DuraSpace Mission

We are committed to providing open source technologies and services that promote durable, persistent access to the scholarly record.
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 required
• Internal Policies
• Sustainability over time
Why cloud?

Massively scalable compute and storage offered as a web based service
Higher Ed survey, 211 responses

<table>
<thead>
<tr>
<th>Answer</th>
<th>Very likely</th>
<th>Likely</th>
<th>Unlikely</th>
<th>Very unlikely</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online primary storage of digital content</td>
<td>18.0% (34)</td>
<td>24.9% (47)</td>
<td>28.6% (54)</td>
<td>20.1% (38)</td>
</tr>
<tr>
<td>Online back-up of digital content: a remote, secondary store of digital content</td>
<td>29.0% (54)</td>
<td>46.8% (87)</td>
<td>14.5% (27)</td>
<td>4.3% (8)</td>
</tr>
<tr>
<td>Preservation support: content replication, auditing, repair</td>
<td>19.9% (37)</td>
<td>43.5% (81)</td>
<td>16.1% (30)</td>
<td>5.9% (11)</td>
</tr>
<tr>
<td>Shared infrastructure for managing digital content within the institution</td>
<td>16.8% (31)</td>
<td>38.4% (71)</td>
<td>23.8% (44)</td>
<td>9.2% (17)</td>
</tr>
<tr>
<td>Shared infrastructure for managing digital content across institutions</td>
<td>16.8% (31)</td>
<td>39.5% (73)</td>
<td>18.4% (34)</td>
<td>10.3% (19)</td>
</tr>
<tr>
<td>Ability to provision compute services on stored digital content</td>
<td>12.6% (23)</td>
<td>30.8% (56)</td>
<td>25.3% (46)</td>
<td>5.5% (10)</td>
</tr>
</tbody>
</table>
Digital archiving by media type

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

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

- Optical
- Cloud
- Tape
- External disk
- Internal disk

<table>
<thead>
<tr>
<th>Year</th>
<th>Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>2%</td>
</tr>
<tr>
<td>2011</td>
<td>3%</td>
</tr>
<tr>
<td>2012</td>
<td>5%</td>
</tr>
<tr>
<td>2013</td>
<td>7%</td>
</tr>
<tr>
<td>2014</td>
<td>9%</td>
</tr>
<tr>
<td>2015</td>
<td>12%</td>
</tr>
</tbody>
</table>

ESG white paper, Feb 2011
What is DuraCloud?

Platform and service based on cloud infrastructure
Across multiple cloud providers
DuraCloud apps

Archiving and Preservation focused-

- Online Backup(s)
- File Format Identification
- File health check
- Synchronization of content to multiple clouds

...more on the roadmap
Archiving and Preservation support

• Duracloud provides
  ✓ Easy back up to multiple cloud providers
  ✓ Keep backups in sync
  ✓ Check health of backups
  ✓ Ability to view and download files
  ✓ Retrieve and restore files
  ✓ Web accessible
Using DuraCloud for Archiving & Preservation

Bryan Beecher
Director, Computer & Network Services
ICPSR
About ICPSR

- Inter-university Consortium for Political and Social Research
- Located at the University of Michigan
- World’s largest archive of social science research data
- In operation for 50 years
- About $15m in revenues
Archival holdings

• Lots of little files
  – text/plain
  – application/pdf
  – text/xml
  – other stuff

• 2m files; 6TB of storage
Strategy

- Bit-level for original (SPSS + Word)
- Normalize into more durable formats (plain text data + XML metadata + PDF/A documentation)
- Transform for better delivery
- Retain transform and derivatives
- Lots of copies
Data archiving, 1 BC
Geographic Diversity, 1 BC
Geographic Diversity, 1 BC
Geographic Diversity, 1 BC
Maybe disk instead of tape?

- Synchronize content to other locations
- Fixity checking lets us know when we need to “fix” something
Get by with a little help from our friends
And they are friends

- Based on relationships
- No SLA
- No scale up/down
- Idiosyncratic interface
- *Contracts? We don’t need no stinkin’ contracts!*
A copy in the cloud
Are you crazy?

- FISMA Low
- Not encrypted
- Machine room open access
- Firewalled
- Professional IT staff + others

- FISMA Medium
- Encrypted
- Machine room controlled access
- Firewalled
- Professional IT staff
Honeymoon period

- Automated monthly billing for usage (storage, computer, network I/O)
  - Small EC2 instance + 6 x 1TB EBS volumes bound together as a RAID
- Easy to scale up and down
- Easy to synchronize
And best of all...
So what’s not to like?

• Cloud diversity
  – Location
  – Technology platform
  – Operational processes
  – Business viability

• Vendor lock-in
Who can save us?
What we like

- Single interface to “the cloud”
- Single billing contact
  - Single relationship
- Value-added services
  - Fixity checking
What we would change

• Filesystem semantics would work better for us
  – rsync v. synctool
  – files v. objects
• Support for big files/objects
• Tools suitable for automated batch use (i.e., out of cron)
Takeaways

• Cloud is a viable option for additional archival copies
• Physical infrastructure may be at least as good as your own
• Encrypt the sensitive stuff
• Not the low-cost solution; but may be the low-hassle solution
More info

- Bryan Beecher
  - bryan@umich.edu
  - http://techaticpsr.blogspot.com/

Thank you for attending this talk
Upcoming DuraCloud Webinars

- Technical Overview of DuraCloud
  November 16 at 1pm ET

- DSpace and DuraCloud
  November 30 at 1pm ET

- Fedora and DuraCloud
  January 11 at 1pm Et
Try DuraCloud Free for One Month: Trial or Subscription
Where can I find out more?

• Web site: www.duracloud.org

• Email: csmith@duraspace.org