Digital Preservation in Production

CNI Spring 2017
Who are we?

Bill Branan
Services Technical Director
DuraSpace
bbranan@duraspace.org

Dave Pcolar
Chief Technology Officer
DPN
dave@dpn.org
What is DPN?

- Network of distributed preservation systems (nodes)
- Built for truly long-term preservation
- Both technical and legal

Mission: Create a system that ensures the survival of the digital scholarly record for the next generation
DPN Nodes

- DuraCloud Vault
- Texas Digital Library
- Stanford Digital Repository
- APTrust
- HathiTrust
Moving into Production

- **Legal**
  - Member Deposit Agreement
  - Node Service Agreement

- **Technical**
  - Node software - version 1 and 2
  - DuraCloud Vault deposit workflow

- **Operational**
  - DPN-Node demarcation and ingest
Deposit in 7 easy steps

1. Become a DPN member
2. Sign Deposit Agreement
3. Create DuraCloud Vault account
   a. No additional fees
   b. No additional agreements
4. Attend DuraCloud Vault orientation
5. Select content to deposit
   a. 5 TB per year is included in membership
6. Transfer content into DuraCloud Vault space
7. Submit Snapshot
### Snapshot Detail

**vault_1063_sample-collection-1_2017-03-29-20-42-59**

<table>
<thead>
<tr>
<th>Details</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>Sample Snapshot 1</td>
</tr>
<tr>
<td><strong>Alternate IDs</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Snapshot Date</strong></td>
<td>Wed Mar 29 2017 16:42:59 GMT-0400 (Eastern Daylight Time)</td>
</tr>
<tr>
<td><strong>Source Host</strong></td>
<td>vault.duracloud.org</td>
</tr>
<tr>
<td><strong>Source Space</strong></td>
<td>1003</td>
</tr>
<tr>
<td><strong>Source Collection</strong></td>
<td>sample-collection-1</td>
</tr>
<tr>
<td><strong>Status</strong></td>
<td>WAITING_FOR_DPR</td>
</tr>
<tr>
<td><strong>Preservation Network Member ID</strong></td>
<td>3164687-5925-4ca-e036-R5326f42226b</td>
</tr>
<tr>
<td><strong>Content Item Count</strong></td>
<td>251</td>
</tr>
<tr>
<td><strong>Total Size</strong></td>
<td>1.1 MB (113326 bytes)</td>
</tr>
</tbody>
</table>

**Snapshot History**

<table>
<thead>
<tr>
<th>Date</th>
<th>Property</th>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>snapshot-action</td>
<td>Value</td>
<td>snapshot-initiated</td>
</tr>
<tr>
<td></td>
<td>initiating-user</td>
<td><a href="mailto:library@DuraSpace.org">library@DuraSpace.org</a></td>
<td>vault_1063_sample-collection-1_2017-03-29-20-42-59</td>
</tr>
</tbody>
</table>
Deposits per month

- GB
- Bags

Month

4-2016 to 1-2017
Active DPN Members using DuraCloud Vault

- Indiana University - Bloomington
- University of Iowa
- University of New Mexico
- Figshare
- Iowa State University
- University of Tennessee
- Indiana University Purdue University Indianapolis
- University of Kentucky
- California Institute of Technology
- University of Washington
- Brigham Young University
- Kansas State University
- Pennsylvania State University
- State University of New York at Buffalo
- Tufts University
- University of California Los Angeles
- University of California San Diego
- University of Illinois at Chicago
- University of Nebraska
- University of Wisconsin
- Utah State University
- University of Utah
- Vanderbilt University
- University of Kansas
Deposit Best Practices

1. Start early
2. Plan content deposit strategy
3. Prefer stable content sets
4. Group files into collections
5. Avoid very small snapshots (< 50 GB) where possible
6. Avoid very large snapshots (> 1 TB) where possible
   a. Snapshot size is not limited by DPN bag size
7. Use archive files (tar or zip) to group very small files, but otherwise use sparingly
   a. Archive files over the bag size limit must be split
Collections

- University of Iowa - Van Allen Belt Discovery
  - Original data tapes from the Explorer Missions

- University of Kentucky - Civil Rights History
  - Robert Penn Warren - oral history interviews

- Indiana University Purdue University Indianapolis (IUPUI)
  - Indianapolis Motor Speedway Collection
Challenges and Lessons Learned

1. Legal and structural framework is just as important to preservation as technology
2. Selecting the highest priority content is difficult
3. Heterogeneous systems: preservation benefits, coordination challenges
4. Data transfer takes time
5. Large data sets
6. System upgrades and software releases
7. Metadata creation and workflows
What is next?

- Expanding Membership
- Expanding Ingests
- Member/Vendor workflow automation
- Large Transfer Optimization
- Pilots
  - Bags over 250GB
  - Restoration Workflows
  - Peer to Peer Transfers - Heavy Users
  - Transaction Audits and Reporting
Questions?

Bill Branan
Services Technical Director
DuraSpace
bbranan@duraspace.org

Dave Pcolar
Chief Technology Officer
DPN
dave@dpn.org