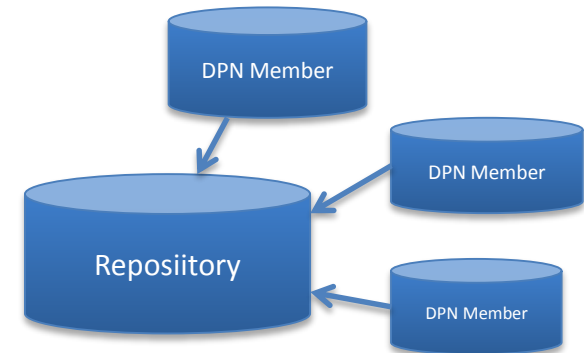# The Digital Preservation Network at UT Austin
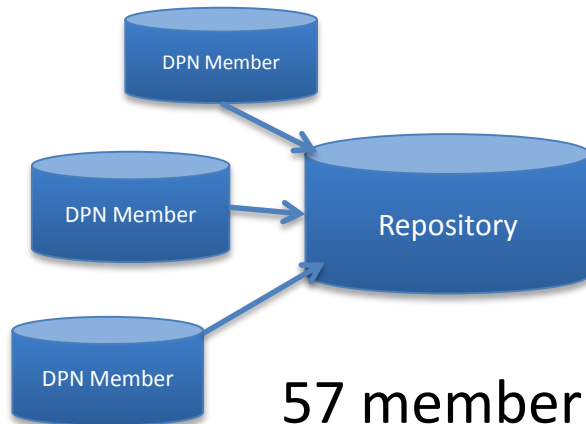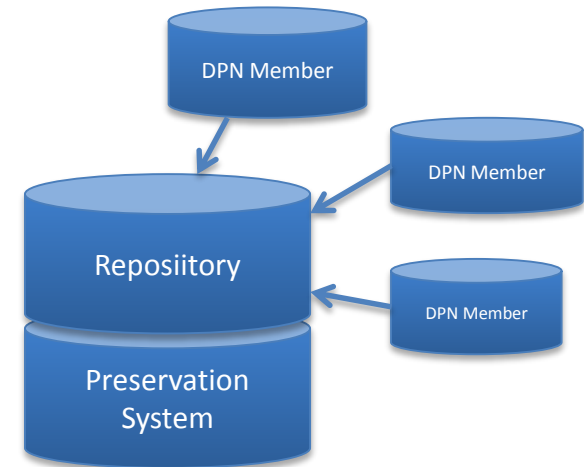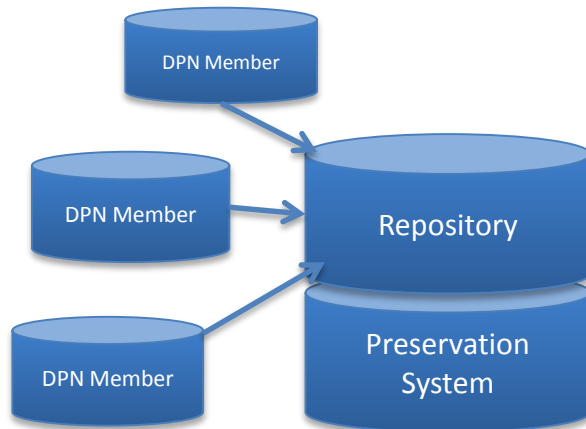
Chris Jordan

Texas Advanced Computing Center

What Is DPN?
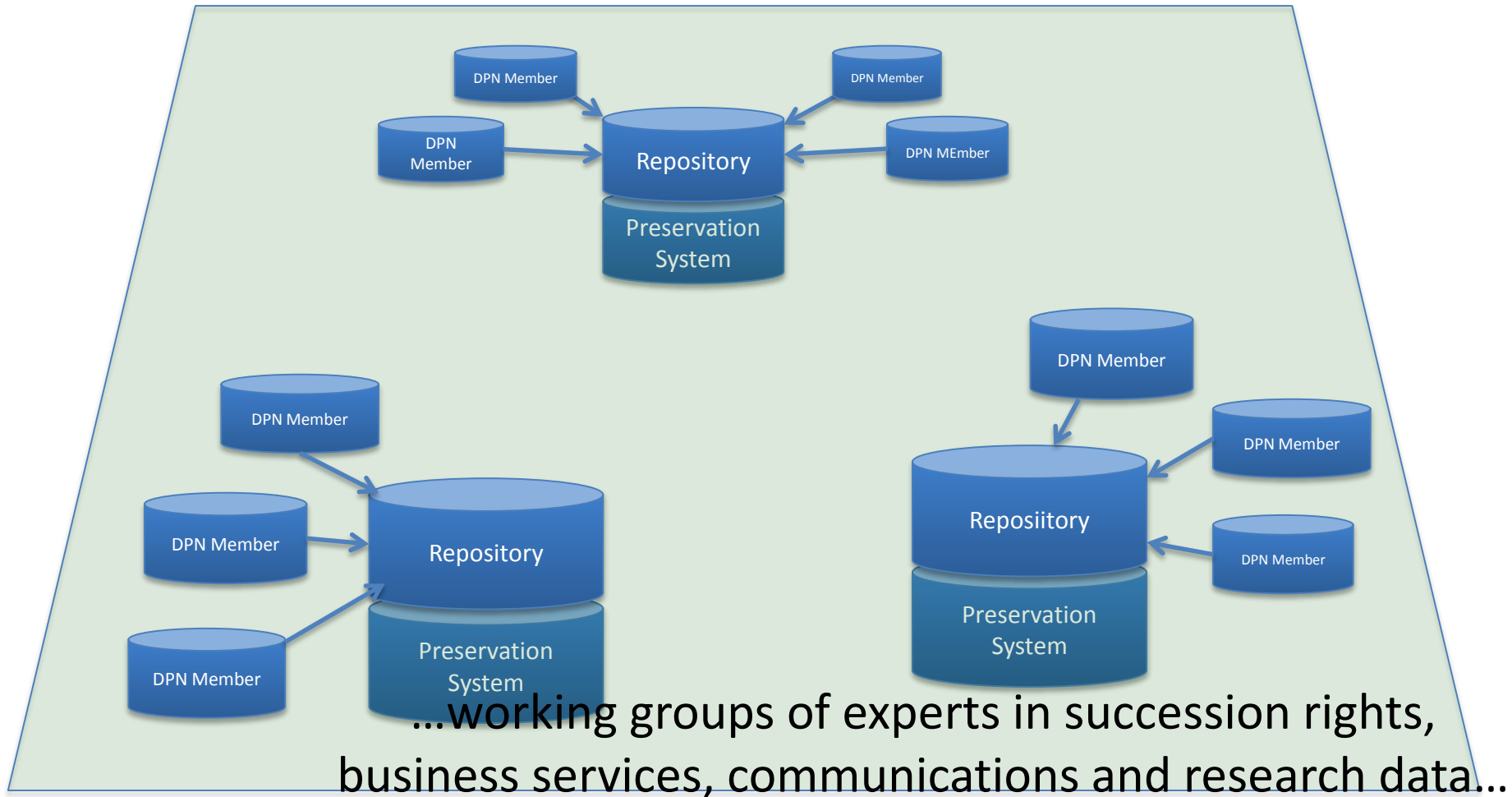

57 member organizations cooperatively investing in long-term, scalable, digital preservation

What Is DPN?
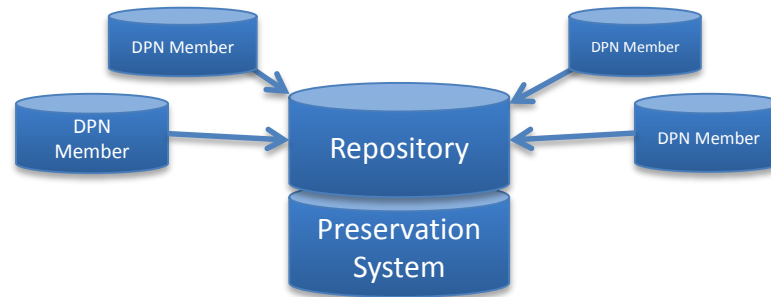
technical staff and systems from 5
large scale preservation repositories

What Is DPN?

...working groups of experts in succession rights, business services, communications and research data...

What is DPN?
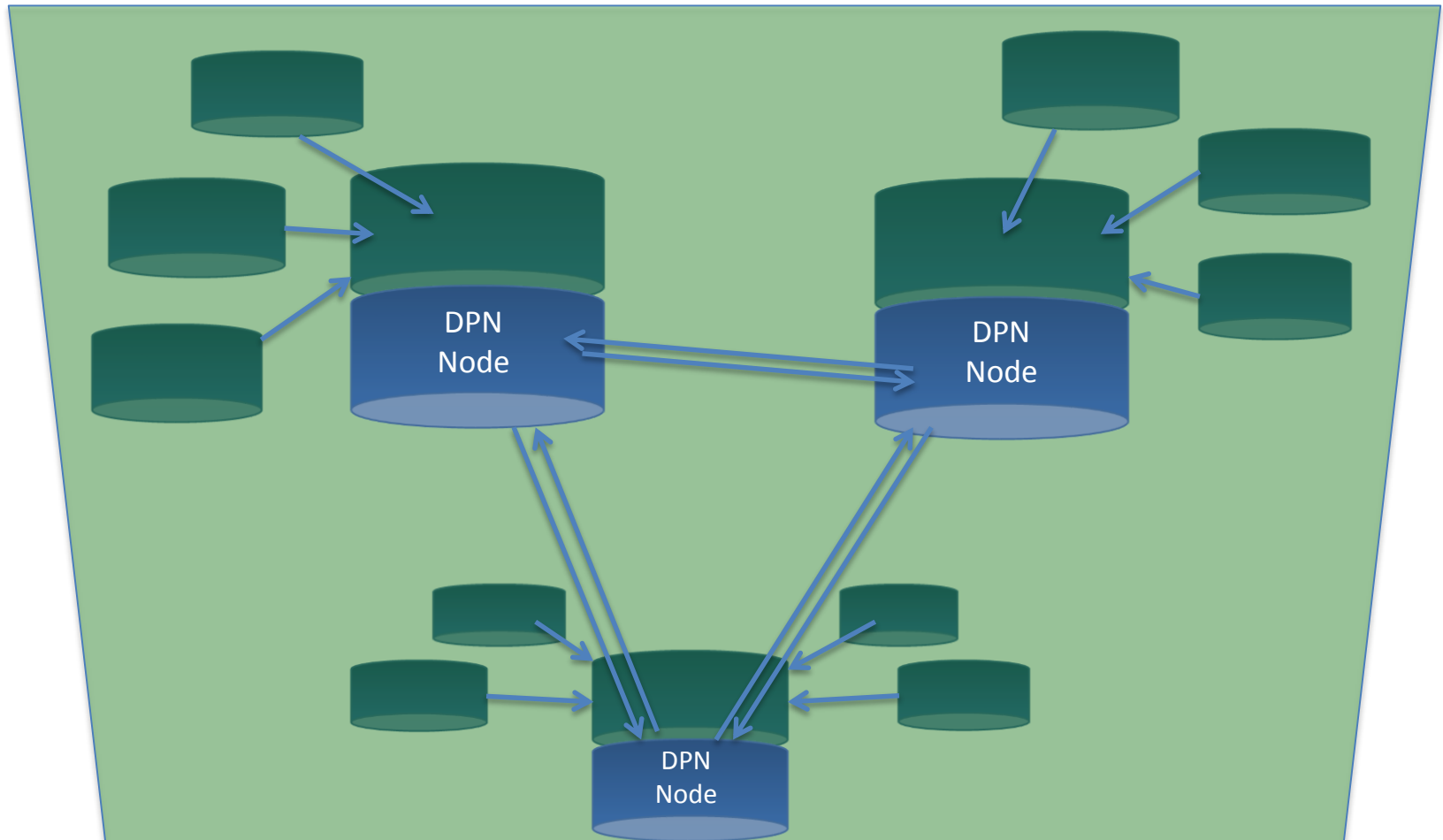
All building a digital preservation backbone for the academy

What Does DPN Do?

1. Establishes a network of heterogeneous, interoperable, trustworthy, preservation repositories (Nodes)

2. Replicates content across the network, to multiple nodes

3. Enables restoration of preserved content to any node in the event of data loss, corruption or disaster

4. Ensures the ongoing preservation of digital information from depositors in the event of dissolution or divestment of depositors or an individual repository

5. Provides the option to (technically and legally) "brighten content" preserved in the network over time

Initial DPN technical partners

Initial DPN launch will feature five nodes:

- Academic Preservation Trust (APTrust)

- Chronopolis

- HathiTrust

- Stanford Digital Repository (SDR)

- University of Texas Data Repository (UTDR)

And a participating partner:

- DuraSpace

DPN, UT and TDL

- TACC & TDL have an established partnership
- TACC also collaborates with UT Library on:
  - Data Management Planning
  - Local research support
  - HPC for Digital Libraries
- DPN extends these efforts to include design and implementation of a replicating node

What is UTDR?

- UT Research Cyberinfrastructure Initiative

- Supports all 15 UT System Schools with:
  - High Performance Computing
  - 10Gb Research Network
  - 5PB Replicated Data Repository

- Deployed in early 2012, now over 100 investigators, 100s of users, over 1PB allocated

TACC Capabilities

- Corral UTDR System – 5PB, geographically replicated online storage

- iRODS Data Management, Databases, Web applications

- Ranch – 100PB+ Tape Archive capacity

- Additional data-intensive systems this year

- Stampede/Lonestar/Longhorn
  - World-Class Supercomputing and Visualization

DPN Network Concepts

- "First Nodes" submit data packages

- "Replicating Nodes" hold copies of data

- Messaging framework and Registry track data submissions and replicas

- "Bags" are used to package data for preservation – contents are opaque to DPN

- Each node provides its own interfaces

DPN Design Principles

- Nodes should be as independent as possible

- Content owners should have control over format of data

- Network should be flexible – easy to add and remove nodes

- Diversity of implementation is crucial to successful long-term preservation

Components in Technical Architecture

- Messaging infrastructure to support federated services
- Registry to track objects within the federation, including copies, version, rights, brightening information
- Transfer mechanisms (rsync, https, gridFTP, etc.)
- Private PKI for securing transport layers
- Logging and reporting
- Other components we implement separately, but may be common, for example a secure transfer area.
- DPN objects that hold administrative content such as DPN framework agreements, DPN bagit profiles, versioned Brightening information for a collection/repository

TDL and DPN

- In DPN terms, TDL is a content provider and "first node"

- TDL retains primary responsibility for data

- DPN provides a backup function for institutional, technical, or other failures

- TACC provides storage for both TDL and DPN
  - Data packages will be separate
  - Content packaging will be different

UT DPN Implementation

- UT Library, TACC have significant presence in DPN leadership teams

- Participation in technical, sustainability, other DPN working groups

- Library will provide interfaces to TDL and other local repositories

- TACC will provide back-end storage and other implementation components

Other Repositories and DPN

- DPN is effectively a "dark archive"
- Repositories still must have their own solutions for access/data management/etc
- But DPN can provide preservation functions
- If you are a DPN member and can generate "bags" you can deposit data into DPN
- Many institutions are already DPN members
- Membership is open but fee-based

The DPN Technical Team

APTrust

Scott Turnbull

Tim Sigmon

Adam Soroka

Chronopolis

David Minor

Mike Smorul

Don Sutton

DuraSpace

Andrew Woods

HathiTrust

Sebastien Korner

Bryan Hockey

Stanford

Tom Cramer

James Simon

Texas Data Repository

Ladd Hanson

Christopher Jordan