Hydra/Fedora

August 4, 2014

Library IT
Brief History of YUL Digital Collections

- CONTENTdm
- Greenstone
- Custom systems (Luna, Portfolio, dbtext, Filemaker Pro, Excel, etc.)
- ODAI
- Fedora (standalone collections, e.g., AMEEL, YFAD)
Fedora is...

- Flexible Extensible Digital Object Repository Architecture
- Open Source
- Used by hundreds of organizations
- Originally developed at Cornell, now led by Fedora Project Steering Group under stewardship of DuraSpace.org
- [http://www.fedora-commons.org](http://www.fedora-commons.org)
- Currently engaged in development of Fedora 4
Hydra is...

• A Repository Solution
• A Community (25 partners now, including us)
• 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 “Heads”

- Blacklight (for viewing)
- Ladybird (de facto)
- Avalon (A/V)
- Sufia (ScholarSphere)
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
- University of Oregon
- 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
- Case Western Reserve Univ.
Benefits of ongoing 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
- Provide a consistent user experience across many collections and content types, along with discoverability
- 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)
Software Architecture
Current/FY2015 Implementation
Hydra Project
Hydra Stack

- Fedora
- Blacklight
- Ladybird
- Active Fedora
- Apache Solr
- Media Server
- Internet Archive Book Reader
- Ingest applications
Ladybird
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 Goals

• Centralize image cataloging into a single tool
  • Luna, Portfolio, DB Text, Excel, FileMaker Pro, CONTENTdm

• Provide vocabularies that could be shared across the library
  • Potential for integrating Getty vocabularies and Linked Data

• Simplify the ingest of assets into the DAM hosted by YDC2

• Migrate content off Rescue Repository

• Simplify IT Support by having One System to manage
Ladybird

- Started June 2010
- Version 1.0 December 2013
- 20 background applications
- 4 desktop applications
- 3 web applications
- C# .Net 4.0
- 575,000 lines of source code
- 2,449,839 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

- Potential partners with:
  Columbia, Princeton, MIT, Northwestern
- Release Ladybird as Hydra Head
- Collection migration 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
- Sufia – Faculty Self Archiving
- Avalon – A/V support
- Spotlight – Exhibitions
- Auditing – Statistics and Audit Trails
Preservation
Preservation Tools
“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
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
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 Services

**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
Proposed Digital Repository and DPS Architecture

Fedora

Storage Management
(Based on storage policies specified at ingest)

Hydra interface

Curation / “Ingest” Services

Disk 1
Disk 2
Tape 1
Tape 2
Cloud 1 (e.g. Glacier)

Active Preservation Interface

Bit Preservation Interface

Checksum Checking
Manage Storage policies
Manage Physical Infrastructure
Preservation Planning/Scheduling/Risk Analysis
Migration/Normalisation
Manage Emulators/Software Archive
Automated Metadata Creation/Extraction
Format Validation

Hydra interface

Hydra interface

Access Services
Proposed Basic Digital Preservation Services

▲ **Bit Preservation**
At least 4 copies, stored in at least 3 locations with different risk profiles, regularly monitored, with seamless media & software management (refreshment, replacement, etc)

▲ **Secure Storage with Managed Access**
Audited secure storage with authorized, timely access and clear exit strategies

▲ **Obsolescence Monitoring**
Identify technical characteristics of files, associate with interaction software and hardware, software and hardware obsolescence monitoring, informing content owners when content is becoming inaccessible

▲ **Provenance and Authenticity Assurance**
Logging & preserving all provenance events, ability to report on history of activities, checksum creation, independent storage and regular validation

▲ **Standards Compliance**
Compliance with ISO 14721:2012: Open archival information system (OAIS)
Reference model & with ISO 16363:2012: Audit and certification of trustworthy digital repositories
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
Infrastructure
Storage Infrastructure
Proposed FY2015

- Fedora
  - Akubra Low Level Storage Fedora Plug-in
  - Akubra Blob Store API
    - /FedoraStore
      - NFS
      - ITS NetApp Storage
        - 10TB
    - /RSS
      - NFS
      - ITS RSS Storage
        - 100TB
    - /StorHouse
      - NFS
      - Sgi StorHouse
      - Dell MD Series 7TB
        - Spectra Logic T950
          - 100TB

= new component
= new service
Proposed Trusted Edge Policy

Policy 1:
- Any object >= 0.5 GB MOVE from /FedoraStore to /RSS
- All objects COPY from /FedoraStore to /RSS

Policy 2:
- All objects COPY from /RSS to /StorHouse

Note:
- Sgi Trusted Edge Storage Management (see note)
Proposed FY2015 Staging
Proposed Staging Trusted Edge Policy

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>
Possible Future Paths

- Research Data support
- Support for A/V via Avalon
- Support for self-archiving of materials via Sufia (and later via Hydramata project)
- Active preservation tools
- Embedding content in LMS systems via LTI
- Support for exhibitions via Spotlight
- GeoBlacklight
- ORCID support
- Fedora 4 – active storage management, migration path
Yale UNIVERSITY LIBRARY