Preservation in the Cloud: Three Ways

Michele Kimpton – CEO, DuraSpace
Richard Rodgers, Mark Leggott, & Simon Waddington

Open Repositories 2012 ❖ Edinburgh, Scotland ❖ July 12, 2012
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**

  ✓ **Review integrated health reports from ongoing health checking services**

  ✓ **Access anywhere from any Internet connected device**
DuraCloud @ MIT Libraries

Richard Rodgers
Open Repositories 2012
Edinburgh, Scotland
MIT Libraries Use Case

• Core: replication for preservation in our production IRs – heterogeneous types (video)
• System administration practices only address HW or admin failures – other errors unsecured
• Service should be automatic yet visible
• Geared towards collection and system admins
• Must be cost-effective, user friendly, etc
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
Challenges/Solutions

• Need content containers: AIPs over loose files
• Repository managers involvement: admin UI integration, in addition to batch tools
• Big data files: queued v synchronous operation
• Reduce mistakes: automatic replication
• Deletion is forever: logical deletes + GC
• Service visibility: ‘ad hoc’ auditing
• Preserve what matters: copy masters (video)
Preservation in the Cloud: Islandora and Duracloud

Mark Leggott, University of PEI/DiscoveryGarden
Open Repositories 2012 - Edinburgh, Scotland
Islandora 101

• Drupal+Fedora framework from UPEI
• Flexible UI on top of Fedora + other apps
• Deployed in 100+ institutions, growing
• Desire to provide stronger preservation features and services - e.g. PREMIS coming
• DuraCloud a natural extension
Approach

• Leverage DuraCloud(DC) + DuraSync(DS)
• Maintain context of individual objects and/or complete collections
• “Single Button” restore of damaged assets
• Integrate with standard or private DC
Ist Release

• Admin panel for setup of DC and DS
• “Vault” component on Manage Tab
• Access to DS/DC functions
• Enable sync & restore from specific copy
• View DC and DS reports
Admin Panel
Vault Functions
Collection Restore
Object Health
Object Restore
Backup Task started. Please check back later. When the task is complete the state will be reported as idle and you will be able to view the log in the Cloudsync Reports section. You will not be able to run this task again until it is completed.

Run Cloudsync Task
Create Cloudsync Tasks
Delete Cloudsync Tasks
Cloudsync Reports
Duracloud Reports

Cloudsync Task for islandora:bookCollection
This task will run in the background. When this task is completed it's state will be reported as idle.

Current state is Starting

Copy from Fedora to Duracloud

› ADVANCED
Add a datastream

<table>
<thead>
<tr>
<th>VAULT</th>
<th>ID</th>
<th>LABEL</th>
<th>TYPE</th>
<th>MIME TYPE</th>
<th>SIZE</th>
<th>OPERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>DC</td>
<td>DC Record</td>
<td>Inline XML</td>
<td>text/xml</td>
<td>761 B</td>
<td>download</td>
</tr>
<tr>
<td>Match</td>
<td>MEDIUM_SIZE</td>
<td>MEDIUM_SIZE</td>
<td>Managed</td>
<td>image/gif</td>
<td>75.33 KB</td>
<td>download</td>
</tr>
<tr>
<td>Match</td>
<td>OBJ</td>
<td>Belgium.gif</td>
<td>Managed</td>
<td>image/gif</td>
<td>2.52 KB</td>
<td>download</td>
</tr>
<tr>
<td>Match</td>
<td>RELS-EXT</td>
<td>Fedora Object to Object Relationship Metadata</td>
<td>Inline XML</td>
<td>text/xml</td>
<td>558 B</td>
<td>download</td>
</tr>
<tr>
<td>Mismatch</td>
<td>TN</td>
<td>TN</td>
<td>Managed</td>
<td>image/plain</td>
<td>16.87 KB</td>
<td>download</td>
</tr>
<tr>
<td>Match</td>
<td>CLOUDSYNC</td>
<td>CLOUDSYNC</td>
<td>Managed</td>
<td>text/plain</td>
<td>1.98 KB</td>
<td>download</td>
</tr>
</tbody>
</table>

In Collections
- a test collection
Next Steps

• Tighter integration and more UI functions
• Automated recovery
• Full Fedora/Collection restore
• Support for private DuraCloud instances
Kindura:
Hybrid Cloud Repository

12th July 2012

Simon Waddington
Jun Zhang
Gareth Knight

Jens Jensen
Roger Downing
Cheney Ketley

Centre for e-Research
King’s College London

Science and Technology Facilities Council
Kindura Project

• Project in JISC Flexible Service Delivery Programme

• Kindura was an exploratory project
  • Determine requirements
  • Proof-of-concept content repository
  • Case study to document issues (technical, cost, legal etc.)
  • 9 months 01/02/11 to 30/11/11
Problem description

- Ad hoc storage of research outputs (data, documents)
  - Data stored on PCs, portable drives, USB sticks, local servers, ...
  - Little or no provision for backup and disaster recovery
- Changing landscape
  - UK funders requiring retention of data for 10+ years
  - Journals need supporting data to be available for scrutiny
- Limited support from central IT departments for research data management tools
  - Long lead times to deploy storage
Hybrid cloud solution

- Commercial cloud offers many potential benefits
  - Elasticity, rapid deployment, transparent costs
  - Risks of commercial cloud
  - Data sensitivity, data protection laws, service availability or loss
- In house storage
  - Fully under internal control, inflexible
- Hybrid solution – combine commercial cloud and internal storage
  - Provides elasticity as well as flexibility to retain certain data types in house
  - Allows reuse of existing storage resources
  - Higher complexity
Kindura

- Proof-of-concept repository for research data combining
  - Commercial cloud
  - Internal storage – iRODS
- Based on Fedora Commons
- DuraCloud provides a common storage interface
  - Deployed DuraCloud from source code
- Developed a storage management framework
  - Based on policies, legal and technical constraints, cost
  - Automated decisions for storage and migration
  - Content replication across storage providers for resilience
  - Storage providers transparent to user
  - Provide cost optimisation
Kindura system

Web browser
- Log on
- Upload
- Add Project
- Search
- View

Management server
- Ingest
- Workflow engine
- Migrate
- Interfaces and tools
- Tomcat

Data cache

DuraCloud
- DuraStore
- DuraService
- AWS plug-in
- Rackspace plug-in
- Azure plug-in
- iRODS plug-in

Fedora repository
- Fedora object

Storage providers
- AWS
- Rackspace
- Data
- Azure
- iRODS
- Castor
Further Information

Project blog:

http://kindura.cerch.kcl.ac.uk

Contact:

simon.waddington@kcl.ac.uk
Where can I find out more?

Web site: www.duracloud.org

Email: csmith@duraspace.org

Contact Carissa Smith if interested in a live DuraCloud demo session.