Preserving Digital Content at Scale

CNI
Albuquerque, NM
April 4, 2017

Carolyn Caizzi
Head of Repository and Digital Curation
Northwestern University Libraries
State of the Northwestern Libraries’ Local Digital Content
Data Storage

Isilon EMC2 at 2 data centers
Daily syncs & 45-day snapshots
Tape backups
Different geographic region
## How Much

<table>
<thead>
<tr>
<th>Date</th>
<th>Actual TB</th>
<th>Increase TB</th>
<th>Percentage Increase from previous year</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/2010</td>
<td>2</td>
<td>2</td>
<td>100%</td>
</tr>
<tr>
<td>9/2011</td>
<td>21</td>
<td>19</td>
<td>950%</td>
</tr>
<tr>
<td>9/2012</td>
<td>38</td>
<td>17</td>
<td>81%</td>
</tr>
<tr>
<td>9/2013</td>
<td>75</td>
<td>37</td>
<td>97%</td>
</tr>
<tr>
<td>9/2014</td>
<td>125</td>
<td>50</td>
<td>67%</td>
</tr>
<tr>
<td>9/2015</td>
<td>200</td>
<td>75</td>
<td>60%</td>
</tr>
<tr>
<td>9/2016</td>
<td>267</td>
<td>67</td>
<td>34%</td>
</tr>
</tbody>
</table>
What makes up the 267 TB

- Preservation and Mezzanine Master Files managed by Fedora repository: 52 TB
- Streaming Derivatives A+V Repository: 19 TB
- Digitization & Born Digital Project Files Manually Managed: 172 TB
- Misc. and Outdated Content: 24 TB
What makes up the 267 TB

• Preservation and Mezzanine Master Files managed by Fedora repository: 52 TB

• Streaming Derivatives A+V Repository: 19 TB

• Digitization & Born Digital Project Files Manually Managed: 172 TB

• Misc. and Outdated Content: 24 TB
Repository (ies)
Active Repository(ies)

- Images Repository
- A+V Repository
- Self-Deposit IR “Arch”
- Books Repository
- Output of digitization & born digital projects
Inactive Repositories and Websites

• Special collections websites
• Faculty project websites
• Books repository
Digital Preservation Scorecard
# Active Preservation

<table>
<thead>
<tr>
<th>NDSA Levels</th>
<th>Storage &amp; Geo Location</th>
<th>File Fixity &amp; Data Integrity</th>
<th>Info Security</th>
<th>Metadata</th>
<th>File Formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protect</td>
<td>B</td>
<td>F</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>Know</td>
<td>A</td>
<td>F</td>
<td>C</td>
<td>F</td>
<td>A</td>
</tr>
<tr>
<td>Monitor</td>
<td>C</td>
<td>F</td>
<td>F</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Repair</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>C</td>
</tr>
<tr>
<td>NDSA Levels</td>
<td>Storage &amp; Geo Location</td>
<td>File Fixity &amp; Data Integrity</td>
<td>Info Security</td>
<td>Metadata</td>
<td>File Formats</td>
</tr>
<tr>
<td>------------</td>
<td>------------------------</td>
<td>-----------------------------</td>
<td>--------------</td>
<td>----------</td>
<td>--------------</td>
</tr>
<tr>
<td>Protect</td>
<td>B</td>
<td>F</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>Know</td>
<td>A</td>
<td>F</td>
<td>C</td>
<td>F</td>
<td>B</td>
</tr>
<tr>
<td>Monitor</td>
<td>C</td>
<td>F</td>
<td>F</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Repair</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>C</td>
</tr>
</tbody>
</table>
Tipping Point
Next Generation

Before and Now
Before: Staffing and Structure

2 Departments

• **Digital Collections**: production staff; project managers; product owners

• **Enterprise Systems**: developers; sysadmins; devops; tech support

• **Missing**: metadata coordinator for digital content; preservation focused positions
Now: Staffing and Structure

Library Wide Re-Organization 2014-15

- 2 new AULs; 1 new department head
- 1 department: Repository & Digital Curation
- Digital preservation & digital curation librarians; 6 developers with 1 focused on digital preservation; metadata coordinator
Better Management

• Published a library digital preservation policy
• Created an inventory of our Isilon shares
• Created plans for migration of legacy content
Before: Repository Architecture

• Focus on showing content, not on management or digital preservation actions
• Based on format or for a specific collection
• Added on functionality through time
Now: Repository Architecture

Current project to **re-imagine** our repository

- Hyrax/Fedora 4 with focus on digital preservation & staff management tools
- De-couple front-end from staff tools
- Bring together objects of all formats
- Focus on batch import and batch editing
- Multiple tiers for storage, including sending most valued content to DPN
Now: IMLS Planning Grant
Lg-72-16-0135-16

• How does one curate objects to ingest into a long-term dark preservation system?

• How does versioning of objects and metadata play out in a long-term dark preservation systems and how to automate these actions?

• How can systems that store data differently be made more interoperable?
Now: Enterprise Level Digital Preservation Strategy

- Consultant engagement: digital preservation assessment of Northwestern

- Recommendations will be presented to the Provost this spring

- Recommendations include: firm organizational structure; infrastructure; education of PIs, faculty, and graduate students
Summary: Planning for More Active Preservation

+ Strategic re-imagining of vacancies
+ Review policies: hit publish
+ Storage solutions: tier
+ Repository architecture: staff tools & decoupled front end design
+ Workflows: start fresh & design with batch in mind
Thank You!

carolyn.caizzi@northwestern.edu