Building community clouds to support access to scholarship

Michele Kimpton
CEO, DuraSpace

Jonathan Markow
CSO, DuraSpace
Who are we?

We are committed to providing open source technologies and services that promote durable, persistent access to the scholarly record.
Community challenges

• Libraries being asked to steward increasing amounts of digital content
• Difficult to easily provision storage, without advanced planning and capital
• Limited technical resources
• Need more flexible and scalable solutions dependent on type of content they are managing
Why cloud?

Massively scalable compute and storage offered as a service
Survey Higher Ed, 145 responses

Key Benefits

- Scalability
- Remot off campus storage
- Ease of implementation
- Flexibility
- Lack of Local staff
- Cost
- Elasticity
- Pay for use

Responses
Survey Higher Ed, 145 responses

Key Challenges

- Trust 3rd party: 60 responses
- Long term reliability: 50 responses
- Data security: 40 responses
- Performance: 35 responses
- Loss of control: 30 responses
- Admin burden SLA's: 25 responses
- Data lock-in: 20 responses
- Transparency: 15 responses
Advantages of a community cloud solution

• Develop cloud expertise within our own community
• Mitigate the risks of working with a single cloud vendor
• Develop solutions to solve needs of our community
• Aggregate demand across our community to come up with best cloud infrastructure and best pricing
• Consolidate efforts across community to produce a diverse and robust platform and service
DuraCloud Project History

- **Release 0.1**: Cloud Storage Mediation (11/2009)
- **Initial Pilot Program Begins** (10/2009)
- **Release 0.2**: Service Infrastructure (02/2010)
- **Release 0.3**: 05/2010
- **Release 0.4**: Media Streaming (06/2010)
- **Release 0.5**: Open Source Release (07/2010)
- **.9 release Beta production**: 06/2011
- **.7 release Bulk services**: 12/2010
- **Service Launch**: 10/2011
- **Second Pilot program starts**: 09/2010
- **Release 1.0**: 08/2011
## Pilot Partners

<table>
<thead>
<tr>
<th>University</th>
<th>Use Case</th>
<th>Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rice U</td>
<td>Preservation</td>
<td>DSpace, meta archive</td>
</tr>
<tr>
<td>Hamilton College</td>
<td>Access/international collaboration</td>
<td>Fedora</td>
</tr>
<tr>
<td>Northwestern U</td>
<td>Preservation books, audio, image</td>
<td>Fedora</td>
</tr>
<tr>
<td>U of PEI</td>
<td>Image access</td>
<td>Fedora/Islandora</td>
</tr>
<tr>
<td>ICPSR</td>
<td>Preservation</td>
<td>Fedora</td>
</tr>
<tr>
<td>IUPUI</td>
<td>Preservation</td>
<td>DSpace, Content DM</td>
</tr>
<tr>
<td>Rhodes College</td>
<td>Image Access</td>
<td>DSpace</td>
</tr>
<tr>
<td>North Carolina State U</td>
<td>Preservation</td>
<td>DSpace</td>
</tr>
<tr>
<td>CARL</td>
<td>Preservation and Services</td>
<td>Fedora</td>
</tr>
<tr>
<td>MIT</td>
<td>Preservation</td>
<td>Dspace</td>
</tr>
<tr>
<td>Columbia</td>
<td>Preservation and Services</td>
<td>IA, Fedora</td>
</tr>
</tbody>
</table>
What is DuraCloud?

Digital archiving solution run on cloud infrastructure

Multiple copies in multiple clouds
DuraCloud Platform

Unique URL:
http://yourinstitution.duracoud.org

Organization/Administrator using DuraCloud

Repository, Content Management System, File System
(via web, REST APIs, utilities)

DuraCloud platform running on a compute cloud

Programming interfaces

UI
System Administration

API
Service Management

API
Storage Management

Services:
Backup
Health check
Synchronization
Replication

Amazon S3
Rackspace Cloud Files
Microsoft Azure

Other Clouds
Use cases explored

Digital archiving
Image serving
Collaboration
Video streaming
Data management
What partners are saying

• “The ease-of-use of DuraCloud is its most impressive feature.”
  ~Columbia

• “DuraCloud delivers the benefits of a diverse network of storage locations, but without the overhead of managing different vendors.”
  ~ICPSR

• DuraCloud provides a robust infrastructure to support digital preservation of content in our institutional repository, along with services that enhance the delivery and integrity of the content.
  ~Geneva Henry, Rice University
Benefits of working with Community Partner, DuraSpace

• Trusted non profit from academic community
• Community source software
• We have developed expertise cloud technology
• We can aggregate volume across institutions and negotiate a better price and better terms
• We can monitor all cloud providers looking out for your best interests
Internet 2 implementation community cloud

Duracloud platform

Storage cloud 1

Storage cloud 2
Advantages

- Shibboleth authentication and integration
- No additional bandwidth charges
- Cloud providers are part of Internet 2 partner network
- Community managed and supported service
On the Roadmap

• Services
  – Shibboleth authentication
  – DSpace as a service application
  – Video management
  – Data curation and management

• Storage providers
  – UCSD cloud store, Chronopolis
  – SHI cloud, Net plus offering
  – Local Eucalyptus integration
Free DuraCloud Trial Accounts
www.duracloud.org
Why DuraCloud DfR?

• Protect vulnerable research data
• Enable archiving, access, and preservation
• Facilitate cooperation between researcher and institutional data managers
• Provide services to support the research process
Advisors and Participants

- Cornell
- George Washington
- Georgia Tech
- Harvard
- ICPSR
- Johns Hopkins
- MIT
- NCAR/UCAR

- Oregon State
- Rice
- Smithsonian
- U of Oregon
- U of Prince Edward Island
- U of Virginia
- Fluid Project
- DuraSpace
Top Five Priorities

- #1: Connect the operational and archival phases of the data management lifecycle.
Top Five Priorities

• #2: Create simple workflows across the data management lifecycle that automatically capture metadata and provenance.

• (…and create incentives for additional metadata creation)
Top Five Priorities

• #3: Ensure confidentiality, security, privacy, and predictability of data in the cloud. (Trust and Control)
Top Five Priorities

• #4: Automate basic metadata creation and “catalogue” creation.
Top Five Priorities

• #5: Create interoperability of operational systems, archiving solutions, and discovery systems used by specific research communities.
DfR Principles

- Open source, enterprise software solution
- Capture data close to the source
- Don’t interfere with researchers’ processes
- Provide incentives, added value for metadata creation
- Easy to use; workflows for collaboration, hand-off to institution
Architecture

• Take advantage of open source reuse wherever possible, e.g.:
  • Authentication (Enterprise SSO, SAML2)
  • Cloud-based repository components
  • Encryption, service bus, messaging
  • DMP tools, persistent identifiers, citations, publishing references,
  • etc.
Components

- Authentication:
  - Pluggable
  - Shibboleth support in 1.0 release
Components

- DuraCloud Storage Service (Shibboleth SP)
  - Space A
    - Copy of Monitor/Sync Settings
    - Copy of Source Data
  - Space B
    - Copy of Fedora Objects
      - FOXML Files
      - "M" Datastreams for Metadata (data in space B, immutable)
      - "E" Datastreams for Content (data in space A, immutable)
    - DuraStore Messaging

- Fedora CloudSync Service
  - Has a scheduled "Sync" task to keep space B up to date with Fedora.

- Object Creation Service
  - INPUT:
    - Conversion Rules
    - Space A Credentials

- Fedora Repository
User eXperience

- Reuse existing tools for visualization, manipulation of research data
- Create seamless experience for researcher
- Utilize best practices for interaction design
Partnerships

- Smithsonian Institution Office of Research Information Services – UI for the management and visualization of research project data
- Internet2/InCommon (in discussion)
Project Schedule

- Iterative development
- Evolving prototypes
- User Focus
- Expanding group of participants
- First production release at end of 2012 (Four interim releases throughout the year)
Questions

DuraSpace: www.duraspace.org
DuraCloud: www.duracloud.org

Michele Kimpton
Chief Executive Officer
mkimpton@duraspace.org

Jonathan Markow
Chief Strategy Officer
jjmarkow@duraspace.org