



## Preservation in the Cloud: Three Ways

Michele Kimpton – CEO, DuraSpace

Richard Rodgers, Mark Leggott, & Simon Waddington

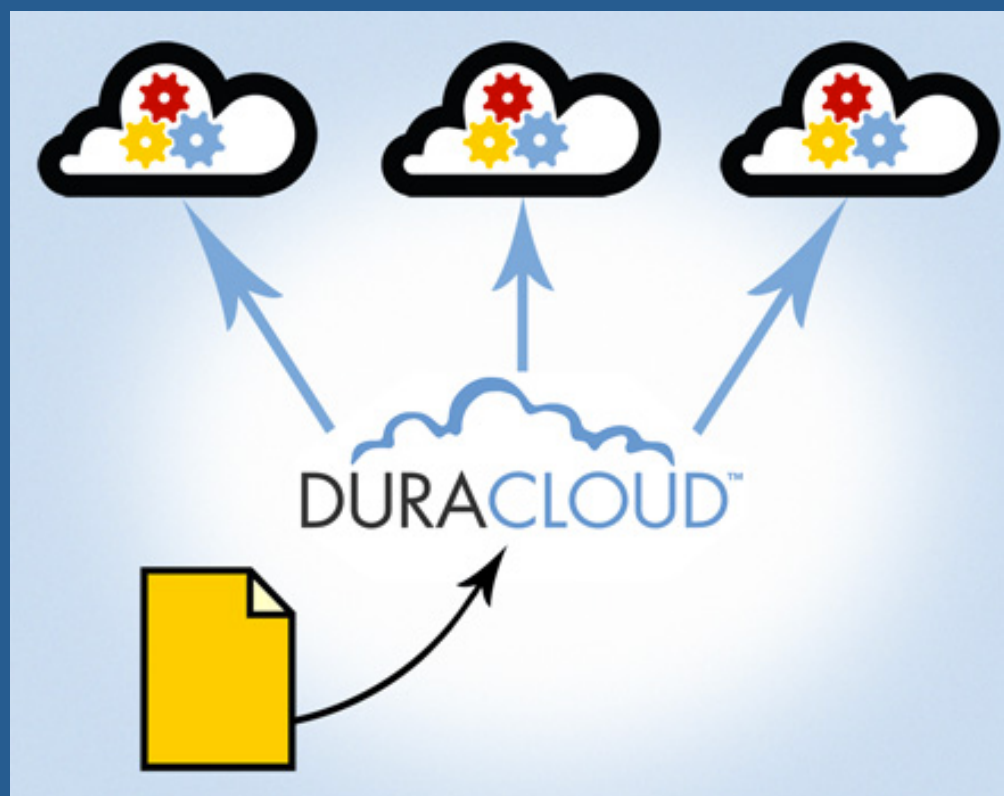


Open Repositories 2012 ❖ Edinburgh, Scotland ❖ July 12, 2012

# 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: ALPs 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





# 1st 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





DashboardContentStructureAppearancePeopleModulesConfigurationIslandoraReportsHelp

dd contentFind contentEdit shortcuts

My accountLog out

Islandora codesprint-centos.islandora.ca

Home » Administration

> Configuration

Configure settings for Islandora.

> Islandora Vault

Configure settings for Islandora Duracloud cloudsync.

> Basic Image collection

Configure the Basic Image solution pack.

> Core collection

Configure the core Islandora collection functionality.

> Form associations

Associate metadata ingest forms with Islandora content models.

> Solr client

Configure Solr search settings.

> XML Form Builder

Manage XML-based Forms.

# Admin Panel

DiscoveryGarden

Add contentFind content

Edit shortcuts

My accountLog out

Islandora Vaultcodesprint-centos.islandora.ca

Home » Administration » Islandora

Duracloud space \*

upeid-dev

The Duracloud space that where you will copy items to/from

CloudSync Url \*

http://184.72.239.57:18081/cloudsync/api/rest

The URL to use when connecting to Duracloud spaces

Cloudsync User \*

fedoraAdmin

The username to use when connecting to Duracloud

Cloudsync password \*

Password \*

Password strength:

Confirm password \*

Duracloud store url

http://184.72.239.57:18081/cloudsync/api/rest/objectStores/2

The url for the duracloud store. This url will be provided by your cloudsync administrator.

Save configuration

Vault Functions

DiscoveryGarden



The screenshot displays the Islandora Repository web interface. The top navigation bar includes links for Dashboard, Content, Structure, Appearance, People, Modules, Configuration, Islandora, Reports, and Help. A user is logged in as 'admin' with a 'Log out' option. The main header shows the repository name 'codesprint-centos.islandora.ca' and tabs for 'VIEW' and 'MANAGE'. The 'MANAGE' tab is active, showing sub-tabs for 'Datastreams', 'Properties', 'Collection', and 'Vault'. The 'Collection' sub-tab is selected, leading to the 'Run Cloudsync Task' page. On the left, a sidebar lists options: 'Run Cloudsync Task', 'Create Cloudsync Tasks', 'Delete Cloudsync Tasks', 'Cloudsync Reports', and 'Duracloud Reports'. The main content area is titled 'Cloudsync Task for islandora:testCollectionPP' and states 'This task will run in the background. When this task is completed it's state will be reported as idle.' and 'Current state is Idle'. There are two buttons: 'Copy from Fedora to Duracloud' and 'Copy from Duracloud to Fedora'. The 'Copy from Duracloud to Fedora' button is highlighted. Below this, an 'ADVANCED' section is expanded, showing 'Run the restore task' with a description: 'This will copy items from Duracloud to Fedora, Make sure you want to restore Fedora objects before you run this task. This task will run in the background.' and 'Current state is Idle'. The interface is framed by decorative corner ornaments.

Dashboard Content Structure Appearance People Modules Configuration Islandora Reports Help

Hello admin Log out

add content Find content Edit shortcuts

Repository codesprint-centos.islandora.ca

My account Log out

VIEW MANAGE

Home » Repository » Manage

Datastreams Properties Collection Vault

Run Cloudsync Task

Create Cloudsync Tasks

Delete Cloudsync Tasks

Cloudsync Reports

Duracloud Reports

Cloudsync Task for islandora:testCollectionPP

This task will run in the background. When this task is completed it's state will be reported as idle.

Current state is Idle

Copy from Fedora to Duracloud

ADVANCED

Run the restore task

This will copy items from Duracloud to Fedora, Make sure you want to restore Fedora objects before you run this task. This task will run in the background.

Current state is Idle

Copy from Duracloud to Fedora

belgium

# Collection Restore



[Run Cloudsync Task](#)[Create Cloudsync Tasks](#)[Delete Cloudsync Tasks](#)[Cloudsync Reports](#)[Duracloud Reports](#)

Cloudsync Task already Created and configured for islandora:testCollectionPP

To recreate this task you must first delete it.

## Backup Task Data

KEY	VALUE
queryType	ITQL
queryText	select \$object from <#ri> where (\$object or \$object or \$object ) order by \$object
setUri	http://184.72.239.57:18081/cloudsync/api/rest/objectSets/27
queryStoreUri	http://184.72.239.57:18081/cloudsync/api/rest/objectStores/4
sourceStoreUri	http://184.72.239.57:18081/cloudsync/api/rest/objectStores/4
destStoreUri	http://184.72.239.57:18081/cloudsync/api/rest/objectStores/2
overwrite	true
includeManaged	true
name	ISLANDORA MANAGED Task for object - islandora:testCollectionPP - drupal - 137.149.200.66 - backup
state	Idle
type	copy
uri	http://184.72.239.57:18081/cloudsync/api/rest/tasks/23

## Restore Task Data

KEY	VALUE
queryType	ITQL
queryText	select \$object from <#ri> where (\$object or \$object or \$object ) order by \$object
setUri	http://184.72.239.57:18081/cloudsync/api/rest/objectSets/28
queryStoreUri	http://184.72.239.57:18081/cloudsync/api/rest/objectStores/2

# Object Health





DashboardContentStructureAppearancePeopleModulesConfigurationIslandoraReportsHelp

contentFind content

Repository codesprint-centos.islandora.ca

My accountLog out

VIEWMANAGE

Home » Repository » Manage

DatastreamsPropertiesCollectionVault

Run Cloudsync Task

Create Cloudsync Tasks

Delete Cloudsync Tasks


Cloudsync Reports


Duracloud Reports

Cloudsync Logs for islandora:testCollectionPP

RESULT	STARTED	FINISHED
Succeeded	2012-07-04T19:47:25.150Z	2012-07-04T19:47:28.229Z

RESULT	STARTED	FINISHED
Succeeded	2012-07-04T19:48:45.171Z	2012-07-04T19:48:49.574Z

  
belgium

  
test image for duracloud

# Object Restore

DiscoveryGarden

- ✓ Task and logs deleted for *islandora:576* backup task
- Task, set and logs deleted for *islandora:576* restore task
- CLOUDSYNC datastream deleted for *islandora:576*

## Run Cloudsync Task

## Create Cloudsync Tasks

## Delete Cloudsync Tasks

## Cloudsync Reports

## Duracloud Reports

## Create a Cloudsync tasks for islandora:576

This will create two tasks, one for copying objects to Duracloud and one for restoring objects from Duracloud

## Duracloud datastore

`http://184.72.239.57:18081/cloudsync/api/rest/objectStores/2`

All chosen content will be replicated to the datastore *DuraCloud Space upei-dev at dgi.duracloud.org (AMAZON\_S3)*

## Choose a Fedora datastore

Fedora Repository at codesprint-centos.islandora.ca:8080

All chosen content will be replicated to the selected datastore

## ▼ ADVANCED

## Query to get a list of pids for this collection

```
select $object from <#ri>
where ($object <fedora-rels-ext:isMemberOfCollection> <info:fedora/islandora:576>
or $object <fedora-rels-ext:isMemberOf> <info:fedora/islandora:576>
or $object <mulgara:is> <info:fedora/islandora:576>)
order by $object
```

This query will be used to generate a cloudsync set to be used by a cloudsync task. In most cases you can use the default

## Query type

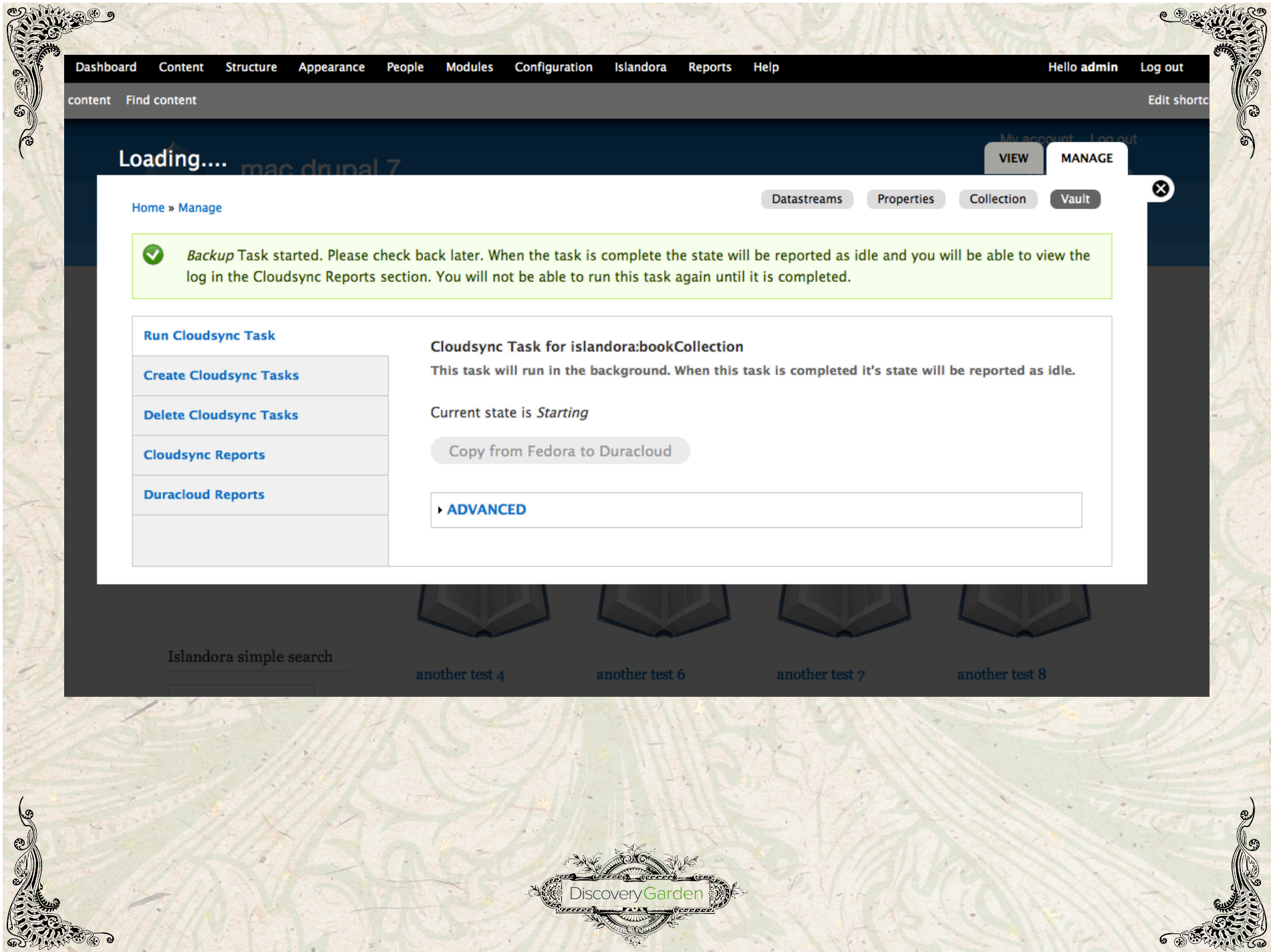
☒ ITQL

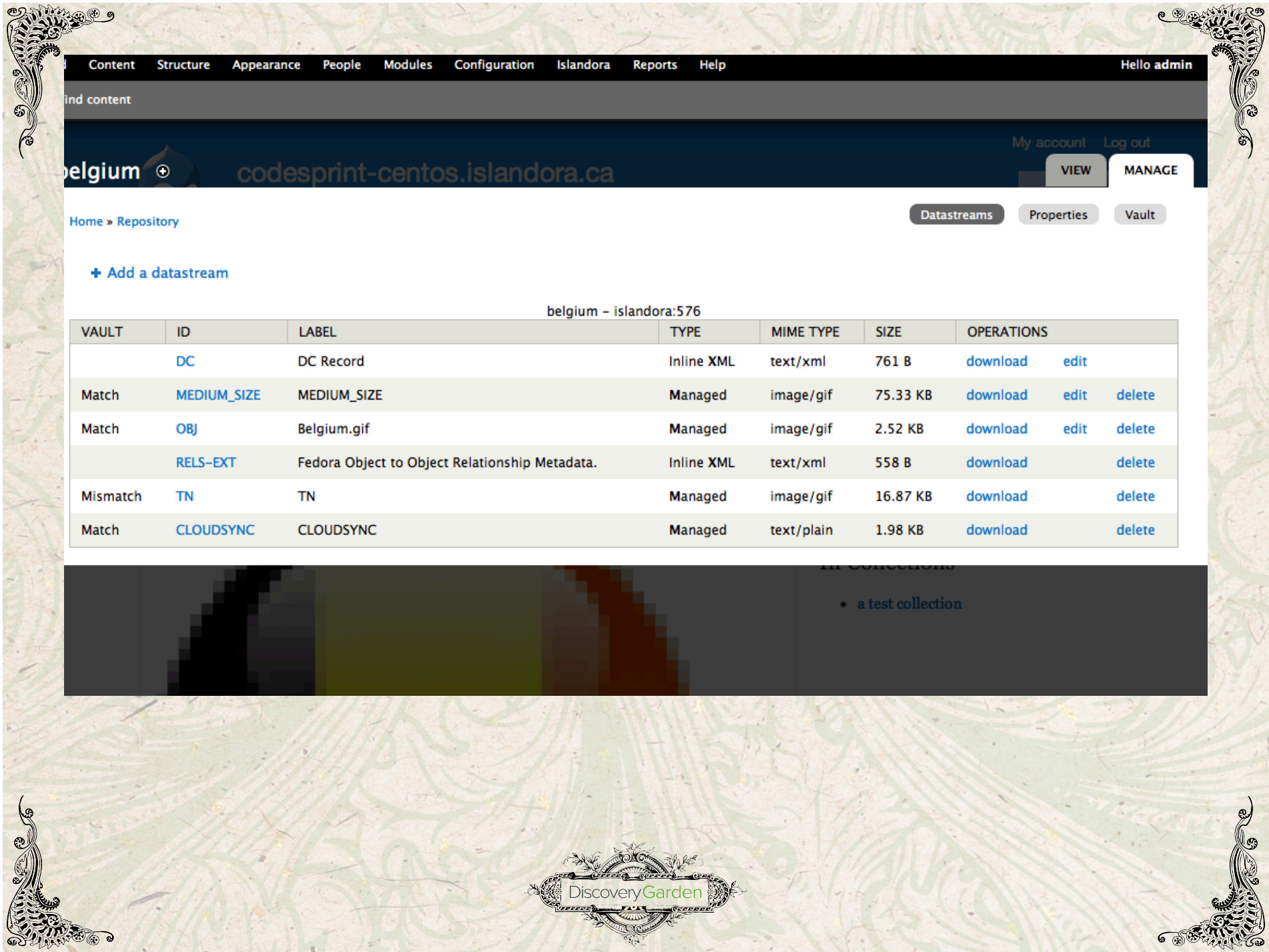
☐ SPARQL

# Reports









Find content

Home » Repository Datastreams Properties Vault

+ Add a datastream

belgium - islandora:576

Table with 7 columns: VAULT, ID, LABEL, TYPE, MIME TYPE, SIZE, OPERATIONS. It lists various datastreams including DC Record, MEDIUM\_SIZE, Belgium.gif, RELS-EXT, TN, and CLOUDSYNC.

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

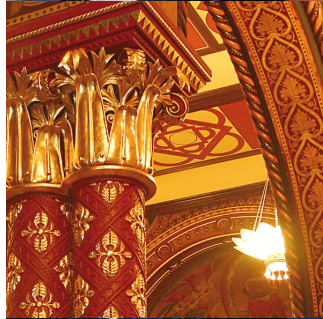
12<sup>th</sup> July 2012

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

JISC

- 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



Science & Technology  
Facilities Council



# 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

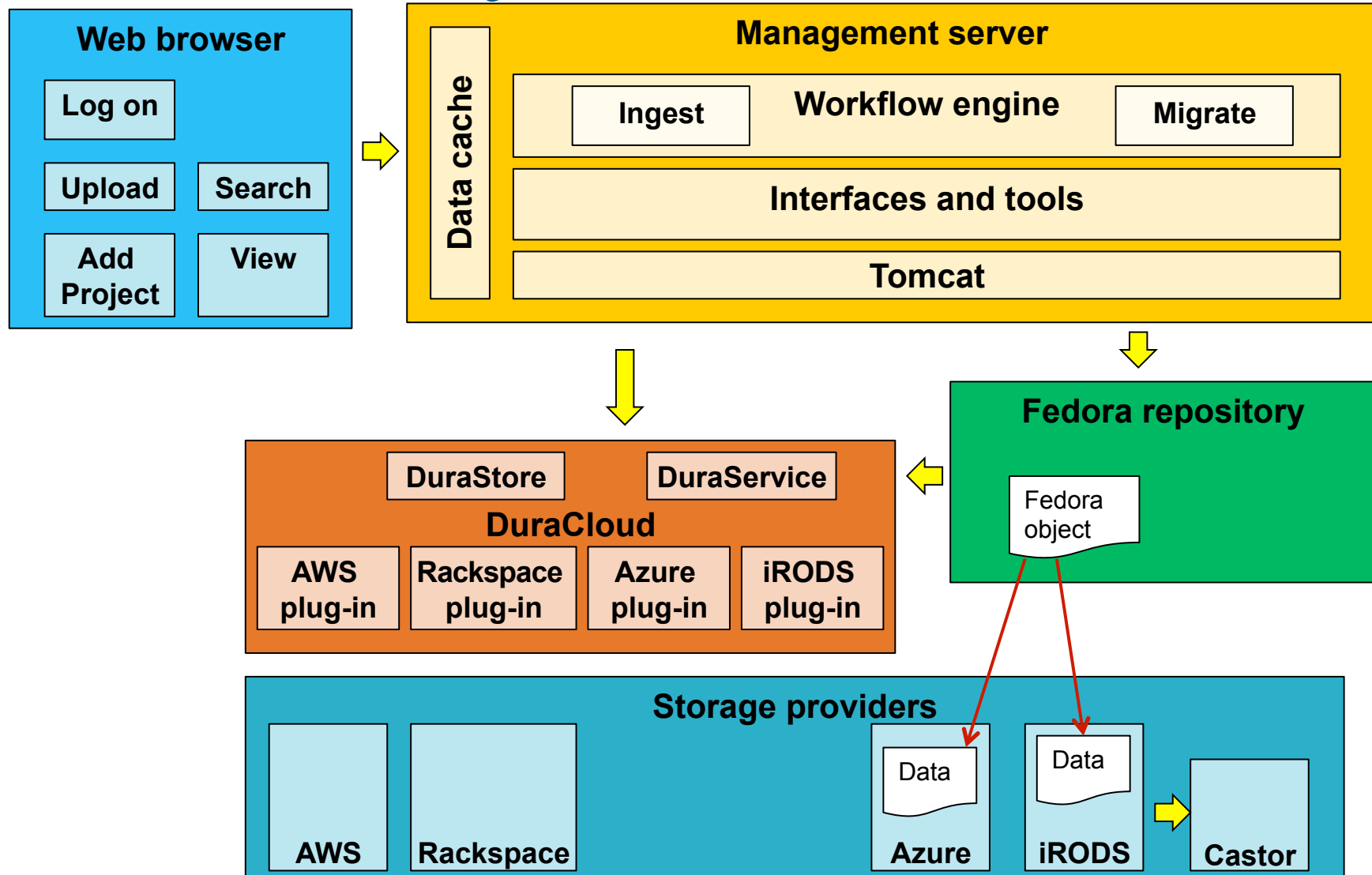
- Commercial cloud offers many potential benefits
  - Elasticity, rapid deployment, transparent costs
- Risks of commercial cloud
  - Data sensitivity, data protection laws, service availability or loss
- In house storage
  - Fully under internal control, inflexible
- Hybrid solution – combine commercial 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



# Further Information

Project blog:

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

Contact:

[simon.waddington@kcl.ac.uk](mailto:simon.waddington@kcl.ac.uk)

# Where can I find out more?

The screenshot shows the DuraCloud website homepage. At the top, there's a navigation bar with links: MY ACCOUNT | DOCUMENTATION | SUPPORT | LOGOUT | and social media icons for Twitter and RSS. Below this is a search bar and a menu with links: Overview, Solutions, Features, and Learn More. The main banner features the text "ONE CLICK TO THE CLOUD: MULTIPLE COPIES, LOCATIONS AND PROVIDERS" and "Preservation PRESERVE AND ARCHIVE AS MANY COPIES AS YOU WANT" over a background image of a person's head. Below the banner, there are three buttons: "FIND OUT MORE", "TRY IT", and "SUBSCRIBE" (with a "get two months FREE" badge). To the right of these buttons is a video player showing the DuraCloud logo. At the bottom, there's a footer with four columns: "Organization" (About, Pricing, News, Press Mentions, Events, Contact), "Open Source" (Overview, Software, Documentation, Listserv, Forum, Tracker), "Connect" (Twitter, RSS, Wiki, Newsletter, Site Map), and "DuraCloud Says" (DuraCloud: CERNET, Indiana U & Internet2 demo data transfer, global video++ to show value of 10G China & US Internet link, http://t.co/Ac0rDkmj #lod).

Web site:

[www.duracloud.org](http://www.duracloud.org)

Email:

[csmith@duraspace.org](mailto:csmith@duraspace.org)

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