Developing and Implementing an Online Research Data Repository for Your University or College Campus

Raymond J. Uzwyshyn
Texas State University, ruzwyshyn@txstate.edu

Follow this and additional works at: https://digital.sandiego.edu/symposium

Part of the Data Storage Systems Commons

Uzwyshyn, Raymond J., "Developing and Implementing an Online Research Data Repository for Your University or College Campus" (2017). Digital Initiatives Symposium. 1.
https://digital.sandiego.edu/symposium/2017/2017/1

This 45-minute concurrent session is brought to you for free and open access by Digital USD. It has been accepted for inclusion in Digital Initiatives Symposium by an authorized administrator of Digital USD. For more information, please contact digital@sandiego.edu.
Developing and Implementing an Online Research Data Repository for Your University or College Campus

Presenter 1 Title
Developing and Implementing an Online Research Data Repository for Your University or College Campus

Session Type
45-minute concurrent session

Abstract
Data-driven research is becoming increasingly important on university and college campuses. Most US federal and many international granting agencies mandatorily require that researchers applying for public grants possess a data management plan and make their research and data publically available through online access. This presentation overviews online research data repositories and implementation strategies for university and college campus libraries. The presentation pragmatically surveys this newer technology landscape and how organizations can begin to think about and implement an online data research repository. This session will survey the landscape but also makes use of practical example from Texas State University and the Texas Data Repository, a large state consortial data repository customizing and utilizing Harvard’s open source Dataverse infrastructure.

Location
KIPJ Theatre

Keywords
Research Data Repository, Data Management, Open Data, Federal Grant Compliance, Dataverse, Texas State University, Open Educational Resources

Creative Commons License
This work is licensed under a Creative Commons Attribution 4.0 License.

This 45-minute concurrent session is available at Digital USD: https://digital.sandiego.edu/symposium/2017/2017/1
Developing and Implementing An Online Research Data Repository for Your University or College Campus

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

- Online Way to Manage a Researcher’s Data/Metadata
- Permalinking Strategy for Online Data Citation/Access
- Way to Manage Federal Grant Compliance
- Long Term Data Archiving, Preservation, Sharing Strategy
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, 2013 USDA)
What makes Data Management Repositories useful?

• Leverage and make available faculty, departmental and institutional research
• Allow publication of negative data (less research replication)

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

- Becoming Integral Part of Evolving Science, Social Science and Humanities Research Process (Promote accuracy, efficiency, sharing)
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
Types of Research Data Repositories

1) Project specific
   (usually large single faculty/faculty team projects)

2) Discipline specific
   (i.e. Purdue Nanohub/Nanotechnology,
    Archeological Data from Academic Center, etc.)

3) Institutional Repository
   (either institution wide or consortial)
Academic Research Libraries
Environmental Scan
Online Data Research Repositories (CNI)

Specific and All-Purpose Data Repository Platforms

<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

• May be hosted or installed on a university’s server
• Each software contains different ranges of management, collaborative options
• Open source and proprietary options
• Ingestion of Various Data Types
  (from Excel to SPSS to more esoteric disciplinary specific formats)
Environmental Scan of Current Possibilities for Your Institution

- System Performance
- Robustness
- Usability
- platform availability
- an active open source community

Conclusion: The group recommends that TDL adopt Harvard’s Dataverse to facilitate the discovery of research data.

Data Repository Working Group Report
(August 28, 2015)
Dataverse
Harvard’s Open Source Research Data Solution

Software framework that enables institutions to host research data repositories

Allows data sharing, control, persistent 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 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
Dataverse Data Citation and Metadata Example
Dataverse Metadata Example
(From the Simple to Complex)

<table>
<thead>
<tr>
<th>Citation Metadata</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td>Data from “Social determinants of unmet hospitalisation need amongst the poor in Andhra Pradesh, India: A cross-sectional study.”</td>
</tr>
<tr>
<td><strong>Author</strong></td>
<td>Name</td>
</tr>
<tr>
<td></td>
<td>Nagulapalli, Srikant</td>
</tr>
<tr>
<td></td>
<td>Affiliation</td>
</tr>
<tr>
<td></td>
<td>Andhra University</td>
</tr>
<tr>
<td><strong>Identifier</strong></td>
<td>Identifier Scheme</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Subject</strong></td>
<td>Medicine, Health &amp; Life Sciences</td>
</tr>
<tr>
<td><strong>Keyword</strong></td>
<td>unmet hospitalisation need</td>
</tr>
<tr>
<td><strong>Production Date</strong></td>
<td>2013-06-01</td>
</tr>
<tr>
<td><strong>Depositor</strong></td>
<td>Privileged, admin</td>
</tr>
<tr>
<td><strong>Deposit Date</strong></td>
<td>2013-08-03</td>
</tr>
</tbody>
</table>

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

DC

Data Citation Principles

FORCE

The Future of Research Communications and e-Scholarship

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

Main Working Group & Subcommittees:
• Policy and Governance
• Workflows and Outreach
• Budget/Business Model
• Technology
• State Data Repository Symposium

Final Report October, 2016
TDL Consortium of 22 universities across Texas leveraging technological cooperation among academic libraries,
The Many Planning Aspects of the New World of Data Research Repositories
Texas Data Repository Consortial Architecture

Research Study Data

- Data Set Files
- Metadata (Data Describing the data)
- Paratextual Research Material (Methodology, Field Notes, Multimedia, Graphs, Programs etc.)
Texas Digital Library Test 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 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
UT Medical Branch Dataverse
University of Texas Dataverse
Texas State University Dataverse
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!

http://data.tdl.org
(UX Usability Focus)
Repository Service Models

1) Mixed
2) Mediated
3) Unmediated (Direct)

Texas Data Repository

Member Libraries (service & outreach)

Researchers (deposit, search, publish)
Texas State Data Repository Architecture

1. Texas State Academic Research

2. Research Data

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

3. Reports and Publications
One Size Does Not Fit All Data Project Needs

Types of Data Projects (Sizes)

1) Normal range
Files/Data Fit on Server/Cloud, may be uploaded, Dataverse, Purr)

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

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

Overview Video

Customizable
Plan Outline Tool
Resource Links
Supports All
Major Funders

Connections with
Office of Sponsored
Research and
Other Relevant
University Offices

https://dmptool.org/
California Digital Library
Data Management Plan Support

The Library Supports:

Publication repositories
D-Space

Data repositories, Texas Data Repository

Human Resource Infrastructure

Data Repository Liaison
Publication Repository Liaison
Specialized Metadata Liaison
Subject Liaisons (Outreach)
Committee for Workflows, Standard & Policies

Current/Future Hires
Data Visualization and Analytics Librarian (Tableau, Bayesia)
Digital Collections Librarian (TDR Dataverse/D-Space)

http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research
Institutional Repository Connections
(MIT, D-Space)

Faculty publications, white papers, preprints, theses, dissertations, working projects

Larger Idea, Grant Compliance, Enabling Faculty Research Online, Raising Research Visibility,
Electronic Thesis and Dissertations (ETD) Repository (D-Space) Possibilities

Co-publish data sets in ETD (D-Space) and Data Repository, Links in metadata in D-SPACE and DATA REPOSITORY

Future Possible ETD (D-Space), VIREO, DATA REPOSITORY CONNECTIONS
Data Repository Adoption Lifecycle (2017)

- Research Libraries with Data Specific Repositories
- Early Adopters 13.5%
- Early Majority 34%
- Late Majority 34%
- Laggards 16%

2.5% Innovators
Further Links/References

- ARL NSF Data Sharing Policy and Resource Links, [http://www.arl.org/focus-areas/e-research/data-access-management-and-sharing](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](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](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*

- California Digital Library DMT Tool: [https://dmptool.org/](https://dmptool.org/)
- Chronopolis: [http://www.digitalpreservation.gov/partners/chronopolis.html](http://www.digitalpreservation.gov/partners/chronopolis.html)
- Dataverse (Data Science Site). [http://datascience.iq.harvard.edu/dataverse](http://datascience.iq.harvard.edu/dataverse)
- Data Information Literacy Guide. [http://www.datainfolit.org/dilguide/](http://www.datainfolit.org/dilguide/)
- Duracloud: [http://www.duracloud.org/](http://www.duracloud.org/)
- Purr. (Purdue Institutional Data Repository). [https://purr.purdue.edu/](https://purr.purdue.edu/)
- Hubzero. [https://hubzero.org/](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”. pp. 6-7 and pp 24 – 45.
• 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)
Comments/Questions
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.

Data Sharing

Currently, 80% of researchers do not share their data

Research Data Reproducibility Crisis

(Nature. 2016)

IS THERE A REPRODUCIBILITY CRISIS?

1,576 researchers surveyed

- 38% Yes, a slight crisis
- 3% No, there is no crisis
- 7% Don’t know
- 52% Yes, a significant crisis

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 ›](#)