Digital Repository Initiative

01 April 2014

Ray Frohlich
Director, Enterprise Systems and Infrastructure
with

Euan Cochrane
*Digital Preservation Manager, Preservation*

Michael Friscia
*Manager of Digital Library & Programming Services, Library IT*
What is a digital repository?

A combination of people, processes and technologies, which together provide the means to capture, preserve and provide access to digital objects.

What is a digital preservation service (DPS)?

A combination of people, processes and technologies, which together provide the means to preserve digital objects.
Current state

- Existing content is curated, ingested, and accessed through a variety of systems, many custom built for the specific collection or project

- Within the next year YUL will have nearly a petabyte of digital objects that will need to be preserved long term

- Digital storage infrastructure is not robust, objects are at risk of loss, and content is spread across several disparate storage infrastructures

- Digital preservation software systems and services are mostly non existent
Benefits of further investment

• Alignment with the Yale University Library’s commitment to the stewardship of digital collections and content

• Unified, consistent, and efficient approach to long term access and retention

• Low risk of information loss
  – 4 copies of an object across 3 locations (New Haven, West Haven, Glastonbury) on 2 storage platforms
  – Internal integrity validation (checksum)
  – Media refreshing and replacing

• Low cost (compared to non-Yale service providers)

• Meet user and systems access requirements
What will the repository contain?

- Digitized content
- Collections based born digital content
- Collections based metadata
- Vendor content
- Research data
Reference architecture

The OAIS Reference Model (ISO Standard 14721)

DIP: Dissemination Information Package
AIP: Archival Information Package
SIP: Submission Information Package
Current/FY2015 Implementation
Hydra Project
Hydra is...

• A Repository Solution
• A Community
• A Technical Framework
• Open Source Software
• www.ProjectHydra.org

If you want to go fast, go alone.

If you want to go far, go together.
Hydra Partners

- Duraspace
- Stanford University
- University of Hull
- University of Virginia
- MediaShelf
- University of Notre Dame
- Northwestern University
- Columbia University
- Penn State University
- Indiana University
- London School of Economics
- Rock and Roll Hall of Fame
- Royal Library of Denmark
- Data Curation Experts
- WGBH
- Boston Public Library
- Duke University
- *Yale University*
- Virginia Tech
- University of Cincinnati
- Princeton University
- Cornell University
Hydra Stack

• Fedora
• Blacklight
• Ladybird
• Active Fedora
• Apache Solr
• Media Server
• Internet Archive Book Reader
• Ingest applications
Your localized Web Applications

Ruby on Rails

Primary Gems

Hydra-Head
creating and managing objects (CRUD)

Hydra Model Logic

Hydra Access Controls

Supporting Gems

OM
ActiveFedora
Rubydora

Solrizer
(Solrizer-Fedora)

System Services & Datastores

Fedora REST API

Fedora

Solr REST API

Solr
What is Ladybird

LadyBird is a Hydra-compliant group of web-based and client applications designed to process digital collections including metadata management and digital media for both reformatted items and born-digital content across the Yale University Libraries.

LadyBird routes content to the Hydra/Fedora repository which in turn exposes content through our public discovery/access system, Blacklight.
Ladybird

- Started June 2010
- Version 1.0 December 2013
- 17 background applications
- 4 desktop applications
- 3 web applications
- C# .Net 4.0
- 575,000 lines of source code

- 2,067,198 assets
- 2.5 mil on deck
- Growth: 1,500 assets per day
- 3 Microsoft SQL databases
- 360GB of raw data
- 20 TB files staged
- 40 TB to import
- A Jazz song by Tadd Dameron
Ladybird with Hydra

Import, Curate, Ingest, Publish
Ladybird Roadmap

- Partnership with Columbia
- Potential partners with Princeton, MIT, Northwestern
- Release Ladybird as Hydra Head this fall
- Platform migration to Java 8, MySQL
Hydra Roadmap

• Blacklight 5.x
• Fedora 4
• Open Archival Information System (OAIS) ingest model
• Workflow System Architecture
• Digital Preservation Interfaces
Preservation Tools

[Diagram with preservation tools and relationships]
“Digital Information lasts forever or 5 years, whichever comes first”

Digital Preservation Challenge: Bit Rot
Digital Preservation Challenge: Hardware Failure
Digital Preservation Challenge: Hardware Obsolescence
Digital Preservation Challenge: Software Obsolescence
Digital Preservation Challenge: Natural Disasters
Digital Assets Degrade Without Maintenance

Modern software alters this data:

- Changing its meaning
- Reducing the asset’s value

Original digital asset includes visual data
Inaction will Reduce Asset Value

Modern software alters these equations:

- Changing their meaning
- Removing trust in information
- Destroying the asset’s value

Original digital asset includes important equations.
## Addressing the Challenges

<table>
<thead>
<tr>
<th>Risks</th>
<th>Create &amp; Maintain Multiple Copies</th>
<th>Store Copies in Risk-Diversified Locations</th>
<th>Employ Diverse Systems &amp; Vendors</th>
<th>Preserve Original Software</th>
<th>Preserve Original Hardware</th>
<th>Implement Emulation solutions</th>
<th>Migrate Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Disasters</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User Error</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Media Failure</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bit Rot</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Hardware Evolution</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>OS Evolution</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Software Evolution</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Long Term Digital Preservation Uses Tools and Techniques from Related Disciplines

**How is Digital Preservation different to Digital Asset Management?**

- Implementing Digital Preservation processes, tools and services enables trust in the integrity and authenticity of digital assets throughout long-term technological change.

- Unlike Digital Asset Management, Long Term Digital Preservation requires a comprehensive understanding of the formats and software dependencies of the objects being preserved and more comprehensive storage and access policies than DAMSs usually require.

**How is Digital Preservation Different to Data Migration?**

- Data Migration techniques can form a part of more comprehensive digital preservation workflows. Some digital archives transfer content from owners using systems migration processes.

**How is digital Preservation different to Disaster Recovery?**

- Digital Preservation storage solutions have a unique risk profile and normally require a lower risk of loss of information than standard disaster recovery solutions allow for.

- Digital Preservation storage solutions often have access profiles that can be significantly different to disaster recovery requirements.

**How is digital Preservation different to Digital Forensics?**

- Digital Preservation uses the tools of digital forensics to achieve different aims, such as using disk imaging tools to snapshot data in place and using forensic analysis tools to identify and characterise the technical properties of digital files.
Digital Preservation Tools & Services

Intellectual Management and Access Module

Intellectual Management and Access Services
(Inventories, Catalogues, Finding Aids, Single Point of Access)

Content Acquisition Modules

- Digitization Services
- Digital Transfer Tools and Services
- Digital Archaeology Tools and Service

Core Digital Archive Modules

- Ingest Tool(s) and Services
- Storage Services
- Metadata Extraction & Creation Tools Services
- Bit Preservation Tools and Services

Active Preservation Modules

- Preservation Planning Services
- Software & Environment Archive
- Format Transfer Services (Migration / Normalization)
- Emulation Services
Digital Preservation Tools Roadmap

- Programming team formed
- Gathering use cases and user stories
- Platform selection

Simplest use case:
- Validate file: 17 sec average
- Validate current repository: 883 days
- Target: 1 day
Digital Preservation with Hydra
Parallel, Multi-Threaded Applications Spanned Across Virtual Servers
Storage Infrastructure
Proposed FY2015

Fedora

Akubra Low Level Storage Fedora Plug-in

Akubra Blob Store API

/FedoraStore

/RSS

/StorHouse

ITS NetApp Storage
-10TB

ITS R55 Storage
100TB

Sgi StorHouse

Dell MD Series
7TB

Spectra Logic T950
100TB

NFS

= new component

= new service
Proposed Trusted Edge Policy
Proposed FY2015 Staging
Proposed Staging Trusted Edge Policy

[Diagram showing Samba connected to Sigi Trusted Edge Storage Management (see note) which connects to /RSS and /StorHouse. /RSS connects to ITS RSS 250TB and Dell MD Series 7TB. /StorHouse connects to Sigi StorHouse and Spectra Logic T950 250TB.]

Note
Policy 1
- All objects COPY from /RSS to /StorHouse
Storage Roadmap

Fall 2014

• Transition from NetApp storage to ITS RSS 2
• Stand-up Fedora 4 for testing. Configure and exercise new storage management layer (ModeShape/Infinispan).

Opportunities to explore

• Migration to Yale ITS Sgi StorHouse implementation
• ITS RSS 2 and/or HPC storage
• Out-of-region location for data replication
• Continue exploring external storage providers
## A note about external storage providers

<table>
<thead>
<tr>
<th>Service Provider</th>
<th>Cost per GB/Year</th>
<th>Endowment cost</th>
<th>Endowment Period</th>
<th>Content types accepted</th>
<th># of Copies</th>
<th>Bit preservation?</th>
<th>Active Preservation?</th>
<th>Curation?</th>
<th>Access?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronopolis</td>
<td>$2.15</td>
<td>N/A</td>
<td>N/A</td>
<td>all</td>
<td>3</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Digital Preservation Network (DPN)</td>
<td>$0.83</td>
<td>$4.88/GB</td>
<td>20 years</td>
<td>all</td>
<td>3</td>
<td>y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Dspace Direct</td>
<td>$33.00</td>
<td>N/A</td>
<td>N/A</td>
<td>Limited</td>
<td>2 - 4</td>
<td>y</td>
<td>N</td>
<td>P</td>
<td>Y</td>
</tr>
<tr>
<td>DuraCloud</td>
<td>$1.11</td>
<td>N/A</td>
<td>N/A</td>
<td>all</td>
<td>2 - 4</td>
<td>y</td>
<td>N</td>
<td>N</td>
<td>P</td>
</tr>
<tr>
<td>HathiTrust</td>
<td>N/A</td>
<td>N/A</td>
<td>Permanent</td>
<td>Limited</td>
<td>3</td>
<td>y</td>
<td>P</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>LOCKSS</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Limited</td>
<td>N/A</td>
<td>y</td>
<td>P</td>
<td>N</td>
<td>P</td>
</tr>
<tr>
<td>OpenICPSR</td>
<td>$6</td>
<td>$60/GB</td>
<td>10 years</td>
<td>Limited</td>
<td>6</td>
<td>y</td>
<td>P</td>
<td>P</td>
<td>Y</td>
</tr>
<tr>
<td>Portico</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Limited</td>
<td>&quot;multiple&quot;</td>
<td>y</td>
<td>P</td>
<td>P</td>
<td>Y</td>
</tr>
<tr>
<td>Preservica (Tessella)</td>
<td>$2.74</td>
<td>N/A</td>
<td>N/A</td>
<td>all</td>
<td>&quot;multiple&quot;</td>
<td>y</td>
<td>P</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>DPS - Steady Growth</td>
<td>$0.97</td>
<td>TBD</td>
<td>TBD</td>
<td>all</td>
<td>4</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>DPS - Medium Growth</td>
<td>$0.82</td>
<td>TBD</td>
<td>TBD</td>
<td>all</td>
<td>4</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>DPS - High Growth</td>
<td>$0.72</td>
<td>TBD</td>
<td>TBD</td>
<td>all</td>
<td>4</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>