DuraCloud pilot program

Michele Kimpton, CEO DuraSpace

Richard Rodger,
Dept Head Software development,
M.I.T. Libraries

Claire Stewart
Dept Head Digital Collections,
Northwestern University University

OR2011 June, 2011
What is DuraCloud

Open technology and managed service that utilizes cloud infrastructure for preservation support and access services

Upload once and make multiple copies to various storage providers
Access anywhere from any internet connected device
Features:
- **Simple**: upload once and make multiple copies to various storage providers
- **Hosted service**: web-enabled, no client software needed
- **Fully managed solution**: one unified interface across multiple providers
- **Flexible**: built on highly scalable cloud platform
- **Turnkey services**: preservation support and media access
- **Open source platform**
Underlying Providers

(still in testing)
Services available

Archiving and Preservation
- Online backup and replication
- Bit Integrity Checking
- Duplicate on Demand

Multimedia Access
- Advanced Image Viewing
- Media Streaming
- Image Transformation
Interest for Repositories

• Take advantage of cloud at less risk
• Simple preservation support
• Take advantage of the other services
Goals for Project

• Community driven
• Open Source and Transparent
• Easy pathway to the cloud

• Produce solutions not just technology
• Mitigate risk of using commercial cloud
• Simple and cost effective
Goals of Pilot

- Use Cases
- Feedback
- Validation
- Testing
- Issues
How DuraCloud Works

Transfer Content via:
-- Web User Interface
-- Sync Utility
-- REST API
## Pilot Partners

<table>
<thead>
<tr>
<th>University</th>
<th>Use Case</th>
<th>Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rice U</td>
<td>Digital archiving</td>
<td>DSpace, MetaArchive</td>
</tr>
<tr>
<td>Hamilton College</td>
<td>Access, international collaboration</td>
<td>none</td>
</tr>
<tr>
<td>Northwestern U</td>
<td>Preservation books, audio, image</td>
<td>Fedora</td>
</tr>
<tr>
<td>U of PEI</td>
<td>Digital archiving</td>
<td>Fedora/Islandora</td>
</tr>
<tr>
<td>ICPSR</td>
<td>Multiple online backups</td>
<td>Fedora</td>
</tr>
<tr>
<td>IUPUI</td>
<td>Dark online archive</td>
<td>DSpace, CONTENTdm</td>
</tr>
<tr>
<td>Rhodes College</td>
<td>Image Access</td>
<td>DSpace</td>
</tr>
<tr>
<td>North Carolina State U</td>
<td>Digital archiving</td>
<td>DSpace</td>
</tr>
<tr>
<td>Colorado Alliance</td>
<td>Preservation and Services</td>
<td>Fedora</td>
</tr>
<tr>
<td>Chronopolis</td>
<td>Integration with Grid</td>
<td>IRODS</td>
</tr>
<tr>
<td>MIT</td>
<td>Online replication and easy file restore</td>
<td>DSpace</td>
</tr>
<tr>
<td>Columbia</td>
<td>Online archiving</td>
<td>IA, Fedora</td>
</tr>
<tr>
<td>WGBH</td>
<td>Media access and archiving</td>
<td>DAM</td>
</tr>
</tbody>
</table>
DuraCloud Integrations tested

- DSpace
- Fedora Commons
- CONTENTdm
  Digital Collection Management Software
What we learned about the cloud

- Durability - No loss of data over 2 year period, 15+ accounts, 15+ TB
- Data availability - Internet Latency is a factor
- Transfer rates - highly dependant on user premise
- Monetary risks - runaway jobs, user attacks
- Crude access controls - S3 and cloudfront
Key Findings from Pilots

• Importance of “Archival Information Package” for Integration with existing systems
• Many of the pilots preferred fixed allocations vs. pay as you go, few could use the cloud directly
• Biggest value to users was simplicity
• Biggest need was increased reporting and management tools
DuraCloud @ MIT Libraries

Pilot Program Experience
Richard Rodgers
Software Development and analysis, MIT
Open Repositories 2011
MIT Libraries Use Case

• Core: replication for preservation in our production IRs
• System administration practices only address HW or admin failures – repo errors unsecured
• Service must be both automatic yet visible
• Geared towards DSpace and system admins
• Must be cost-effective, user friendly, etc

OR 2011 – DuraCloud Pilot MIT Libraries
DuraCloud Value Proposition

- Provides convenient geo-distributed copy management
- Multi-vendor model buys abstraction and lock-in prevention
- Tools and APIs for DSpace integration
- High-bandwidth access to developers
- Platform for preservation services
- Institution-friendly service terms

OR 2011 – DuraCloud Pilot
MIT Libraries
Pilot Challenges/Solutions

- Need content containers: AIPs over loose files
- Repository managers involvement: admin UI integration, in addition to batch tools
- Bandwidth = cost control: checksum verification
- Cost visibility: activity ‘dashboard’
- Service visibility: ‘ad hoc’ auditing

OR 2011 DuraCloud Pilot
MIT Libraries
Pilot Challenges/Solutions

• Not all replication operations interactive: queueing system
• Automatic handling of content updates: change listeners
• Deletions require special handling
DuraCloud Integration (1.8.x)

"Replication Task Suite":
- Suite of Curation Tasks
- One step Sync process
- Via UI or CLI

Package for each Community, Collection & Item

Local Temp Folder (Cache)
Future Activity

• Based on success of pilot, proceeding to production in 2011
• Has triggered an evaluation of preservation policy and practices – formats, extent
• Looking into specific needs and requirements around video content

OR 2001 DuraCloud Pilot
MIT Libraries
Thank you!

- Web site: Duracloud.org
- Demos at the DuraSpace Table during break
- Contact us!
  - Michele Kimpton: mkimpton@duraspace.org
  - Carissa Smith: csmith@duraspace.org

*Sign up at DuraCloud.org for a trial account*