Annual Report 2009
(January 2009 – December 2009)

By Sandy Payette, Chief Executive Officer
On Behalf of the Executive Team

Fedora Commons (DuraSpace) Annual Report 2009
By Sandy Payette, CEO, on behalf of the Executive Team

1 Executive Summary

The Executive Team of DuraSpace, also known as the "Triad," is comprised of the Chief Executive Officer (Sandy Payette), the Chief Business Officer (Michele Kimpton), and the Chief Technology Officer (Brad McLean). Together, the Triad has been crafting the strategic agenda for the newly constituted DuraSpace organization, bringing forth the assets and experiences from both Fedora Commons and the DSpace Foundation. Their focus has been on strategies to better position our open source repository platforms, DSpace and Fedora, while also expanding the DuraSpace technology portfolio to respond to the Web and cloud computing.

Our technology portfolio inherently addresses the issue of *durability* of digital content. We consider *durability* to be a necessary pre-requisite to the process of digital preservation. We are especially interested in providing technologies and services that ensure that digital content *accessible*. Thus durability translates to long-term or perpetual *access* to digital content. We have adopted a new organizational byline, "open technologies for durable digital content."

The Triad believes that the DuraSpace organization is well positioned to play a significant role in providing open source technologies for the academic and scientific sectors, particularly by continuing to serve universities, libraries, and research institutions. The cultural sector is showing an increased interest in open source for digital asset management, with notable interest by public broadcasting, small archives, museums, and NGOs. We also see increased opportunity in the government sector, particularly national archives, libraries, and research-oriented agencies.

Currently, there are nearly 1000 installations of the Fedora and DSpace repository solutions worldwide. The new DuraCloud service is now at the Alpha stage and it supports cloud-based storage and replication, with services supporting preservation and access to data. Pilot partners funded by the Library of Congress NDIIPP program are currently testing the DuraCloud service, and 15 new pilot partners will be brought on in 2010.

The DuraSpace organization works with the Board of Directors to developer strategies to capitalize on the strengths inherent in its technology portfolio and to develop viable business models for the long-term sustainability of the organization. The organization must also be very clear about its strengths and its points of vulnerability, ensuring that we position strategically for the future and avoid potential pitfalls. A key focus of our work with the Board in the upcoming year (2010) will be on refining our strategic positioning, implementing business models, and securing revenue sources to fulfill our mission and achieve our financial goals for the five year forecast and beyond.

2 Highlights in 2009

There are many notable highlights during the startup period of DuraSpace:

- The joining of Fedora Commons with the DSpace Foundation became operational on July 1, 2009.
    - Fedora Commons, Inc as taken the name DuraSpace (registered assumed name in NYS)
    - DSpace Foundation dissolved
    - DSpace copyright was transferred to Fedora Commons
    - All DSpace employees were hired by Fedora Commons as of 7/1/09
- New organization website unveiled to public: http://duraspace.org/
- Strong team effectively transitioned into new roles: http://duraspace.org/people.php
- Open Repositories 2009: Major public announcement of DuraSpace with very positive response from repository communities on the joint mission and the prospects of the organization. Excellent attendance at conference and user groups.
    - See: http://smartech.gatech.edu/dspace/handle/1853/28076
- Paving the way for the new organization, Fedora Commons, DSpace Foundation, and Sun Microsystems collaborated on a successful new Webinar series, "All About Repositories." The series ran from February through June 2009 and was very well attended, on average over 100 attendees per session. We will continue to use Webinars as an ongoing outreach vehicle.
- Library of Congress provides grant to DuraSpace to kick off DuraCloud Pilot Partners Program
    - See: http://www.digitalpreservation.gov/partners/duracloud/duracloud.html
- DSpace uptake continues to grow at an impressive rate. The worldwide user community is robust and the new DSpace Ambassador's Program is being launched as part of the global community outreach strategy. The DSpace registry is now approaching 900 installations worldwide:
    - See: http://www.dspace.org/index.php/DSpace-Instances/Repository-List.html
- Fedora has been collaborating with key community members to produce Fedora-based solutions for e-research and e-scholarship. Notable is the Islandora solution (UPEI) that provides an integration of Fedora with the Drupal web content management system. Sun Microsystems entered the collaboration and is introducing a new product based on Islandora/Fedora and Sun Open Storage (Sun Open Archive for Virtual Research Environments and BioScience)
    - See: http://www.sun.com/storage/archive/oasolutions.jsp#4
- The DuraCloud initial architecture is complete and an alpha version was released to pilot partners this Fall 2009. We have engaged in strategic partnership discussions with four commercial cloud providers (Amazon, EMC, Rackspace, and Microsoft) and have built a storage interface that connects to four commercial cloud providers (Amazon, EMC, Rackspace) with Microsoft Azure forthcoming. We have reallocated significant staff to the DuraCloud project to fast track the development and launch. We have actively engaged 3 pilot partners with funding from the Library of Congress NDIIPP program. These three partners are the Biodiversity Heritage Library (BHL), WGBH (public broadcasting archives), and the New York Public Library.

We have successfully loaded 10TB of data for each partner and are currently testing the core DuraCloud services – storage, replication, integrity checking, image transformation, JPEG2000 image viewing, video steaming, and local-to-cloud synchronization. The initial pilots projects will be complete in June 1010, after which time we plan to transition to a beta pilot phase with 15 additional pilot partners.

- The Mulgara semantic store has increased its performance more this year than any time in the past 5 years. Tests reveal an increase query performance by over 100 times in some cases. The majority of these improvements came directly out of work done addressing the needs of the Public Library of Science (PLoS), which is the major production user of Mulgara.

2 DuraSpace Organization Identity and Mission

At the heart of the organizational identity question is our not-for-profit mission. While we believe that the prior mission statements declared in the Fedora Commons and DSpace Foundation articles of incorporation are on target, they are notably sweeping in their breadth of purpose, covering every part of the information lifecycle from the creation of digital content through preservation, access, sharing, and knowledge linking. We have created a new mission statement that conveys the essence of our purpose:

> *DuraSpace is committed to providing leadership and innovation in the development and deployment of open technologies that promote durable, persistent access to digital data. We collaborate with academic, scientific, cultural, and technology communities in creating practical solutions to help ensure that current and future generations have digital access to our collective heritage.*

The DuraSpace organization has staked out its position as providers of "*open technologies for durable digital content*." We use this as a byline on our website and marketing materials as a means to convey the essence of the DuraSpace identity. There are different ways that "open technologies" can be interpreted, for example open source, open core, open interfaces, open standards, open access, open content. As we introduce new products, such as DuraCloud, we will ensure that openness is part of our offering, while selecting the appropriate open orientation for the product and its target market. For example, as we enter the terrain of hosted solutions and the cloud, there are some decisions to be made and boundaries to be drawn to ensure that we live by our principles of open technologies, but also provide solutions with viable business models. DuraCloud is emerging as a hybrid solution with both open source and proprietary components as appropriate. While the core components of the DuraCloud service will be open source, DuraCloud also has connectors to commercial cloud providers, and some parts of the DuraCloud hosted web application will not be open source. We believe that this is an example of something that is strategically sound both in terms of value to our communities and in moving towards a businesses model for the DuraSpace organization. Nevertheless, the Triad and the Board will continue discussions of the best strategies to be true to our mission and while achieving our financial goals.

3 Technology Portfolio 2009

Below, is a summary overview of the DuraSpace technology portfolio with highlights of the key features and strategic implications for each technology. Detailed reports of each technology are provided later in the report.

DSpace

DSpace is an out-of-the-box open source repository application for delivering digital content to end-users. There are over 750 digital repositories using DSpace software. Globally it is the most widely used open source repository software for institutional repositories and open access repositories. DSpace has been installed all over the world by organizations, especially libraries, as a way to provide access to research output, scholarly publications, library collections, and more. The DSpace application has many features and tools for managing digital content and enabling digital preservation. DSpace stores any type of content and offers built-in workflows for content submission and review. Organizations can easily make their digital collections available on the Web using DSpace's customizable end user interfaces along with many community-developed features and utilities.

Fedora

Fedora is a robust, modular repository system for the management and dissemination of digital content. It is especially suited for digital libraries and archives, both for access and preservation. It is also used to provide specialized access to very large and complex digital collections of historic and cultural materials as well as scientific data. Fedora's flexibility enables it to integrate gracefully with many types of enterprise and web-based systems, offering scalability (e.g., millions of objects) and durability (e.g., all of the information is maintained in files with no software dependency, from which the complete repository can be rebuilt at any time). It also provides the ability to express rich sets of relationships among digital resources and to query the repository using the semantic web's SPARQL query language. Fedora has a worldwide installed user base that includes academic and cultural heritage organizations, universities, research institutions, university libraries, national libraries, and government agencies.

DuraCloud

DuraCloud is a hosted service and open technology that enables organizations and end users to use cloud services. It is a cloud-based service that leverages existing cloud infrastructure to enable durability and access to digital content. The service is particularly focused on meeting requirements for academic, scientific, and cultural heritage. DuraCloud leaves the basics of pure storage to those who do it best (storage providers) and overlays storage solutions with additional functionality that is essential to ensuring long-term access and ease of use. DuraCloud offers cloud storage, plus replication of content across multiple providers. Once digital content is stored in the cloud, compute services are the key to unlocking its value. DuraCloud provides services that enable digital preservation, data access, transformation, and data sharing. DuraCloud offers customers an elastic capacity with a "pay as you go"

approach. It is appropriate for individuals, single institutions, or for multiple organizations that want to make use of cross-institutional infrastructure. DuraCloud has been in a pilot phase during Fall 2009. Plans are to release DuraClound as a service hosted by the DuraSpace not-for-profit organization following completion of the pilot phase.

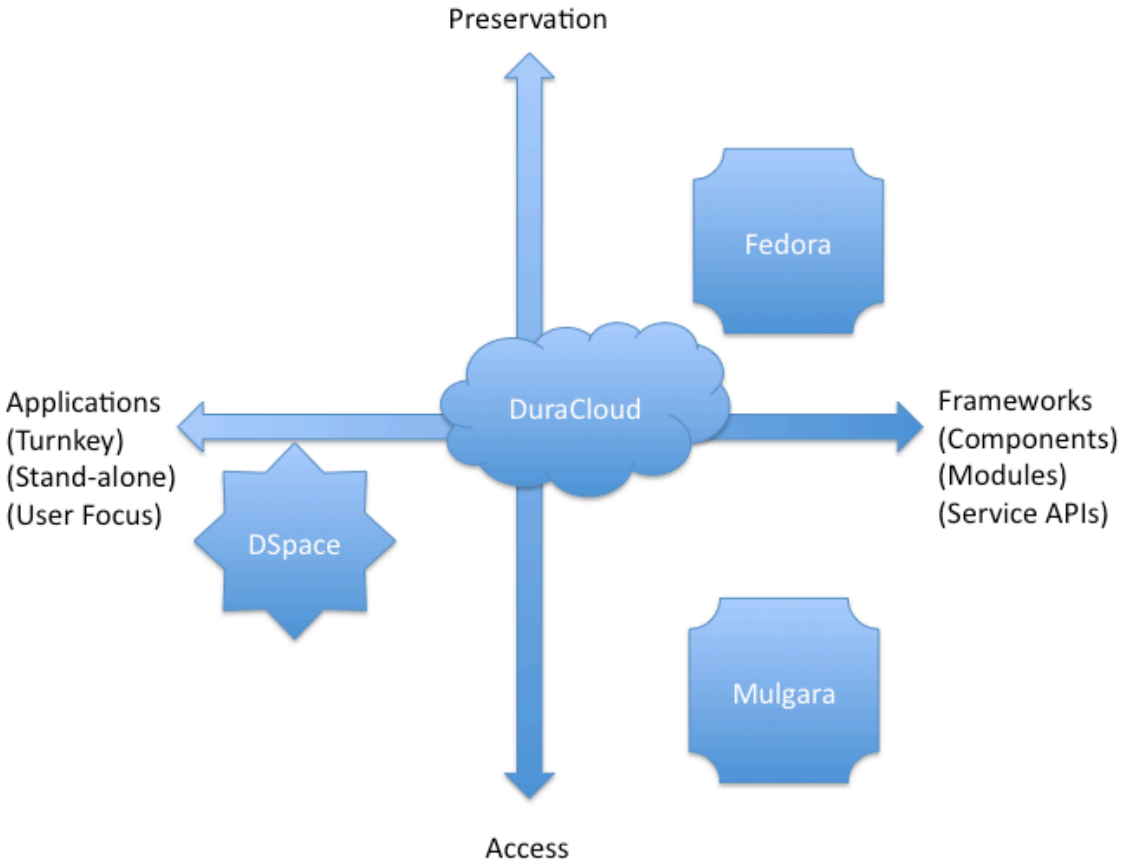
Mulgara

Mulgara is a database for storing and retrieving data that conforms to the Resource Description Framework (RDF). RDF is associated with future trajectories of the World Wide Web such as the Semantic Web and the Linked Data Web. Mulgara is an important component for knowledge-intensive systems that enables the interconnection of information entities in accordance with ontologies that describe such entities, their properties, and their relationships. It also provides powerful query capabilities using the SPARQL query language, the official RDF query language of the World Wide Web Consortium (W3C). Mulgara has many applications in institutions engaged in research, scientific publishing, digital libraries, data-intensive science, and humanities computing. The DuraSpace organization deploys the Fedora repository system with Mulgara. This integration of Fedora and Mulgara provides a semantically-enabled repository solution with the ability to interconnect "digital objects" and query the repository using SPARQL.

4 Product Positioning

The Triad has been analyzing the DuraSpace product portfolio to identify dimensions on which the products vary. We provide a preliminary analysis that shows how our current and emerging products fit across two notable dimensions: (1) product use case (dominant orientation towards preservation vs. access) and (2) product deployment style (turnkey solutions vs. modules or services to be embedded in other systems). Figure 1 depicts the relative positioning of the current products in the DuraSpace portfolio.

Figure 1: Product Positioning



The matrix depicts a "realistic" positioning for each product, as opposed an "idealized" positioning of each product. Particularly with DSpace and Fedora, we have seen some "idealized" positioning of these products over the past several years, resulting in the blurring of the key distinctions of these products. Sometimes this has been healthy as each community has been motivated to extend the reach of its product. However, at a time when technologies are evolving quickly and "grand integrated systems" are pushing up against the Web and more distributed approaches to systems, it is worth our pausing to think about the appropriate scope and role of our repository solutions. Moving forward we must think about how to best position DSpace and Fedora in the changing technology landscape where both cloud computing and "cyberinfrastructure" are emerging alongside existing institutional and enterprise solutions.

Our initial thinking on the "realistic" positioning for DSpace and Fedora is that there exists a "sweet spot" for each brand. For example, DSpace was originally conceived with a heavy branding for digital preservation, yet it appears that the dominant use case is more around a turnkey solution for providing open *access* for research output. Meanwhile, Fedora was originally conceived as an system offering a unique way to provide flexible access to content with its "disseminator" architecture; however, it has had more traction as a modular service for management of digital assets offering both scalability (i.e., millions of objects) and durability (i.e., rebuild-from-files). The point is that we believe the time is right

to more clearly position these two products to decrease brand confusion and *more clearly target specific markets* to enable higher impact.

In thinking about the positioning of new and emerging products and services, the Triad believes the new DuraSpace organization has the responsibility to be innovative and forward-looking. Our goal is that new products fit naturally into the emerging Web infrastructure and that they naturally enable cross-institutional collaboration. DuraCloud is our first new product initiative that directly addresses these motivations.

We are positioning DuraCloud as a trusted service, hosted by a not-for-profit organization that leverages existing cloud providers to create a general-purpose cloud infrastructure for "durable digital content." The business model will be predicated on a value proposition of DuraCloud as an easy, low-barrier-to-entry cloud solution, offering services that enable digital preservation while breaking down barriers to content access and sharing. DuraCloud will offer customers an elastic capacity with a "pay as you go" approach, and the potential for cross-institutional infrastructure. Here are the key assumptions that underpin the value proposition:

- Replicating digital content across multiple cloud providers can mitigate risk of data loss.
- Monitoring and auditing via DuraCloud can help ensure durability.
- Making digital content available for discovery and access is highly desirable for ensuring longevity. Content that is "dark" or hidden away in offline storage runs the risk of lapsing into obscurity.
- Services are key to unlocking the value of digital content. DuraCloud can provide an easy-entry cloud platform for running a set of services, provided by DuraCloud, or by our technology partners. We have also entertained a future prospect of enabling an "application exchange" and plug-in environment in a manner similar to iPhone apps and the "AppExchange" of Salesforce.com.

We are currently in discussions with cloud providers to scope out the potential of "preferred partnerships" that would be valuable to our business model by virtue of the partner offering us free or lower cost services (e.g., storage and compute), revenue-sharing, technical support, connections to new customers, joint marketing, or other benefits. In 2009, we completed initial work on the DuraCloud business plan, including market research, product positioning, and pricing models.

In Q1-Q2 2010, the Triad will work out the details of the DuraCloud business plan, with key inputs from the Thanos Partners whom we have engaged for DuraCloud business consulting and market research. The final business plan will be presented to the Board of Directors for review, input, and approval.

5 Markets and Competition

The DSpace and Fedora brands are well rooted, particularly in university libraries and are also gaining traction in other types of organization, especially research institutes, academic departments, and IT organizations. The Fedora and DSpace brands are very closely associated with the notion of digital repositories, which is very well understood in the library and scholarly communities. However, repositories cannot be deemed a universally understood construct. Other types of systems have similar and overlapping feature sets, for example systems marketed under classifications such as enterprise content management, digital asset management, web content management, and digital archives. There are many commercial and open source products in these areas that represent our competition.

Thus, we must be aware that the DSpace and Fedora brands may not resonate clearly outside of our existing communities. If we want to try for expanding the DSpace and Fedora market outside of the library and university community, we must figure out how to position alongside other similar technologies, or find partnerships that naturally bring us into other sectors. The Sun Microsystems partnership with the Islandora/Fedora/OpenStorage integration is an example of how we can connect with other sectors, as the intent is to market the product to the BioSciences and government organizations.

The market opportunities for DuraCloud are more expansive. We are still scoping out the competition, but we believe we are well positioned as a value-add player upon commercial cloud providers. Thus, we see many potential partnership opportunities with commercial cloud providers, rather than direct competition. Currently, DuraSpace is uniquely positioned as a not-for-profit offering a cloud mediation/brokering solution (especially appealing to our existing communities). We currently see the following markets in play:

- **Public Institutions of Knowledge:** interestingly, our first two pilot partners are not with universities but with public institutions. NYPL and BHL (Smithsonian and Missouri Botanical Gardens) have shown particularly high enthusiasm and interest in innovation with the cloud.

- **University Libraries:** University Libraries have shown notable interest in DuraCloud as a replication solution, and are especially intrigued with the integration of DSpace and Fedora repositories with DuraCloud. The scenario presents libraries with a solution that creates a hybrid architecture including institutional systems and the cloud.

- **University IT**: The opportunity is ripe to target DuraCloud at University IT organizations. We have given several presentations to University CIOs, which have generated interest in the DuraCloud proposition. There are also early discussions about a possible pilot with Microsoft, DuraCloud, and University CIO organizations. We currently have engaged the Thanos partners to do market research by interviewing key University CIOs on their needs and reactions to DuraCloud. Preliminary data from Thanos indicates that University CIOs report being out

capacity and are looking for secondary storage solutions for low risk content. Also, CIOs are validating the need for brokers to cloud solutions (like DuraCloud).

- **Pubic Broadcasting:** We have been exploring a DuraCloud pilot with WGBH and have gleaned from our discussions with them that the public broadcasting community, in general, is interested in open source and may be a market opportunity for DuraCloud. We will further explore this market area.

- **Not-for-Profit Organizations:** We believe DuraCloud will be attractive to a variety of small institutions and not-for-profit organizations that are constrained both in IT infrastructure and staff. Particularly for institutions that are aware of the need to address digital preservation in some form, DuraCloud can provide an "instant infrastructure" at low cost, and services that they would not have resources to develop themselves. We will further explore this market area.

- **Government:** In terms of our new DuraCloud product, we are looking to both serve our existing markets and to expand our market into new sectors. There are news reports about the Obama administration directing funds for government agencies using the cloud. We have already had inquiries from government organizations, including NASA's Goddard Space Center, suggesting that there may be opportunities in markets that we have not yet explored.

6 Future Revenues and Funding Sources

In terms of revenue-generating potential around the DuraSpace products, the Triad believes our best strategy is to pursuing a multi-pronged approach. Table 2 provides a summary of the currently projected revenue sources and how they relate to the DuraSpace product portfolio.

Table 2: Planned Revenue Sources

Revenue Source	**Assumed in 5 year forecast?**	**Products**	**Explanation**
Hosted Services	yes	DuraCloud	Unlike our open source software distributions, this strategy is built around a customer paying for a subscription to a service that we host. DuraCloud will be first up. We are currently working on the DuraCloud product packaging and pricing models.
Sponsorship Program	yes	All	Community-focused donation program. Benefits are provided for different levels of

			sponsorship.
Partners Programs	yes	DSpace Fedora DuraCloud	Service Provider Partner Program
Grants	yes	All	To date, the bulk of our funding has been from grants. We continue to secure new grants, but we are aware that these funds will be tied to particular projects and partnerships, with particular goals. In the long-term we cannot depend solely on grants to sustain the organization.
Consulting	no	?	We have experimented with consulting for revenue, but have observed that it involves heavy focus and is not scalable with our current employees who have multiple responsibilities. We believe this is still something to consider in future, but we'd have to staff up for it, thus it has a chicken and egg for starting it up (upfront money to staff it).
Service Contracts	no	?	We are not currently planning on the DuraSpace organization offering support contracts on the DSpace and Fedora software. As we have noted previously, we are skeptical of the willingness of our current communities to pay. Also, the DuraSpace organization is not staffed properly to run a support organization. Also there are active debates by well-known analysts and business strategist on service contracts for o/s software as a long-term business model.
Venture Capital	no	?	This is an option that we have not discussed in any detail. This would likely present a conflict of interest to our not-for-profit orientation, however, since return on investment is key driver with VCs.
Build Endowment	no	?	We should discuss this more seriously. Is it possible to encourage a group of universities to make a substantial investment in DuraSpace so that we could build a modest endowment? Can we endow

			one or two of the key leadership positions in the organization? Other ideas?

In 2009, we designed and developed the Community Sponsorship Program and the Service Providers Partner Program. Both will be launched in 2010.

Sponsorship Program (launch May 2010)

This is an annual sponsorship program modeled closely after the program launched by the former DSpace Foundation. The target audience is universities, libraries, and other organizations that benefit from the work of DuraSpace. This includes the current users of DSpace and Fedora (approximately 1000 institutions worldwide). The sponsorship funds will be used to support the not-for-profit and continue its work in the development of open source repository platforms, cloud services, community education, outreach programs, and infrastructure to run the organization. The funds will be undirected and will be at three levels of annual giving: Gold ($10,000), Silver ($5000) and Bronze ($2500). We will provide a range of benefits to the sponsor at these different levels.

Service Provider Partner Program: (launch Q2 2010)

Service providers are eligible for the partner program if they are currently providing commercial services to the community using our open source software. Services can range from development, hosting, consulting, training, customization, and service contracts. Service providers in the partner program have an annual agreement with DuraSpace to provide a specific level of monetary and in kind contribution to the not-for-profit organization in return for benefits of being a partner. The level of in-kind or monetary contribution can be determined on a case-by-case basis to match the fiscal and personnel strengths of each organization. Depending on the level of contribution, benefits to the service provider can include:

- Recognition on the DuraSpace website as service provider partner
- Higher status on the DuraSpace web site
- Monthly group calls with non profit organization to update on strategy, programs, new releases
- Letters of support from non-profit on grant proposals, extension developments when applicable
- Direct communication channel into non profit for consultation on technical strategy where it directly impacts programs and business at service provider

7 Technology Strategy

Open Source Repositories (DSpace and Fedora)

From a technology strategy perspective, there is the reasonable question as to what is the useful lifespan the DSpace and Fedora technologies in their current manifestations. Prior to the joining of the two organizations, Fedora Commons and DSpace Foundation were on paths towards more modern technical architectures that would begin to blur the distinctions between DSpace and Fedora. DSpace kicked off the DSpace 2 project in a move towards a next generation architecture that provided more flexibility and modularity. Fedora was beginning to plan for a more lightweight plug-in architecture, and a DSpace-like front end. The Triad now believes there is an opportunity to take a two pronged approach to the DSpace and Fedora question: (1) influence the developer communities to converge on a common underlying modular architecture and move towards a libraries of core shared modules, and (2) unabashedly re-focus on the "essence" of each brand (i.e., the Fedora brand means… and DSpace brand means…) and re-direct our brand marketing towards promoting each for what it does best.

The Triad believes we should evolve the current DSpace and Fedora technologies to be consistent with emerging technologies and new paradigms for managing and publishing digital content. This can be a tricky proposition since it involves the not-for-profit organization providing leadership to two established open source developer communities. The Triad recognizes that "path dependence" is a force to be reckoned with in the DSpace and Fedora technologies and existing communities. What this means is that decisions that were made in the past (e.g., data models, designs, use cases), tend to have undue influence on the ability to make significant change to the technology in the present – even if new requirements and competing technologies are emerging.

Nevertheless, there is hope and expectation in these communities that the new DuraSpace organization can provide a context for DSpace and Fedora to gracefully adapt to new technological and social forces, both at the institutional level and the Web at large. The Triad believes that the right direction for the Fedora and DSpace technologies is to move in the direction of a shared vision of modularization, essentially deconstructing the core parts of each system into fundamental components that can be shared by both. Ideally, the two projects would begin to converge technically, but diverge in terms of the meaning of each brand. Ideally, DSpace would continue to provide turnkey offering focused closer to the end user workflows and experience. Fedora would continue to offer core modules focused lower down in the architecture (mediation and management functions upon storage systems). The tour de force could be Fedora core modules bundled inside the DSpace solutions distributions.

Many DSpace installations, and most of the Fedora installations involve *integrations* with other web applications and resources. For example, Fedora core modules appear in 3rd party solutions such as Islandora and Hydra. There is continuous interest in positioning our digital repository systems alongside digital content applications, asset management systems and archives. As part of our

technology strategy, we will continue to foster integrations that will both strengthen our position in our existing sector, and open possibilities in the adjacent sectors.

New and Emerging Technologies (DuraCloud)

Moving on to the opportunities in new and emerging technologies, we anticipate that a significant opportunity lies ahead with the new DuraCloud project. DuraCloud provides a clean slate technologically, freedom from path dependence, and an opportunity fresh positioning in emerging cloud infrastructure. A key part of achieving our overall technology strategy is allocating our limited software development resources to new and strategic work. Accordingly, we have re-directed the effort of two of the developers from the Fedora project to work on DuraCloud starting in Q3 2009. While this move demonstrates our decision to invest in new and emerging technologies, we are committed to ensuring the health and progress of Fedora and DSpace by providing a Lead Developer for each of the repository projects and by providing the context to enable ongoing community participation in these projects.

With a grant from the Library of Congress NDIIPP program we have launched the DuraCloud Pilot Partners Program to test the DuraCloud technologies and demonstrate key functionality. Our initial pilot partners are the New York Public Library, the Biodiversity Heritage Library, and WGBH. We are currently engaging with several others to participate as technology partners to provide interesting cloud compute services within the DuraCloud platform. An expanded pilot program will commence in Q2 2010. Successful pilot phases will lead to a public launch of the DuraCloud service, which include the following capabilities:

- Ability to replicate content to multiple cloud providers through a single web interface (multiple copies, multiple geographic areas, multiple administrations)
- Data integrity checking and monitoring across multiple cloud providers
- Ability to get, put, move content across multiple cloud providers through the DuraCloud web interfaces
- Ability to connect DuraCloud to digital repositories through extensions to the DSpace and Fedora software.
- Ability to run a set of compute services on top of content in the cloud through the DuraCloud web interface (such as data transformations, data mining, indexing, media services, and hosting)

8 Open Source Projects – 2009 Accomplishments

Fedora

In May 2009, Fedora 3.2 was released and in December 2009 Fedora 3.3 was released. These two releases provide many significant advanced to the Fedora repository platform including:

- **Web Client:** Fedora Web Administrator provides a new web-based user interface to perform repository administration. All coordination with the Fedora repository is performed using the newly completed Fedora web-based REST API. The Fedora Web Administrator can be used to discover, display, ingest, modify, and purge Fedora objects and datastreams.

- **Knowledge Networks:** Improvements to Fedora object model to enable relationships among inner components of digital objects, permitting creation of enhanced knowledge networks.

- **Web Interoperability:** SWORD 1.3 compatibility was fully tested, providing the ability to deposit content into Fedora using SWORD's specialization of the popular Atom Publishing Protocol.

- **Improved Security Architecture:** Initial integration of the results of a community-driven project that re-engineered of the Fedora repository security architecture (authentication and authorization) - the "Fedora Enhanced Security Layer (FeSL).

- **Storage Abstraction:** Integration of Akubra for improved low level storage abstraction and better plug-ability of commercial and community-developed storage modules.

- **Performance:** Rigorous performance testing to 100+ million objects with Sun Microsystems

- **Modernization and Quality Control:** instituted modern software builds using Maven; launched a strategic re-factoring project to move Fedora into a modern modular framework using the Spring Framework. Ongoing quality improvements with 55 bug fixes.

- **Fedora Create Forum:** The Fedora Create Community exists to coordinate and integrate the community of developers working with the Fedora software. There is a Registry that provides information on the diverse software being created by this community to work with Fedora. See: http://fedora-commons.org/confluence/display/FEDORACREATE/Create+Forum

Fedora 3.3 marks a new milestone in the process of developing the Fedora open source software. For the first time, the Fedora community came together under the leadership of a Community Release Manager who facilitated the software development process and the integration of community contributions. The core Fedora open source committer team now has 11 members with the majority of the developers (9 people) being community members, with 3 developers from DuraSpace. This development team demographic, plus the fact that the Fedora project is now on its second community-

managed release, indicates that Fedora has reach a stage of healthy maturity as an open source project and is well positioned for sustainability.

Thornton Staples, Director of the Fedora Project from DuraSpace, observed, "The process of developing open-source software with a community-based process requires dedicated effort by many community participants, both developers and users. Though there is a perception that open-source software gets written by hobbyist programmers working on their own late at night, it is more commonly written by programmers working for institutions that are committed to the software and understand its shared benefits."

Fedora Users:

1. **Number of registered Fedora installations:** approximately 200
2. **Interesting Fedora users in 2009**
   - The Islandora project (Drupal on Fedora) has been gaining traction and there are many new "Virtual Research Environments" using this integrated solution.
   - Hydra Project is making significant advances in providing flexible platform for building applications on top of Fedora. The project is gaining interest in the Fedora community.
   - Data Conservancy, a NSF Datanet project, is using Fedora as a core component in its initial prototype for a data archive system.
   - VMware is developing a prototype system that is using a Fedora repository to register virtual appliances and record the provenance of machine images

DSpace

In 2009 DSpace 1.5.2 was released providing many bug fixes. The big work in 2009 was in the development of DSpace 1.6, which was completed during the year, with final testing and release in Q1 2010. DSpace 1.6 provides many significant advancess to the DSpace repository application including:

- **Web Interoperability:** SWORD 1.3 compatibility was fully tested, providing the ability to deposit content into DSpace using SWORD's specialization of the popular Atom Publishing Protocol.

- **Language translation:** new language translations for the 1.5.x platform.

- **Statistics Reporting –** new statistics package with enhanced statistical reports about repository usage. Out-the-box simple statistical views are available for each item, collection, and community in the user interface. Information is given about item views, bitstream downloads, and user metadata such as the location the users of the repository came from.

- **Embargo functionality -** embargo functionality supports the need to manage content that cannot be made public for a certain period of time. For example some open access journal articles may be under a 6, 9 or 12-month embargos. DSpace 1.6 provides out-the-box rules that allowing an embargo lift date to be set during the submission of the item. The bitstreams (but not item metadata) are locked from public view until that date has passed. DSpace 1.6 also makes it possible for users to write their own embargo rules.

- **Improved Repository Management:** bulk metadata editing and distributed administration
- **Improved Authentication:** new authentication methods using hierarchical LDAP and Shibboleth
- **Modernization and Quality Control:** Upgraded the DSpace software to work with Apache Cocoon 2.2; Ongoing quality improvements with numerous bug fixes.
- **DSpace Add-ons and Extensions**: DSpace users are continually adding new functionality to the DSpace platform to suit their organization's needs. Some developers are willing to share their work with the larger DSpace community.. See: (http://www.dspace.org/add-ons-and-extensions/addons/).

Also in 2009, the DSpace Committers Team began scheduling regular, weekly meetings, formally adopted Apache voting practices (http://www.apache.org/foundation/voting.html) and added four new members to its team. Along with other developers, the Committers Team began wide adoption of JIRA (http://jira.dspace.org) as the new issues and feature tracking system for DSpace software. They also migrated DSpace's code repository from SourceForge to the Oregon State University Open Source Lab (http://osuosl.org).

The DSpace 2.0 funded project came to a close, demonstrating a new service framework. This work is being considered for inclusion in forthcoming DSpace releases. DSpace 1.5.2 already included the Spring Framework and Cocoon upgrades resulting from this 2.0 project, and 1.6.0 included the initial release of the 2.0 service framework. In addition, DSpace took part in its third Google Summer of Code student internship program, sponsoring four successful student projects.

DSpace Users:

1. **Number of DSpace installations:** 831 (up from 537 in 2008) . New DSpace instances for the first time in the following 27 countries: Argentina, Azerbaijan, Bangladesh, Belarus, Botswana, Bulgaria, Cameroon, Cape Verde, Croatia, Cyprus, Ecuador, Equador, Ethiopia, Ghana, Iran, Jamaica, Kenya, Lithuania, Malawi, Malta, Mexico, Namibia, Pakistan, Qatar, Romania, Uganda, Uruguay
2. **Dynamic Database:** designed and began development on an expanded instance list to provide the community with more and better useful information about the community of DSpace users. The data from the above Community Networking Survey was used to seed database, provide for faceted and open search of database. The database will be launched in Spring 2010.
3. **Interesting DSpace users in 2009**
    - Harvard University, Countway Library Repository (http://repository.countway.harvard.edu)
    - DASH - Digital Access to Scholarship at Harvard (http://dash.harvard.edu/)
    - Encyclopedia of Iranian Architectural History (http://eiah.org:8084/arch/?locale=fa) - has many customizations, including Persian translation, menu orientation, calendar,
    - Hospitais da Universidade Coimbra (http://rihuc.huc.min-saude.pt/) - Hospital/Medical center in Portugal

DuraCloud

The purpose of DuraCloud is to provide trusted cloud mediation with different levels of service aimed at making digital content (1) durable, meaning it is accessible for long periods of time and (2) usable, meaning that it can be re-exposed or dynamically transformed to fit within in a variety of application and system contexts. DuraCloud provides a simple, open API with backend connectors to multiple cloud storage providers. Mediating multiple cloud providers is a strategy to hedge risks and overcome obstacles of storing data at any one provider, such as single point of failure for data storage and data lock-in. Currently, there are DuraCloud connectors to four commercial cloud services (Amazon, EMC Atmos, Rackspace Cloud), with a forthcoming connector to Microsoft's Azure. Currently, 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.

Market Research

In 2009, DuraSpace engaged Thanos Partners, Inc. to conduct market research to gain greater understanding of higher education institutions' needs in managing digital collections, and to solicit feedback on elements of the DuraCloud service offering. Nineteen phone interviews were conducted in September and October 2009 as part of this effort. One hundred forty-five participants from higher education institutions completed an electronic survey in January 2010. Some of the key findings in the market analysis were focused on the market perception of DuraCloud as a service. A portion of the executive summary from the final report stated the following:

There is a great deal of enthusiasm about DuraCloud services. One of the most compelling elements of the service is third-party management of the services by a knowledgeable, trusted partner. Participants look forward to additional detail about the services, and note the functionality they need is those preservation services most difficult to develop locally.

Trusting an external entity to manage institutions' critical assets is still of significant concern to the market. DuraSpace can best manage this by:

- *Selecting pilot institutions that are inspirational leaders of its target market*
- *Identifying and developing services institutions cannot offer locally, and*
- *Completing all the certifications, requirements and subsequent communications to be seen as a credible player in the market*

The results of the study showed that our community was most interested in using DuraCloud for preservation support, this is an area identified as highly important, but needs are not being met within the institution. An excerpt of this part of the analysis is below:

When evaluating the importance and extent to which preservation support needs are met, participants were asked to consider services such as content replication, auditing and repair. The following table details participant responses.

Category		Subcategory	Importance	Extent Needs Met	Difference
Non-US			3.14	2.15	0.99
US Institutions	Institution Size	Large, very large	3.57	2.38	1.19
		Medium	3.38	1.83	1.55
		Small, very small	3.14	2.07	1.07
	Enrollment Profile	RU/VH	3.63	2.29	1.34
		RU/H, DRU	3.25	1.50	1.75
		Master's S, M and L	3.18	2.30	0.88
		Bac and Assoc	3.14	1.86	1.29

With an overall mean rating of 3.35 (1=low importance, 4= high importance), preservation support is considered very important. It is also the service area that shows the greatest gap between its importance and how well institutions' needs are met. Only one participant rated his institution as meeting preservation support needs very well.

The most frequently cited reason for this gap is the lack of an overall preservation support strategy at institutions.

DuraCloud Technology:
In 2009, the following goals have been achieved in the development of the DuraCloud platform:

- Solidify cloud strategic partnerships
- Hold pilot partner meeting to review web app, workflow and large dataset movement
- DuraCloud Alpha service released to pilot partners
- Enlist beta testers for DSpace and Fedora repository plug-ins
- Demonstrate web-based interface to move content between storage providers

- Demonstrate first compute service over content via a web-based interface

DuraCloud v.1 is the first version of the DuraCloud software, currently being tested as part of the NDIIPP pilot partner program. The initial release allowed users to do the following:

1. Ingest content into a DuraCloud instance through a web browser and/or through a REST API.
2. View content within their DuraCloud instance through web based interface
3. Store their content in up to two Cloud commercial providers, Amazon and Rackspace
4. Download their content from DuraCloud through web browser or REST API
5. Replicate their content in up to two cloud providers upon ingest

DuraCloud v.2 included the following additional capabilities, also being tested by NDIIPP partners:

1. Ability to ingest content, either over the wire or via hard disk, into DuraCoud platform of up to 10 terabytes per user account.
2. Ability to ingest and view multiple types of files including audio, video, image, and text
3. Ability to add metadata and tag files and spaces in which files are stored
4. Ability to do bit integrity checking and auditing upon ingest, replication and downloading.
5. Ability to manage the following services through the DuraCloud web interface
    a. Replication across multiple cloud providers(up to three)
    b. Running of Djatoka image server for viewing JP2000 image files
    c. Running of Imagemagick file conversion service for file format transformation
6. Ability to synchronize your DuraCloud storage with another local file directory utilizing a client side utility
7. Added security layer to the application for login and authentication.

From a cloud services perspective, the core DuraCloud modules for the storage, integrity checking, and replication were completed. Also, the basic architecture for compute services in DuraCloud was designed and implemented providing the ability to install, remove, start, and stop services running in the DuraCloud platform.

Mulgara:

Mulgara 2.1.3 was released in August 2009 and 2.1.6 in December, continuing a pattern of releasing new versions of the software every 2-3 months. This year also saw Mulgara implementing a series of new data and interface standards. Most notable is that Mulgara has become compliance with the SPARQL query language. This positions Mulgara for seamless integration into many types of systems and consistency the directions of the W3C and the semantic web community. Mulgara inspires to be a robust triplestore that can enable exposure of data as part of the emerging web of linked and open data.

9 Marketing and Communication

DuraSpace has established the Marketing and Communications (MarComm) group that is responsible for communications, marketing, and educational services required meet the strategic objectives of the DuraSpace organization. The key objective of MarComm is to create messages and vehicles that will encourage global communities to use DuraSpace technologies, participation in DuraSpace communities, and support the DuraSpace mission (though in-kind or financial contributions).

Combining Fedora Commons and DSpace expertise in this area has enabled the rapid creation of an integrated Communications and Marketing program. The MarComm planning Matrix, established in July 2009, is a synergistic schedule of strategic messages coordinated with events, meetings, conferences, publications and materials. The MarComm Matrix provides benchmarks for accomplishments, and is a dynamic tool for adjusting efforts based on organizational priorities and work plans. In the first two months of operation, we have improved service to the community while growing participation, interest and awareness.

Key goals of the program are to define, target and engage all current and potential DuraSpace audience and market segments, to conduct related campaigns to reach those markets while building a corporate persona. All of vehicles will contain consistent messaging about the DuraSpace organization to maintain strong brand awareness for flagship products—DSpace, Fedora, DuraCloud and Mulgara–while at the same time linking these brands to the new organization. MarComm collaborates with all parts of the DuraSpace organization in developing a unique organizational identity that has as its hallmarks innovation, openness, transparency, and trust.

DuraSpace MarComm seeks and develops open access advocacy opportunities and produces related content including serial publications, news reporting and dissemination on a regular schedule, professional product specification sheets, and product briefs.

MarComm aims to promote the collaborative nature of open source community software development by amplifying and expanding the activities around the new Solution Communities Program. The program will reach across all DuraSpace product lines by providing tools, methods and communications vehicles to enable solution-focused communities to share their ideas and work towards common goals.

A sampling of the activities in 2009 include:

Campaign: "DuraSpace is a key player in the open access movement"

- Launch of DuraSpace Open Access Week Contest
- Release DuraSpace Open Access Package through SPARC and on DuraSpace web site. Will include product spec sheets and web seminar content (Oct. 2009)

Materials

- DuraSpace organizational press release and distribution

- NDIPP Pilot Program - press release and distribution
- DuraSpace initial web site – developed and launched
- Corporate logo and tag line – developed and released
- Initial business materials: designed business cards, stickers, and handouts
- UPCOMING: Product spec sheets for DSpace, Fedora and Mulgara (Oct. 2009)

Events

- MarComm organizational meeting, July 2009
- DuraSpace organization launch – publicity
- OR09 - publicity (DLib article, press releases, blog posts and tweets)
- "All About Repositories" web seminar series – developed/launched with Sun and SPARC
- Solution Communities – re-launch with embedded tools and services (late fall 2009)

Publications, news and information

- RSS news system – revamped for feeding DuraSpace web properties from a single source
- DuraSpace Blog - launched with highlights of news, events, releases, opportunities, and tweets from the past month. Subscribe: http://feedburner.google.com/fb/a/mailverify?uri=DuraSpace
- DSpace newsletter (ongoing)
- New Web metrics reporting

10 Community Strategy and Outreach

Building Solutions Communities

As part of its not-for-profit mission, DuraSpace launched its new Solution Communities Program. DuraSpace believes that *communities* are essential to the creation of shared and enduring solutions. To address the challenges inherent in digital information, the power of community can be harnessed towards solutions for preserving our digital heritage, promoting open access, enabling scholarly communication, managing scientific data, and more.

Our goal is to promote a bottom-up approach to community organizing and enable grassroots efforts to flourish. DuraSpace helps guide motivated community leaders and practitioners in organizing their

communities. Together, we hope to advance our collective knowledge and develop solutions to ensure that our digital data and digital collections are durable and accessible.

Developing community activities around DuraSpace's priority areas is one of our central functions as an organization. We are well placed to organize communities in a variety of areas, especially around application development that has a natural affinity with repositories. Other community activities can focus on standards development, issue advocacy, or simple information sharing in a particular area.

Rather than take a top-down approach to organizing communities, there is good reason to believe that we can act as the catalyst for self-organizing communities. Taking lessons from others who are applying research about emergence in complex systems and "small world" theory (AKA, "the six degrees of Kevin Bacon"), we have begun to develop a methodology to get started. The basic idea is that communities of interest develop around knowledge bases and good communication infrastructure. By recruiting some motivated individuals in a set of specific roles, we can get a starting team in each area that can work with us on growing the community. More information is available at: http://fedora-commons.org/confluence/display/FCCWG/Home.

DuraSpace solutions communities are initially targeted at the broader “open repositories” community to provide a context for collaboration on new ideas, application, tools, and services for repositories. We see these communities as being connected to our mission in the most general sense, particularly the durability of digital information, and we encourage participation by all, whether they are using DuraSpace software products or not. Active communities are:

* Data Curation

* Preservation and Archiving

* Scholars Workbench

* Small Archives

* Media Archives (under development for 2010)

* Open Access Publishing (under development for 2010)

DuraSpace offers a suite of tools and services to support Solution Communities:

- Wiki for creating community knowledge bases
- Group email lists for regular communications
- Social networking tool
- Source code repository for qualifying software projects
- Webinar support
- Help in coordinating collaborative projects

Visit Solution Communities:
http://www.fedora-commons.org/confluence/display/FCCWG/Home

Visit Solution Communities Crowdvine:
http://duraspace.crowdvine.com/

Ambassadorship: A Model for Outreach to New Repository Users

Initially in the context of the DSpace community, we have been exploring new models for reaching out to new repository users. There are currently over 600 DSpace instances in over 70 different countries with a variety of use cases. While there are many helpful resources available to the community, barriers due to geography and language can make it challenging for new or potential DSpace users to connect with the DSpace community. To aide those new to DSpace, the DSpace Global Outreach Committee, with the support and coordination from DuraSpace, will launch the DSpace Ambassador Program in late September.

The main focus of a DSpace Ambassador Program is to identify a DSpace user in every country or region who is willing to be a point of contact for organizations just getting started. These Ambassadors will help new or potential users with general questions about DSpace, as well as providing guidance on additional resources available in the larger DSpace community. The Ambassador is not intended to be a technical expert and instead will point individuals in the right direction for more technical questions. Ideally the Ambassador will also help to identify resources, such as new helpful references or new service providers in their country or region. Some Ambassadors may also be involved in planning regional DSpace user group meetings or trainings. Ambassadors can also leverage their efforts along with other networks, associations and groups that may already exist in their region/country.

At the end of 2009 there were 29 Ambassadors in 21 different countries (Australia, Bangladesh, Belgium, Colombia, Ethiopia, France, Germany, Ghana, Greece, Hungary, India, Iran, Kenya, Nigeria, Norway, Peru, Saudi Arabia, Singapore, Tanzania, Thailand, United States). The volunteers are enthusiastic about helping new and potential DSpace users as well as building user network in their region/country. We had our first meeting in February 2010 to hear how each Ambassador was doing and how DuraSpace can continue to support their efforts. The Ambassadors continue to share information and help each other through their mailing list.

We hope that the DSpace Ambassadors program can serve as a model for our other open source communities, particularly Fedora, in the future. More information is available at: http://wiki.dspace.org/index.php/DSpace_Ambassador_Program.