Diving into Data: Implementing a Data Repository at the Texas Digital Library

TDL Dataverse Implementation Working Group
## Panel Outline

<table>
<thead>
<tr>
<th>The Introduction</th>
<th>The Demo</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Kristi Park, Texas Digital Library)</td>
<td>(Ryan Steans, Texas Digital Library)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The Need</th>
<th>The Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Bruce Herbert, Texas A&amp;M)</td>
<td>(Elizabeth Quigley, Harvard IQSS)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The Design</th>
<th>Q&amp;A</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Santi Thompson, University of Houston)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Santi Thompson, University of Houston)</td>
</tr>
</tbody>
</table>
Part 1:
The Introduction
“Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research…”

TDL Mission Statement

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.
Step 1: Put our heads together

**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)
...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.
Step 2: Make it happen

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)
Part 2:
The Need
“Jim Gray described his vision of the fourth paradigm of scientific research.

He outlined a two-part plea for the funding of tools for data capture, curation, and analysis, and for a communication and publication infrastructure.

He argued for the establishment of modern stores for data and documents that are on par with traditional libraries.”

Use Case: Make Research Data Publicly Available

Primary Actors:

- PIs of federally funded research
- Researchers working on unfunded research or funded research with no retention requirements
- Graduate students working on theses, dissertations, or other data-generating projects.
Federal Mandates For Public Access to Research

The Library Supports:

- Publication repositories
- Tools to create data management plans
- TDL Data repository
- Workflows, standards, & policies

http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research
Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Pivovar, Roger S. Day, Douglas B. Frisina
Published: March 21, 2007 • DOI: 10.1371/journal.pone.0000308 • Featured in PLOS Collections

Abstract

Background
Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available.

Principal Findings
We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly (p = 0.006) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression.

Significance
This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.

Figures

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000308
ETDs in Institutional Repositories

Enhance the Impact of our ETDs:

- Co-publish data sets
- Vireo ingestion
- Links in metadata
- Workflows, standards, & policies
Use Case: Share Data within a Trusted, Collaborative Network

Primary Actors:

Researchers involved in collaborative teams or networks
Fig. 2. The relative impact of teams. (A to D) Mean team size comparing all papers and patents with those that received more citations than average in the relevant subfield. (E to H) The RTI, which is the mean number of citations received by team-authored work divided by the mean number of citations received by solo-authored work. A ratio of 1 indicates that team- and solo-authored work have equivalent impact on average. Each point represents the RTI for a given subfield and year, whereas the black lines present the arithmetic average in a given year.
Collaboration Across Institutions

Fig. 1. The rise in multi-university collaboration. By comparing the incidence of papers produced by different authorship structures, we see that the share of multi-university collaborations strongly increases from 1975 to 2005. This rise is especially strong in SE (A) and SS (B), whereas it appears weakly in AH (C), in which collaboration of any kind is rare. The share of single-university collaborations remains roughly constant with time, whereas the share of solo-authored papers strongly declines in SE and SS.
Use Case: Seek Data to (Re)Use

Primary Actors:

- Researcher is interested in conducting a meta study reusing data developed in earlier studies
- Public using data for personal needs
- Organizations seeking data for their needs.
Ontogeny in the tube-crested dinosaur *Parasaurolophus* (Hadrosauridae) and heterochrony in hadrosaurids

Andrew A. Farke, Derek J. Chok, Annisa Herrera, Brandon Scolieri, Sarah Werning

Published October 22, 2013
PubMed 24167777

https://peerj.com/articles/182/#supplemental-information
Open Data
Open Sharing of the Paper and the Data

Part 3:
The Design
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>I.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>I.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>I.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>
TDL Dataverse Implementation Working Group

Policy and Governance
- Sean Buckner
- Santi Thompson
- Ray Uzwynshyn

Workflows and Outreach
- Jeremy Donald
- Wendi Kaspar
- Cecilia Smith
- Chris Starcher

Budget/Business Model
- Bruce Herbert
- Kristi Park
- Ryan Steans
- Santi Thompson

Technical Configuration
- Nick Lauland
- Todd Peters
- Denyse Rodgers
- Ryan Steans

Marketing and Coordinator Extraordinaire: Laura Waugh
Policy and Governance

Internal and external policies creation
Workflows and Outreach

Develop and test workflows for researchers and librarians
Budget and Business Model

Assess costs and identify potential funding models
Technical Configuration

Setup, configure, and test system and its features
User Guide

Texas Research Data Repository Pilot Project

Participants in the Texas Research Data Repository Pilot Project are asked to complete the Required Tasks (steps 1 – 8) below. Participants are also welcome to spend as much time as they like in the repository and complete the Optional Tasks (steps 9 – 15).

After completing the Required Tasks (and Optional Tasks, if applicable), participants are asked to complete the follow-up survey and provide feedback and observations about using Dataverse for the Texas Research Data Repository: SURVEY LINK

Required Tasks:
1. Create a user account
2. Prepare data, code, and additional documentation files
3. Create metadata
4. Create a Dataverse (i.e., collection)
5. Upload information about the dataset
6. Share dataset
7. Publish dataset
8. Download dataset

Optional Tasks:
9. Use mapping and statistical analysis tool
10. Alter default Terms of Use
11. Make your dataset restricted
12. Create multiple versions of a dataset
13. Deaccession a dataset
14. Turn on the Guestbook feature
15. Add a logo to your Dataverse

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
## Pilot Survey Demographics

<table>
<thead>
<tr>
<th>Type of Respondent</th>
<th>Participation Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researchers</td>
<td>31%</td>
</tr>
<tr>
<td>Librarians</td>
<td>69%</td>
</tr>
<tr>
<td><strong>Overall Rate of:</strong></td>
<td>%</td>
</tr>
<tr>
<td>Response</td>
<td>59%</td>
</tr>
<tr>
<td>Completion</td>
<td>89%</td>
</tr>
</tbody>
</table>
## Completing Required Tasks

<table>
<thead>
<tr>
<th>Answer</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create a user account</td>
<td>100%</td>
</tr>
<tr>
<td>Create a Dataverse (i.e., collection)</td>
<td>88%</td>
</tr>
<tr>
<td>Upload at least one dataset</td>
<td>100%</td>
</tr>
<tr>
<td>Provide metadata information for dataset(s)</td>
<td>100%</td>
</tr>
<tr>
<td>Publish dataset(s)</td>
<td>94%</td>
</tr>
<tr>
<td>Download a dataset</td>
<td>100%</td>
</tr>
</tbody>
</table>
## Completing Optional Tasks

<table>
<thead>
<tr>
<th>Answer</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilize the mapping analysis tool</td>
<td>17%</td>
</tr>
<tr>
<td>Utilize the statistical analysis tool</td>
<td>33%</td>
</tr>
<tr>
<td>Request access to a restricted dataset</td>
<td>17%</td>
</tr>
<tr>
<td>Utilize versioning of data</td>
<td>17%</td>
</tr>
<tr>
<td>Turn on the Guestbook feature in Dataverse</td>
<td>50%</td>
</tr>
<tr>
<td>Add a logo to the Dataverse instance that you created</td>
<td>17%</td>
</tr>
</tbody>
</table>
# Meet Disciplinary Data Needs?

<table>
<thead>
<tr>
<th>Answer</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely well</td>
<td>13%</td>
</tr>
<tr>
<td>Very well</td>
<td>56%</td>
</tr>
<tr>
<td>Moderately well</td>
<td>6%</td>
</tr>
<tr>
<td>Slightly well</td>
<td>25%</td>
</tr>
<tr>
<td>Not well at all</td>
<td>0%</td>
</tr>
</tbody>
</table>
Future Repository Services?

<table>
<thead>
<tr>
<th>Answer</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistance describing data</td>
<td>40%</td>
</tr>
<tr>
<td>Assistance setting up a location in the repository for research</td>
<td>20%</td>
</tr>
<tr>
<td>projects</td>
<td></td>
</tr>
<tr>
<td>Assistance finding data in the repository for reuse</td>
<td>20%</td>
</tr>
<tr>
<td>Assistance managing data prior to submitting it to the repository</td>
<td>47%</td>
</tr>
<tr>
<td>Assistance applying digital preservation best practices with</td>
<td>53%</td>
</tr>
<tr>
<td>research data</td>
<td></td>
</tr>
</tbody>
</table>
## Future Repository Features?

<table>
<thead>
<tr>
<th>Answer</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linking research data with an existing publication</td>
<td>100%</td>
</tr>
<tr>
<td>Linking supplemental data with an electronic theses or dissertation</td>
<td>50%</td>
</tr>
<tr>
<td>Management of collaborative teams within the data repository</td>
<td>13%</td>
</tr>
<tr>
<td>Customizable submission screen with instructions</td>
<td>6%</td>
</tr>
<tr>
<td>Development and growth of interdisciplinary research data related to</td>
<td>13%</td>
</tr>
<tr>
<td>Texas geographic regions and topics</td>
<td></td>
</tr>
</tbody>
</table>
### Most Important Benefits?

<table>
<thead>
<tr>
<th>Answer</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fulfill federal mandates for sharing publications and research data</td>
<td>56%</td>
</tr>
<tr>
<td>Make your research data more widely available</td>
<td>50%</td>
</tr>
<tr>
<td>See statistics on downloads and citations of my data</td>
<td>31%</td>
</tr>
<tr>
<td>Make my data citeable through the assignment of a DOI (digital object identifier)</td>
<td>44%</td>
</tr>
<tr>
<td>Save versions of your dataset</td>
<td>31%</td>
</tr>
<tr>
<td>Collecting all my data in one place</td>
<td>63%</td>
</tr>
</tbody>
</table>
Part 4:
The Benefits
Dataverse

Texas Digital Library Prod/Test Dataverse
A statewide collaboration of higher education institutions in Texas

Add Data
Share, publish, and archive

Find Data
Search across disciplines

Get Recognition
Obtain a citation and unique identifier

Share, publish, and archive your data. Find and cite data across all research fields.
Welcome to the Texas Digital Library Test Dataverse!
IMPORTANT: This Dataverse server does NOT include the TwoRavens add-on.
Because of this, you may receive errors when ingesting certain datasets and the “explore” button will not work.

Search this dataverse...  Find  Advanced Search

Dataverse Category
Organization or Institution (19)
Researcher (5)
Research Project (2)

Publication Date

1 to 10 of 49 Results

Peters, Todd. 2016: "size test 6.1 GB", http://dx.doi.org/10.5072/FK2/8FSKOE, Texas Digital Library Prod/Test Dataverse, V1
 upload test 6.1 GB mp4

New file for multi-gig files
  Jan 15, 2016 - nicktest Dataverse
Part 5:
Texas Research Data Repository Demonstration
What we’ll cover

1. What is a Dataverse versus a Dataset?
2. Uploading Data
   a. Creating a Dataverse
   b. Creating a Dataset
3. Visualizations
4. Re-Using and Sharing your data
Dataverse and Datasets

● A **Dataverse** is a home for your research project, your community, etc...
  ○ You can easily build a Dataverse **within** a Dataverse (Ex: University Dataverse > Archaeology Dataverse)
  ○ You can **stack Dataverses within one another** (Ex: University Dataverse > Archaeology Dataverse > Roman Archaeology Dataverse) to create sub-sub Dataverses

● **Datasets live within a Dataverse** - you can associate multiple datasets within a Dataverse (Ex: Different dig site **datasets** from different locations may all be under one **Dataverse** for a single researcher archaeologist)
A Dataverse Demo (Uploading Data)
Add a New Dataverse (a home for your datasets)
Add Metadata and Settings to your Dataverse

Basic Information

Metadata Fields

Search Facets
Now - add a dataset to your Dataverse
First, add Metadata for each dataset
Adding data files requires only a simple upload.
Your files will appear inside the dataset
Users can manage Metadata and Terms
Add to and Edit the full Metadata record

<table>
<thead>
<tr>
<th>Citation Metadata</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dataset Persistent ID</strong></td>
<td>dot:10.5072/FK2/QL4Q3T</td>
</tr>
<tr>
<td><strong>Title</strong></td>
<td>TCDL 2016 Test Dataset</td>
</tr>
<tr>
<td><strong>Author</strong></td>
<td>Admin, Dataverse (TDL.org)</td>
</tr>
<tr>
<td><strong>Contact</strong></td>
<td>Use email button above to contact: Admin, Dataverse (TDL.org)</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>This is a test Dataset for TCDL 2016 Demo (2016-03-03)</td>
</tr>
<tr>
<td><strong>Subject</strong></td>
<td>Earth and Environmental Sciences: Other</td>
</tr>
<tr>
<td><strong>Keyword</strong></td>
<td>Earth Science Stuff</td>
</tr>
<tr>
<td><strong>Producer</strong></td>
<td>TDL Science Labs (Texas Digital Library) <a href="http://tdl.org">http://tdl.org</a></td>
</tr>
<tr>
<td><strong>Depositor</strong></td>
<td>Admin, Dataverse</td>
</tr>
<tr>
<td><strong>Deposit Date</strong></td>
<td>2016-05-09</td>
</tr>
</tbody>
</table>
Specialized metadata can be expanded

<table>
<thead>
<tr>
<th>Geospatial Metadata</th>
<th>Geographic Coverage</th>
<th>Country / Nation</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>City</td>
<td>Austin</td>
<td></td>
</tr>
<tr>
<td></td>
<td>State / Province</td>
<td>Texas</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geographic Unit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geographic Bounding Box</td>
<td>West Longitude</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>East Longitude</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>North Latitude</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>South Latitude</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Users can manage Terms of Use
Users may alter their license and terms of access.
Publish Data with one button
You have now published your data!
Visualization of Data

TCDL 2016 Test Dataset

Admin. Dataverse, 2016, "TCDL 2016 Test Dataset", http://dx.doi.org/10.5072/FK2/QL4Q3T, Texas Digital Library Test Dataverse, V1

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Description
This is a test Dataset for TCDL 2016 Demo

Subject
Earth and Environmental Sciences; Other

Keyword
Earth Science Stuff

Files Metadata Terms Versions

Search this dataset...

3 Files

- catdata.tab
  Tabular Data - 21.5 KB - May 9, 2016 - 0 Downloads
  3 Variables, 1600 Observations
  UNF:5.0WmGvPw=0VwsEggG+gig==

- climate_data.xls
  MS Excel - 183.0 KB - May 9, 2016 - 0 Downloads
  MD5: c7f43a6a1597cf46e6a85a1561d1;

- german.data-numeric.txt
  Plain Text - 69.6 KB - May 9, 2016 - 0 Downloads
  MD5: did51be69bfce45685b90000c003554c32;

Explore
Download
TwoRavens allows users to dynamically view data
Maps and Shape Data may work with WorldView
World Map
Reuse of Data via Download
Some files have multiple download options

- Original File Format (Comma Separated Values)
- Tab-Delimited
- RData Format
- Variable Metadata
- Data Subset
- Data File Citation
Easy to share on social media
Citation Options

TCDL2016_Test_Dataverse Dataverse (TDL.org)

Texas Digital Library Test Dataverse > TCDL2016_Test_Dataverse Dataverse > TCDL 2016 Test Dataset

Admin, Dataverse, 2016. "TCDL 2016 Test Dataset", http://dx.doi.org/10.5072/FK2JL4Q3T, Texas Digital Library Test Dataverse, V1

[UNF68/hwGJvPw+2MrSe9gG+Vg==]

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Description
This is a test Dataset for TCDL 2016 Demo

Subject
Earth and Environmental Sciences; Other

Keyword
Earth Science Stuff

Download Citation
EndNote XML
RIS Format
Part 6:

Upcoming Dataverse Developments
Community & Future Work

Elizabeth Quigley
User Experience Lead
The Dataverse Community

- Dataverse Community Meeting
- Universities Using Harvard Dataverse
- Community & Working Groups
- Code & Issue Contributors
- Dataverse Installations
- Dataverse Advisory Board
- Biweekly Community Calls
- Grant Funded Collaborations
Future Work

Upcoming Releases:
- 4.4: Widgets Updates & Remote Authentication
- 4.5: Metadata Harvesting & Exporting, Private URL

File Level Metadata:
- Provenance
- Richer support for file level metadata

Sensitive Data Support:
- Secure Data Storage for Harvard Dataverse (hosting set up)
- DataTags compliant version of Dataverse

Large Data Support:
- Streaming social sciences data, e.g.- billions of GeoTweets
- Biomedical large scale data (SBGrid Repository)

Funding Agencies
Welcome to the Dataverse Community!

Texas Digital Library

@dataverseorg | http://dataverse.org | Code: github.com/IQSS/dataverse
Questions and Discussion