THE DARK SIDE OF DIGITAL PRESERVATION:

Distributed Digital Preservation

Sibyl Schaefer, Mary Molinaro, Katherine Skinner, Chip German, David Pcolar
Sam Meister

Archives Records 2016
Session 606
Saturday August 6, 2016
Atlanta, Georgia
Briefing for SAA, August 6, 2016

Chip German, Program Director, APTrust and Senior Director, Content Stewardship, at the University of Virginia Library
chip.german@aptrust.org

APTrust Web Site: aptrust.org
Sustaining Members:

Columbia University  University of Connecticut
Indiana University  University of Maryland
Georgetown University  University of Michigan
Johns Hopkins University  University of North Carolina
North Carolina State University  University of Notre Dame
Pennsylvania State University  University of Virginia
Syracuse University  Virginia Tech
University of Cincinnati
Quick description of APTrust:

- Provides simple, collaboratively designed and built preservation environment (bit-level with fixity checking every 90 days) for the digital scholarly and cultural record (NOW)
- Will provide collaboratively developed services for that content (FUTURE), which may include:
  - choice of preservation assurance levels/cost plans
  - access
  - format migration, emulation, software preservation
  - analysis tools for research
- Currently uses Amazon Web Services (as back-end platform only)
Ingest Services

Three content copies in East Coast data center (three separate availability zones)

S3 service (preservation)

for metadata and events
Three content copies in West Coast data center (three separate availability zones) + Fedora copy
Money

- Sustaining Members pay $20,000 per year in dues
  - Engage in governance, strategic direction-setting, development, etc.
  - Get allocation of 10 TB of content (with replication, total of 60 TB)
- University of Virginia provides $380,000 per year toward staffing costs
- Institution wishes to deposit more than 10 TB of content?
  - Pass-through of incremental costs in 5 TB blocks of content
    - $2,750 per year
  - Approximately $600,000 cash reserve
  - Serves as DPN ingest/replicating node (full cost-recovery from DPN)

Current metrics in our 2nd year of production

- As of August 2, we have 64,111 content objects from 8 institutions in production environment
- 17.2 TB of content, 3 million (as of June) PREMIS events
- Additional institutions with content in current testing
Future
● Much more volume in deposits
● Develop collaborations for additional services
● Develop sophisticated reporting for depositors
● Services for smaller cultural heritage organizations, libraries, museums; new categories of APTrust members

Challenges
● **Drive costs downward** to encourage preservation
  ■ Balance between preservation principles and cost (if not, just storage)
  ■ Technical architecture enables interchangeable back-end platforms
● Balance between
  ■ importance of Trusted Digital Repository certification
  ■ cost and workload of TDR audit (potentially $10-$15K every three years, although Sibyl is educating me about this)
APTrust as a DPN ingest and replication node

- APTrust members can designate their deposit to DPN (individually sign a separate agreement with DPN)
- APTrust “unbags” a verified copy of deposit from APTrust bag and “rebags” it to DPN standard into AWS Glacier service in the AWS VA data center (where three copies are kept, each in a separate availability zone)
- APTrust then collaboratively replicates the data to two other DPN replication nodes
CHRONOPOLIS

Sibyl Schaefer // @archivelle
University of California, San Diego
Digital preservation dark storage network spanning multiple institutions and geographic regions.

Partners:

University of California, San Diego (UCSD)

National Center for Atmospheric Research (NCAR)

University of Maryland Institute for Advanced Computer Studies

Chronopolis Network Partners
● Funded by a series of NDIPP grants designed to develop the technological infrastructure needed for digital preservation.

● First ingest date: 2008

● Nodes are all non-profit research centers.

● Content-agnostic - depositors form SIPS for ingest, Chronopolis runs minimal packaging processes on data

● Majority of data is UCSD digital library and research data
Focused on *active* preservation – constant checking of items.

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Type</th>
<th>Total Files*</th>
<th>Last Audit</th>
<th>Next Audit</th>
</tr>
</thead>
<tbody>
<tr>
<td>rontest</td>
<td>local</td>
<td>1</td>
<td>Apr 30 2015</td>
<td>Unknown</td>
</tr>
<tr>
<td>acadis</td>
<td>local</td>
<td>12215 / 125.7 GB</td>
<td>Jun 03 2015</td>
<td>Unknown</td>
</tr>
<tr>
<td>acadis_database_02-02-2013</td>
<td>local</td>
<td>6</td>
<td>Jun 03 2015</td>
<td>Unknown</td>
</tr>
<tr>
<td>acadis_datasets_01-31-2013</td>
<td>local</td>
<td>12209</td>
<td>Jun 03 2015</td>
<td>Unknown</td>
</tr>
</tbody>
</table>
DuraCloud as a pathway to Chronopolis
● Provides an existing hosted user interface, reduced need for new development
● Simplifies the process of moving content in and out of the systems
● Shared institutional values

Chronopolis as a storage provider for DuraCloud
● Extends the DuraCloud network to include a non-commercial, highly distributed, dark archive option

DuraCloud Vault
● Chronopolis + DuraCloud + DPN
Chronopolis and DPN

Same:
- dark storage
- 3 copies at geographically different nodes

Different:
- business model
- stewardship succession
- administration
- membership involvement
DPN is a 50+ member cooperative organization investing in long-term, scalable digital preservation for the academy.
As a project of Internet2, DPN is able to leverage established Research and Education Networks to facilitate large scale data movement into preservation environments across the United States.
Why DPN?

- Despite institutions’ individual efforts only a fraction of what should be preserved is being preserved.

- How will we ensure that the digital assets created and collected in our institutions are available to scholars in the future?

- A solution is needed to save at risk content in a way that protects against technical failure, natural disaster, and institutional failure.
What does a solution look like?

1. Geographically distributed
2. Content replicated
3. Audit and repair
4. Technological diversity
5. Meet best practices for repositories
6. Succession agreements
The business model

Currently:
- Members pay an annual fee of $20,000
- Members receive 5TB of preservation storage per year
- Pay once / stored in perpetuity
- Members can purchase additional TB

Now working on solutions for smaller and larger institutions - available next year
DPN is in production and accepting deposits!
What is next?

- Address problems we know about
  - getting started
  - workflow at the local level
- Continuing to examine cost
- Increase technical diversity within DPN
  - evaluating new nodes
- Open the door to new members
Meeting the DPN Mission

- To save the most valuable content from the academy in a dark preservation system that is geographically dispersed and is supported by replication and audit.

- The content is protected by succession agreements made at the time of deposit.

- This content is being preserved for humanity.
Preserving the historical record for this and future generations.

Learn more about our Network

The Digital Preservation Network

The Digital Preservation Network (DPN) is the only large-scale digital preservation service that is built to last beyond the life spans of individuals, technological systems, and organizations. DPN provides members of the academy and their successors with assurance that future access to their scholarly resources will be available in the event of

*DPN is filling a vital strategic need for the long-term preservation of scholarship and research like no other service or program can. DPN is key to secure the intellectual capital of our
Mary Molinaro  
Chief Operating Officer  
mary@dpn.org

Dave Pcolar  
Chief Technical Officer  
dave@dpn.org
MetaArchive Cooperative: Case Study in Collaboration

Sam Meister
Educopia Institute

www.metaarchive.org
Why collaborate?
To preserve our digital collections
Maintain integrity for continued access
To preserve our institutional missions
Outsourcing core mission is dangerous
Collaboration Benefits

Cost effective

Scalable

Transparent

Control

Responsive

Standards-based
MetaArchive History

- 2004 - Founded as part of NDIIPP funding and activities
- 2006 - Educopia Institute founded to serve as administrative home for MetaArchive
- 2007 - MetaArchive becomes initial Affiliated Community
- 2014 - Celebrated 10 years of successful preservation network
EDUCOPIA INSTITUTE

Community-building

Network management expertise

Administrative, legal, and financial services
Hallmarks

**Distributed** digital preservation

Institutions maintain *control* over their own content

Preservation as a *process*, not a push-button exercise

*Simplicity* in ingest, management
Cooperative Preservation

MetaArchive is a *cooperative*, not a vendor:

All hardware and software assets are owned by members

Membership fees and storage fees go to a central pool of support for members’ co-op activities
Philosophy in Practice

- Compatible with any repository system
  - E.g., Dspace, Fedora, Archivalware, ETD, CONTENTdm, BePress, Digital Commons, etc

- Member institutions determine their own curatorial practices

- MetaArchive is a community of support to help them make informed decisions
MetaArchive Practices

Basic processes

“Producer” (OAIS) determines curation practices; brings SIPs to MetaArchive
MetaArchive Practices

Basic processes

“Producer” (OAIS) determines curation practices; brings SIPs to MetaArchive

Multiple copies of AIPs dispersed across geographical, political, and environmental lines
MetaArchive Practices

Basic processes

“Producer” (OAIS) determines curation practices; brings SIPs to MetaArchive

Multiple copies of AIPs dispersed across geographical, political, and environmental lines

Checks and repairs automated across network
MetaArchive Practices

Basic processes

“Producer” (OAIS) determines curation practices; brings SIPs to MetaArchive

Multiple copies of AIPs dispersed across geographical, political, and environmental lines

Checks and repairs automated across network

Deaccession cycle versus data deletion
Membership

- Auburn University
- Boston College
- Cal Poly San Luis Obispo
- Carnegie Mellon University
- Consorci de Biblioteques Universitaris de Catalunya
- Florida State University
- Isabella Stewart Gardner Museum
- Greene County Public Library
- HBCU Library Alliance
- Indiana State University
- Oregon State University
- Penn State University
- Pontificia Universidade Catolica do Rio de Janeiro
- Purdue University
- Rockefeller Archives Center
- University of Louisville
- University of North Texas
- University of South Carolina
- Virginia Tech University
Membership Levels

Collaborative members: $2.5K/year

Preservation members: $3K/year

Sustaining members: $5.5K/year

Server cost: <$5K/term
Storage cost: $585/TB/year
Membership Responsibilities

• Undertake a **3-year** membership term
• Take responsibility for content preparation, evaluation, staging, and ingest testing
• *Monitor* collections to ensure accurate long-term preservation
• Host and maintain a MetaArchive cache (server) or pay in a technology support fee
• Join and contribute to Committees
Governance Model

Independent

Nonprofit

Member owned, operated, governed
Governance Model

Charter
Outlines mission, goals, and organizing principles of cooperative

Governance Procedures
Outlines roles and responsibilities for officers and member participants

Member agreements
Outlines terms and responsibilities of membership
Leadership

Steering Committee

Primary governing body responsible for overall management, coordination, communication, and reporting efforts

One representative from each Sustaining Member

Strategic direction-setting, policy setting, including membership costs and storage fees
Community Building

Common, neutral center

Distribution of work

Community of engagement
Building knowledge
Accomplishing preservation

Concentrated effort toward unified goals
Community Building

Communicate!

Monthly discussion-topic focused community calls

Two Steering Committees per year

Community Listserv

MetaArchive Wiki
Sustainability

Cooperative framework

Strong organizational center

Limited dependence on any one member

Collaborative model for long-term preservation

Geographic diversity/distribution

Expertise diffusion

Maintain cost-effective, in-house options
Thank you!

www.metaarchive.org
@metaarchive
sam@educopia.org
@samalanmeister
Preservation and Community Action
QUESTIONS?
Questions?
QUESTIONS?
Questions?
questions?