

Texas Data Repository Information Sheet

For Administrators

The **Texas Data Repository** is 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 an open-source application called [Dataverse](#), originally developed and used by Harvard University.

The repository is hosted by the 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 (<http://www.tdl.org/>).

Benefits of a Texas Data Repository

- **Compliance with funding requirements.** The Texas Data Repository helps researchers comply with funder mandates for data archiving and sharing, and supports research grant-seekers by having infrastructure available at the time of proposal.
- **Reliable, managed access for data.** The Texas Data Repository provides a convenient and reliable place to collect and share data. And by depositing data there, researchers benefit from the Texas Digital Library's focus on long-term access and preservation of scholarly content.
- **Increase scholarly impact.** By publishing their data in the Texas Data Repository, researchers give their data credibility through a unique, citable, and persistent online identifier (i.e. a Digital Object Identifier), which makes it easy for others to cite reliably.
- **Collaboration with research teams.** Some situations may necessitate restricting access to data, at least for a period of time. The Texas Data Repository allows researchers to share their data with a select group of colleagues, version the data, and publish it when they're ready.
- **Access to local support through their institution's library.** Along with robust technical support from the TDL, users of the Texas Data Repository can rely on trained librarians at their home institution to assist with multiple phases of the research cycle, including data management planning, preparation for data publishing, and long-term curation.
- **Efficient use of resources.** By pooling resources across multiple institutions, the Texas Data Repository realizes cost savings through a shared infrastructure while showcasing local contributions through university-branded data collections and local library services. Each institution can focus its resources on unique services that meet local research community needs.

The Basics: How a Texas Data Repository Will Work

- The Texas Data Repository is designed for regular to mid-sized data sets (individual file sizes up to 2 GB), which comprises the majority of research data. These data can include:
 - Data from any scholarly discipline and in any file type
 - Materials such as codebooks and other supplementary documentation
 - Data that does NOT contain confidential or sensitive information (like social security numbers or other identifiers)
- The Texas Data Repository does not accept datasets (which can contain multiple individual files) larger than 10 GB.

- Researchers affiliated with participating TDL member institutions will be able to:
 - Store and organize data sets and upload files
 - Maintain multiple versions data sets
 - Share data sets online with trusted colleagues OR release data for public access online
 - Get recognition and proper academic credit for their scholarly work through a data citation with a persistent identifier (i.e. a DOI, or digital object identifier)
- Library faculty or staff at each of TDL's participating member institutions will provide local assistance to researchers at their institution as they prepare and deposit their data.
 - Each participating university will have its own branded "dataverse" within the overall repository, which it can use to showcase its researcher contributions.

Facts about Research Data and the Importance of Reproducibility

- Texas institutions of higher learning produce vast amounts of research data. In 2014, research expenditures at Texas institutions of higher education collectively was \$4,522,990,861. Nearly half of those expenditures (>\$2 billion) came from federal agencies.¹
- Funding agencies and institutions increasingly require that products resulting from funded research (both articles and the underlying data) be made publicly accessible:
 - In February 2013 the White House's Office of Science and Technology Policy 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.²
 - Besides US Agency funders, 42 funder institutions worldwide require data archiving.³
- Transparency of methods and reproducibility of results are key values of science, as they enable testing and validation.
 - By providing technical infrastructure to enable data sharing and robust description of methodology -- and by training up a generation of library personnel able to assist researchers in managing their research data effectively -- the Texas Data Repository will help facilitate a culture of reproducibility within the sciences.⁴

More about the Texas Digital Library

- The Texas Digital Library (TDL) is a consortium of academic libraries in Texas that provides shared services in support of research and learning. The TDL currently has 22 academic library members, including the state's largest research institutions.
- Since 2005, TDL member institutions have worked together to develop a set of services that support discoverability, access, and preservation of the unique research and archival collections of its member institutions. These include hosted online repositories, online scholarly journals, and thesis and dissertation publishing.

¹ "Research Expenditures Summary, September 1, 2013-August 31, 2014: Texas Universities and Health-Related Institutions." Texas Higher Education Coordinating Board, March 2015. <http://bit.ly/2blUZ25>

² John Holdren. "MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES." Executive Office of the President Office of Science and Technology Policy, February 22, 2013. <http://bit.ly/2bCxDTt>

³ "Sherpa/Juliet -- Some Juliet Statistics." (2015, December 7) Retrieved from <http://www.sherpa.ac.uk/juliet/stats.php?la=en&mode=simple>.

⁴ Julia Belluz, Brad Plumer, and Brian Resnick. "The 7 biggest problems facing science, according to 270 scientists" Vox.com. July 16, 2016. <http://bit.ly/2bqVTtt>

Technical Infrastructure and Security Policies

Systems security. The Texas Digital Library actively works 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 Texas Data Repository is hosted in Amazon Web Services, which provides cloud security services and support (<https://aws.amazon.com/security/>) including segmentation and firewalls, API endpoints allowing HTTPS access, SSL encryption, network monitoring and protection, and identity management and authentication.
- TDL updates its Operating Systems (OS) quarterly at a minimum, and immediately when urgent security patches are available.
- The Texas Data Repository uses a federated authentication application (Shibboleth) to ensure that only faculty and staff from TDL member institutions can deposit or administer data. Only users that log into the repository using their university credentials (signifying their connection to that TDL member institution) have permission to deposit data in the repository, unless granted that permission explicitly by TDL staff.
- Only TDL staff maintain root access to the repository. A controlled number of library personnel -- one user account per participating member institution -- will have special privileges for administering an institutional collection of research data deposited by users from their university.
- The TDL backs up data in the Texas Data Repository according to its organization-wide backup policy, maintaining copies of data in three distinct locations.

Data Access policies. The TDL encourages the use of the Texas Data Repository for open publication of data, in order to maximize its re-use, but provides flexible options for restricting access to a few or no other individuals.

- All data published in the repository is published by default under a “no rights reserved” (CC0) license; however, depositors can choose other licensing options or create a custom re-use policy for their published data. They may also choose not to publish their data at all, but simply to archive it for safe-keeping or limited sharing on a case-by-case basis.

Policy on Confidential and Sensitive Data. The Texas Data Repository does NOT accept content that contains confidential or sensitive information (even if it remains unpublished in the system), and requires that contributors remove, replace, or redact such information from datasets prior to upload.

- Users that wish to deposit data into the repository must assent to a usage agreement that includes these terms and conditions. The full usage agreement is available on the Data Repository website at <http://data.tdl.org/terms-of-use/>.