

Research Data Repositories

Developing and Implementing Infrastructures for Institutional and Consortial Environments



Ray Uzwyshyn, Ph.D. MBA MLIS
Director, Collections and Digital Services,
Texas State University Libraries

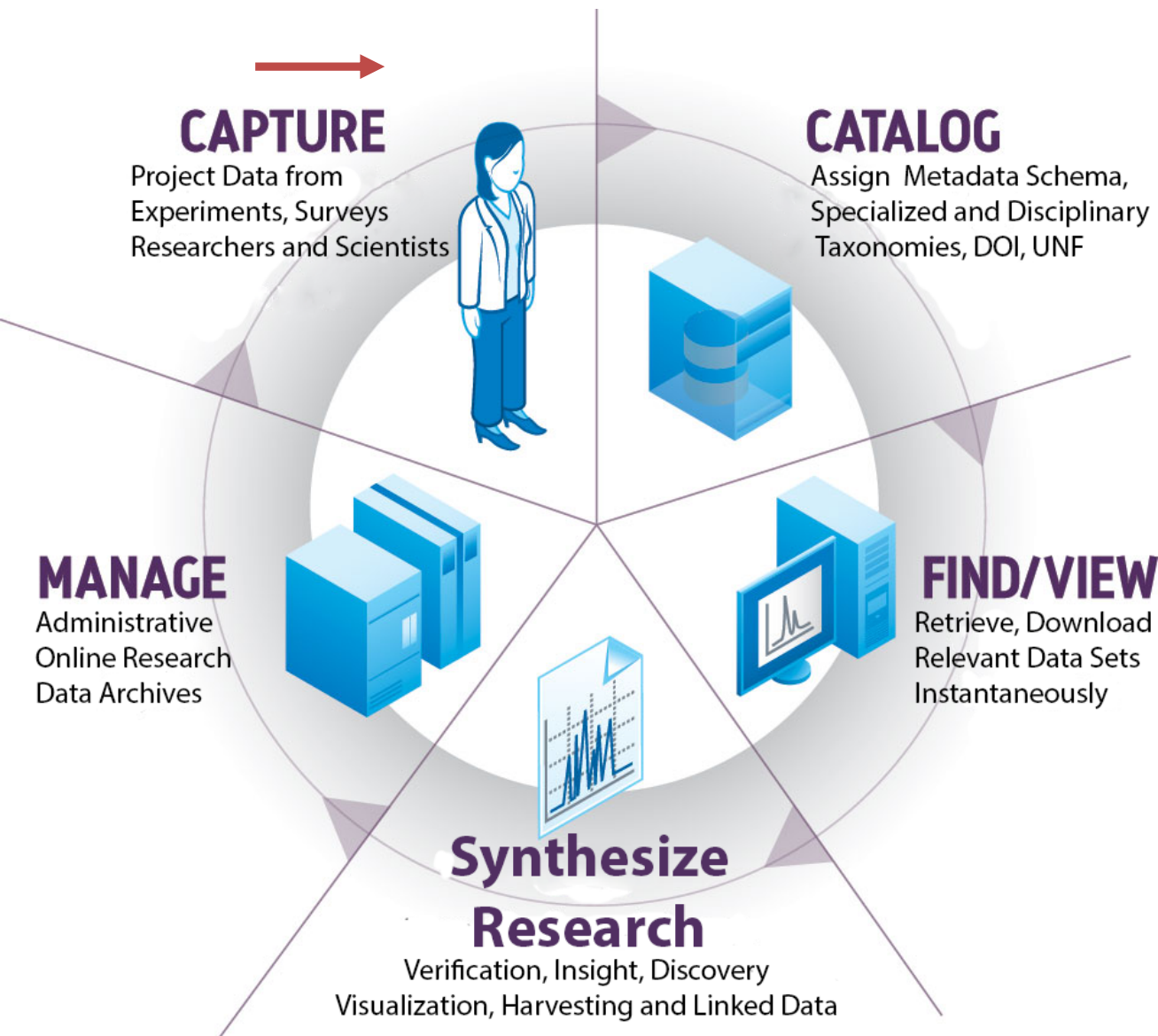
Online Data Research Repositories

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



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)



Wordle of the Final NIH Statement on Sharing Research Data, Mandatory 2003

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

Inst. Repository w/ Data (top 5)

Dspace

Fedora

BePress Digital Commons

Hydra

Drupal

Data-specific Repository

Dataverse

Chronopolis

HubZero (customized)

DataConservancy

Custom repository

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



TDL Data Management Working Group Report
Published August 28, 2015

Table of Contents

Introduction	1
Methodology	2
Evaluation of Dataverse	3
Recommendation	5
Next Steps	5
Appendices	7

Introduction

The need for Data Management services is one of two large-scale needs consistently expressed by Texas Digital Library (TDL) members, a need driven in part by the February 2013 mandate from the White House's Office of Science and Technology Policy to make the results of federally funded research publicly accessible.¹ For more information on how federal agencies plan to implement this policy, please see Appendix D.

The TDL Data Management Working Group convened in Fall 2013 to begin to address this gap, with a particular focus on finding solutions for making research data accessible and reusable.

The charge of the group was to help the Texas Digital Library determine what kinds of data management services it could provide at a consortial level.

Its objectives included:

- Articulating criteria for selecting pilot projects
- Evaluating proposed projects based on that criteria
- Selecting no more than three projects to implement
- Investigating issues related to storage and accessibility of data sets
- Documenting findings and recommendations for services

¹ The February 2013 OSTP directive, entitled "Increasing Access to the Results of Federally Funded Research" mandated that, each Federal agency with over \$100 million in annual research and development expenditures develop a plan to support increased public access to the results of research.

[Data Repository](#)
[Working Group Report](#)

(August 28, 2015)

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

The
**Dataverse
Network**TM
Project



A Web Application for Publishing, Citing,
Analyzing and Preserving Research Data

Data sharing, data citation, data publishing and versioning
management



The Institute for Quantitative Social Science
HARVARD UNIVERSITY

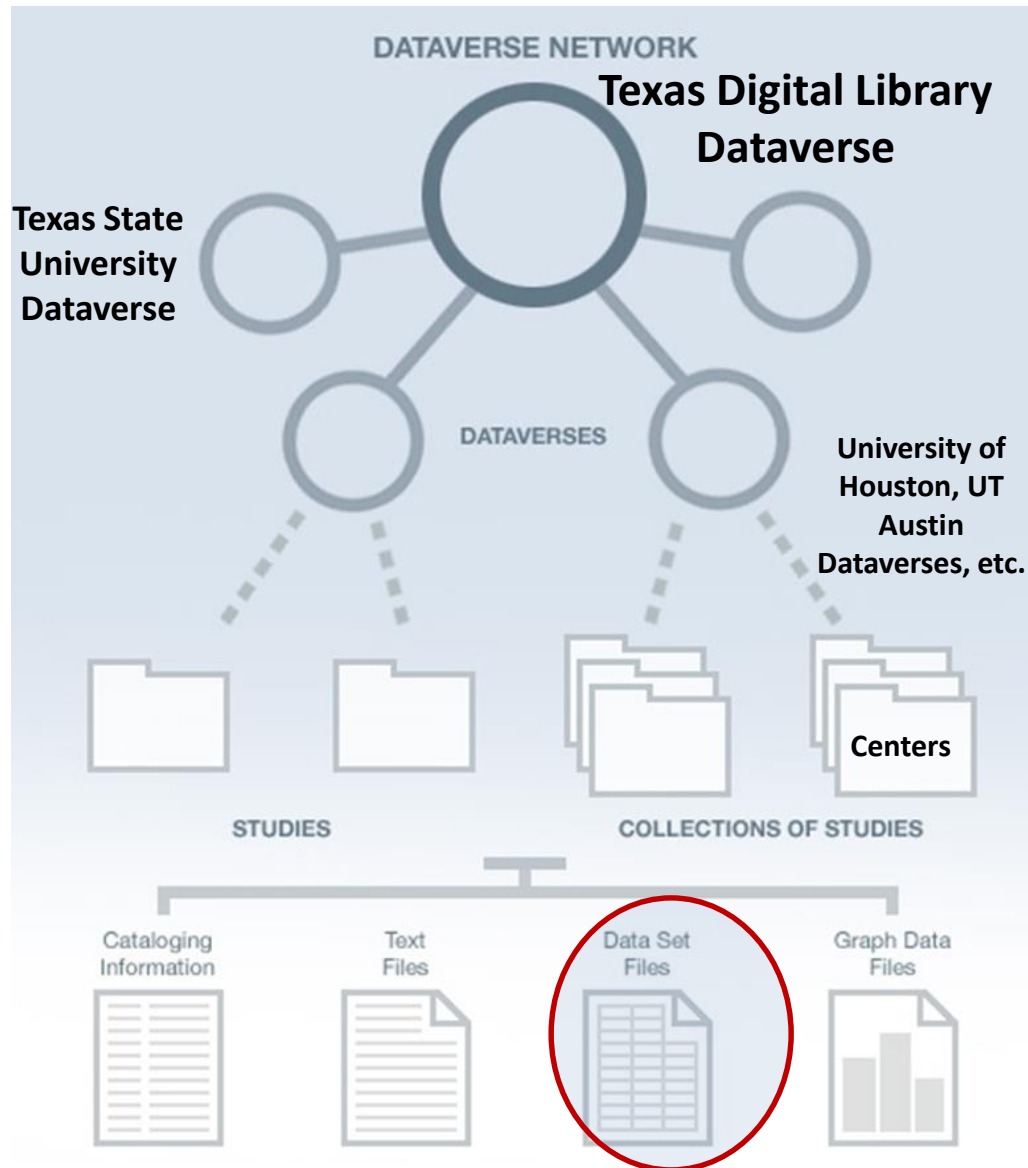
Social Sciences Beginnings (IQSS)

[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

Search Information Comments Create Account

REPLICATION DATA FOR: A MULTIVARIATE MODEL OF STRATEGIC ASSET ALLOCATION

hdl:1902.1/QBXRSFLBQJUNF:3:ZnYhHkZe2veTJAWaBDpPKA==

Version: 2 – Released: Thu Oct 03 16:46:32 EDT 2013

CATALOGING INFORMATION

Data & Analysis

Comments (0)

Versions

Data Citation

i If you use these data, please add the following citation to your scholarly references. [Why cite?](#)

John Y. Campbell; Yeung L. Chan; and Luis Viceira, 2007, "Replication data for: A Multivariate Model of Strategic Asset Allocation", <http://hdl.handle.net/1902.1/QBXRSFLBQJUNF:3:ZnYhHkZe2veTJAWaBDpPKA==> The Harvard Dataverse Network [Distributor] V2 [Version]

Citation Format

Original Publication

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

Campbell, John Y.; Chan, Yeung Lewis; and Viceira, Luis M., 2003, "A multivariate model of strategic asset allocation," Journal of Financial Economics, Elsevier, vol. 67(1), pages 41-80: [article available here](#)

Publications

John Y. Campbell & Yeung Lewis Chan & Luis M. Viceira, 2001. "A Multivariate Model of Strategic Asset Allocation," NBER Working Paper National Bureau of Economic Research, Inc. [article available here](#)

Campbell, John Y & Chan, Yeung Lewis & Viceira, Luis M, 2001. "A Multivariate Model of Strategic Asset Allocation," CEPR Discussion Paper 3070, C.E.P.R. Discussion Papers. [article available here](#)

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  HARVARD
Faculty of Arts and Sciences
DEPARTMENT OF ECONOMICS

Production Date 2003

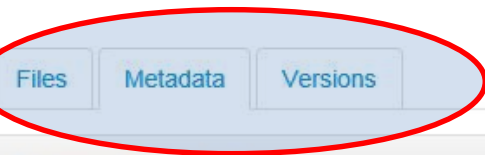
Funding Agency

National Science Foundation; Hong Kong RGC Competitive Earmarked Research Grant (HKUST 6965/01H); Division of Research of the Business School

Dataverse Metadata Example

(From the Simple to Complex)

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



Citation Metadata ^

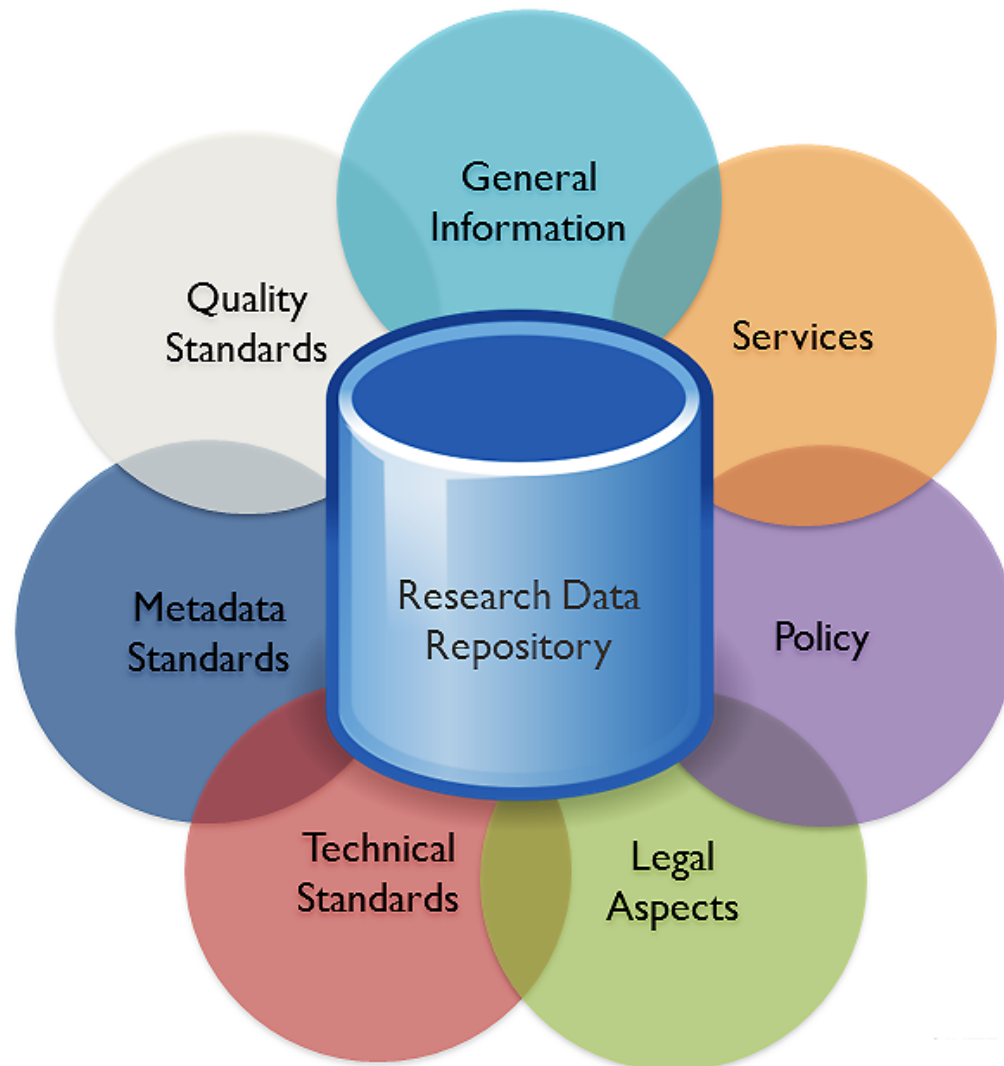
Title	Data from "Social determinants of unmet hospitalisation need amongst the poor in Andhra Pradesh, India: A cross-sectional study."	
Author	Name	Affiliation
	Nagulapalli, Srikant Identifier	Andhra University Identifier Scheme
Description	The dataset is of a health survey amongst the 21.5 million 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](#)

14

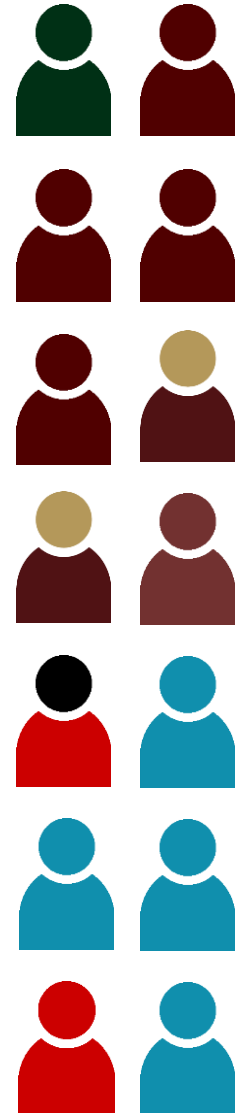
Working
Group
members

7

Texas
Universities

5

Sub-Committees



Interface Design & Usability

Search the Texas Data Repository

FIND



Add a Dataset



Create a Dataverse



Explore Data
Repository



Learn More



Get Help

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

TEXAS RESEARCH DATA REPOSITORY



Texas Digital Library Test Dataverse

A statewide collaboration of higher education institutions in Texas

Metrics

26 Downloads



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.



Trinity University Dataverse



Working together to work smarter™

UT Medical Branch Dataverse



TEXAS
University of Texas Dataverse



Texas State University Dataverse



Find

[Advanced Search](#)



Add Data

Service Models

Texas Data
Repository



Member University Libraries
(service & outreach)



Researchers
(deposit, search,
publish)

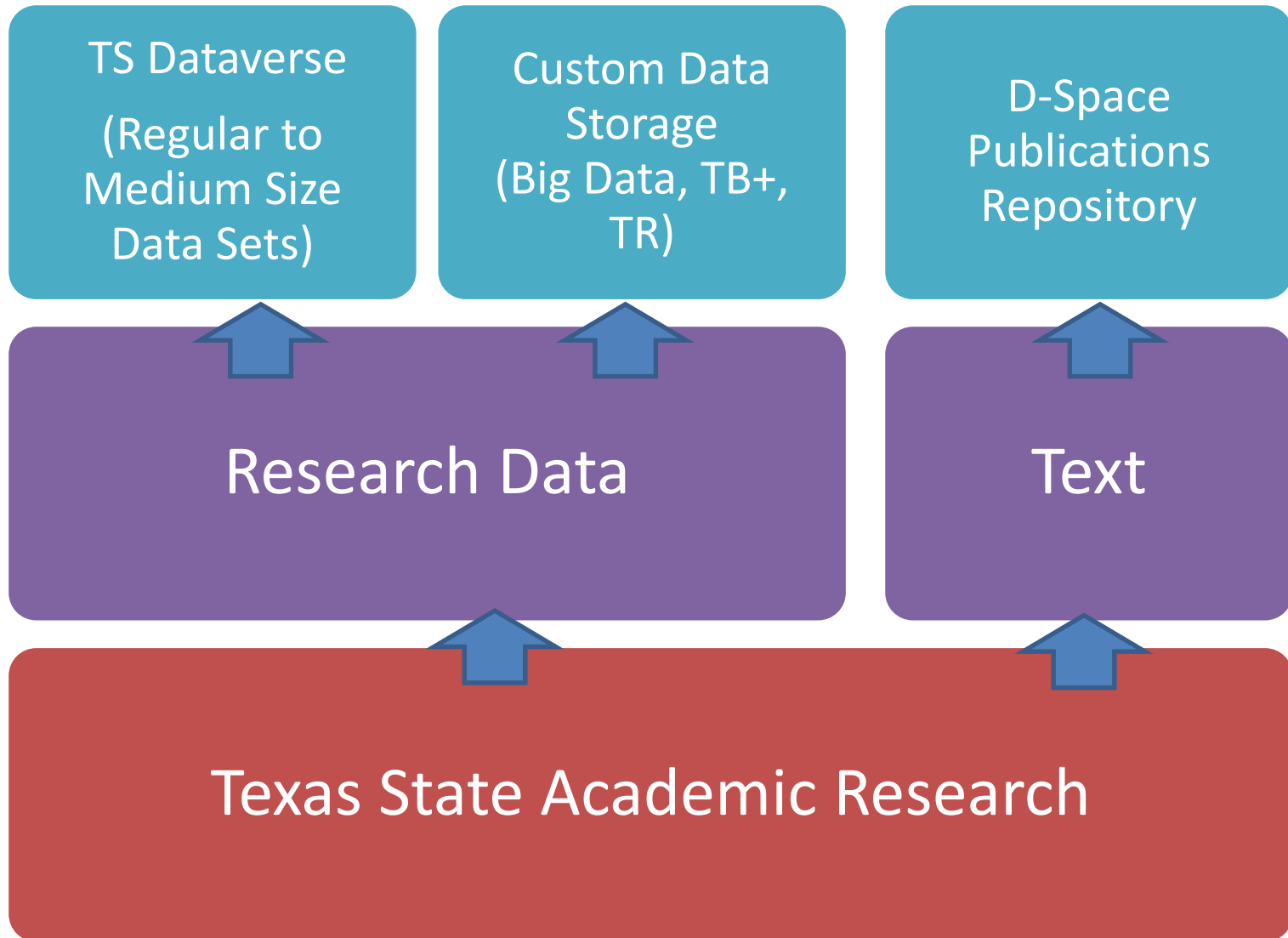


1) Mixed

2) Mediated

3) Unmediated
(Direct)

Texas State Repositories Architecture



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

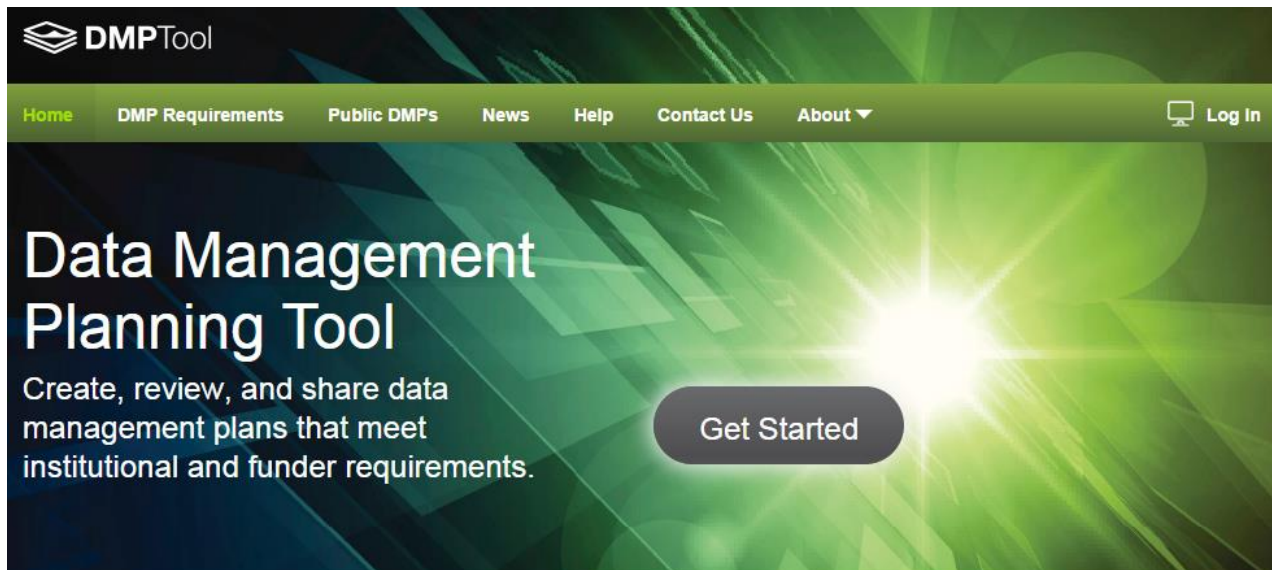
(Data may require specialized university IT Support, i.e. terabyte/petabyte drives, Pointers etc.)

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



[Overview Video](#)

Customizable
Plan Outline Tool
Resource Links
Supports All
Major Funders

 PUBLIC DMPs	 DMP TOOL NEWS	 DMP TOOL HELP
List of sample data management plans provided by DMPTool users. » CAREER: Parietal Cortex and the Transformation of Spatial Cognition into Action	Latest information about data management and the DMPTool. » US Dept of Energy data management requi... » MRC & RCUK Research Council	Overview of how to use the tool, plus resources and guidance on data management. » Frequently Asked Questions » Create a DMP

Connections with
Office of Sponsored
Research and
Other Relevant
University Offices
Library/Dataverse
Templates

<https://dmptool.org/>
California Digital Library

Part III: Human Resource Infrastructures (Working Teams)

Texas State University Dataverse

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

Dataverse

TEXAS STATE
UNIVERSITY LIBRARIES

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.

Texas State University Dataverse

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



Electronic Thesis and Dissertations (ETD) Repository (D-Space)

Future Possibilities: VIREO, DATA REPOSITORY CONNECTIONS

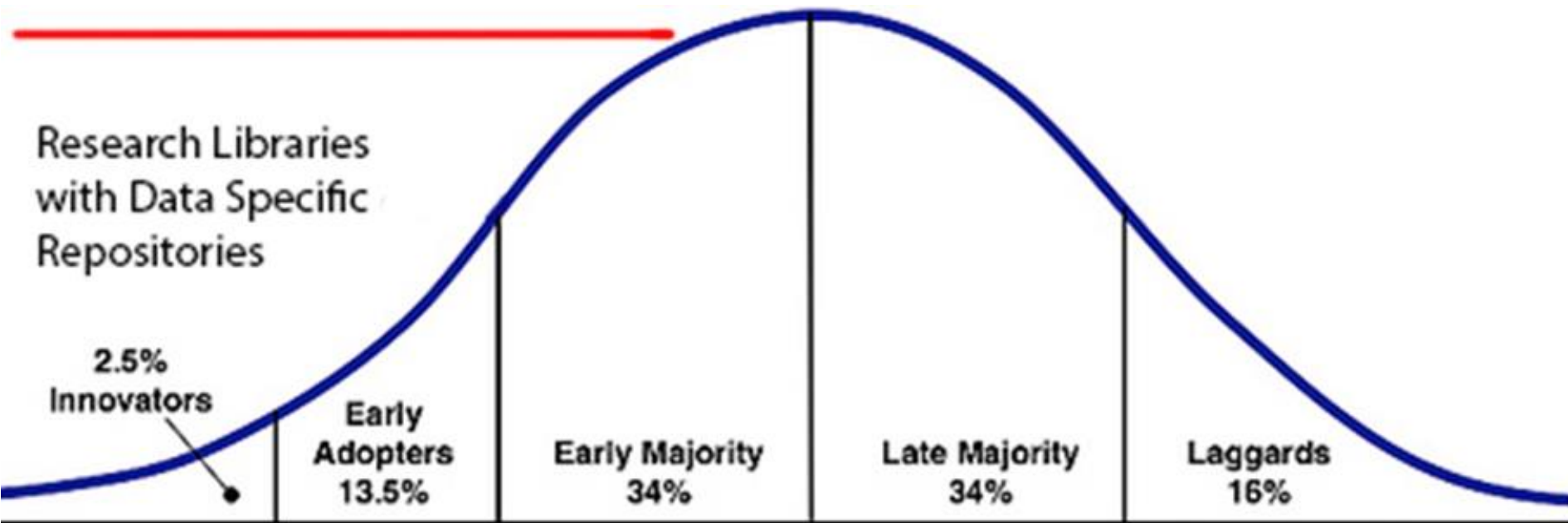
Working with the Data – Support Mechanisms

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>
- Borgman, C. 2015. *Big Data, Little Data, No Data. Scholarship in the Networked Age*. MIT Press
- 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>
- Data Reproducibility Crisis. Nature. <http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970>
- 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/>
- Force 11. Data Citation Principles. <https://www.force11.org/group/joint-declaration-data-citation-principles-final>
- Purr. (Purdue Institutional Data Repository). <https://purr.purdue.edu/>
- Hubzero. <https://hubzero.org/>

Further Links/References

- Figshare. <http://figshare.com/>
- ICPSR Data Management & Curation. <http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/>
- Research Data Management. Principles, Practices, and Prospects (November 2013). *Council on Library and Information Resources*. <http://www.clir.org/pubs/reports/pub160>
- Cox, A. and Pinfield, S. Research Data Management and Libraries. *Journal of Librarianship and Information Science*. June 2013.
- Fearon, D & Sallans, A. C. (January 2014). Institutional Research Data Management: Policies, Planning, Services and Surveys. Coalition for Networked Information. <https://www.youtube.com/watch?v=rvbrW7S2fes> (video presentation)
- Data Management for Libraries: (LITA Guide) <http://www.alastore.ala.org/detail.aspx?ID=10737>
- *NMC Horizon Report: 2014 Library Edition*. <http://cdn.nmc.org/media/2014-nmc-horizon-report-library-EN.pdf>
- “Research Data Management”. pp. 6-7 and pp 24 – 45.
- Holden, J. Memorandum for Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Research (2013).
http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
- Green, A. Macdonald, S and Rice, R. Policy-making for Research Data in Repositories: A Guide. DISC-UK.
<http://www.disc-uk.org/docs/guide.pdf>
- Research Data Management in the Arts and Humanities (2013). University of Oxford.
<http://www.dcc.ac.uk/events/research-data-management-forum-rdmf/rdmf10-research-data-management-arts-and-humanities> (Conference Presentations)
- **Texas Data Repository**. TDR Final Report (October, 2016), Selection Process, Aug. 2015, Peace Williamson et al. UT Arlington, Data Competencies. TDL Texas Data Repository Presentation. Video., Kristy Park, Santi Thompson et al (October, 2016)
- **Uzwyszyn, R.** 2016. Research Data Repositories: The What, When, Why and How of Data Research Repositories *Computers in Libraries*.

Comments/Questions

Contact Information:

Ray Uzwyszyn, Ph.D. MBA MLIS

Director, Collections and Digital Services

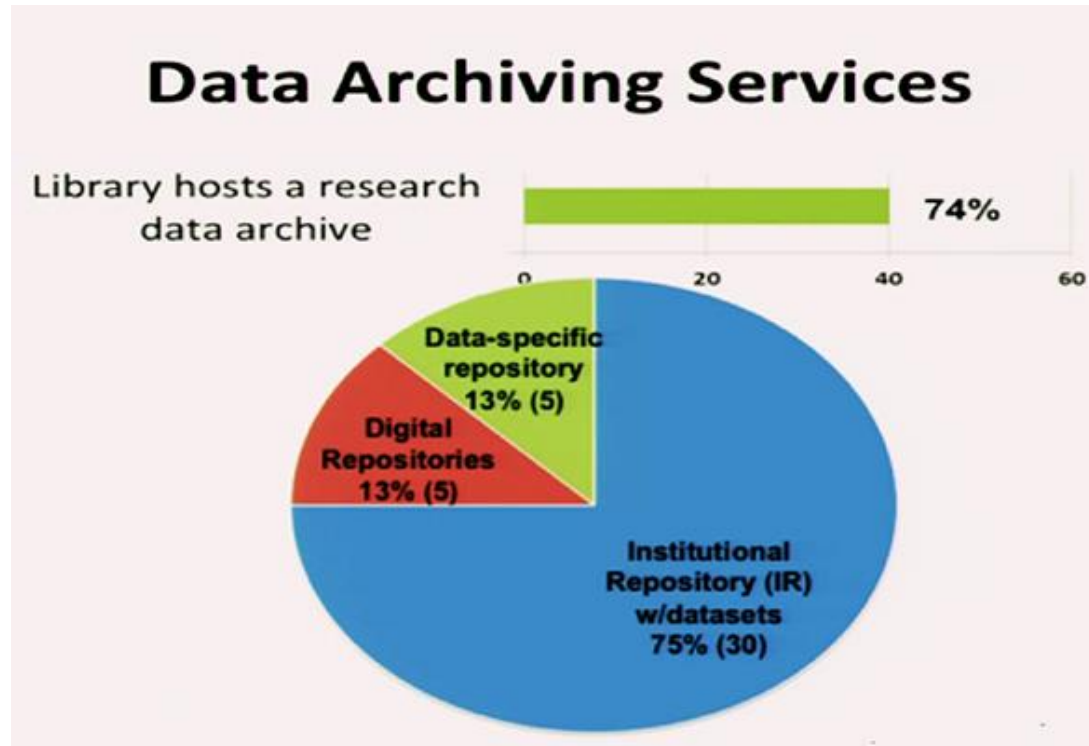
Texas State University Libraries

ruzwyszyn@txstate.edu (512)245-5687

Academic Research Libraries

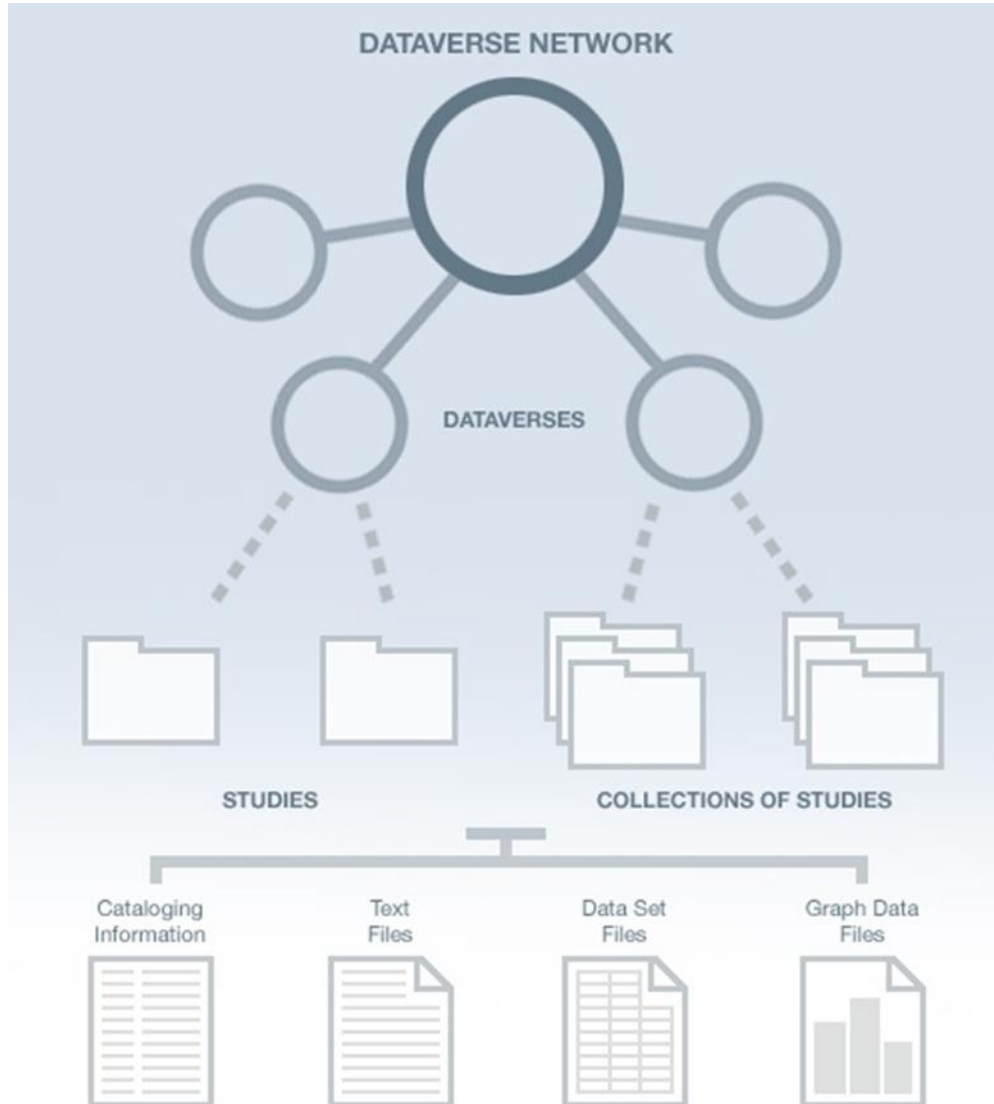
Environmental Scan

Online Data Research Repositories (CNI)



Fearon, D & Sallans, A. C. (January 2014) Institutional Research Data Management: Policies, Planning, Services and Surveys. Coalition for Networked Information. <https://www.youtube.com/watch?v=rvbrW7S2fes> (54 ARL Libraries currently offer data management services_)

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.)

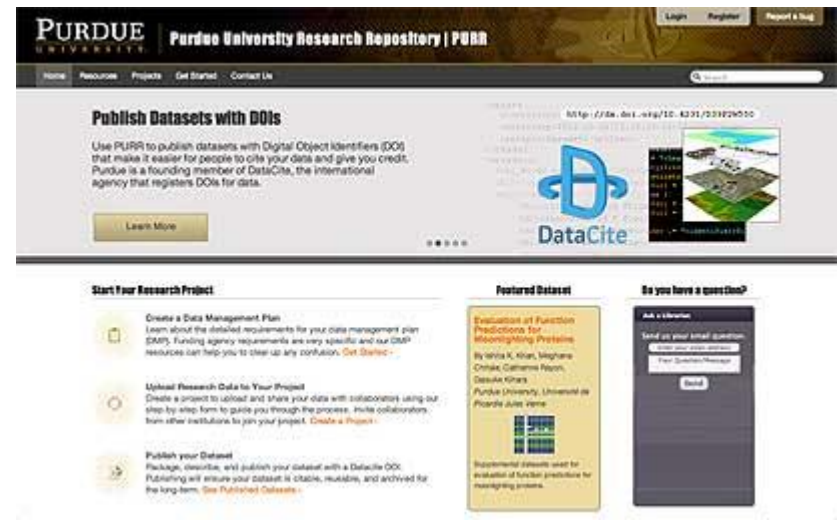
Graph Data Files

PURR and Hubzero: Purdue's Data Management System

- 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)
- 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/>
- [Getting Started](#), [Downloadable](#) and [Hosted Options](#)
- [Hubzero Video](#), [Hubzero2](#)

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)



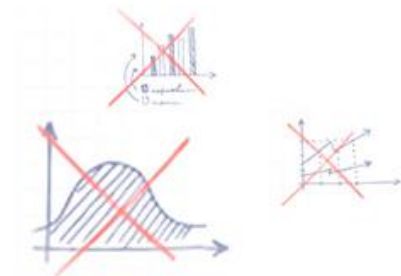
1GB of private space

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



<https://www.force11.org/group/joint-declaration-data-citation-principles-final>

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)

The screenshot shows the Texas State University Digital Collections Home page. The header features the university's logo and name, along with the tagline 'The rising STAR of Texas' and the name of the library, 'Albert B. Alkek Library'. The main content area is divided into two columns. The left column contains a search bar with a 'Go' button, a link to 'Advanced Search', and a 'Browse' section with links to 'All of Digital Collections', 'Communities & Collections', 'By Issue Date', 'Authors', 'Titles', and 'Subjects'. Below this is an 'Author's Corner' section with links to 'About Digital Collections', 'First-time Users', 'Submission Types', 'License and Agreements', 'FAQs', and 'Login'. The right column has a 'Digital Collections Home' link, a 'Texas State University' heading, a 'Digital Collections Repository' link, and a paragraph describing the repository's purpose. Below this is a 'Communities in Digital Collections' section with a link to 'Select a community to browse its collections.' and a list of five categories: 'Departments, Schools, Centers & Institutes', 'Dissertations & Theses', 'Journals & Peer Reviewed Series', 'The Wittliff Collections', and 'University Archives'.

TEXAS STATE UNIVERSITY
The rising STAR of Texas
Albert B. Alkek Library

Search Digital Collections

Go

Advanced Search

Browse

All of Digital Collections

Communities & Collections

By Issue Date

Authors

Titles

Subjects

Author's Corner

About Digital Collections

First-time Users

Submission Types

License and Agreements

FAQs

Login

Digital Collections Home

Texas State University

Digital Collections Repository

The Digital Collections repository is a service that provides free and open access to the scholarship and creative works produced and owned by the Texas State University community. The Digital Collections centralizes, preserve and makes accessible the knowledge generated by the university community, which includes faculty publications, theses & dissertations, plus digitized materials from The Wittliff Collections, the University Archives, and other materials unique to Texas State University. It is a professionally maintained archive that gives the university's intellectual and creative output increased visibility and accessibility over time.

Communities in Digital Collections

Select a community to browse its collections.

- Departments, Schools, Centers & Institutes
- Dissertations & Theses
- Journals & Peer Reviewed Series
- The Wittliff Collections
- University Archives

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

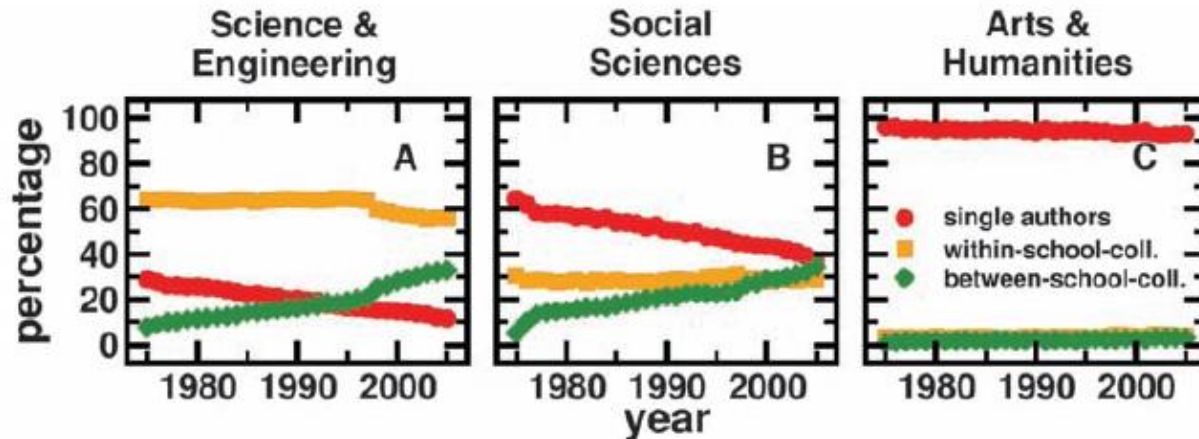


Fig. 1. The rise in multi-university collaboration. By comparing the incidence of papers produced by different authorship structures, we see that the share of multi-university collaborations strongly increases from 1975 to 2005. This rise is especially strong in SE (A) and SS (B), whereas it appears 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.

Jones et al. (2008). *Science* 322: 1259-1262.

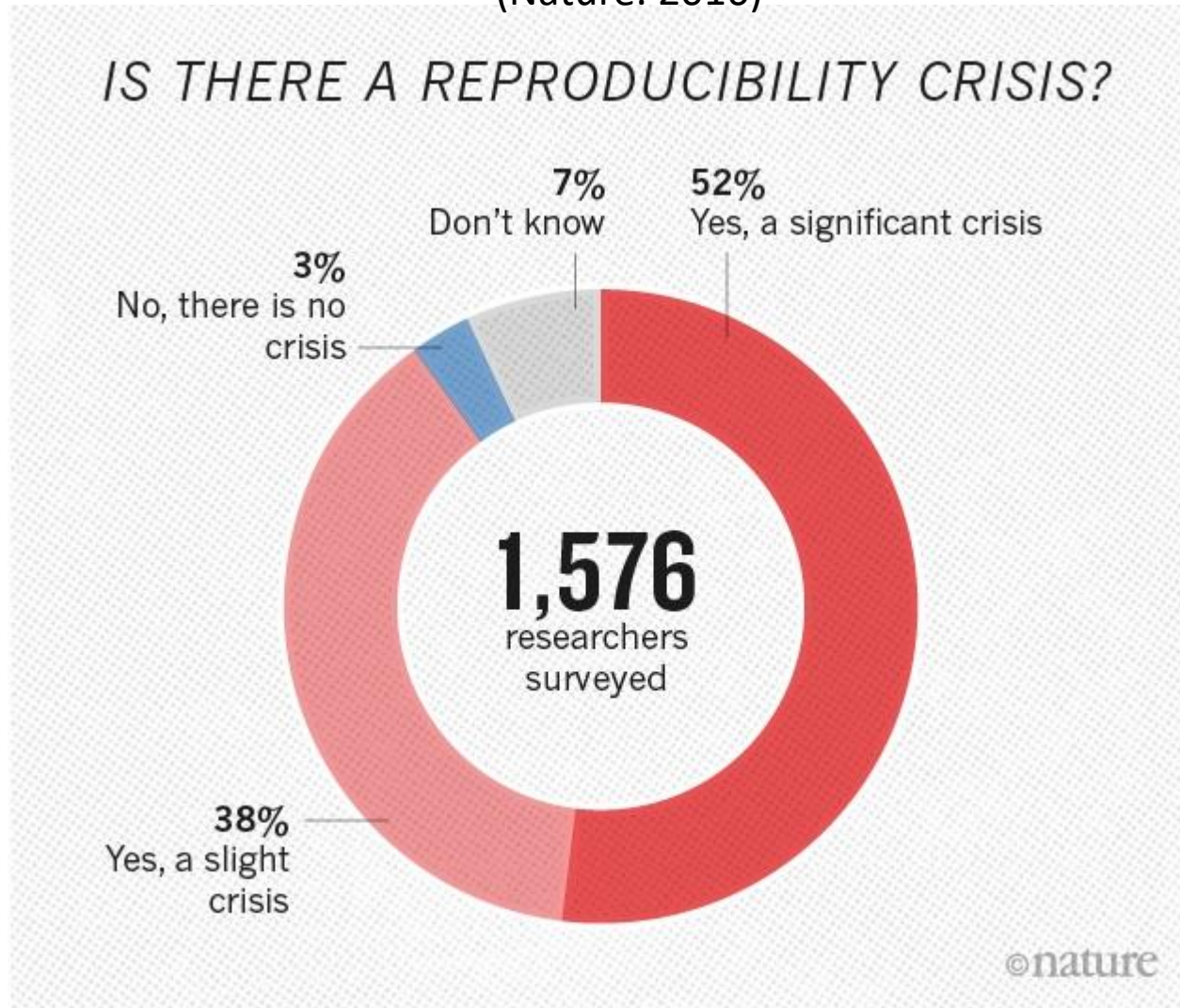
Data Sharing



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 >](#)

