A Consortial Model for Research Data Services Using Dataverse

Kristi Park, Director, Texas Digital Library
Ryan Steans, Assistant Director, Texas Digital Library

On behalf of the TDL Dataverse Implementation Working Group
Outline

Background
Evaluation and Selection of Dataverse
TDL’s Dataverse Implementation
Pilot Testing
Lessons Learned and Future Plans
The **Texas Digital Library** is a consortium of Texas higher education institutions that builds capacity for preserving, managing, and providing access to unique digital collections of enduring value.
Every member is unique.

Nativi, Sarah. "The fact that one of the girls I assist for came to princess day as a hotdog made my week" May 10, 2016. Tweet.
Technology hosting on shared infrastructure

- 23 DSpace repositories
- 17 Vireo installations (ETD submission, management, and publication)
- ~40 open access journals (Open Journal Systems)

- DPN’s Texas Preservation Node (with UT Austin)
- Open source DuraCloud
Texas Research Data Repository

- Compliance with mandates.
- Recognition for researchers.
- Support for the evolving academic library.
- Better research
Selection of Dataverse
TDL Data Management Working Group

**Charge**: Help the TDL determine what kinds of data management services it could provide at a consortial level.

- Develop criteria
- Evaluate proposed projects
- Investigate issues
- Document findings
- Make recommendations for services

Bruce Herbert, Texas A&M (Chair)

Maria Esteva (TACC); Colleen Lyon (UT Austin); Jeremy Donald (Trinity); Martha Buckbee (UT Southwestern); Christie Peters, Santi Thompson (UH); Kristi Park, Ryan Steans (TDL)
Researcher Use Cases

**Title:** Researcher needs to make their research data publicly available

**Primary Actor**
Primary actors may include PIs of federally funded research, researchers working on unfunded research or funded research with no retention requirements, and graduate students working on theses, dissertations, or other data-generating projects.

**Title:** Researcher needs a virtual research environment to share active data, which may or may not be publicly accessible, within a prescribed collaborative network

**Primary Actor**
Researchers involved in collaborative networks.

**Title:** Researcher seeks data to (re)use

**Primary Actor**
Researcher is interested in conducting a meta study reusing data developed in earlier studies.
<table>
<thead>
<tr>
<th>ID</th>
<th>Function</th>
<th>Use Case #</th>
<th>Evaluation Factor</th>
<th>How important is this feature? (Average Score 0-3)</th>
<th>How well does the system perform this function? (Average Score 0-3)</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Ingest</td>
<td>1</td>
<td>Upload -- the system offers a simple ingest option for user</td>
<td>3</td>
<td>2.5</td>
<td>Platform offers user the ability to drag and drop files from their desktop. Unclear how it would interact with other file destinations (including drop box).</td>
</tr>
<tr>
<td>1.2</td>
<td>Ingest</td>
<td>1</td>
<td>Controlled vocabulary -- the system provides users with standardized lists of terms to describe their data (using drop down menus or other interfaces)</td>
<td>2.33</td>
<td>1.4</td>
<td>Controlled vocabulary terms are offered only as a broad list at the subject level.</td>
</tr>
<tr>
<td>1.3</td>
<td>Ingest</td>
<td>1</td>
<td>Copyright Permissions Verification/Notification -- the system requires the user to agree to a series of statements regarding</td>
<td>3</td>
<td>0</td>
<td>The system does not alert user of copyright issues or policies prior to the ingest</td>
</tr>
</tbody>
</table>
The Dataverse provides the best combination of system performance and robustness, user ease, platform scalability, and an active open source community that responds to the evolving needs of the user community.

The group recommends that TDL, through its membership, adopt Dataverse to facilitate the discovery of research data and its associated metadata.
Why Dataverse?

(Relatively) simple **self-service** ingest workflow for researchers

Ability to share with **trusted group** of researchers prior to publication

Creation of **DOIs** for data citation

Helps researchers fulfill **Data Management Plan** requirements

Comes with an **active open source community** led by Harvard IQSS
About Dataverse

Open Source
Harvard IQSS
17 installations worldwide
**dataverse**: a container for datasets, or other dataverses

**dataset**: contains descriptive metadata and data files
TDL’s Dataverse Implementation
TDL Dataverse Implementation Working Group

**Charge:** Pilot test, assess, and launch a consortial repository for research data archiving and management.

**Members:** Denyse Rodgers (Baylor); Bruce Herbert, Sean Buckner, Wendi Kaspar, Cecilia Smith (TAMU); Ray Uzwyshyn, Todd Peters (Texas State); Christopher Starcher (Texas Tech); Jeremy Donald (Trinity); Kristi Park, Ryan Steans, Nick Lauland, Laura Waugh (TDL)

**Working Group members**

- 14

**TDL member institutions**

- 7

**Santi Thompson (Chair)**
TDL
- Technology Hosting
- Helpesk (tech support)
- Training
- Coordination & Governance

Member Libraries
- Data Curation Support, Mediated Deposit, etc.
- Campus Outreach
- Active Management of Institutional “dataverse”

Researchers
- Deposit and publish data
- Share data with research group
Institutional dataverses

Dataverse
Texas Research Data Repository Dataverse

A statewide collaboration of higher education institutions in Texas

Share, publish, and archive your data. Find and cite data across all research fields.

Welcome to the Texas Research Data Repository, a repository for research data collected by researchers at the Texas Digital Library (TDL) member institutions. Select your TDL member institution below.

Search this dataverse...  Find  Advanced Search

Dataverses (47)  Datasets (22)  Files (115)

Spatial extents of ICA projects
May 9, 2016 - Institute of Classical Archaeology Dataverse
Institute of Classical Archaeology, 2016: "Spatial extents of ICA projects" http://dx.doi.org/10.18738/T2C83Y
Infrastructure Components

1. Amazon Linux server
2. Glassfish (a Java EE application server)
3. Apache
4. PostgreSQL
5. Solr search engine with a Dataverse-specific schema
6. SMTP server for sending mail for password resets and other notifications.
7. Persistent identifier service (DOIs)
Additional integrations

EZID account for DOIs
TwoRavens: Harvard-built package for statistical data exploration and analysis (utilizes R, rApache, Zelig)
Shibboleth authentication system (requires Apache)
Amazon Infrastructure

AWS M4 X-Large
  a. 16GiB Memory
  b. Intel Xeon E5-2676 v3 @2.4 GhZ

Back Ups
  a. Full backup in Amazon S3 (US West Coast)
  b. Daily changes/snapshot (US East Coast)
Scaling up
Security Considerations

Etc.

User Guide
Policies and Terms of Use
Promotional materials
Minimal site branding

Still working on...

- A customized front page
- Member MoUs
- API Guide

Resources

- About the Repository
- User Guide
- Licensing
- Policies
- Frequently Asked Questions
- Texas Research Data Repository Metadata Guidelines
- Data Management Working Group Report
- Pilot Project Promotional Flyer

Texas Digital Library
Pilot Testing
TEXAS RESEARCH DATA REPOSITORY
Pilot Survey Demographics

- Overall Rate of:
  - Response: 59%
  - Completion: 89%

- Pie Chart:
  - Librarians: 69%
  - Researchers: 31%
Future Repository Services

- Assistance finding data in the repository for reuse
- Assistance setting up a location in the repository for research projects
- Assistance describing data
- Assistance managing data prior to submitting it to the repository
- Assistance applying digital preservation best practices with...
Lessons Learned and Future Plans
Plans for launch

Do away with institutional dataverses? Still an open question.
Re-work the homepage to provide clarity, especially for first-time users.
Integration of senf (secure number finder) tool.
Data Symposium in November 2016 to launch and educate.
Need for a central/TDL services manager for outreach, training, and management.
Beyond Launch

Explore integrations
- OSF
- Vireo/DSpace

Ensure scalability

Financial sustainability and governance

Continuous improvement
Contact information

Kristi Park
kristi.park@austin.utexas.edu

Ryan Steans
rsteans@austin.utexas.edu