Scan the QR Code and Finish the Love Data Week Pre-Survey

go.txstate.edu/unq

Open Data Resources and Data Management Skills in File Naming

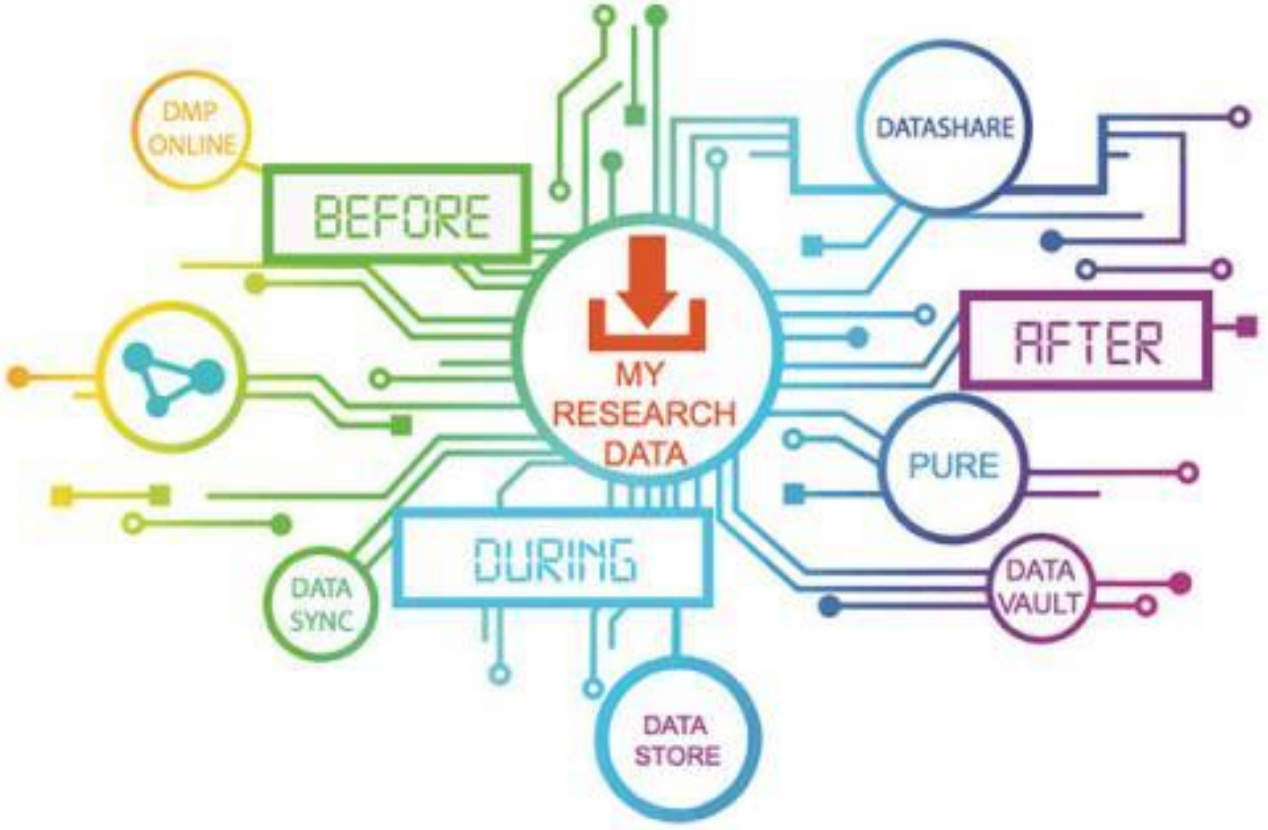
2024 TXST Love Data Week Workshop Series

XUAN ZHOU, PHD
DATA CURATION SPECIALIST
UNIVERSITY LIBRARIES
FEBRUARY 13, 2024

Goals for Today

- Understand the research data life cycle
- Get Resources of Open Access Data
- Know good practices in data file naming and backup

What do we mean when we talk about research data?



What data do you use and create?

Research Data is recorded, factual material commonly accepted in the scientific community as necessary to validate research findings. (Awasthi & Tripathi, 2019)

Numeric data
Spreadsheets
Binary files
Code

PDFs
Image files
Audio files
Physical specimens
Archival materials
Geospatial data

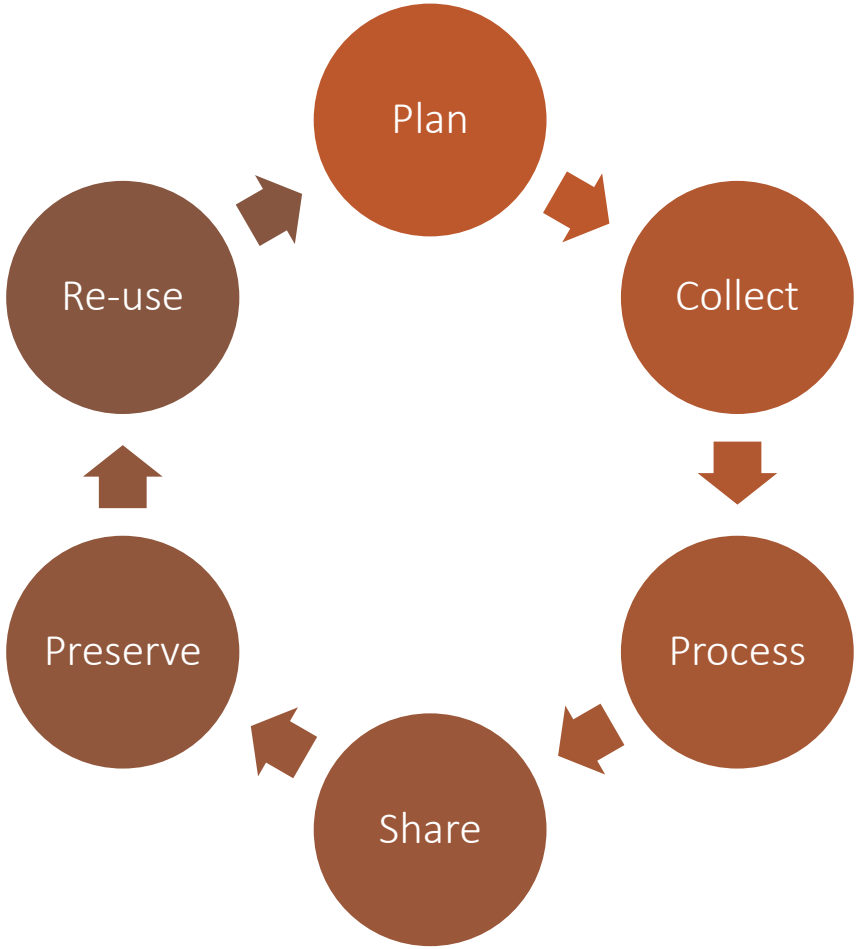
Or something else

What is research data management?

Research Data Management (RDM) is the organization, management, publication, and preservation of the products of research.

Mandate	Facilitate	Reuse	Impact
Meet requirements and expectations set by funding agencies, publishers and domain associations	Ensure that your data is complete, documented, and accessible to you and to future researchers	Encourage the discovery and reuse of your data to further discoveries in your field of research	Receive credit for your data and increase its impact and visibility

Beneficial to you and your research in a long run!



Research Data Lifecycle

Data Collection

Data collection is the methodological process of gathering information about a specific subject. It's crucial to ensure your data is complete during the collection phase and that it's collected legally and ethically. If not, your analysis won't be accurate and could have far-reaching consequences.

- **First-hand data**, which is collected directly from users by your organization
- **Secondary data**, which is data shared by another organization about its customers (or its first-party data)
- **Third-party data**, which is data that's been aggregated and rented or sold by organizations that don't have a connection to your company or users

Open Data Resources

What is Open Data?

Open data is data that can be freely used, re-used, and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike. To summarize the most important:

- **Availability and Access:** the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Re-use and Redistribution:** the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.
- **Universal Participation:** everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, ‘non-commercial’ restrictions that would prevent ‘commercial’ use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

TXST Dataverse Repository

A research data management system

Add, share, publish, and manage your data

Find datasets from across Texas institutional Dataverse collections.

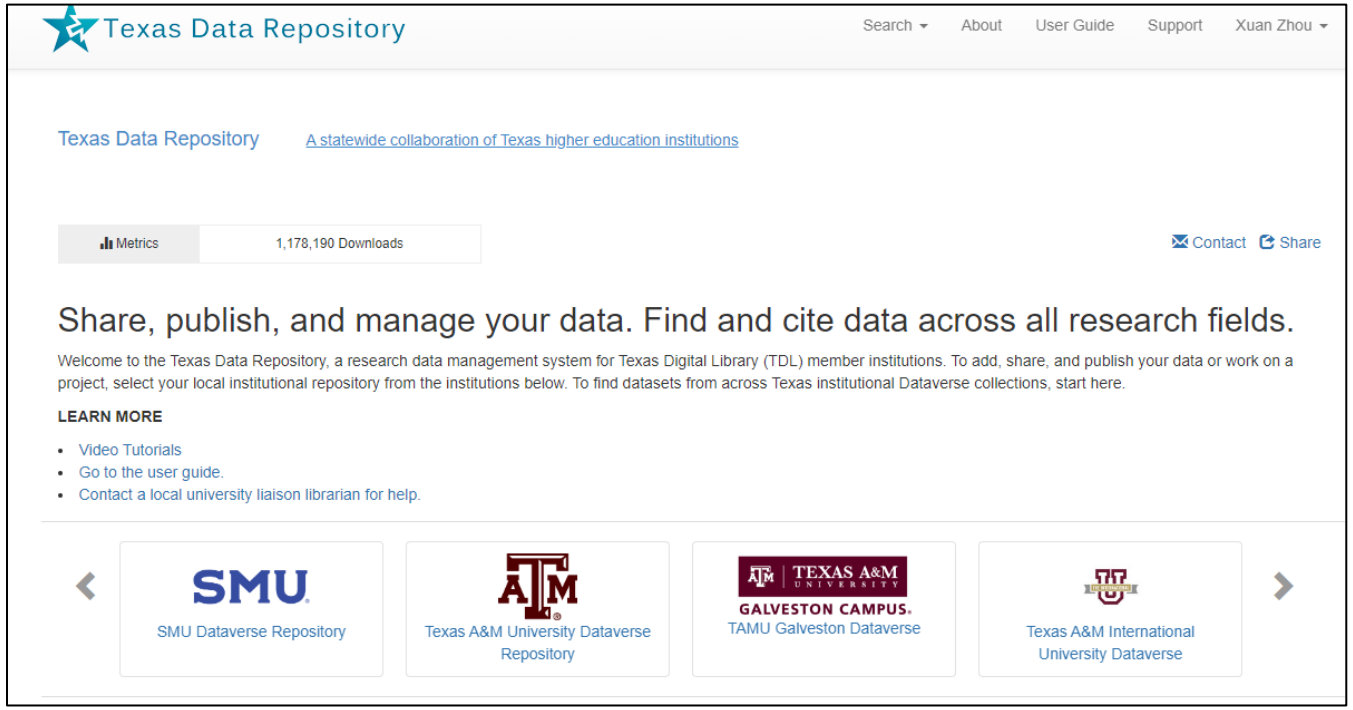
https://dataverse.tdl.org/

Texas Data Repository

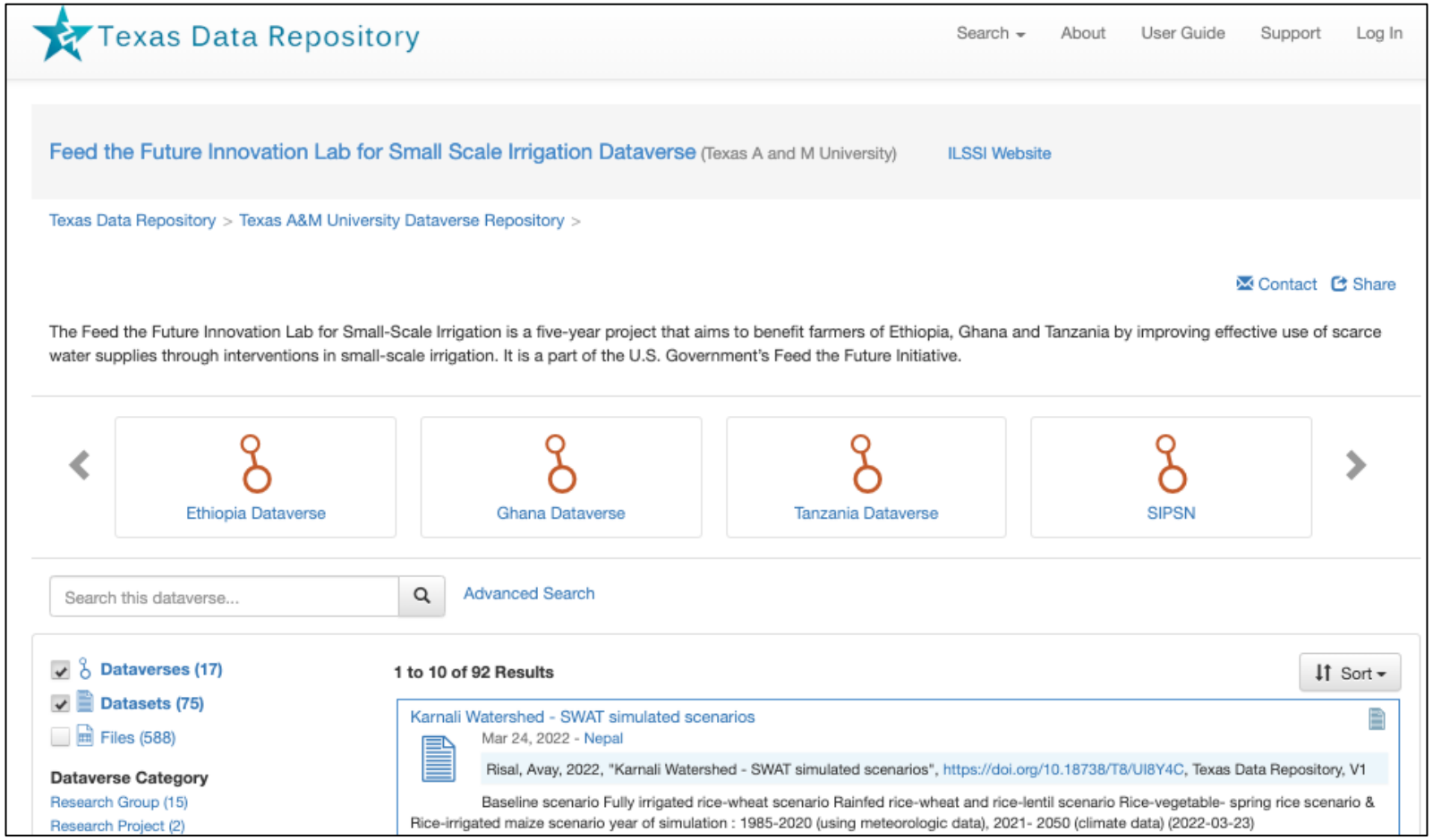
Why TXST Dataverse Repository?

- Provides a platform for archiving and publishing the data developed or used in support of research at Texas State University
- An open access data repository for researchers affiliated with TXST
- Served by RDM service team: help with DMP and preparing data to deposit
- University libraries offer advice on appropriate file formats, metadata, and licensing options
- Provide consultation services or training workshops for users to upload and manage their own data collections

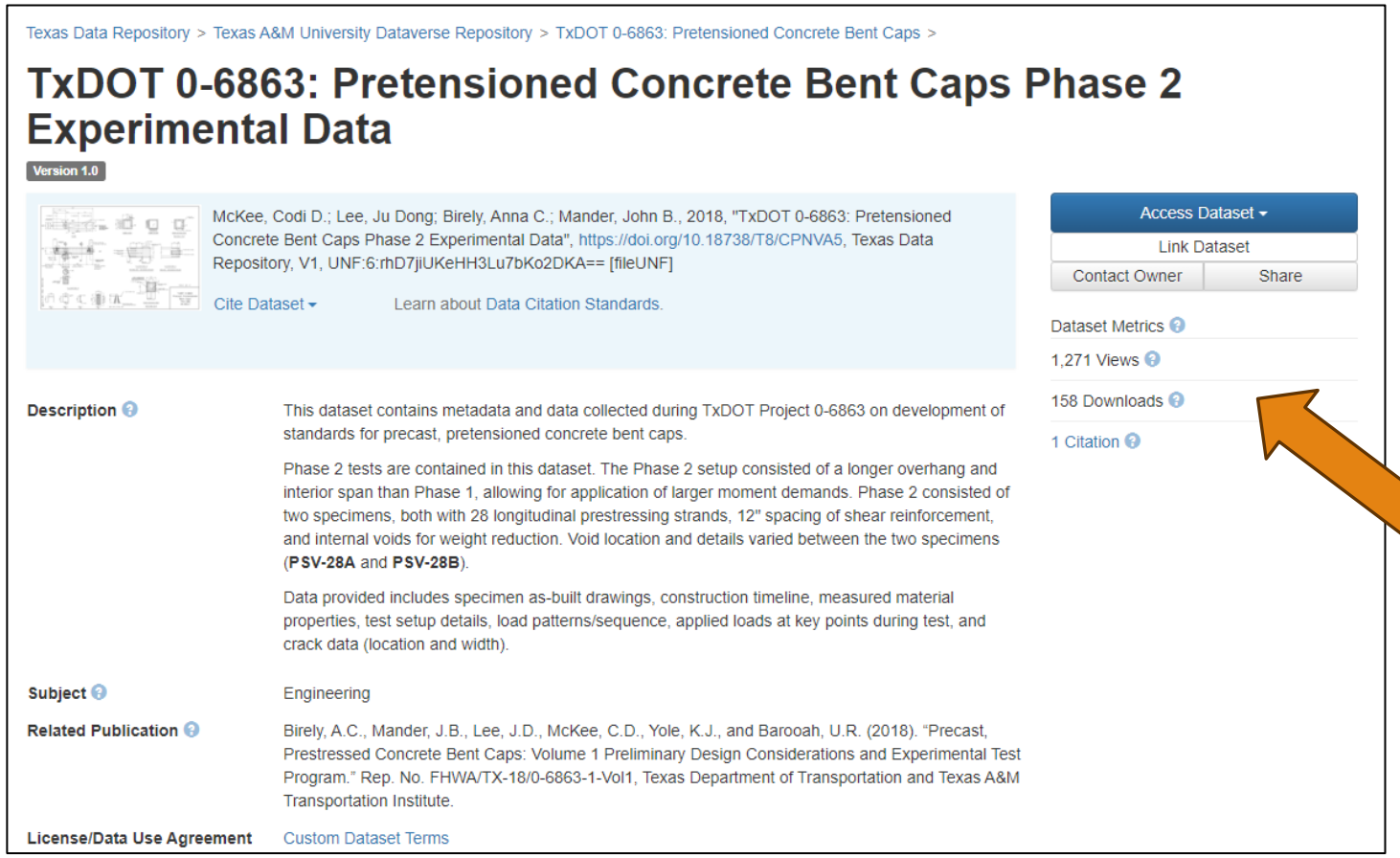
Find Open Data



Facilitate Collaboration



Increase scholar impact



Texas Data Repository > Texas A&M University Dataverse Repository > TxDOT 0-6863: Pretensioned Concrete Bent Caps >

TxDOT 0-6863: Pretensioned Concrete Bent Caps Phase 2 Experimental Data

Version 1.0

McKee, Codi D.; Lee, Ju Dong; Birely, Anna C.; Mander, John B., 2018, "TxDOT 0-6863: Pretensioned Concrete Bent Caps Phase 2 Experimental Data", https://doi.org/10.18738/T8/CPNVA5, Texas Data Repository, V1, UNF:6:rhD7jiUKeHH3Lu7bKo2DKA== [fileUNF]

Cite Dataset Learn about Data Citation Standards.

Access Dataset
Link Dataset
Contact Owner | Share

Dataset Metrics
1,271 Views
158 Downloads
1 Citation

Description This dataset contains metadata and data collected during TxDOT Project 0-6863 on development of standards for precast, pretensioned concrete bent caps.

Phase 2 tests are contained in this dataset. The Phase 2 setup consisted of a longer overhang and interior span than Phase 1, allowing for application of larger moment demands. Phase 2 consisted of two specimens, both with 28 longitudinal prestressing strands, 12" spacing of shear reinforcement, and internal voids for weight reduction. Void location and details varied between the two specimens (**PSV-28A** and **PSV-28B**).

Data provided includes specimen as-built drawings, construction timeline, measured material properties, test setup details, load patterns/sequence, applied loads at key points during test, and crack data (location and width).

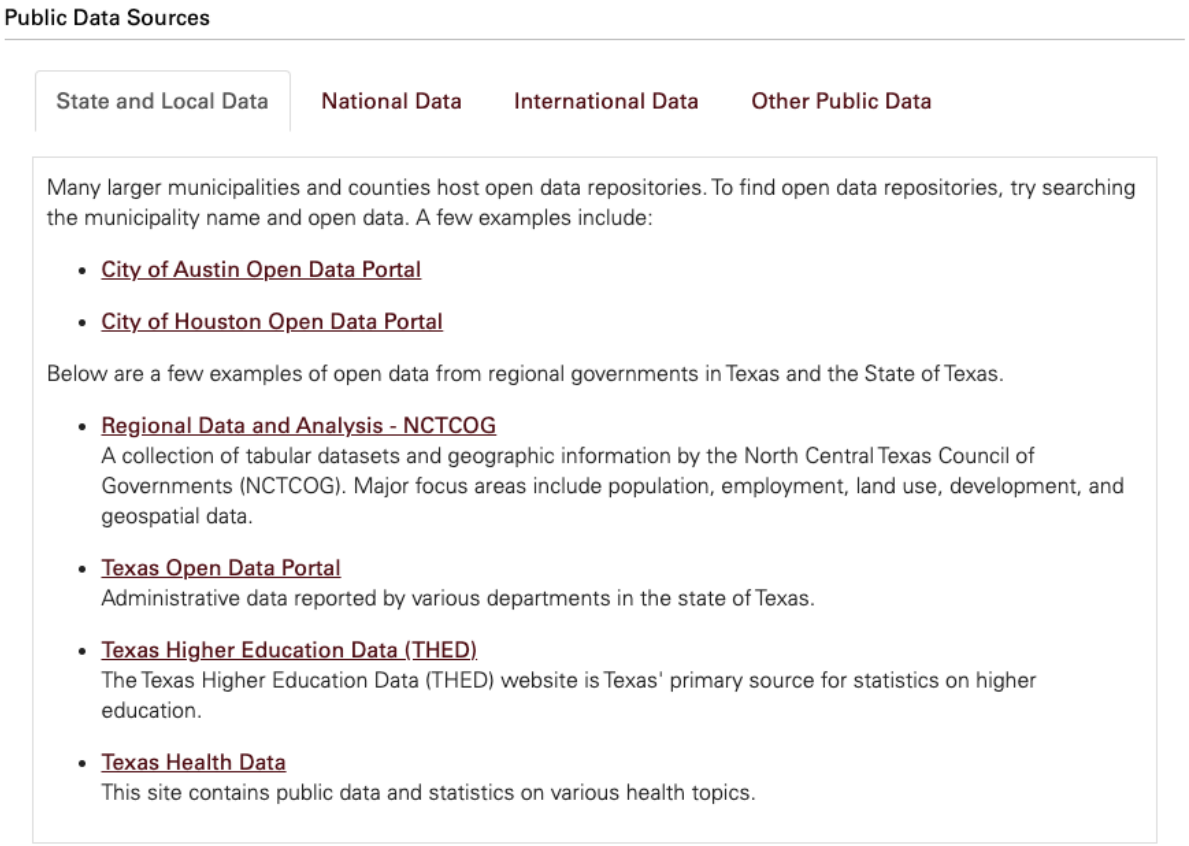
Subject Engineering

Related Publication Birely, A.C., Mander, J.B., Lee, J.D., McKee, C.D., Yole, K.J., and Barooah, U.R. (2018). "Precast, Prestressed Concrete Bent Caps: Volume 1 Preliminary Design Considerations and Experimental Test Program." Rep. No. FHWA/TX-18/0-6863-1-Vol1, Texas Department of Transportation and Texas A&M Transportation Institute.

License/Data Use Agreement Custom Dataset Terms

Other Open Data Resources

- University Libraries Research Databases
- National Center for Education Statistics (NCES)
- Registry of Research Data Repositories:
- NAHDAP
- Zenodo
- ICPSR
- kaggle
- LibGuide



Public Data Sources

State and Local Data | National Data | International Data | Other Public Data

Many larger municipalities and counties host open data repositories. To find open data repositories, try searching the municipality name and open data. A few examples include:

- City of Austin Open Data Portal
- City of Houston Open Data Portal

Below are a few examples of open data from regional governments in Texas and the State of Texas.

- Regional Data and Analysis - NCTCOG
  A collection of tabular datasets and geographic information by the North Central Texas Council of Governments (NCTCOG). Major focus areas include population, employment, land use, development, and geospatial data.
- Texas Open Data Portal
  Administrative data reported by various departments in the state of Texas.
- Texas Higher Education Data (THED)
  The Texas Higher Education Data (THED) website is Texas' primary source for statistics on higher education.
- Texas Health Data
  This site contains public data and statistics on various health topics.

FILE NAMING STRATEGIES

DATA STORAGE

Good Practices in RDM

SAR_090320.doc

What does this mean?

- Survey Analysis Results?
- Survey of Agriculture Research?
- Sam A. Rodriguez, a researcher?
- September 03, 2020?
- March 09, 2020?
- March 20, 2009?

File Naming

Two main criteria: **Context & Consistency**

Good File Naming Practices

- Use descriptive file names
- Use a standard date system
- Use leading zeros
- Use basic characters and avoid (/ , # ?)
- Version files
- Be consistent

Good File Naming Practices

- Use descriptive file names
- Use a standard date system
- Use leading zeros
- Use basic characters and avoid (/ , # ?)
- Version files
- Be consistent

- ❖ Date (YYYY-MM-DD)
- ❖ Project name/Grant #
- ❖ Type of data
- ❖ Location/site/spatial coordinates
- ❖ Researcher info
- ❖ Version

Good File Naming Practices

- Use descriptive file names
- Use a standard date system
- Use leading zeros
- Use basic characters and avoid (/ , # ?)
- Version files
- Be consistent

SAR_090320.doc ?

in YYYY-MM-DD format (2023-09-19)

Sort, with standard dates

2023-03-16_Code_descriptions.docx
2023-05-24_Code_descriptions.docx
2023-11-03_Code_descriptions.docx

Sort, without standard dates

11-3-23_Code_descriptions.docx
3-16-23_Code_descriptions.docx
5-24-2023_Code_descriptions.docx

Good File Naming Practices

- Use descriptive file names
- Use a standard date system
- Use leading zeros
- Use basic characters and avoid (/ , # ?)
- Version files
- Be consistent

Sort, with a leading zero

Test01_RDM assessment.xlsx
Test02_RDM assessment.xlsx
Test03_RDM assessment.xlsx
…
Test10_RDM assessment.xlsx
Test11_RDM assessment.xlsx

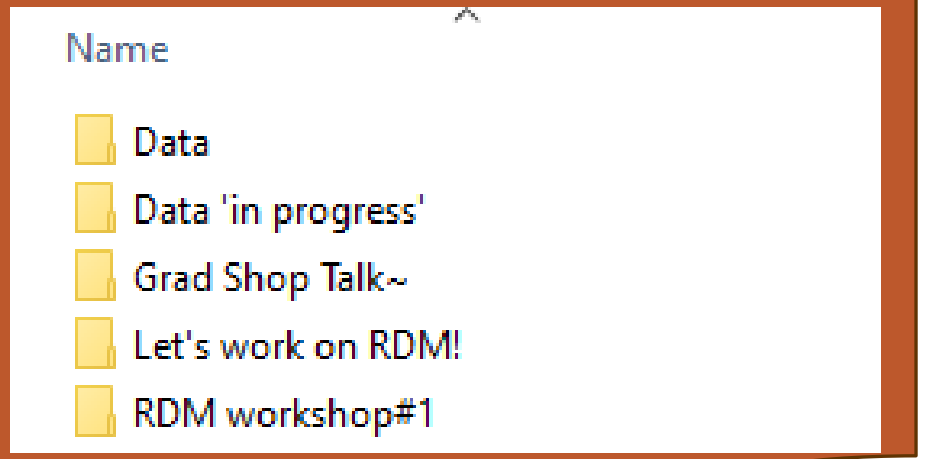
Sort, without a leading zero

Test1_RDM assessment.xlsx
Test10_RDM assessment.xlsx
Test11_RDM assessment.xlsx
…
Test2_RDM assessment.xlsx
Test3_RDM assessment.xlsx

Good File Naming Practices

- Use descriptive file names
- Use a standard date system
- Use leading zeros
- Use basic characters and avoid (/ , # ?)
- Version files
- Be consistent

- **Use_as delimiters**
  - **Data_v01**
- **Avoid Special Characters**



Good File Naming Practices

- Use descriptive file names
- Use a standard date system
- Use leading zeros
- Use basic characters and avoid (/ , # ?)
- Version files
- Be consistent

❖ **Using consecutive numbering for major version changes**

Code_descriptions_**20230919_v01**.docx

❖ **Using decimals for minor changes**

Code_descriptions_**20230919_v01.1**.docx

Good File Naming Practices

- Use descriptive file names
- Use a standard date system
- Use leading zeros
- Use basic characters and avoid (/ , # ?)
- Version files
- Be consistent

❖ **Consistency with Spaces**

Data_projectname **v03**.docx

Data_**project name v01**.docx

Data_**projectname_v02**.docx

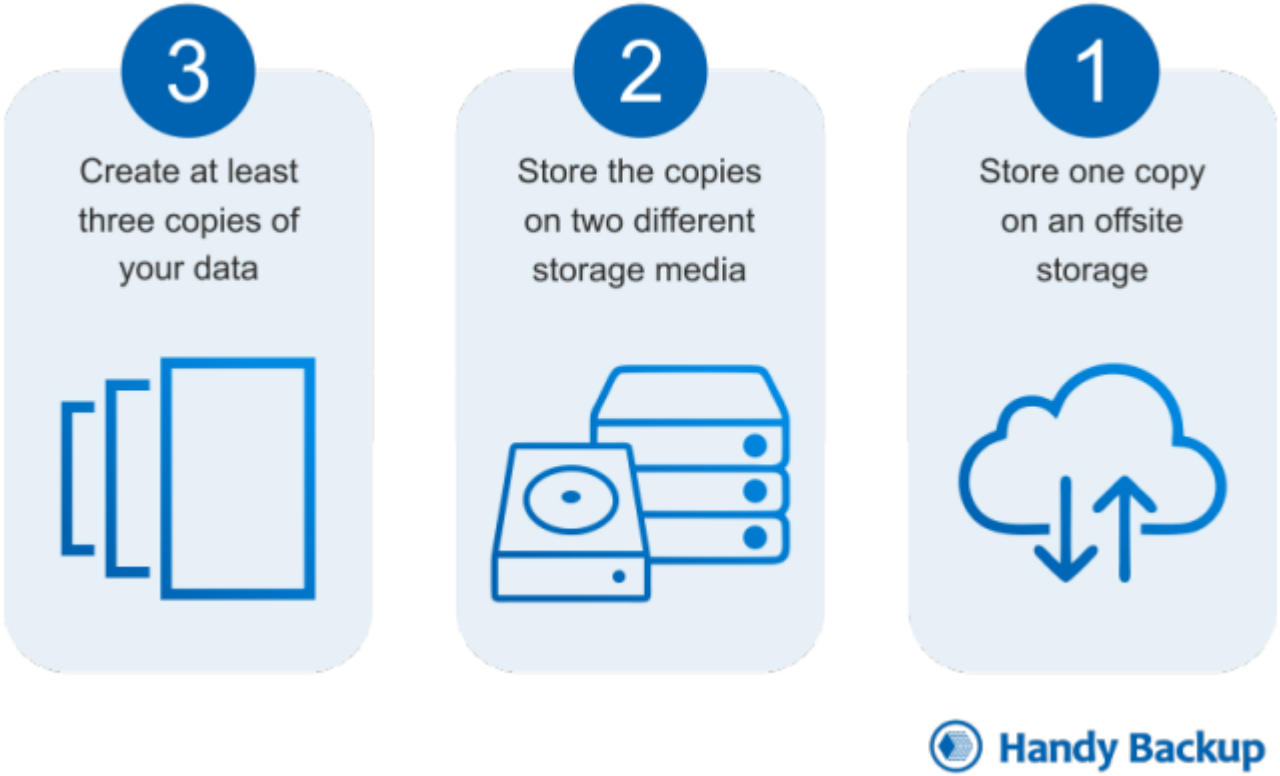
Data Documentation: README File

README files are plain text documents that sit at the top level of project folders and describe the purpose of the project, contact details, and organization of files.

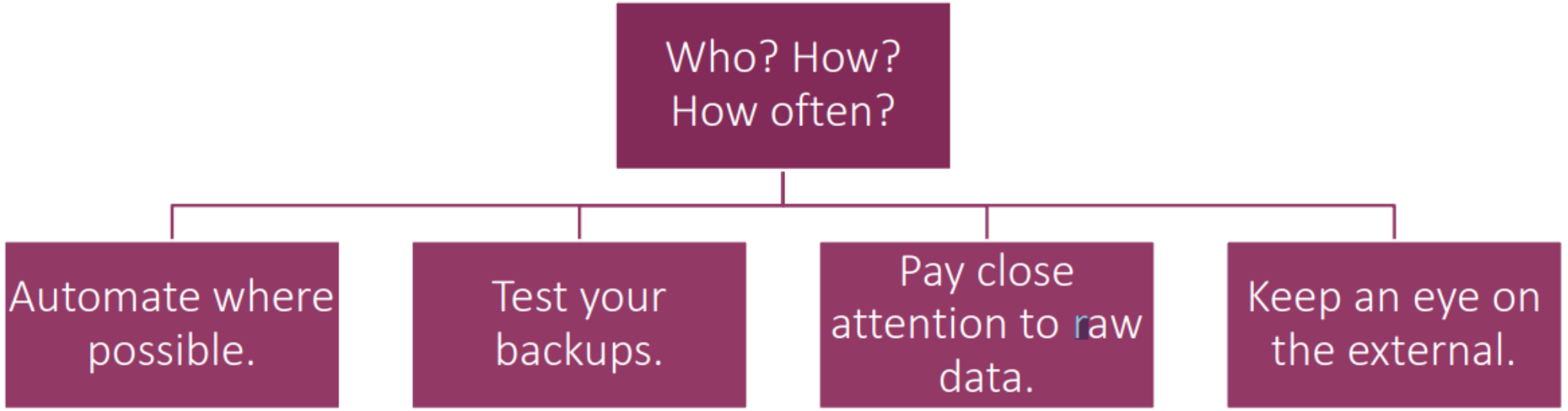
A standard document detailing information about the documents:

- Title of dataset
- Name/institution/contact information for
- Principal Investigator (or person responsible for collecting the data)
- File name structure and the description of the attributes used to name the files.
- Descriptions of every folder, file, format, data collection method, instruments, etc.
- Codes: Provide a complete list of any codes/abbreviations used.
- Dates/Locations of data collection
- Funding information
- People involved

Data Storage and Backup **3-2-1 Rule**

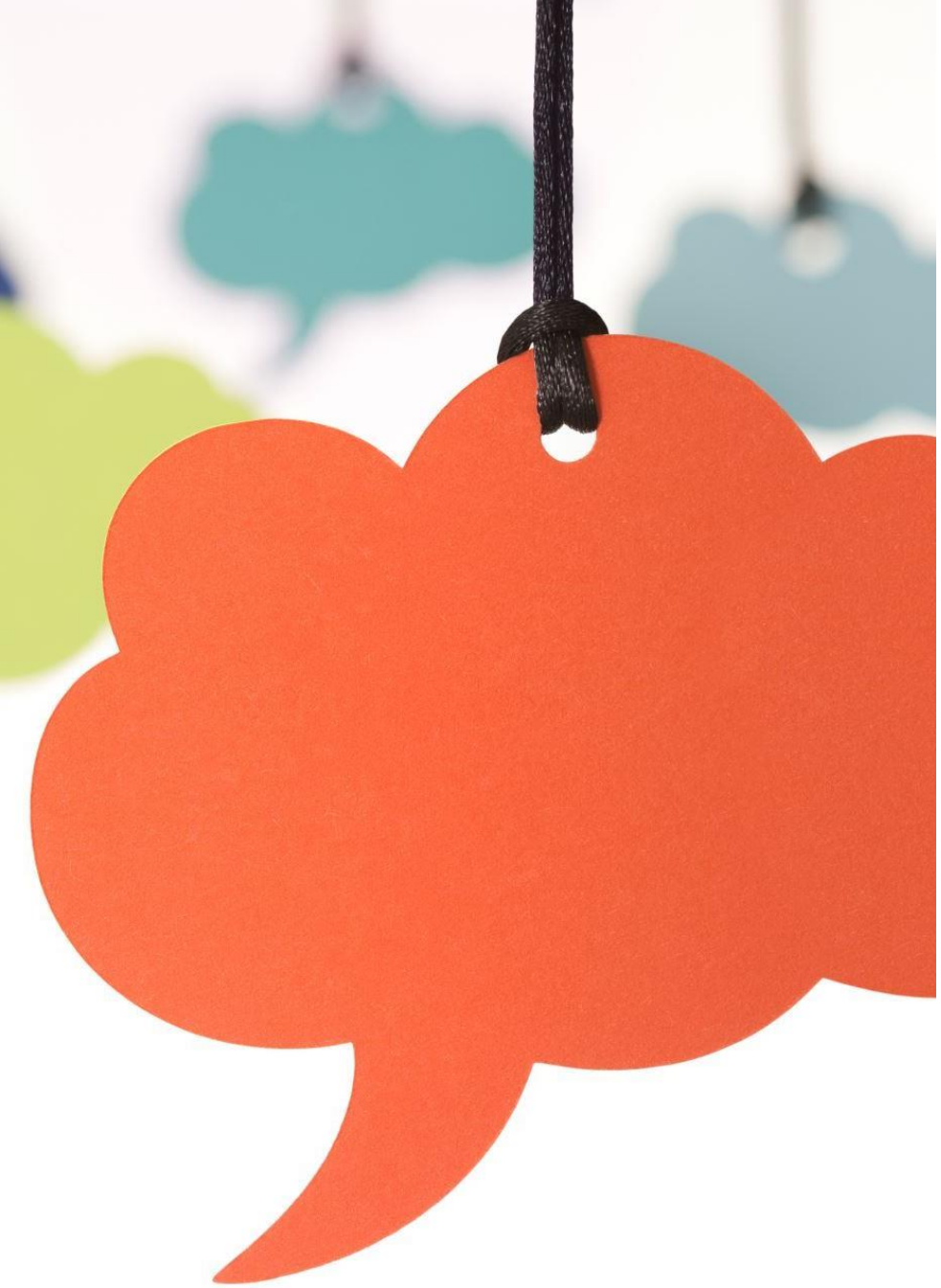


Consideration for back-up



TXST Data Classifications

	Confidential Information	Sensitive Information	Public Information
Level of Sensitivity	High	Moderate	Low
Legal Requirements	Protection of data is required by law (e.g., TPIA, FERPA, and HIPAA data) or contractual agreements.	Often considered "public" in the sense it is releasable under the Texas Public Information Act, some assurance is required so release of information is both controlled and lawful.	Public information by its very nature is designed to be shared broadly, without restriction, at the complete discretion of the owner.
Disclosure Risk	Confidential information presents the most serious risk of harm if improperly disclosed.	Unauthorized disclosure of Sensitive information could adversely impact the University, individuals or affiliates.	From the perspective of confidentiality, public information may be disclosed or published by any person at any time.
Examples of Information	• Social Security numbers • Credit card info • Personal health info • Student records • Crime victim info • Library transactions • Court sealed records • Access control credentials	• Performance appraisals • Employee DOB • Employee email addresses • Donor information • Voicemail records • Email contents • Unpublished research	•Job posting • Service offerings • Published research • Directory information • Degree programs • General information about university products and services



Final Tips and Reminders

- Know your data
- Choose file formats that last
- Remember the documentation
- Don't forget backups
- Consider ownership and privacy

May & Summer (2024)

The Carpentries Workshop

Software Carpentry (R for Reproducible Scientific Analysis)

Our more introductory R lesson. In addition to our **standard content**, this workshop covers data analysis and visualisation in R, focusing on working with tabular data and other core data structures, using conditionals and loops, writing custom functions, and creating publication-quality graphics. As our more introductory R offering, this workshop also introduces learners to RStudio and strategies for getting help. This workshop is appropriate for learners with no previous programming experience. For audiences with some experience with R or other programming languages, we recommend our **Programming with R** lesson.

Software Carpentry (Programming with R)

Our more advanced R lesson. In addition to our **standard content**, this workshop covers data analysis and visualisation in R focusing on working with core data structures, using conditionals and loops, writing custom functions, and running R programs from the command line. This is the more advanced of our two R offerings for Software Carpentry and is appropriate for learners with some previous programming experience, in R or other languages. For audiences with no previous programming experience, we recommend our **R for Reproducible Scientific Analysis** lesson.

THE CARPENTRIES

Innovate for Impact: Open Data Solutions for the Bobcat Community

What: The Bobcats Community is growing as we are gaming at running into R1 institutions. Of course, growth brings changes and creates opportunities and challenges. In celebration of Love Data Week, University Libraries is proud to be hosting our inaugural TXST Open Datathon challenges for our Texas State students to immerse themselves in the world of open data resources and data science. The theme of this Datathon activity encompasses the commitment of TXST to innovation, community engagement, and leveraging open data for meaningful impact. Participants will work collaboratively to make a meaningful impact on real-world issues through open data and share final digital artifacts to support TXST university communities aligned with the university's mission.

Who can join?

- Teams of up to 4 undergraduate and/or graduate students from all academic disciplines are invited to participate.
- Teams may form in advance, or we will organize teams after registration closes.
- No expertise in data science is required; we will guide you through the resources and processes you need to explore the open data
- Please consider attending an event orientation and all other scheduled hours on Friday, February 16th, 2024, and staying through the awards notifications.

When: Feb. 12-16, 2024.

Participants will gather on Monday, Feb. 12 for the open talk and kickoff event and will work independently with their teams during the week. Final submissions will be due by 1 p.m. on Friday, Feb. 16, and in-person team presentations will begin at that time. (See full event schedule below).

Thank you!

Xuan Zhou, PhD

Data Curation Specialist

x_zhou@txstate.edu

Research Data Services

Digital Scholarship & Research

Texas State University Libraries

https://www.library.txstate.edu/