What happens if an archive fails?
What happens if an archive fails?
The Problem

Preservation repositories offer long-term stewardship environments for valued information. But what happens when they fail?

- Can another organization take the content? Or will it be lost?
- If another organization takes the content, will they have the necessary rights and control to preserve and provide access to it?
What is DPN?

- A federation of networked preservation repositories, with diverse...
  - Software architectures
  - Organizations
  - Geographic regions
  - Political environments (longterm)
- Provides bit-level preservation of dark content
- Succession option for archives
Benefits

- Resilience
  - DPN serves as the “archive for archives”
  - Safety net for failure of any single preservation system due to technical, economic, legal or catastrophic causes

- Succession
  - Contracts among DPN Members ensure other nodes in the network can legally brighten content if orphaned
Current DPN Members

Arizona State University
Brown University
California Digital Library
California Institute of Technology
Columbia University Library
Cornell University
Dartmouth College
Duke University
Emory University
Harvard University
Indiana University Bloomington
Iowa State University
John D. Evans Foundation
Johns Hopkins University Libraries
Kansas State University
MIT
Michigan State University
New York University
North Carolina State University
Northwestern University
Ohio State University
Pennsylvania State University
Purdue University
Stanford University
Syracuse University
Texas Digital Library
Texas Tech University
Tufts University
Tulane University
University of Alabama
University of Arizona
University of California, San Diego
University of Chicago
University of Florida
University of Illinois at Chicago
University of Illinois Urbana-Champaign
University of Iowa
University of Kansas
University of Kentucky
University of Maryland
University of Miami
University of Michigan
University of Minnesota
University of Nebraska-Lincoln
University of North Carolina
University of North Texas
University of Notre Dame
University of Tennessee
University of Texas
University of Virginia
University of Washington
University of Wisconsin
Vanderbilt University
Virginia Tech
Yale University
Initial DPN Replication Nodes

- Stanford Digital Repository
- HathiTrust
- Chronopolis
- University of Texas Digital Repository
- Academic Preservation Trust
Some Definitions...

• The **First Node** initiates replication and controls changes in registry/database holding information regarding content that it is the First Node for

• A DPN First Node is the authoritative node for its content

• **Replicating Nodes** replicate content from First Nodes.

• All Nodes have a registry that holds information about content held by DPN, including which node is the First Node for content and which nodes are replicating nodes.

• A **DPN identifier** will be unique across all of DPN nodes

• Nodes that are First Nodes can also be Replicating Nodes

• **Member Nodes** contribute to DPN via First Nodes
Current tasks

- Define requirements
  - E.g. metadata, messaging format, packaging, auditing, transfer
  - Mostly done, but need confirmation from leadership team

- Design initial system – High level done
  - Many details need to be ironed out
  - Federated Messaging using AMQP
  - Transfer using Rsync/HTTPS and transfer area
  - Registry entries yet to be defined
  - DPM UUID and Fixity need to be defined to minimize impact on archives
  - Determine modality of versioning, is it possible for all archives?

- Prototype this Winter

- Begin testing
  - Starting with Messaging and transfer protocols
  - Spring 2013
SDR to DPN Node Process Flow

SDR

Move to Transfer area

Notify

Confirm

DPN Transfer Sync

Pull/Push

Validate

update

Registry

Logs

DPN Staging area

AIP

DPN

AIP

AIP

DPN

DPN
### DPN registry Entry (proposed)

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUID</td>
<td>DPN Universal Unique Identifier</td>
</tr>
<tr>
<td>DOMAIN</td>
<td>Repository Domain identifier of First Node e.g. SDR</td>
</tr>
<tr>
<td>Fixity</td>
<td>Digital Object Fixity Information</td>
</tr>
<tr>
<td>Copy Domain</td>
<td>Domain of current copy, may be different than First Node. If different than first node, indicates copy at site is a replicating copy.</td>
</tr>
<tr>
<td>Date/Timestamp</td>
<td>Time and Date that copy was made at the current replicating node.</td>
</tr>
<tr>
<td>Brighting Metadata UUID</td>
<td>DPN UUID of DO that holds technical metadata data useful for brightening the DO in this Registry entry.</td>
</tr>
<tr>
<td>Version</td>
<td>Current version of the DO represented by the UUID.</td>
</tr>
<tr>
<td>Version UUID’s</td>
<td>All of the UUID’s associated with this DO, in order.</td>
</tr>
<tr>
<td>Rights</td>
<td>Right pertaining to current copy at this Domain, i.e. the holding copy.</td>
</tr>
<tr>
<td>Encryption Metadata</td>
<td>Information pertaining to encryption, including for example, key management, escrow holder, digital signatures, algorithms, etc.</td>
</tr>
<tr>
<td>TBD</td>
<td></td>
</tr>
</tbody>
</table>
Systems architecture showing Clustering and Federation (RabbitMQ)

Clustered Messaging Brokers with Heartbeat

Broker

Broker

Clustered brokers share state and are highly available.

Contain Upstream Exchange(s)

Clients can be producers and consumers

Federated Exchange publishes to Upstream Exchanges

Federated exchanges link to other exchanges, called upstream exchanges. Messages published to upstream exchanges are copied to the federated exchange as though they were published directly to it.

All configuration is in the broker of the federated exchange.

Client

Client

Client

Client

Client

Client

Client

Client

Client

Client
Difficult questions

- Brightening data after succession
- Versioning across different architectures
- Auditing DPN nodes’ operations
- Managing (and smoothing) cost differences across heterogeneous replicating nodes
- Encryption???