Research Data Repositories

Developing and Implementing Infrastructures for Institutional and Consortial Environments

Texas State University Dataverse
A platform for publishing and archiving Texas State University's research data.

Dataverse

Texas State University Libraries

Ray Uzwyshyn, Ph.D. MBA MLIS
Director, Collections and Digital Services,
Texas State University Libraries
Online Data Research Repositories
What are They?

- Way to Manage a Researcher’s Data/Metadata
- Permalinking Strategy for Data Citation
- Way to Manage Federal Grant Compliance
- Middle-Term Data Archiving and Sharing Strategy
The Research Data Repository Lifecycle

**CAPTURE**
Project Data from Experiments, Surveys Researchers and Scientists

**CATALOG**
Assign Metadata Schema, Specialized and Disciplinary Taxonomies, DOI, UNF

**MANAGE**
Administrative Online Research Data Archives

**FIND/VIEW**
Retrieve, Download Relevant Data Sets Instantaneously

**Synthesize Research**
Verification, Insight, Discovery Visualization, Harvesting and Linked Data

Becoming part of Science, Social Science and Humanities Research Process

Promotes: accuracy, efficiency, sharing
Why are Data Management Repositories Necessary?

Most major Federal grant agencies require data access as mandatory part of the grant proposal/oversite process. (NIH, NSF, NEH, USDA)
What makes Data Management Repositories useful?

- Makes available faculty, departmental and institutional research
- Allows publication of negative data (lessens research replication)

Wordle of the National Science Foundation’s Award and Administration Guide. Chapter VI.D.4, Mandatory 2011
Types of Research Data Repositories

1) Project specific
   large single faculty/team projects

2) Discipline specific
   i.e. Purdue Nanohub/Nanotechnology

3) Institutional or Consortial
   (either institution wide or consortial repositories)
All-Purpose and Specialized Data Repository Platforms

Data Archiving Infrastructure

Primary platform choice

<table>
<thead>
<tr>
<th>Inst. Repository w/ Data (top 5)</th>
<th>Data-specific Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dspace</td>
<td>Dataverse</td>
</tr>
<tr>
<td>Fedora</td>
<td>Chronopolis</td>
</tr>
<tr>
<td>BePress Digital Commons</td>
<td>HubZero (customized)</td>
</tr>
<tr>
<td>Hydra</td>
<td>DataConservancy</td>
</tr>
<tr>
<td>Drupal</td>
<td>Custom repository</td>
</tr>
</tbody>
</table>

Research Data Repository
Software Characteristics

• Hosted or on a server
• Software contains management and collaborative options
• Open source or proprietary software
• Wide Variety of Data Types
  (Excel to SPSS to various disciplinary specific formats)
Part I: Planning Your Repository

Environmental Scan of Needs for Your Institution or Consortium

Evaluation Criteria
- System Performance/Robustness
- Usability
- an active open source community

Gather Finalists:
Harvard’s Dataverse, Purdue’s Hubzero Figshare

Make Final Choice: Harvard’s Dataverse
Dataverse
Harvard’s Open Source Research Data Solution

Data sharing, data citation, data publishing and versioning management

Social Sciences Beginnings (IQSS)
Data Science (site)
http://thedata.org
Dataverse Open Source Download (Github), Software Background
Dataverse Architecture (Consortial)

Research Study Data

Original Data Set Files
Metadata
Paratextual Materials (Methodology, Field Notes, Multimedia, Graphs, Programs etc.)
Data Citation and Metadata

Harvard Dataverse Network

REPLICATION DATA FOR: A MULTIVARIATE MODEL OF STRATEGIC ASSET ALLOCATION
hdl:1902.1/QBXRSFLBQJUNF.3.ZnYHkkZe2vTJAWaBdpPKA==

CATALOGING INFORMATION

If you use these data, please add the following citation to your scholarly references. Why cite?


Results found in this publication can be replicated using these data.

Original Publication

Publications


Data Citation Details

Title
Replication data for: A Multivariate Model of Strategic Asset Allocation

Study Global ID
hdl:1902.1/QBXRSFLBQJ

Authors
John Y. Campbell (Harvard University); Yeung L. Chan; and Luis Viceira

Producer
John Y. Campbell

Production Date
2003

Funding Agency
National Science Foundation; Hong Kong RGC Competitive Earmarked Research Grant (HKUST 6965/01H); Division of Research of th Business School
Dataverse Metadata Example
(From the Simple to Complex)

Schemas Supported: GeoSpatial, Life Sciences, Astronomy and Physics, Georeferenced Data

Title
Data from “Social determinants of unmet hospitalisation need amongst the poor in Andhra Pradesh, India: A cross-sectional study.”

Author
Name: Nagulapalli, Srikant

Affiliation: Andhra University

Identifier

Description
The dataset is of a health survey amongst the 21.5 million poor families of the Indian state of Andhra Pradesh conducted during April and May 2013. The dataset captures individual characteristics and household characteristics of the past 365 days and was used to analyse the unmet hospitalisation need in the Indian State of Andhra Pradesh. Data was collected by 2022 trained field staff of Aarogyasri Health Care Trust (AHCT) of Government of Andhra Pradesh using a questionnaire modelled after that used for the health surveys by National Sample Survey Organisation of India.

Subject
Medicine, Health & Life Sciences

Keyword
unmet hospitalisation need

Production Date
2013-06-01

Depositor
Privileged, admin

Deposit Date
2013-08-03
The Many Planning Aspects of Data Research Repositories

Planning Principles
Wide Flexibility on Institutional Levels.
Guiding Consortial Templates which can be customized on institutional levels
Part II: Developing Your Data Repository
TDL Dataverse State Working Group
(August 2015 – December 2016)

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

**Main Working Group (14)**
(4 Subcommittees)
- Policy and Governance
- Workflows and Outreach
- Budget/Business Model
- Technology

State Data Repository Symposium Group (Baylor)

[Final Report October, 2016](#)
Interface Design & Usability

Search the Texas Data Repository

Search...
FIND

Add a Dataset
Create a Dataverse
Explore Data Repository
Learn More
Get Help

Publish and Track Your Data, Discover and Reuse Others' Data!

POWERED BY Dataverse

http://data.tdl.org
Service Models

Texas Data Repository

Member University Libraries (service & outreach)

Researchers (deposit, search, publish)

1) Mixed
2) Mediated
3) Unmediated (Direct)
Texas State Repositories Architecture

- **TS Dataverse** (Regular to Medium Size Data Sets)
- **Custom Data Storage** (Big Data, TB+, TR)
- **D-Space Publications Repository**

---

**Research Data**

---

**Text**

---

**Texas State Academic Research**
One Size Does Not Fit All

Types of Data Projects (Sizes)

1) Normal Range Projects
Files/Data Fit on Server, may be uploaded, Dataverse, Hubzero)

2) Large Projects
(Data may require specialized university IT Support, i.e. terabyte/petabyte drives, Pointers etc.)

3) Huge Projects
(Projects require consortial possibilities, national models, Texas Advanced Computer Center TAAC, DEEPN, Duracloud, AWS, Custom Solutions)
Faculty Data Management Plan
Documentation/Policy Tool

https://dmptool.org/
California Digital Library
Part III: Human Resource Infrastructures (Working Teams)

Full or Part Time

- Data Repository Liaison
- Publication Repository Liaison
- Metadata Liaison
- Subject Liaisons (Outreach)
- Committee for Workflows & Policies

Current Hires

- Digital Collections Librarian
  (Texas State Data Repository Dataverse/Publications Repository: D-Space)

- Data Visualization and Analytics Librarian
  (Tableau, Bayesia)

Future Hires

- Machine Learning/Neural Networks/AI Librarian (working with the data)
Marketing and Other Possibilities

Texas State University Dataverse

Why deposit your data with TXST Dataverse?

Comply with funding requirements. The Texas State University Dataverse Repository can help you meet data sharing and archiving requirements from federal (or other) funding agencies and publishers. We want to hear from you! Contact the Digital Collections Librarian [digitalcollections@txstate.edu] for guidance on how to include the Texas State University Dataverse in your data management plan.

Get credit and increase your scholarly impact.
Data published within the Texas State University Dataverse Repository are widely indexed in search engines, making it more likely to be found. The Texas State University Dataverse assigns published datasets a persistent “digital object identifier” (DOI), which makes it easier for others to reliably cite your work and facilitates reproducibility. You can also track user access and have the option to ask for people to sign a “guestbook” when downloading your data.

Collaborate.
Flexible access controls mean you can decide when, with whom, and how much of your data to share. Version control helps you track your progress and keep things up-to-date with collaborators and publish your work when you’re ready.

Ensure long-term access.
The repository is built on Dataverse, a robust, open-source software application developed by Harvard University. It is hosted by the Texas Digital Library, whose focus on long-term preservation and access can help ensure secure, reliable, and persistent access to digital data collections.

Take advantage of local support.
The Texas State University Dataverse Repository is hosted by the Texas Digital Library, which provides robust technical support and is committed to long-term access and preservation. You can also rely on trained librarians here at the Alkek Libraries to assist you throughout your research, from the earliest planning stages, through publication and long-term archiving.

Electronic Thesis and Dissertations (ETD) Repository (D-Space)
Future Possibilities: VIREO, DATA REPOSITORY CONNECTIONS

Working with the Data – Support Mechanisms
Data Literacy (Workshops/Education)
Data Visualization, Data Analytics
Machine Learning/Neural Networks/ AI
Research Data Repository Adoption Lifecycle (2018)
Further Links/References

• ARL NSF Data Sharing Policy and Resource Links, http://www.arl.org/focus-areas/e-research/data-access-management-and-sharing
• ARL (White House Directives and Funded Research Data) http://www.arl.org/focus-areas/public-access-policies#.VoaV0I-cFzo
• Baker, Monya. 1500 Scientists Lift the Lid on Reproducibility. www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
• Harris, Richard. (April 2017). Rigor Mortis How Sloppy Science Creates Worthless Cures
• California Digital Library DMT Tool: https://dmptool.org/
• Chronopolis: http://www.digitalpreservation.gov/partners/chronopolis.html
• Dataverse. http://thedata.org/
• Dataverse (Data Science Site). http://datascience.iq.harvard.edu/dataverse
• Data Information Literacy Guide. http://www.datainfolit.org/dilguide/
• Data Information Literacy Competencies (Purdue). http://blogs.lib.purdue.edu/dil/the-twelve-dil-competencies/
• DPN (Digital Preservation Network) http://www.dpn.org/
• Duracloud: http://www.duracloud.org/
• Purr. (Purdue Institutional Data Repository). https://purr.purdue.edu/
• Hubzero. https://hubzero.org/
Further Links/References

- ICPSR Data Management & Curation. [http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/](http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/)
- “Research Data Management”. pp. 6-7 and pp 24 – 45.
Comments/Questions

Contact Information:

Ray Uzwyshyn, Ph.D. MBA MLIS
Director, Collections and Digital Services
Texas State University Libraries
ruzwyshyn@txstate.edu (512)245-5687
Dataverse Network Architecture

Why the Dataverse Network?
(silent video overview)

Open Journal Systems
Dataverse Integration

Research Study Data
Data Set Files
Metadata (Data Describing the data)
Paratextual Research Material
(Methodology, Field Notes etc.)
Graph Data Files
1.) Create Data Management Plans
2) Collaborate with other Researchers
3) Publish Data Sets (Purdue can publish a DOI: Digital Object Identifier for Data Sets)
Useful For Citation
4) Archive Data Sets

Boilerplate text for data management proposals available

Purr is part of Hubzero platform for scientific collaboration (Originally Nanohub)

• Purr: Purdue University Research Repository (video)
• Purr Site (Proprietary to University)
• Purr Background
Hubzero: Open Source Platform for Scientific Collaboration

Research Collaboration and Data Management Solution

Research Data Types
Spreadsheets
Instrument or Sensor Readings
Software Source Code
Surveys
Interview Transcripts
Images and Audiovisual Files

- [https://hubzero.org/](https://hubzero.org/)
- [Getting Started, Downloadable](https://hubzero.org/) and [Hosted Options](https://hubzero.org/)
- [Hubzero Video, Hubzero2](https://hubzero.org/)
Figshare/Cloud based/Proprietary

Repository where users make their research available in citable, shareable and discoverable manner

Figures, datasets, media, papers, posters presentations and file sets can be disseminated In a way that the current scholarly publishing Model does not allow

Open Source Platform for Sharing Research

Figshare (video)

Figshare for Institutions (Video)
Figshare Features
(Cloud Based/Proprietary)

1GB of private space
- taggable and easily filtered, your research data is better managed and easy to locate

Unlimited public space
- upload to your heart's content, the more - the better

Publish negative data
- all published research is citable

Upload all formats

Quick & simple upload

Cloud based
Data Citation Principles

DC¹

Data Citation Principles

Texas Data Repository
Texas Digital Library Initiative, 2014 -2016

TDL Consortium of 22 universities across Texas leveraging technological cooperation among academic libraries
Institutional Repository (MIT, D-Space)

Faculty publications, white papers, preprints, theses, dissertations, working projects, reports, grey literature

Larger Idea, Grant Compliance, Enabling Faculty Research Online, Raising Research Visibility, 

https://digital.library.txstate.edu/
Pilot Study Responses
Perceived Benefits of Data Repository

• Fulfill federal mandates for sharing publications and research data
• Make research data more widely available
• Statistics on downloads and citations of my data
• Make my data citeable through the assignment of a DOI (digital object identifier)
• Saving various versions of the dataset (data lifecycle)
• Collecting all my data in one place
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.

Currently, 80% of researchers do not share their data

Research Data Reproducibility Crisis

(Nature. 2016)

http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
Harris, Richard. (April 2017). *Rigor Mortis How Sloppy Science Creates Worthless Cures*
Hubzero/Purr Customization

Start Your Research Project

Create a Data Management Plan
Learn about the detailed requirements for your data management plan (DMP). Funding agency requirements are very specific and our DMP resources can help you to clear up any confusion. Get Started

Upload Research Data to Your Project
Create a project to upload and share your data with collaborators using our step-by-step form to guide you through the process. Invite collaborators from other institutions to join your project. Create a Project

Publish your Dataset
Package, describe, and publish your dataset with a Datacite DOI. Publishing will ensure your dataset is citable, reusable, and archived for the long-term. See Published Datasets