DuraCloud and Flexible Digital Preservation at the Texas Digital Library
(or: A Series of Tubes)

MONDAY, APRIL 27, 2015

Gad Krumholz
Sr. Software Engineer, TDL

Debra Hanken Kurtz
CEO- DuraSpace
Former Exec. Director, TDL

Ryan Steans
Director of Operations, TDL
Summary

- History of Preservation at TDL
- DuraCloud™@TDL Infrastructure
  - Duracloud™
  - Digital Preservation Network
  - Amazon Web Services
  - TACC and iRods
  - TDL Preservation Architecture
- Contributions
- Futuretunities
- The DuraSpace perspective
- Cost
- Let’s preserve some stuff!
Saving the Past for the Future

DuraCloud™ is an easy-to-use tool for use by librarians and technical staff to assist in the maintenance of your collections across a robust, reliable infrastructure that will keep your preservation materials secure, retrievable and accessible today and tomorrow.
What are we preserving?
Early Development 2008 - 2011

Texas Preservation Network
- Entirely within Texas
- Living on hardware at member campuses
- Homegrown software developed by TDL
- Essentially 3 – 5 copies
- Ended in 2011-2012

Challenges
- Difficult for Working Group to make suggestions
- Guidelines exceedingly difficult to follow
- Varying requirements from varying customers
Change Afoot 2012-2013

Availability of new technologies
  ◦ Cloud service via Amazon (AWS)
  ◦ DuraCloud™ from DuraSpace

TDL DuraCloud™ Open Source Development
  ◦ TACC experiments

Digital Preservation Network – 2012
  ◦ James Hilton at TCDL 2012
  ◦ Nationwide effort
  ◦ Top-down buy in

Change in leadership of TDL
  ◦ Debra Hanken Kurtz arrived 2013
New Directions 2013- Present

Mandate from membership
- Identified Preservation as key member need
- TDL Governing Board supported work

DPN Established
- 5 nodes developed
  - UT Austin/ TPN*
  - Stanford
  - Hathi
  - Chronopolis
  - AP Trust
- Standards and protocols outlined
- TDL participation – DuraCloud as ingestion point

*work done by Dan Galewsky and Ladd Hanson of UT Libraries
DuraCloud™ Tests at TDL

Shared DuraCloud Service for all members

Storage
- Amazon S3 – high availability
- Amazon Glacier – cost efficient slow retrieval
- iRODS at Texas Advanced Computing Center

Initial tests
- Local
- Houston – Rob Spragg
- Baylor – Darryl Stuhr (implementation of DuraCloud Sync)
DuraCloud™ TDL Working Group

Charter and Scope
- Perform Beta testing of DuraCloud™
- Review Ingest methods
- Develop Suggestions for Best Practices
- Serve as focus group for structuring basic service model
- Make recommendations for next steps

Participants
- Darryl Stuhr – Baylor
- Annie Wu & Robb Spragg – University of Houston
- Mike Bolton – Texas A&M
- Heidi Winkler – Texas Tech
- Todd Peters – Texas State University
- Jennifer Lee, Wendy Martin & Ladd Hanson – UT Austin
- Carissa Smith – DuraSpace.org
DuraCloud™@TDL as a Service

- Available in January 2015 (it’s here!)
- Consistent Upgrades - at DuraCloud 3.3
- Webinars offered and now available online
- Visit the Website
- Charge-back model – Members pay for what they use
  - Members receive a discount on yearly usage
  - Cost varies depending upon storage option
  - TDL Membership covers cost of running the basic service, technical support, etc...
Introduction
Duracloud

Duracloud Features

● REST API’s
  ○ Storage system REST API
  ○ Reporting system REST API
  ○ Security requiring authentication on all DuraCloud applications

● Core Services
  ○ Media Streaming
  ○ Bit Integrity Checking
  ○ Duplication

● Tools
  ○ Sync Tool
  ○ Retrieval Tool
  ○ Chunker Tool
  ○ Stitcher Tool
Digital Preservation Network

DPN Nodes

- Academic Preservation Trust (APTrust)
- Chronopolis
- Hathi Trust
- Stanford Digital Repository (SDR)
- Texas Preservation Node (TPN)
Amazon Web Services

AWS Features

- S3
  - Designed for 99.99% availability
  - Designed to provide 99.999999999% durability of objects over a given year
  - Within a region, your objects are redundantly stored on multiple devices across multiple facilities

- Glacier
  - “Dark” Storage
  - Designed to provide average annual durability of 99.999999999% for an archive

- CloudFront
  - Content Delivery Network
  - Streaming (HTTP / RTMP)

- Elastic Cloud Compute (EC2)
  - Virtual Machines
  - Scalability
  - The Amazon EC2 Service Level Agreement commitment is 99.95% availability for each Amazon EC2 Region.
Texas Advanced Computing Center

Features

- Corral is a collection of storage and data management resources primarily located at TACC, with 5 petabytes of storage installed in the UT data centers at TACC and in Arlington, and an additional petabyte of unreplicated storage for low-latency applications.
Architecture

TDL Preservation Features

- EC2 Computing
- S3
  - Content Delivery Network
- Glacier
  - “Dark” Storage
- CloudFront
  - Streaming
- iRODS @ TACC
  - “Dark” Storage
- DPN

This work is licensed under a Creative Commons Attribution 4.0 International License.
Contributions

Duracloud
- https://github.com/duracloud/duracloud
- https://github.com/duracloud/snapshot
- https://github.com/duracloud/mill

University of Maryland Institute for Advanced Computer Studies
- https://github.com/msmorul/irods-api

Texas Digital Library
- https://github.com/TexasDigitalLibrary/duracloud
- https://github.com/TexasDigitalLibrary/irods-api
Archivematica+Duracloud Features

- **aip-store**: Where completed AIPs will be stored.
- **dip-store**: Where completed DIPs are stored.
- **transfer-source**: Space to sync digital objects for transfer into Archivematica. You may also create multiple spaces with other names for this purpose.
- **transfer-backlog**: Space used by Archivematica for material sent to [backlog](#).
Streaming

CloudFront Features

- S3
  - Used as the “origin server”
- CloudFront
  - Content Delivery Network
  - Streaming (HTTP / HLS / RTMP)
  - Supports FLV and MP4 (VP6 and H.264 codecs)
Static File Hosting

Duracloud Static File Hosting Features

- Can host static files through duracloud and S3
  - The example uses static HTML and JavaScript to show video players (JWPlayer) for all of the videos within a duracloud space
  - Could be used to create rich and full-featured JavaScript web user-interfaces
  - Does not support running server-side code
Ecosystem

“...we are all together in this single living ecosystem called planet earth.”

Sylvia Earle
American Marine Biologist
National Women’s Hall of Fame
1998 National Conservation Achievement Award
“Stand together, join arms, take a step together!”
Spaces and Storage Options

- Can have multiple spaces on any TDL storage platform (except TPN)
  - S3
  - Glacier
  - Irods/TACC
  - Right now, one space at DPN/TPN
- Manage multiple collections
- Allow for multiple contributors from your campus
What’s this going to set me back?

**Amazon S3**
- Storage: $360 per TB/year
- Ingestion (Data In): $0
- Retrieval (Data Out): $120 per TB

**Amazon Glacier**
- Storage: $120 per TB/year
- Ingestion (Data In): $0
- Retrieval (Data Out): $120 per TB retrieved

**Texas Advanced Computing Center (iRODS @TACC)**
- Storage: $205 per TB/year
- Ingestion (Data In): $120 per TB uploaded
- Retrieval (Data Out): $120 per TB retrieved

**Allowance**
Example

A member institution wishes to store 5 TB at Amazon S3, Amazon Glacier and TACC.

<table>
<thead>
<tr>
<th>Type</th>
<th>Storage Cost/ TB</th>
<th>TB Stored</th>
<th>Years Stored</th>
<th>Upload Cost/ TB</th>
<th>TB Uploads</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>360</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>1800</td>
</tr>
<tr>
<td>Glacier</td>
<td>120</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>600</td>
</tr>
<tr>
<td>TACC</td>
<td>205</td>
<td>5</td>
<td>1</td>
<td>120</td>
<td>5</td>
<td>1625</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4025</td>
</tr>
</tbody>
</table>

*Total Cost for all three storage options concurrently*

*Cost $4025 – $250 member discount = $3775*
Let’s Get Preserving!

• What collections do you need to preserve?
• Put your third copy in DuraCloud™@TDL
• Let’s pick some materials and give it a shot.
• It’s easy (and fun!) to try out.
• TDL is happy to consult with you on
  • Cost
  • Space types
  • Long range planning
• DuraCloud™@TDL today, DPN tomorrow!
Questions and Discussion