Texas Digital Library Dataverse Implementation Working Group Final Report

Released September 30, 2016

Texas Digital Library

Table of Contents

Introduction

Since September 2015, the Texas Digital Library (TDL) Dataverse Implementation Working Group (DIWG) has worked with Texas Digital Library staff to pilot and implement a consortial repository for small to medium-sized research data, as well as to develop policies and workflow documentation associated with a data repository service.

Comprised of 14 librarians and technical staff across six universities and the TDL, the DIWG's charge was to "pilot test, assess, and launch a consortial repository for research data archiving and management," addressing costs and possible funding models, technical configuration of the repository, workflows and outreach, policy and governance, and metadata needs.[1]

The DIWG built upon the work of a predecessor group – the TDL Data Management Working Group – which evaluated multiple available data management platforms and recommended the use of Dataverse as a consortial service.[2] The result of the group's work is the Texas Data Repository (https://data.tdl.org), a platform for publishing and archiving datasets and other data products created by faculty, staff, and students at Texas higher education institutions. The repository is built in Dataverse and is intended for sharing small- to medium-sized datasets from any discipline.

The need for research data management services

The primary driver behind the TDL's work to implement a research data repository is the increased need -- derived from funders, governments, advocacy groups, and others -- to improve accessibility and usability of research data (as well as other research outputs). This increased focus on data sharing and re-use was famously accelerated by the 2013 OSTP Directive that required plans from federal agencies to support increased public access to the results of the research they fund.[3] The Texas Digital Library has since its inception, supported greater access to scholarly communication through a number of hosted services; infrastructure support for open data is a natural extension of its mission to increase access to scholarly work, thus supporting academic research and enhancing scholar recognition.

Along with increased focus on openness, a parallel and related consensus about the value of data as "first class objects" of scholarly research has grown, and, as a result, so has the need to effectively curate these research products, make them citable, and preserve them for future generations.[4] Academic libraries see themselves as key stakeholders with relevant expertise and skills to support the management of research data; the Association of Research Libraries has identified RDM as an essential

[1] See Appendix A of this report for a full list of TDL Dataverse Working Group members.
[2] Herbert, B., Buckbee, M., Donald, J., Esteva, M., Lyon, C., Peters, C., Park, K. L., Steans, R., Thompson, S. (2015) TDL Data Management Working Group Report. Retrieved from TDL Repository. http://hdl.handle.net/2249.1/68438.
[3] United States White House Office of Science and Technology Policy. (2013, February 22). Expanding public access to the results of federally funded research. Retrieved from https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
[4] Clement, G. P., & Schiff, L. R. (2015). Mapping the Landscape of Research Data: How JLSC Contributors View this Rapidly Emerging Terrain. *Journal of Librarianship and Scholarly Communication*, 3(2). http://dx.doi.org/10.7710/2162-3309.1279

component of next-generation libraries through multiple initiatives, including SHARE (http://www.share-research.org/). Four-year academic libraries continue to see data management services as important services they can (and do) offer, although many struggle to implement them effectively because of lack of expertise, resources and/or institutional will.[5] TDL member institutions – representing four-year institutions of widely varying sizes and means – have consistently expressed through member surveys and informal assessments a need for infrastructure to support research data management services, though many of them do not have the wherewithal to support that infrastructure on their own.

It is within these interrelated contexts – increased advocacy for open data and the establishment of research data management services as an important role for research libraries – that the Texas Digital Library undertook an effort to implement a consortial research data repository. By working together across institutions, with the greater resources and expertise available through our collective effort, we can collectively address the barriers to the establishment of research data management services at our individual campuses. We believe a consortial implementation best supports the development of new data management services and programs at individual institutions and provides curation services at a lower cost per institution.

A unique service model for research data management

The Texas Data Repository (TDR) seeks to respect the spirit of a "loose federation" that has guided the TDL's deployment of other services over its history, honoring the need for local control over workflows while creating a meaningful shared service.

In this spirit, the TDR will be operated and governed according to a hybrid service model, with TDL staff hosting a single repository for all research data content statewide (as represented through the TDL membership) and representatives at TDL member libraries providing service to their local university constituencies. This hybrid model marries the benefits of a single repository (i.e. operation of information technology at scale) with local institutional control of associated services and programs.

This model offers flexibility to each member institution, which can use the repository in multiple ways, according to its own needs and resources. An institution can choose, for instance, to build a robust suite of services around support for data management planning and curation with the Texas Data Repository at its core. Alternately, a smaller academic library may choose to offer the service to faculty as a fully "self-service" option for research data sharing and publication. Either model is supported by the Texas Data Repository as it currently exists.

Methodology

The working group organized into subgroups that were focused on addressing four objectives: budget and business plan, policy and governance, technical configuration, and workflows and outreach. They

[5] Tenopir, C., Hughes, D., Allard, S., Frame, M., Birch, B., Baird, L., Sandusky, R., Langseth, M., and Lundeen, A. (2015). Research Data Services in Academic Libraries: Data Intensive Roles for the Future? *Journal of eScience Librarianship*, 4(2). http://dx.doi.org/10.7191/jeslib.2015.1085

divided their work into two broad phases, with the completion and assessment of a pilot project serving as a unifying activity across the phases. In the first phase, the group devised draft workflows and policies and configured the repository software in preparation for a pilot project. See Table A, Phase One Activities and Outputs for more information on phase one work.

After compiling a preliminary user guide, drafting policy documentation, and building a production-level repository, the working group launched a pilot project in May 2016. The goal of the pilot study was to ensure that Dataverse implementation and associated services met the needs of stakeholders by assessing the experiences of researchers and librarians. Prior to starting the pilot, the working group submitted and received formal approval for an IRB application.[6] Lasting a month, the pilot project solicited participation from TDL member librarians and researchers. It asked volunteers to conduct a series of required and optional tasks, aimed at testing the system's usability and performance against live users with "real-world" data. Required tasks included key steps in submitting, describing, and making research data accessible in the repository. Optional tasks allowed participants to complete tasks that could enhance a dataset and the repository, including the creation of multiple versions of a dataset, using built in functionality to visualize some types of datasets, and adding institutional logos to brand your dataverse environment.

Upon completion of the required and any optional tasks, participants completed a pilot project survey.[7] Working group members analyzed the results of this survey to assess the current state of the repository, along with its associated documentation. Feedback from the survey suggested that researchers appreciated the extensive amount of documentation but struggled to determine how best to get started with the deposit of data into the repository. Respondents also expressed confusion on key repository terminology and emphasized a need for hands-on assistance with key elements of data management, including planning and describing research data. For more results from the pilot project, see Appendix B: Pilot Project Survey Results.

Table A: Phase One Activities and Outputs/Outcomes

Subgroup	Activities	Output/Outcomes
Budget and Business Plan	Gathered cost data for running a data repository, Explored alternative data repository business models (e.g., Researcher pays, Institution pays, Grant funding, Hybrid Model), Identified external funding sources (grants foundations),	• Cost model based on content acquisition rates at Harvard's Dataverse repository • Decision to operate the Texas Data Repository without additional funds during first year of operation

[6] The working group submitted a complete IRB application to the University of Texas at Austin, which acted as the pilot project's "institution of record." Once approval was obtained from UT-Austin, the group submitted subsequent applications to the other universities and colleges who had representatives on the working group.
[7] See Appendix B: Selected Pilot Project Survey Results.

	Developed a cost assessment plan for a pilot repository.	
Policy and Governance	Generated user and member policies that addressed collection development, data acquisition, metadata, access and rights, long-term storage, and digital preservation.	• TDL Policies (Ver. 1)[8] • TDL Metadata Dictionary (Ver. 1)
Technical Configuration	Configured various facets of the software, including access controls, identity management features, and tools for data visualization, Integrated Dataverse repository with EZID in order to mint DOIs (digital object identifiers) for published datasets, Investigated security and backup protocols, Testing file upload for max file size limits and upload times.	• Pilot project production repository with DOI integration • Determination of 2GB as maximum file size
Workflows and Outreach	Identified user tasks and generated draft documentation for a TDR user workflow, including instructions on the deposit, publication, deaccessioning, and reuse of research data in TDR.	• TDR User Guide (Ver. 1)

Phase two of the Dataverse Implementation Working Group addressed the issues raised by pilot project respondents. It focused on improving the user experience by drafting clearer instructions, creating more user-friendly websites, and generating "talking points" that explained the benefits of the repository for the university, the researcher, and the librarian. The working group organized work in Phase Two around subgroups from Phase One with the addition of several teams to address specific issues. See Table B, Phase Two Activities and Outputs for more information on their work.

Table B: Phase Two Activities and Outputs

Subgroup	Activities	Output
Budget and Business Plan	Crafted the value proposition of the TDR including: • Aggregating data in one place	• Information Sheets for Administrators, Researchers, and Librarians

[8] See Appendix C: TDL Member Policies.

	• Fulfilling Federal mandates • Making research widely available • Citing data in publications through DOI	
Policy and Governance	Created a memorandum of understanding document that outlines the roles and responsibilities of TDL and member institutions, Drafted recommendations around digital preservation and long-term repository governance, and Finalized the TDR policy document.	• TDR Policies • TDR Metadata Dictionary • TDR Memorandum of Understanding[9] • Recommendations for digital preservation • Recommendations for repository governance
Technical Configuration	Implemented secure authentication, including Shibboleth option for integration and local authentication and DUO for two-factor authentication, Developed reporting capabilities using Dataverse API, Alignment of production Dataverse repository with landing page and brochure WordPress site (see "Ad-Hoc Webpage team" below).	• Production version of TDR with Shibboleth and two-factor authentication integrated • CSV reports to be generated regularly by TDL staff and disseminated to data repository librarians • Revised navigation and branding of Dataverse site to align with WordPress site
Workflows and Outreach	Integrated feedback from the pilot project to refine and finalize documentation.	• Final version of the TDR User Guide
Ad-Hoc Webpage Team	Created a new landing page and "brochure" website for TDR with improved user interface design.	• Soft launch of the TDR landing page in WordPress
TDL Dataverse Implementation Working Group	Certain tasks required the input and contributions of the entire working group to resolve. One question focused on whether or not to have distinct folders (dataverses) for each TDL member. Another question asked the formal name of the repository. The working group conducted a naming contest to identify the Texas Data Repository	• Implementation of institutional dataverses in the final version of the TDR • Naming the repository the "Texas Data Repository"

[9] See Appendix D: Memorandum of Understanding.

Outcomes

Phase Two, and with it the work of the Dataverse Implementation Working Group, completed on September 1, 2016. With this end comes numerous products that, collectively, comprise the initial repository software, web page, and supporting documentation. Table C lists the products produced by the working group. http://data.tdl.org/

Table C: TDL Dataverse Implementation Working Group Products

Final Product	**Location**
TDR Homepage	https://data.tdl.org
TDR Repository	https://dataverse.tdl.org
TDR User Guide	https://data.tdl.org/user-guide
TDR Metadata Dictionary	https://data.tdl.org/wp-content/uploads/2016/09/TDR-Metadata-Dictionary.pdf
TDR User Policies	https://data.tdl.org/policies
Researcher Information Sheet	https://data.tdl.org/wp-content/uploads/2016/09/Researcher-Information-Sheet.pdf
Librarian Information Sheet	https://data.tdl.org/wp-content/uploads/2016/09/Librarian-Information-Sheet.pdf
Administrator Information Sheet	https://data.tdl.org/wp-content/uploads/2016/09/Administrator-Information-Sheet.pdf
TDR Member Library Policies	Appendix C: Member Library Policies
TDR Memorandum of Understanding	Appendix D: Memorandum of Understanding

Recommendations

Prior to its completion, the working group identified a set of recommendations that should be addressed by other data-related groups in the future.

1. There is an immediate need for materials introducing the TDR to librarians from TDL member institutions. We recommend that the organizing team for the TDL Data Symposium, a two-day event scheduled for November 2016 that will focus on exploring data management issues and will include an in-depth training session on TDR, should develop a librarian guide for assisting those data repository librarians who will be performing curation and managerial functions for TDR. Additionally, this team should compile a training curriculum for instructing data repository librarians on the use and administration of the repository.

2. Training for a data repository librarian will be critical for the long-term governance of TDR. The working group recommends that the TDL form a statewide steering committee comprised of

one librarian selected from each participating member institution, which we will call data repository librarians. The steering committee should act as the governing body for the data repository. The group's potential activities could include:

- Sharing questions and feedback from local users with the larger committee and formulating responses, including revisions to documentation, workflows, and policies
- Develop best practices for the annotation of data sets to improve the usability of the data to other researchers
- Identifying data repository and/or workflow issues and investigating possible solutions
- Conducting periodic needs assessments to improve user experience and data repository performance
- Developing new features and functionality for the data repository in response to identified needs

Member institutions should develop their own selection criteria for identifying the data repository librarian. The candidate is not required to hold any specific degree or title to perform the data repository librarian duties but should have interests or inclinations in data curation. Any experience with regards to data curation, data visualization, or data intensive disciplines would be beneficial.

3. While TDR offers a robust and rich data archiving application for TDL members, it does not fully address all digital preservation activities needed for long-term access to data. In the future, TDL and the data repository librarian steering committee should also consider addressing the following digital preservation concerns and opportunities.

   A. The curation of data files and content is left entirely to the person uploading content. This presents a series of issues related to file management, including:
      - Redundancy of content
      - Non-standard file naming
      - File format(s) that are proprietary and/or not the current version
      - Files containing viruses and malware

   B. Because the system allows for self-deposit, metadata associated with data may be incorrect or incomplete

   C. The data repository is unable to perform virus/malware checks either at time of upload or subsequent to it. Another mechanism may be needed to perform this task.

   D. The data repository is unable to perform fixity checks at regular intervals and to produce audit report available to administrator(s). Another mechanism may be needed to perform this task.

   E. There is currently no viable method to migrate data files (or request migration) if formats become obsolete or versions superseded

F. There is potential difficulty in maintaining and updating users' and data repository librarian's contact information, such as email addresses and home institutions, needed when addressing changes to the data or repository management, including:
   - The retention and/or deaccessioning of data,
   - The modification of metadata and/or data,
   - Important news and updates about the current and long-term maintenance of the system.

Implementation Guide and Timeline

Roles and Responsibilities

Because of the hybrid nature of the Texas Data Repository service, it is essential that institutions participating in the service understand the roles and responsibilities of all parties. These roles are outlined in general terms in Table D below.

Table D: Texas Data Repository Service Roles and Responsibilities

Group	Roles and Responsibilities
Texas Digital Library staff	• Stewardship, technological oversight, and upgrades of the data repository software infrastructure and associated websites • Assuring access to and secure backup of data submitted to the repository • Coordination of a membership-wide steering committee of data repository librarians • Provision of technical support for all Texas Data Repository users via the TDL Helpdesk, referrals to relevant Data Repository Librarians when needed • Provision of training and professional development opportunities to data repository libraries
Member Library	• Appointment of an individual to serve as data repository librarian (and communication with TDL when a change to the data repository librarian is needed) • Local promotion and support of the Texas Data Repository service within its campus community • Communication with TDL of any institutional requirements necessary to authorize and sustain the Texas Data Repository (e.g., security authorizations through Central IT offices) and serve as liaison to relevant institutional departments • Establishment of institutional policies around copyright inquiries, takedown requests, and rights decisions and inform TDL when necessary of repository actions required

	• Recommendations, in collaboration with TDL, for data repository options should the Texas Data Repository be discontinued
Group	**Roles and Responsibilities**
Data Repository Librarian	• Service as the local liaison/contact person for users and other university community members • Maintenance of institutional collections of data (i.e., an "institutional dataverse") • Participation in TDL-wide data repository librarian committee meetings and work • Optionally, the fulfillment of additional, related duties as assigned by the institution. These might include assisting in other data curation roles such as data ingest, metadata creation and modification, and data management planning, as well as contacting users through the data repository interface. • Optionally, communication with researcher/depositors from their institution who self-deposit data in the TDR
Repository Steering Committee	• Attendance at periodic (virtual) meetings to discuss issues of relevance to the Texas Data Repository • Identification and recommendations for resolution of (non-technical) issues related to repository operation and policy • Identification of collaborative work (via working groups, other committees) related to research data management
TDL Governing Board	• Overall oversight of the Texas Digital Library consortium
UT Austin	• Lead agency and institutional home of the Texas Digital Library and the Texas Data Repository
Researcher/Depositor	• Maintenance of research data, associated materials, and metadata self-deposited in the TDR • Adherence to TDR Terms of Use, including the removal or redaction of any personally identifiable or sensitive information from data deposited in the Texas Data Repository • Optionally, may self-deposit data or seek help from appropriate Data Repository Librarian

Implementation Guidelines

TDL member institutions that wish to offer the Texas Data Repository service to their campuses must take several steps to participate. These include:

- Signing a Memorandum of Understanding outlining the roles and responsibilities of the TDL and the member institution as they relate to the Texas Data Repository.
- Assist TDL staff in integrating the TDR with their campus' local authentication via Shibboleth. (This may require acting as a liaison between TDL staff and campus IT staff).

- Identify local staff to serve as a Data Repository Librarian, will serve as the primary liaison between TDL and the member library on matters related to the Texas Data Repository. Among other things, the Data Repository Librarian will:
    - Act as a point of contact for your institution for researchers who want local assistance with research data management
    - Serve on an advisory committee that will help guide TDL as it addresses on-going needs related to the repository

Support for Implementation

The TDL and the TDL Dataverse Implementation Working Group have developed materials and events designed to help member libraries implement and use the Texas Data Repository effectively. These include:

- Marketing materials:

    - The TDL DIWG has developed a set of Texas Data Repository information sheets for various audiences, available on the TDR website.
        - For Researchers: https://data.tdl.org/wp-content/uploads/2016/09/Researcher-Information-Sheet.pdf
        - For Librarians: https://data.tdl.org/wp-content/uploads/2016/09/Librarian-Information-Sheet.pdf
        - For Administrators: https://data.tdl.org/wp-content/uploads/2016/09/Administrator-Information-Sheet.pdf

- Training and Community Development

    - Webinars, including the "Launching the Texas Data Repository" webinar on October 6 (See Timeline section below for more information)
    - TDL Data Symposium, a two-day event offering training on the Texas Data Repository and opportunities for education on research data management

- TDL Helpdesk and Consulting
    - TDL provides member institutions with on-going, individualized support via the TDL Helpdesk and on-demand consulting services to aid with implementation

Timeline for Launch

September 2016: "Soft Launch"

Announce availability of the service
Solicit interest from member institutions
Begin working with interested members to go through implementation process
Refine website and Dataverse repository in preparation for formal launch

October 2016: Launching the TDR webinar

On October 6, 2016, the TDL will hold a webinar entitled "Launching the Texas Data Repository: How to Implement TDR at Your Institution." In this webinar, Kristi Park (Director of the TDL) and Santi Thompson (Head of Digital Repository Services at the University of Houston Libraries and chair of the TDL Dataverse Implementation Working Group) will give an overview of the repository service, requirements for TDL member participation, and plans for the future of the service.

November 2016: TDL Data Symposium

On November 15-16, 2016, the TDL will hold its first ever Data Symposium at Baylor University Libraries. The two-day Symposium is designed to support the development of a **core community of TDL member librarians** who provide (or intend to provide) research data services on their campuses. Along with opportunities to discuss and learn about research data management generally, the Symposium will offer a half-day training workshop on Dataverse and the Texas Data Repository services.

December 2016: Formal launch

After working throughout the fall of 2016 to onboard and train participating libraries, the TDL will formally launch the Texas Data Repository in December with a press release and a second webinar for TDL members.

Appendix A: TDL Dataverse Implementation Working Group Members

Chair: Santi Thompson, University of Houston Libraries

Members:

- Jeremy Donald, Trinity University Libraries
- Denyse Rodgers, Baylor University Libraries
- Sean Buckner, Texas A&M University Libraries
- Bruce Herbert, Texas A&M University Libraries
- Wendi Kaspar, Texas A&M University Libraries
- Cecilia Smith, Texas A&M University Libraries
- Chris Starcher, Texas Tech University Libraries
- Todd Peters, Texas State University Libraries
- Ray Uzwyshyn, Texas State University Libraries

TDL Staff:

- Kristi Park, Texas Digital Library
- Ryan Steans, Texas Digital Library
- Nick Lauland, Texas Digital Library
- Laura Waugh, Texas Digital Library

Appendix B: Selected Pilot Project Survey Results

Results from the TDL Dataverse Implementation Working Group's pilot project identified repository features and services that meet the needs of the pilot participants as well as require further work before launching the repository. Results also emphasized certain perceived repository/service benefits over others and offered feedback on future areas of focus.

Sixteen people participated in the pilot and survey, eleven coming from member libraries and five from academic departments. A variety of academic disciplines were represented from Physics and Geosciences to Statistics to Retail Management and Archeology. Library positions involved both digital libraries and traditional subject librarian support.

Chart A: Pilot Survey Demographics

Type of Respondent	Participation Percentage
Researchers	31%
Librarians	69%
Overall Rate of:	**%**
Response	59%
Completion	89%

Nearly all participants were able to complete the pilot project's required tasks. A handful of people encountered difficulties when asked to create a Dataverse, suggesting that these instructions could be better articulated.

Chart B: Rate of Completion for Required Tasks

Answer	%
Create a user account	100%
Create a Dataverse (i.e., collection)	88%
Upload at least one dataset	100%
Provide metadata information for dataset(s)	100%
Publish dataset(s)	94%
Download a dataset	100%

The ability to carry out optional tasks by participants was very low. The working group recognized that better documentation or usability features would be needed to assist the user with these functions.

Chart C: Rate of Completion for Optional Tasks

Answer	%
Utilize the mapping analysis tool	17%
Utilize the statistical analysis tool	33%
Request access to a restricted dataset	17%
Utilize versioning of data	17%
Turn on the Guestbook feature in Dataverse	50%
Add a logo to the Dataverse instance that you created	17%

The working group asked participants how well the repository met the needs of data in their respective disciplines? 75% of applicants felt that the data repository met the needs of data from extremely well to moderately well. 25% only thought it managed these needs slightly well, with 0% at not well at all. This data suggested that the data repository generally meets user needs.

Chart D: How Well Does the Repository Meet Your Disciplinary Needs?

Answer	%
Extremely well	13%
Very well	56%
Moderately well	6%
Slightly well	25%
Not well at all	0%

The working group asked participants to provide their input on future actions related to the repository. We asked participants to select the top two services most important to you for a future Texas Research Data Repository? Responses all seem to speak to the need for librarians to be involved in facilitating the submission, description, and preservation of research data.

Chart E: What Potential Future Repository Services Do You Prefer?

Answer	%
Assistance describing data	40%
Assistance setting up a location in the repository for research projects	20%
Assistance finding data in the repository for reuse	20%
Assistance managing data prior to submitting it to the repository	47%
Assistance applying digital preservation best practices with research data	53%

The group asked respondents to identify the top two repository features they saw as the most important. All respondents found the ability to link research data with an existing publication as important and half valued the ability to link supplemental data with an electronic theses or dissertation. This data suggests potential benefits to integrating the repository with the Vireo ETD open source submission system.

Chart F: Most Important Repository Features (Top Two)

Answer	%
Linking research data with an existing publication	100%
Linking supplemental data with an electronic theses or dissertation	50%
Management of collaborative teams within the data repository	13%
Customizable submission screen with instructions	6%
Development and growth of interdisciplinary research data related to Texas geographic regions and topics	13%

Finally, the group asked participants to identify the four most important benefits of using the repository. More than half found that collecting all of their research data in one place, fulfilling federal mandates for sharing research, and making research widely available were critical benefits to the repository.

Chart G: What Repository Benefits Are Most Important To You?

Answer	%
Fulfill federal mandates for sharing publications and research data	56%

Make your research data more widely available	50%
See statistics on downloads and citations of my data	31%
Make my data citeable through the assignment of a DOI (digital object identifier)	44%
Save versions of your dataset	31%
Collecting all my data in one place	63%

Texas Data Repository Policies

Prepared by the TDL Dataverse Implementation Working Group, Policy and Governance Sub-Group

Sean Buckner
Santi Thompson
Ray Uzwyshyn

August 30, 2016

Introduction

This set of policy and governance documents were drafted by the Policy and Governance subgroup of the Texas Digital Library (TDL) Dataverse Implementation Working Group. The documents are divided into two sections (internal and external) with each section further divided into several subsections.

Appendix C focuses on the first section, "Internal Policies." These policies are primarily oriented towards an audience of stakeholders internal to academic libraries such as deans, librarians, and technologists who work with faculty and student researchers to upload, manage, and query research data in the Texas Data Repository application. This section comprises policies on collection development, data submission, metadata, digital preservation, access and use, information security, and a suggested institutional terms of use agreement between individual participating Texas universities and the TDL.

The second section, "External Policies," is primarily oriented towards external researchers and Texas Data Repository users and can be found here: https://data.tdl.org/policies/. It is comprised of subsections to include the general terms of Texas Data Repository use, Texas Data Repository privacy policy, Texas Data Repository community norms, and a Texas Data Repository usage agreement for use between researchers and users of the data available for download. The Texas Data Repository usage agreement should be configured similar to other online 'terms of use' agreements and linked on the site. A final section on Digital Preservation and Security duplicates some of the sections of internal policy and includes them on the public site.

To note, many of these documents for the Texas Data Repository have been freely adapted from various external sources. Where specific sources are used, a citation is given at the beginning of each section citing the general source. Sources include Harvard's Dataverse policy and metadata templates, the Creative Commons BY license 4.0 details, the UK's *DISC Policy-making for Research Data in Repositories Guide*, and the Data Citation Synthesis Group's Joint Declaration of Data Citation Principles.

Hopefully, this set of policy and governance documents provides a larger, initial framework for external and internal users and stakeholders of the Texas Data Repository application. The TDL may revise this document at its own discretion and without prior notice.

Internal Policies

Collection Development

Scope

The Texas Digital Library (TDL) Texas Data Repository is a consortium-level, dataset repository service and virtual research environment that provides support for the research data lifecycle for TDL member institutions, affiliated researchers, and their immediate project collaborators.

A dataset is typically a collection of files, metadata, and ancillary content associated with the data. The Texas Data Repository limits itself in scope to the data that comprise or are associated with the "raw" input/output from research. These include spreadsheets, sensor and instrument data, surveys, imagery, video, etc. (see File Formats section below). Formats such as journal articles and conference papers

resulting from research should be referred to the originating institution's Scholarly Communications division and/or Institutional Repository.

Subjects

The Harvard Dataverse was originally established and designed for social science data as a member (through Harvard's IQSS) of the Data Preservation Alliance for the Social Sciences. However, since its inception in 2006, Dataverse has expanded to other academic subjects and institutions as open source software and includes full support for GIS, astronomy, and biomedical data. Dataverse can also accept datasets from other fields of study, but without subject-specific (geospatial, social science and humanities, astronomy and astrophysics, and life sciences) metadata support. Any Dataverse repository (e.g. the Texas Research Data Repository) is searchable via the Harvard Dataverse Project website (https://dataverse.org/).

Languages

Historically, there has been no support for specifying particular language encoding for deposited data. However, SPSS files can now be tagged with the language in which they were originally coded. This is done by opening Advanced Options during ingest and selecting the language from the list provided.

Types of Research

The Texas Data Repository is configured to accept any particular type and subject of research. Though tabular data is preferred, all file formats are supported regardless of the research type. Potential types of research data include:

- scientific experiments (to include social sciences and humanities)
- input data and simulations results
- derived data (from processing or combining "raw" or other data)
- canonical or reference data (gene sequences, chemical structures, etc.)
- accompanying material, observations, or ephemera

The Texas Data Repository may prioritize certain types of data over others, in regard to long-term preservation and access, dependent upon their perceived value to the academic community. The Texas Data Repository would collect, in order of precedence, 1) datasets associated with journals or other scholarly publications, 2) stand-alone data publications and datasets with high research value, 3) other unpublished data or working files and ephemeral materials.

Status of Research

Inclusion of data into the Texas Data Repository is not determined by its status or stage within the research lifecycle. So long as data is beyond the *creating* step of the process, and has at a minimum the mandatory metadata fields required for submission to the network, it can be included in the collection. Potential stages of research data include:

- raw or preliminary data
- data ready for use by designated users
- data ready for full release, as specified in access policies

- summary/tabular data (potentially associated with a publication)

Versioning

Versioning is an important component of collection management and preservation where updates to metadata and/or files are made after initial submission. Versioning allows for any and all alterations to the data to be tracked and recorded over time.

Version Tracking

The Texas Data Repository:

- keeps the original copies of data, metadata, and documentation as deposited
- allows metadata and file changes to be saved as either draft versions or published sets
- utilizes explicit version numbering to track changes to metadata (small metadata changes; version 1.0 → 1.1) and datasets (file for citation changes; version 1.0 → 2.0)
- display all versions, draft or published, of a particular dataset

Version Control

The Texas Data Repository:

- permits the deaccessioning of a dataset or version of a dataset, though doing so is highly discouraged – "tombstone" landing page with basic citation metadata cannot be removed
- ensures that copies of data files or metadata records held in different formats are subject to the same version controls
- will always link the persistent identifier to the most current published version

File Formats

All file formats that comply with the Texas Data Repository Terms of Use are acceptable for deposit to the Texas Data Repository. Certain file formats may be ingested as tabular data, which can be further examined with TwoRavens, a statistical data exploration application integrated with the Texas Data Repository. GIS data can likewise be analyzed with WorldMap, a geospatial data visualization and analysis tool. The system will extract statistical data and/or metadata from the following preferred formats:

- SPSS (POR and SAV formats)
- STATA
- R data
- CSV
- GIS shapefiles (Esri)
- FITS

Furthermore, the Texas Data Repository automatically unpacks files compressed in ZIP format during the ingest process. All published content is downloadable, but any software requirements for opening and exploring that content are the responsibility of the end user as the Texas Data Repository does not maintain a software library or provide online access and discovery for non-preferred formats.

Volume and Size Limitations

Individual research projects have storage limitations of 10GB per project and there are some file size limitations for ***upload*** (i.e., transfer of files into a Dataverse) and ***ingest*** (i.e., data extracted from uploaded files and archived in an application-neutral text format).

- File uploads can be up to 2GB per file. Please contact support@tdl.org if you need to upload a file that is larger than 2GB in size.
- Research projects are subject to a 10GB maximum limit. Please contact support@tdl.org for information on additional storage options.
- The ingest functionality for tabular data allows for files up to 2GB in size.
- The ingest functionality for R-Data files only allows for files up to 1MB in size.

References

Purdue University Research Repository (PURR), "Policies: Collection Policy," https://purr.purdue.edu/legal/collection-policy

Elizabeth Quigley, IQSS-Harvard University, "The Expanding Dataverse," https://dataverse.org/files/dataverseorg/files/introduction_to_dataverse.pdf?m=1447352697

Data Preservation Alliance for the Social Sciences, "About," http://www.data-pass.org/about.jsp
http://guides.dataverse.org/en/latest/user/index.html

UK Data Archive, "Create and Manage Data: Research Data Lifecycle," http://guides.dataverse.org/en/latest/user/index.html

Harvard Dataverse Project, "User Guide," http://dataverse.org/harvard-dataverse-policies

---

Data Submission

Eligible Depositors

Items may only be deposited by faculty, staff, students, or delegated agents of TDL member institutions who have an active Texas Data Repository account. TDL staff are responsible for approving new user accounts and assigning initial permissions for the user. TDL member institutions are subsequently responsible for managing user accounts. Note that the TDL permits each member institution to impose additional stipulations on eligible depositors depending on local rules, regulations, and resources.

Data Moderation and Quality

The TDL through its member institutions only vet items for the eligibility of authors/depositors, relevance to the general scope of the repository, valid layout & format, and the exclusion of spam. The validity and authenticity of the content of submissions is the sole responsibility of the depositor. Any

copyright violations are entirely the responsibility of the authors/depositors. If the repository receives proof of copyright violation, the relevant item will be removed immediately.

References

DISC-UK DataShare Project, "Policy-making for Research Data in Repositories: A Guide," https://www.coar-repositories.org/files/guide.pdf

University of Edinburgh, "Service Policy: Submission Policy," http://www.ed.ac.uk/information-services/research-support/data-library/data-repository/service-policies/submission-policy

---

Metadata

Metadata in the Texas Data Repository

The Texas Data Repository solicits and generates a variety of descriptive, administrative, technical, preservation, and use metadata as it curates and makes datasets available.

Type of Metadata	Definition	Selected Examples
Administrative	Metadata used in managing and administering collections and information resources	• Deposit date • Depositor
Descriptive	Metadata used to identify and describe collections and related information resources	• Author • Title • Description • Date
Preservation	Metadata related to the preservation management of collections and information resources	• Checksum value • Version
Technical	Metadata related to how a system functions or metadata behaves	• File format • File size
Use	Metadata related to the level and type of use of collections	• Citation information • Number of downloads

More information on specific metadata fields and definitions can be found in the Texas Data Repository Metadata Dictionary.

The Texas Data Repository supports three levels of metadata:

1. Citation metadata: any metadata that would be needed for generating a data citation and other general metadata that could be applied to any dataset. This information can be supplemented with metadata describing a journal in which the dataset is associated.
2. Domain-specific metadata: with specific support currently for social science, life science, geospatial, and astronomy datasets
3. File-level metadata: varies depending on the type of data file

The Texas Data Repository Metadata Dictionary outlines fields that are required for completion before a dataset can be ingested and made accessible (mandatory fields) and other fields that are available for expanded description of particular dataset (option fields).

Metadata Reuse

Depositors submitting items to the Texas Data Repository provide metadata describing the dataset as well as the intellectual content reflected by the data. The Texas Data Repository uses the CC0 option. This option lets others distribute, remix, tweak, and build upon works, even commercially. "The person who associated a work with this deed has dedicated the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law." (Further information: https://creativecommons.org/about/cc0/)

Texas Data Repository metadata may be aggregated by other systems using API applications or the OAI-PMH protocol. Permission from the TDL is not needed to harvest metadata into aggregated discovery or repository platforms unless aggregators intend to harvest on a permanent basis.

By default, users may download portions of the citation metadata for any dataset unless restricted by the depositor. This information includes the author and title of the dataset, the Digital Object Identifier (DOI), and dates of dataset creation and deposit.

Texas Data Repository Metadata Dictionary

The Texas Data Repository Metadata Dictionary is comprised f citation (including journal metadata) and domain specific (geospatial, social science and humanities, astronomy and astrophysics, and life sciences) metadata fields. The dictionary contains a definition for each field, describes any format specifications associated with each field, denotes if the field is mandatory or optional, and establishes whether or not the field is visible to the depositor. Note that TDL members may choose to expand the number of mandatory fields based on the institution's operating procedures.

References

Anne Gilliland, "Setting the Stage, "*Introduction to Metadata,* ed. Murtha Baca, Getty Publications, Nov 11, 2008, p. 9.

Dataverse Project, "User Guide," http://guides.dataverse.org/en/latest/user/index.html

DISC-UK DataShare Project, "Policy-making for Research Data in Repositories: A Guide," https://www.coar-repositories.org/files/guide.pdf

---

Digital Preservation

Policy

It is the mission of the TDL to enable digital initiatives in support of research, scholarship, and learning in Texas. As a part of this mission, the TDL endeavors to collect, preserve, and disseminate scholarly materials for the benefit of both producers and consumers of academic research and scholarship. The TDL's instance of the Dataverse Network, encompassing each of the dataverses of its member institutions, is the digital resource intended to address a consortium-level need for publishing, managing, and providing access to research-generated data sets. The following Digital Preservation Policy describes the extent to which the TDL will support sustainable access to the digital research data and related content deposited in the Texas Data Repository.

The preservation objectives of the Texas Data Repository are:

- to collect, preserve, and disseminate the data sets and related information generated by researchers affiliated with any of the TDL's member institutions who choose to deposit their content therein.
- to enable researchers affiliated with any of the TDL's member institutions to comply with the mandates of funding agencies to manage, preserve, and share their research data.
- to provide the means for users to discover and access the data sets and metadata generated by academics affiliated with any of the TDL's member institutions over the long term.

Part of the TDL's vision in establishing a consortium Texas Data Repository is to make research materials freely available to anyone, anywhere, and at any time. The TDL is an advocate for Open Access to scholarly work and the incentives to researchers for publishing and preserving their research data in the Texas Data Repository are:

- data that might be precariously stored on fragile, random, or unsustainable storage devices can be securely preserved for the long term.
- data that might otherwise become neglected over time can be preserved and made accessible for other interested researchers to use and cite, potentially providing wider visibility and impact for the research.
- many funding agencies and scholarly journals require data management plans that detail how the data will be managed, made accessible, and preserved.

Scope

The TDL accepts the responsibility to preserve and provide access to research data, including associated metadata and documentation that is properly deposited in the Texas Data Repository. This responsibility includes the provision of digital means to preserve and ensure ongoing access to said content for a minimum period of ten years after it is deposited in the Texas Data Repository. Long-term preservation of Texas Data Repository content, beyond the ten-year retention period, is subject to the TDL's selection criteria, appraisal of the content, and budgetary and technical support of resources necessary to meet this goal. Metadata for content removed from the Texas Data Repository, regardless of reason or retention period, may be preserved for an undetermined period of time after said content's removal.

The Texas Data Repository content will be selected and appraised according to the following preservation priorities and levels of commitment:

1. Research data associated with publications – great effort will be made to ensure the long-term preservation of data associated with journal or scholarly publications, so long as the data meets the TDL collection policies and the Texas Data Repository remains the data's hosted or cited repository.
2. Stand-alone data publications with high research value – reasonable effort will be made to ensure the long-term preservation of data and metadata of stand-alone publications that library professionals identify as having high research value to the broader academic community.
3. Other data files and materials – efforts may or may not be made to retain ephemeral materials considered to lack significant or long-term value, although particular files may be preserved on a select basis as appropriate.

Additionally, the Texas Data Repository will accept data submissions of any format. For support of data exploration, analysis, and meta-analysis via the TwoRavens suite of statistical tools, tabular data should be compiled in the following formats:

- SPSS (POR and SAV formats)
- STATA
- R data
- CSV

These files can be in compressed ZIP format at ingest, however, they may not exceed two GB in size. Please see http://guides.dataverse.org/en/latest/user/tabulardataingest/index.html and http://guides.dataverse.org/en/latest/user/dataset-management.html for more specific information on data set and metadata formats.

Texas Data Repository provides basic, bit-level preservation through fixity checks and secure backup of deposited content. Further and more in-depth digital preservation activities and services must be provided by a digital preservation program at the institution where the research data was originally generated.

Strategic Plan

Data Backup

The TDL has an official backup strategy that requires all digital content to be:

- copied nightly with versioning and kept for one year (individual files)
- copied nightly as a snapshot and kept for one month (entire service)

The TDL systems also provide security services key to basic digital preservation, namely access control, network monitoring and protection, encryption, and system updates (see Information Security Policy). Currently, research projects can upload no more than 10GB of data to the repository, with no individual files being larger than 2 GB.

Procedures

Dataverse best practices for data management and preservation include:

- automatic extraction of metadata from tabular files and FITS
- standard descriptive metadata schemas such as OAI DC, DDI (for statistical and social science), ISA-Tab (for biomedical), FITS (for astronomy)
- re-formatting of tabular data to simple open format text files
- data and metadata versioning; database maintenance
- checksum generation upon ingest (UNF for tabular data, MD5 for all other files)
- persistent URL – DOI (minted by EZID)
- deaccessioning of data, but not citation metadata, if necessary

The TDL systems infrastructure includes bit-level fixity checking via Amazon S3 host service.

References

The Dataverse Project, "Harvard Dataverse Preservation Policy," http://best-practices.dataverse.org/harvard-policies/harvard-preservation-policy.html

Purdue University Research Repository (PURR), "PURR Digital Preservation Policy," https://purr.purdue.edu/legal/digitalpreservation

Texas Digital Library, "Our Mission and Vision," https://www.tdl.org/strategic-plan/vision/

Preserving digital Objects With Restricted Resources, "Tool Grid," http://digitalpowrr.niu.edu/tool-grid/

Digital Curation Centre, "DataVerse," http://www.dcc.ac.uk/resources/external/dataverse

Harvard Dataverse, "UCLA Social Science Data Archive Dataverse," http://dataarchives.ss.ucla.edu/archive%20tutorial/archivingdata.html

Harvard's Institute for Quantitative Social Science (IQSS), "About TwoRavens," http://datascience.iq.harvard.edu/about-tworavens

University of North Carolina - The Odum Institute, "Digital Preservation Policies," http://www.irss.unc.edu/odum/contentSubpage.jsp?nodeid=629

Harvard Dataverse Project, "User Guide: Tabular Data File Ingest," http://guides.dataverse.org/en/latest/user/tabulardataingest/index.html

Elizabeth Quigley, IQSS-Harvard University, "The Expanding Dataverse," http://dataverse.org/files/dataverseorg/files/introduction_to_dataverse.pdf?m=1447352697

Access and Use

Types of Access

The accessibility of content in the Texas Data Repository for reuse is determined by the level of access selected by the depositor. These levels include open, controlled, and restricted access.

Open Access

By default, all datasets published in the Texas Data Repository have no rights reserved under the CC0 option. This option lets others distribute, remix, tweak, and build upon works, even commercially. Some data depositors may choose to stipulate alternative reuse terms when appropriate. In such instances, the system articulates these terms to the users of the Texas Data Repository.

Controlled Access

Data depositors may elect to share access to unpublished or published datasets with targeted audiences or individuals. To access this content, a user must be designated as a viewer from the depositor of the data. Alternatively, depositors may restrict access to published data but allow users to request access using the "Request Action" button.

Restricted Access

Through user settings, depositors may elect to restrict access to entire datasets or to portions of data. They may elect to offer access to these restricted files through the "Request Action" button.

The TDL may also choose to restrict access to files determined to be in violation of personal privacy or copyright.

Access Methods

Accessing content in the Texas Data Repository encompasses searching, browsing, and/or downloading data. Users can query items in the repository using basic and advanced search interfaces. Alternatively, users can browse content using targeted facets (including name of depositor, name of collection, and deposit date). Once identified, users have the ability to view metadata about a particular dataset as well as download citation metadata and files within the dataset.

All items made accessible either openly or controlled are done so free of charge to the user.

The Texas Data Repository User Guide outlines methods used to search, identify, and download datasets and citation metadata.

Use and Reuse

Unless otherwise stipulated by the specific terms of use, datasets found in the Texas Data Repository can be reused for a series of purposes, including reproducing, displaying, performing, or giving to third parties in any format or medium. At the same time, the Texas Data Repository Community Norms stipulate that those reusing content should abide by the Creative Commons BY License Attribution 4.0 International. For more information, see the Texas Data Repository Community Norms.

Items in the repository may be harvested by robots transiently. Permission must be sought from the TDL where items are harvested permanently.

References

Dataverse Project, "User Guide," http://guides.dataverse.org/en/latest/user/index.html

DISC-UK DataShare Project, "Policy-making for Research Data in Repositories: A Guide," https://www.coar-repositories.org/files/guide.pdf

---

Information Security

Introduction

Information security is a complex and vital element of maintaining any information system. There are issues that threaten information security and they are generally associated with the areas of systems security, data integrity, and regulatory and legal considerations. Vulnerabilities in web applications, internal processes, and authentication account for most threats to an organization's information assets. These threats need to be constantly addressed and vulnerabilities continuously remediated.

The TDL actively addresses the need to ensure the accuracy, integrity, authenticity, and permanence of the digital content that it manages, as well as the security of the services and platforms that it provides. The TDL ensures the security of its Dataverse instance as follows:

System Security

The TDL systems and services are hosted with Amazon Web Service (AWS), which provides cloud security services and support (https://aws.amazon.com/security/) to include:

- Secure Network Architecture – segmentation and firewalls throughout
- Secure Access Points – API endpoints allowing HTTPS access
- Encryption – connections encrypted by SSL
- Network Monitoring and Protection – against DDoS and MITM attacks, IP spoofing, etc.
- Identify Management and Authentication – secure log-in via password and SSH key pairs

Additionally, the TDL updates its Operating Systems (OS) quarterly at a minimum, and immediately when important security patches are made available.

Data Integrity

The TDL has an official backup strategy that requires all digital content to be stored in three distinct locations for all services including Dataverse. The TDL will retain:

1) the copy of the data residing on the production server (currently an EBS volume),
2) nightly snapshots that hat can be used to restore the entire service to a particular date within the preceding month,

3) a copy of all data files, made nightly with versioning and kept for one year, stored on Amazon S3 (https://aws.amazon.com/s3/); these copies can be used to restore individual files, but not the entire service.

Although the TDL does not curate or conduct preservation planning on content within the Texas Data Repository, it provides some lower-level services to help ensure the integrity of the data it hosts. In addition to the access control and network protection mentioned in the previous section, the AWS S3, where the Texas Data Repository is hosted, performs regular systematic data integrity checks and is built to be self-healing. Also, the TDL ensures the accurate migration and/or transfer of data between storage spaces, servers, and systems wherever such may become necessary.

Regulatory and Legal Considerations

The TDL requires Dataverse contributors to remove, replace, or redact identifying confidential or sensitive information from datasets prior to upload. The Texas Data Repository will not serve this function and takes no responsibility for the inadvertent release of restricted and protected data. Users should contact the TDL and alert them to any data placed into TDL storage and/or infrastructure that requires FERPA, HIPAA, or other federal privacy standards. The TDL can offer dark storage options outside of the Texas Data Repository service for such instances.

The Texas Research Data Repository complies with Texas Administrative Code (TAC) 206.70 as set forth in the University of Texas Web Accessibility Policy (http://www.utexas.edu/cio/policies/web-accessibility).

References

Texas Digital Library Data Security Policy, April 2015, https://tdl.org/wp-content/uploads/downloads/2015/04/Texas-Digital-Library-Data-Security-Policy.pdf

Texas Digital Library Data Management Talking Points, November 2013, http://tdl.org/wp-content/uploads/downloads/2013/11/datamgmt-talking-points-11.19.2013.pdf

University of Texas Web Accessibility Policy, 23 March 2015, http://www.utexas.edu/cio/policies/web-accessibility

Amazon AWS Cloud Security, website, https://aws.amazon.com/security/

Amazon AWS Identity and Access Management (IAM), website, https://aws.amazon.com/iam/

Amazon Web Services: Overview of Security Processes, August 2015, https://d0.awsstatic.com/whitepapers/Security/AWS_Security_Whitepaper.pdf

Digital Preservation Coalition: Information Security, website, http://www.dpconline.org/advice/preservationhandbook/technical-solutions-and-tools/information-security

---

Institutional Terms of Use[10]

Institutional Acceptance of the Texas Data Repository Data Usage Agreement

By setting up an institutional Dataverse with the TDL or utilizing the TDL's Dataverse service, the Institution (name) represents their acceptance of the terms of this Agreement.

Institutional Modification of this Agreement

Institutions utilizing the Texas Data Repository may modify the terms of this Agreement at any time. However, any modifications to this Agreement will only be effective for usage subsequent to such modification. No modifications will supersede previous terms that were in effect at the time of the original agreement (default).

Institutional Use of the Texas Data Repository Data

Uses of the Texas Data Repository include but are not limited to viewing parts or the whole of the content included in the Dataverse; comparing data or content from the Dataverse with data or content in other Dataverses; verifying research results with the content included in the Dataverse; and extracting and/or appropriating any part of the content included in the Dataverse for use in projects, publications, research, or other related work products.

Institutional Representations and Warranties

In the use of the Texas Data Repository, the institution represents that it is not bound by any pre-existing legal obligations or other applicable laws that prevent the institution from utilizing the Texas Data Repository. The Texas Data Repository is provided to participating institutions "as is" and "as available" and without warranty of any kind, express or IMPLIED, including, but not limited to, non-infringement, merchantability and fitness for a particular purpose, and any warranties implied by any course of performance or usage of trade, all of which are expressly disclaimed.

The TDL does not warrant that:

a. the materials are accurate, complete, reliable or correct
b. the materials files will be secure
c. the materials will be available at any particular time or location
d. any defects or errors will be corrected
e. the materials and accompanying files are free of viruses or other harmful components
f. the results of using the materials will meet downloader's requirements. Downloader's use of the materials is solely at the downloader's own risk.

Institutional Integration and Severability

This Agreement supersedes all prior or contemporaneous communications and proposals (whether oral, written or electronic) between the TDL and the Institution. If any provision of this Agreement is found to

[10] The Institutional Terms of Use is adapted from the Harvard best practices terms of use template. For original see, http://best-practices.Dataverse.org/harvard-policies/harvard-terms-of-use.html

be unenforceable or invalid, that provision will be limited or eliminated to the minimum extent necessary so that the Agreement will otherwise remain in full force and effect and enforceable.

Miscellaneous

No agency, partnership, joint venture, or employment relationship is created as a result of the Agreement and neither party has any authority of any kind to bind the other in any respect outside of the terms described within this Agreement. If any provision of this Agreement is found to be unenforceable or invalid, that provision will be limited or eliminated to the minimum extent necessary so that the Agreement will otherwise remain in full force and effect and enforceable.

Deaccessioning Data

Items may be deaccessioned from the repository for the following reasons:

- copyright violation
- legal requirements and proven violations
- national security
- falsified research
- confidentiality concerns, etc.

Items may also be deaccessioned from the repository by the depositor. Deaccessioning a dataset or a version of a dataset is a very serious action that should only occur if there is a legal or valid reason for the dataset to no longer be accessible to the public. If you absolutely must deaccession, you can deaccession a version of a dataset or an entire dataset. To deaccession, go to a dataset you've already published (or add a new one and publish it), click on Edit Dataset, then Deaccession Dataset. If you have multiple versions of a dataset, you can select here which versions you want to deaccession or choose to deaccession the entire dataset. You must also include a reason as to why this dataset was deaccessioned from a dropdown list of options. There is also a free-text box to add more details as to why this was deaccessioned. If the dataset has moved to a different repository or site you are encouraged to include a URL (preferably persistent) for users to continue to be able to access this dataset in the future.

Important Note: A tombstone landing page with the basic citation metadata will always be accessible to the public if they use the persistent URL (Handle or DOI) provided in the citation for that dataset. Users will not be able to see any of the files or additional metadata that were previously available prior to deaccession.

Should a dataset be removed by either the repository or the depositor, TDL reserves the right to retain its citation metadata record in the repository as trace of the dataset. Additionally, the citation metadata of withdrawn items will be searchable.

References

DISC-UK DataShare Project, "Policy-making for Research Data in Repositories: A Guide," https://www.coar-repositories.org/files/guide.pdf

Dataverse Project, “User Guide: Dataset + File Management,”
http://guides.dataverse.org/en/latest/user/dataset-management.html

Appendix D: Texas Data Repository Memorandum of Understanding

Texas Data Repository

DATA REPOSITORY MEMORANDUM OF UNDERSTANDING (MOU) between
[MEMBER INSTITUTION]
and
Texas Digital Library (TDL)

I. Purpose & Scope

The purpose of this MOU is to clearly identify the roles and responsibilities of the member institution and Texas Digital Library (TDL) as they relate to the Texas Data Repository.

The Texas Digital Library is a consortium of Texas higher education institutions, with UT Austin serving as the lead agency.

II. Definitions

- **Backup:** the process of making an exact duplicate of a digital object by copying the bitstream and storing that copy in a separate storage space. It is considered the minimum maintenance strategy of digital preservation.
- **Dataverse:** a platform for publishing and archiving research data developed by Harvard University
- **dataverse:** a collection of datasets (and other dataverses), created by individual researchers
- **institutional dataverse:** a collection of datasets and dataverses organized by member institutions and maintained by data repository librarians
- **Long-term preservation:** the endeavor to preserve digital content over the long term – depending on an institution's policies, anywhere from 10 years to indefinitely. The term generally refers to the application of strategies above and beyond bit-level preservation.
- **Data Repository Librarian:** a librarian selected by a member institution to serve as a liaison to the TDL on matters related to the Texas Data Repository and administer local settings, policies and procedures, and institutional dataverses.

III. Background

The Texas Data Repository is a platform for publishing and archiving datasets (and other data products) created by faculty, staff, and students at TDL member institutions. The repository is built in an open-source application Dataverse, developed and used by Harvard University.

The repository is hosted by Texas Digital Library (TDL), a consortium of academic libraries in Texas with a proven history of providing shared technology services that support secure, reliable access to digital collections of research and scholarship.

IV. [MEMBER INSTITUTION'S] responsibilities under this MOU

[NAME OF MEMBER INSTITUTION] shall:

- Appoint an individual to serve as "data repository librarian" to provide oversight of research data at their respective institutions and to serve on a TDL-wide advisory committee.
- Update TDL when a change to the data repository librarian is needed, in order for TDL to maintain accurate public registry of data repository librarians at member institutions.
- Be responsible for the long-term preservation of data files deposited by faculty/staff from the member institution.
- Inform TDL of the institutional requirements necessary to authorize and sustain the Texas Data Repository, e.g. security authorizations through Central IT offices.
- Serve as a liaison between TDL and institutional departments as necessary (e.g. with Central IT for set up of Shibboleth)
- Establish institutional policies around copyright inquiries, takedown requests, and rights decisions and inform TDL when necessary of repository actions required.
- Recommend, in conjunction with TDL, data repository options should the [name of the repository] be discontinued.
- Promote and support the Texas Data Repository service within its campus community and educate faculty, staff, and student users as necessary of policies and procedures.

IV.A. Data Repository Librarian responsibilities

Data Repository librarian duties include:

- Act as the local liaison/contact person for users and other university community members;
- Link datasets deposited by institutionally-affiliated users to institutional dataverse;
- Participate in TDL-wide data repository librarian committee meeting and work;
- Manage and update institutional dataverse page, including theme (logos, colors) and description;
- Maintain required and optional metadata fields and settings for browseable facets to comply with TDL and local guidelines;
- Control permissions for the institutional dataverse (and other dataverses the data repository librarian creates);
- Fulfill additional, related duties as assigned by the institution. These might include assisting in other data curation roles such as data ingest, metadata creation and modification, and data management planning, as well as contacting users through the data repository interface.

V. TDL's Responsibilities under this MOU

TDL shall:

- Provide access to and secure backup of data submitted to the repository
    - TDL will retain data files, make them accessible (when applicable), and provide secure backup for the duration outlined in the policy
    - TDL will retain metadata for the duration outlined in the policy
- Create and maintain appropriate user profile permissions for data repository librarians
- Be responsible for the stewardship, technological oversight, and upgrades of the data repository software infrastructure
- Provide timely reporting to data repository librarians regarding any system issues, including planned or unplanned outages or other significant changes
- Recommend, in conjunction with member institutions, data repository options should the Texas Data Repository be discontinued
- Coordinate a membership-wide committee of data repository librarians
- Maintain a tech support helpdesk for the data repository librarians
- Maintain a tech support helpdesk for the data repository and for referring requests to the relevant data repository librarians
- Provide training and professional development opportunities to data repository librarians as needed
- Provide periodic reports to data repository librarians detailing information about deposits from their institution

VI. Ownership

UT Austin, as the lead agency for the Texas Digital Library consortium, operates the Texas Data Repository for the benefit of its member institution.

VII. Effective Date and Signature

This MOU shall be effective upon the signature of representatives from [MEMBER INSTITUTION] and TDL. It shall be in force from [START DATE to END DATE]. [MEMBER INSTITUTION] and the TDL indicate agreement with this MOU by their signatures.

_____ _____

[Library Administrator's name] Date

[Title]

_____ _____

[TDL representative's name] Date

[Title]