Who are we?

DSpace
Fedora
DuraCloud
Mission

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

[DSpace logo]

[ Fedora Commons logo ]

[ DuraCloud logo ]
Preservation challenges

• Ability to readily provision online storage (ideally in another geographic area, another administration)
• Synchronize content across storage systems
• Audit integrity of content
• Technical resources
• Internal Policies
Why cloud?

Massively scalable compute and storage offered as a web based service
Digital archiving by media type

Figure 2. Total Worldwide Digital Archive Capacity, by Media Type as Percent of Total, 2010-2015

ESG white paper, Feb 2011
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: 50 responses
- Performance: 35 responses
- Loss of control: 25 responses
- Admin burden SLA's: 15 responses
- Data lock-in: 15 responses
- Transparency: 10 responses
What is DuraCloud?

Digital archiving solution based on cloud infrastructure

Across multiple cloud providers
DuraCloud apps

Archiving and Preservation
- Online Backup(s)
- File Format Identification
- File health check
- Synchronization to multiple clouds

Multimedia Access
- Advanced Image Viewing
- Media Streaming
- Image Transformation

...more on the roadmap
How does DuraCloud work?

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

Content → YourDuracloud.org Virtual server
Amazon S3 (primary) → Azure (beta)

Rackspace
DuraCloud security Architecture

- Instance firewall
- Ports 80 & 443
- Security certificate
- DuraCloud login
- Storage provider credentials
- DuraCloud application security
- Storage provider access control
DuraCloud Project History

- Release 0.1 Cloud Storage Mediation 11/2009
- 10/2009 Initial Pilot Program Begins
- Release 0.2 Service Infrastructure 02/2010
- Release 0.3 05/2010
- 06/2010 Release 0.4 Media Streaming
- 09/2010 Second Pilot program starts
- Release 0.5 Open Source Release 07/2010
- 08/2010 Release 0.7 Beta Production
- 12/2010 Release 0.5 Open Source
- 12/2010 Release 0.7 Beta Production
- 06/2011 .9 release Beta Production
- 08/2011 Release 1.0 Service Launch 10/2011
## 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>Image Access</td>
<td>DSpace</td>
</tr>
<tr>
<td>Rhodes College</td>
<td>Preservation</td>
<td>DSpace</td>
</tr>
<tr>
<td>North Carolina State U</td>
<td>Preservation and Services</td>
<td>Fedora</td>
</tr>
<tr>
<td>CARL</td>
<td>Preservation and Services</td>
<td>DSpace</td>
</tr>
<tr>
<td>MIT</td>
<td>Preservation</td>
<td>IA, Fedora</td>
</tr>
<tr>
<td>Columbia</td>
<td>Preservation and Services</td>
<td></td>
</tr>
</tbody>
</table>
Archiving and Preservation support

- Requirements met
  - Easy back up to primary and secondary cloud
  - Keep backups in sync
  - Check health of backups
  - Ability to view and download files
  - Retrieve and restore files
  - Web accessible
On the roadmap

• Services/Apps
  – Improved video handling
  – File validation
  – Encryption

• Storage providers
  – SDSC Cloud
  – Azure
  – Local Eucalyptus integration

• Simplification
Key Findings from Pilots

• Biggest challenge was transferring data from content owner
• Extremely difficult to compare costs
• Integration with local systems
• Biggest value of DuraCloud was simplicity and abstraction of individual cloud stores
Pilot partner comments

• “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

• “One of the best things that DuraCloud does is bridge the complex gap between normal users and cloud-based services, and it does this job very well.”
  ~Rhodes College
What we learned about the cloud

• Durability- No loss of data over 2 year period, 15+ accounts, 15+ TB BUT no published SLA or policy
• Eventual consistency
• Data transfer rates over http, 1 GB realistic
• Large file handling evolving- 5 GB chunk size
• Verification of content via check sum depends on provider
• No standard API’s yet across providers
• Instability of cloud providers (Atmos,Sun)
On the docket

• Northwestern’s computing environment
• Digital Preservation Strategies
• Duracloud pilot
• Cloud Challenges
NORTHWESTERN UNIVERSITY
LIBRARY

4,930,613  Books
208,757  Maps
601,153  Images
66,977  Audio
33,882  Film and video

98,844  Journals and serials
22 TB  Unique digital content
84  Web sites
750  Workstations
59  Servers
Moving towards a strategy

- Preservation risk assessment, metadata extraction, format migration planning, object identification and validation
- Digital Preservation Services and Workflow
- Scalable Architecture
- Re-allocation of staff expertise and local technology
DuraCloud pilot

• Dark archive
• Locally digitized books
• Fedora repository

For a 300 page book:
  o 300 camera jpegs
  o 300 edited jpegs
  o 300 jp2s
  o 300 pdfs
  o 300 OCR txt file
  o 300 OCR XML files
Cloud Challenges

• Service Provider Security Assessment
• Addressing PII, PHI, or other sensitive data concerns
• More sophisticated access control provisions
• Determining lifecycle costs for preservation; Consortial deals?
• Tiered mass storage strategy; How far do we go?
Upcoming Webinars

- Using DuraCloud for Archiving and Preservation
  November 2nd at 1pm ET

- Technical Overview of DuraCloud
  November 16th at 1pm ET

- DSpace and DuraCloud integration
- Fedora and DuraCloud integration
  November/December days/times TBA
Free DuraCloud Trial Accounts
Where can I find out more?

• Web site: www.duracloud.org

• Email: csmith@duraspace.org
The open source platform

• Available for organizations to run locally
• Requires cloud (commercial or private) infrastructure
• Being tested by several organizations now
• Looking for organizations that want to enable their “software as a service” in DuraCloud
• For more details visit the wiki at: https://wiki.duraspace.org/display/duracloud/