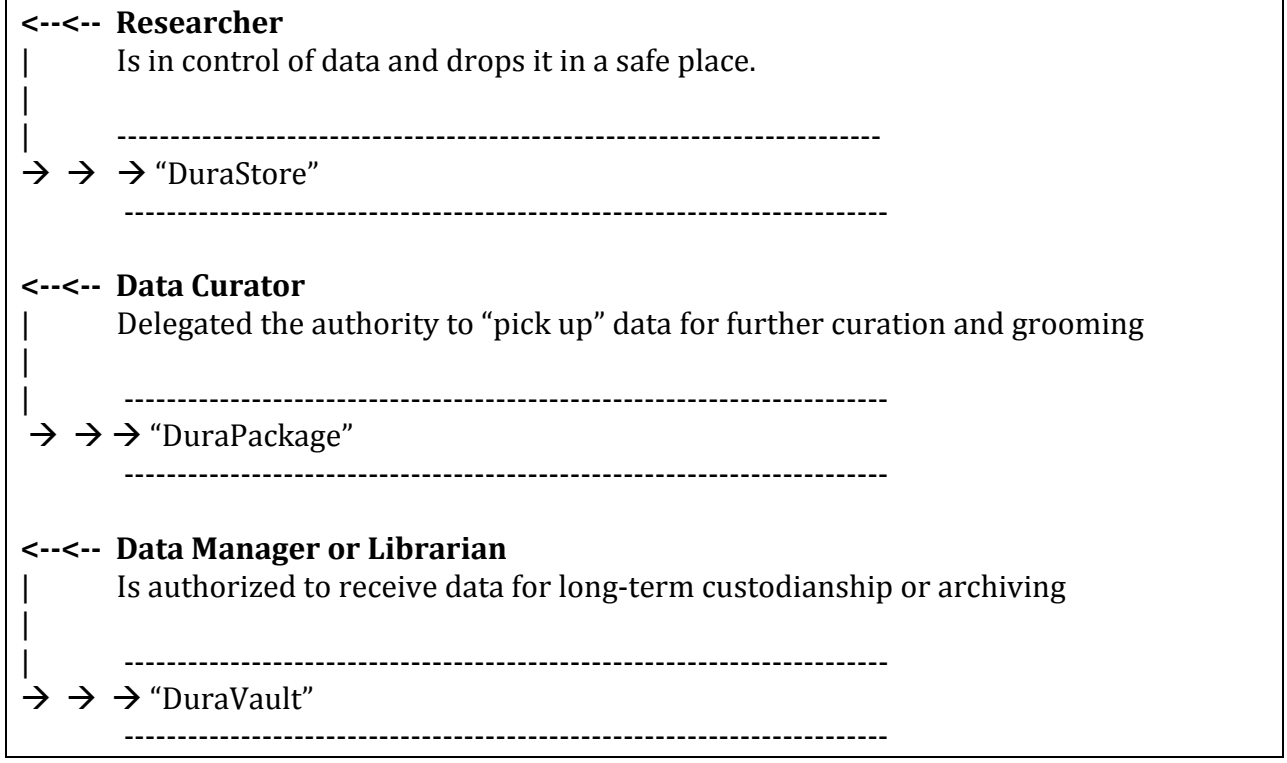
1. Researcher has control over his or her data
2. Researcher has responsibility to deposit data
3. Researcher has a choice of when data is available to others
4. Researcher is owner of credentials to store, manage, and access data
5. Researcher can provide credentials to trusted party to access the data
6. Researcher can delegate responsibility to trusted party to make data fit for long-term management (e.g., curate, groom, and augment with metadata)
7. Researcher can authorize an institution to take long-term custodianship of the data
8. Researcher always has access to the data

4.3 Evolving DuraCloud Software to Support Researcher Requirements

.

DuraCloud Direct-To-Researcher will be conceived of as an end-user service that will not require mediation by a library or data management institution. However, these supporting institutions play an important role in the overall process of managing and preserving research 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. Furthermore, we envision a mechanism by which the researcher can authorize a custodial institution to take responsibility for the data for long-term management or archiving. A notional view of this concept is presented in Figure 1 below.

Figure 1:



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 to evolve the platform. This approach was highly effective in developing the initial versions of the DuraCloud platform. We were able to quickly identify user reactions, translate them into new requirements, then design and implement the new functionality quickly for use in the next iteration.

Pilot Cycle 1 – major theme is data under safe cover

- Recruit – enlist diverse group of pilot testers
- Workshop #1 - with ***researchers***

- Technical work – features in response to Workshop #1
- Test - with ***researchers***
- Wrap – report results and snapshot a new version of the DuraCloud software

Pilot Cycle 2 – major theme is data curation

- Workshop #2 - with researchers + data curation specialists
- Technical work – features in response to Cycle 1 results and Workshop #2.
- Test - with researchers + data curation specialists
- Wrap - report results and snapshot a new version of DuraCloud software

Pilot Cycle 3 – major theme is full data lifecycle (also with tuning, and evaluation)

- Technical work - features in response to Cycle 2
- Test - with researchers + data curation specialists
- Measure - formal scalability and performance measurements
- Evaluate – overall user feedback on DuraCloud for "Direct To Researcher"
- Technical work - final enhancements
- Wrap – report results and snapshot a new version of DuraCloud software

4.5 Developing the Economic Model

An important part of the Sloan project will be to work with researchers and supporting institutions to test out several business models for the DuraCloud Direct-To-Researcher offering. A brief description of several possible models follows below.

- *Individual Subscription:* In this model, an individual signs up for an account with a credit card, and is charged a monthly fee for access to the service, which includes some base level of storage and the ability to purchase more storage as needed. In order for this model to be successful, the pricing should be in line with what is currently available in the marketplace. If the subscription fee is higher than perceived competition, the additional value-add of DuraCloud must be compelling for the researcher to pay a premium. For this model to be economically feasible, we must attract a large number of individual users to bring in enough revenue to offset the costs of running the service. There would need to be an investment upfront to

market the service and gain awareness of individual researchers.

- *Institutional Subscription:* In this model an institution such as a university or library procures a block of accounts and subscribes on an annual basis. Account blocks could be anywhere from 50 to thousands of accounts. The institution would be responsible for the central administration of the accounts, although we anticipate individual researchers would be able to sign up freely from their desktop, once an institutional subscription was established. This model is very attractive to DuraSpace and we are already implementing it with the initial DuraCloud launch.
- *"Write-in" for NSF Data Management Plans:* NSF has required all new grant proposals to contain a data management plan. To date, it is unclear how NSF will enforce this mandate, or help their grantees fulfill this requirement. DuraSpace sees an opportunity to establish DuraCloud as a approved service for meeting the requirements of the data management plan.
- *Corporate Sponsorship:* As the DuraCloud business grows, we anticipate being able to secure corporate sponsorships from our cloud vendors and other businesses who want to provide additional services to our users.
- *Aggregated volume pricing discounts:* For the DuraCloud platform we have set up master accounts at all the cloud providers we are using as part of the service. As a result, through aggregating the volume across all user accounts, DuraSpace gets a reduced storage and compute costs from the underlying providers based on volume aggregations. We plan on passing some of these reduced costs on to the user, but also plan to keep some portion within the DuraSpace organization to contribute to the continual development of the platform.

5. What will be the output from the project?

During the Sloan project, we will work with researchers and other stakeholders to define scenarios, use cases, and technical requirements, and evolve the DuraCloud technical platform to support the Direct-To-Researcher orientation. The outputs of the project are:

1. Reports from two invitational workshops with researchers and data curators
2. Documentation of scenarios and detailed use cases that were the focal point of the Direct-To-Researcher project. This will also include a summary of researcher challenges and opportunities in the area of managing research data.
3. Technical specifications for new development on DuraCloud to support researcher workflows and research data. Initially this will be recorded in the DuraCloud feature tracking system, and ultimately new features/capabilities will be written into the official DuraCloud technical documentation.
4. Implementation of new features, modules, and services for the DuraCloud open source software, coinciding with the each of the three software development cycles in the project. This is consistent with our overall process of agile development.
5. White paper on the DuraCloud Direct-To-Researcher value proposition, focusing on the presenting problem, the benefit of DuraCloud to the researcher, and the economics of the solution.
6. Conference presentations to report results. Relevant conferences include the International Conference on Digital Preservation (iPRES), the International Digital Curation Conference (DCC), the IEEE E-Science Conference, the International

Association of Social Science Information Services and Technology (IASSIST), and domain-specific conferences related to data management.

7. Public release of a new version of the DuraCloud open source software containing the successful features, modules, and services that were developed during the Direct-to-Researcher project.
8. Final report on the DuraCloud "Direct To Researcher" project
9. Launch DuraCloud Direct to Researcher service with appropriate business model

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 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 high quality data that has been successfully curated and archived. Long-term impact for data management systems can be achieved with sustainable software, evolvable systems, open standards, and community-driven processes. As part of its not-for-profit mission, the DuraSpace organization is explicitly committed to these practices. We collaborate with and serve a diverse international user base engaged in research and education, with the shared goal of making important digital information available to future generations.

6. What is the justification for the amount of money requested?

We are requesting $497,433 to investigate and demonstrate the potential of the DuraCloud "Direct To Researcher" proposition. We estimate the project to have duration of 18 months. The Sloan budget contains DuraSpace salaries and benefits to partially fund employees participating in the project. Specifically, we have budgeted 25% FTE for a DuraSpace executive team member to provide overall leadership of the project and to engage strategically with researchers. Related to technology, the budget contains 15% FTE for system architecture and management of the development process. To support open source software development, we budgeted 50% FTE of a DuraCloud senior programmer to focus on evolving the core DuraCloud platform. Additionally, the budget will cover the expense of a software contractor to focus on web development and feature enhancements.

The budget contains travel expenses for DuraSpace employees to meet with researchers and pilot testers at key stages in the project. Also, we have budgeted travel and logistical expenses to host two invitational workshops with researchers and data curators. The expense of a professional facilitator for the workshops is also included. At later stages in the project, we also expect to run virtual meetings using a Webinar platform as part of our agile development process. To pay for infrastructure to conduct the "Direct To Researcher" pilots, we are budgeting the expense of utility cloud storage and compute services. This will cover the cost of cloud storage for data that pilot partners will upload to DuraCloud. Also, it will cover the cost of running compute-intensive services within the DuraCloud platform.

7. What other sources of support does proposer have or applied for to support project?

The DuraSpace organization has current grant funding from the Gordon and Betty Moore Foundation, the National Science Foundation (Data Conservancy), and the Library of Congress (NDIIPP). DuraSpace is also supported by institutional sponsors including universities, research organizations, and libraries. The initial development of DuraCloud has been funded in part by these sources and by directing percentages of DuraSpace employee time towards developing the DuraCloud software and running the NDIIPP DuraCloud pilot program. At this time, we have not applied for other new grants, aside from this Sloan grant, to support the DuraCloud Direct-To-Researcher project.

In terms of the future of DuraCloud, we plan to continue to support the open source software with service revenues paid to the DuraSpace not-for-profit organization. Starting in Q3 2011, we will launch an initial version of the DuraCloud hosted service targeted at institutional users. We are forecasting subscription revenue as part of DuraSpace's 2011 income and we already have the majority of our NDIIPP pilot partners signed up as initial subscribers. In terms of sustaining the output of the Sloan project, we anticipate a future release of the DuraCloud hosted service that will be targeted at researchers and that will yield additional subscription revenue with critical mass of individual DuraCloud users.