Data Management 101

Jessica Trelogan, Data Management Coordinator, UT Libraries
j.trelogan@austin.utexas.edu

https://commons.wikimedia.org/w/index.php?curid=110393
Objectives

• Build confidence and skills
• Create talking points and pathways to better serve your constituents
• Identify gaps
• Assemble resources
• **Build community**
Part 1: Understanding Data and Data Management

Data Basics

• Table discussion: what are data?
• What is data management?
• Associated materials and metadata

Managing data

• Research data lifecycle
• Collecting, organizing, securing, documenting, and sharing

Data Management Planning

• What is a DMP?
• Why is it important (and useful)?
Part 2: Sharing, publishing, and preserving data

Data sharing
- Faculty attitudes
- Barriers, challenges, and incentives
- Reproducibility and impact
- Table discussion: how to talk to about sharing

Publishing and Citing Data
- Why cite data?
- DOIs and ARKs

Data Preservation and Archiving
- Repositories (General vs. Disciplinary)
- Selecting data
Singular? Plural?

Be comfortable. Don’t sweat it.
What are “data”?
What are “data”? 

- Varies widely by discipline 
- Also by context 
- Heterogeneous and situation-dependent 

- For our purposes: 
  - Research data 
  - Digital only (born that way or digitized)
What are “data”?  

<table>
<thead>
<tr>
<th>Natural/Physical Sciences</th>
<th>Social Sciences</th>
<th>Humanities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observational</td>
<td>Qualitative</td>
<td>Raw</td>
</tr>
<tr>
<td>Experimental</td>
<td>Quantitative</td>
<td>Primary</td>
</tr>
<tr>
<td>Simulation</td>
<td></td>
<td>Interpretive/Derived</td>
</tr>
</tbody>
</table>
(3) Research data means the **recorded factual material commonly accepted in the scientific community** as necessary to validate research **findings**, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This “recorded” material excludes physical objects (e.g., laboratory samples). Research data also do not include:

(i) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and

(ii) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study.
National Science Foundation:

“...determined by the community of interest through the process of peer review and program management. This may include, but is not limited to: data, publications, samples, physical collections, software and models.”

From FAQs: https://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp#1
National Endowment for the Humanities

“...materials generated or collected during the course of conducting research.”

**Includes:**
- citations
- software code
- algorithms
- digital tools
- documentation
- databases
- geospatial coordinates
- reports and articles

**Excludes:**
- preliminary analyses
- drafts of papers
- plans for future research
- peer review assessments
- communications
- confidential materials
- information violating privacy

Associated Materials

Questionnaires

Field notebooks

Codebooks

Methodologies

“Background” data

http://nmnh.typepad.com/...
SIMPLY EXPLAINED: METADATA

WHAT'S "METADATA"?

A WORD WITH 8 LETTERS
Metadata

“Structured information that describes, explains, locates, or otherwise represents something else.”

National Information Standards Organization:
http://www.niso.org/standards/resources/UnderstandingMetadata.pdf
Metadata talking points

Data are only useful if understandable

Strive for structured, machine readable, standards-based

Minimum effort is better than none

But even just a story can help

README.txt

Dublin Core 15
Research Data Lifecycle
Data Archiving
- Preservation metadata
- Confidentiality
- Add'l processing

Data Archiving

Study Concept
- Initial concepts
- Questions and answers
- Grant info

Data Gathering
- Questionnaire
- Coded Instrument
- CAI metadata
- Paradata

Data Processing/Analysis
- Data Specs
- Recodes
- Summary descriptive info

Repurposing
- Post-hoc harmonization
- Data transformations

Data Distribution
- Term of use
- Citation
- Packaging Info

Data Discovery
- Catalog Record
- Indexing
- Related publications

Data Analysis
- Replication code
- Publications

Color Key:
Data used by researchers during project
Data archived after project is complete

http://www.lib.utexas.edu/datamanagement/managing
Data Management Plans

A data management plan (DMP) is a written document describing the nature and structure of the data you will likely use or produce in the course of research, along with your strategies for dealing with it throughout and after your project.
Why bother?

Save time and money
Maximize your impact
Allow for reuse
Do better science

"I was close to a breakthrough when the grant money ran out."

Reprinted from Funny Times / PO Box 18530/Cleveland Hts. OH 44118
phone: 216.371.8600 / email: ft@funnytimes.com
Why else?

It’s required.
What goes into one?

It depends.

http://researchsharing.sparcopen.org/
How to write one?

• Sign in with your institution (or create account “Not in List”)
• Templates for most major funding agencies
• Customized for templates for member institutions
• Save, cut/paste, print

[DMPTool](https://dmptool.org/)
Common Elements of a DMP

1. Data description

2. Data documentation

3. Access, sharing, re-use

4. Storage and backups

5. Preservation and archiving

6. Resources and responsibilities
1. Data Description

What data will you gather or create?

- File types, formats, volume
- Methods and context of data collection
- Discussion of data sources
- Structure and organization of data files
- Data validation, quality assurance
- Data transformations or processing steps
- Conditions of access and use; confidentiality
2. Metadata

What documentation will accompany your data?

Type and form

Metadata standards

Basic details

Definitions of variables, units, codes

By Sobebunny (Own work) [CC BY-SA 3.0], via Wikimedia Commons
3. Access, Sharing, and Reuse

Have you gained consent?

Who will have access?
When? How?

Are there any restrictions?

What are the approved uses?

How will you protect sensitive information?

4. Storage and Backups

Do you have enough storage space?

Do you need security measures?

How/how often will you do backups?

What’s your recovery plan?
5. Preservation and Archiving

What is your long-term preservation plan?

What data should be retained?
Shared? Destroyed?

How will you maintain and curate it?

What future uses are there?

Where will data live after the project?
For how long?

Are there any future costs?

6. Resources and Responsibilities

Will you need additional help?

Software? Hardware?

What is this going to cost?

Who is responsible for what?

https://www.tacc.utexas.edu/systems/stampede
Data Management
What is Data Management?

A collection of tasks practiced throughout the lifecycle of research that make it easier to find, understand, navigate, and use your data.
Collecting data

- Test your plan
- Automate where possible
- Create snapshots
- Ensure compliance

Photo by Jessica Trelogan
Re-using data

Find the right data
  • Subject specialists: lib.utexas.edu/subject/index.php
  • re3data.org

Know your sources
  • Restrictions
  • Copyright
  • Data citation datacite.org

Integrate/Normalize
  • OpenRefine
Organize
File Names

Be descriptive, not generic

Include dates

CamelCase vs Pot_hole_case

No funny characters
"/ \ : * ? " < > [ ] & $

Describe your convention

Use a batch re-namer:
www.bulkrenameutility.co.uk (Windows)
www.renamer4mac.com (Mac)
www.powersurgepub.com/products/psrenamer.html (Linux, Mac, Windows)
File Formats

Non-proprietary, open standards

Used commonly in your domain

Encoded with standard characters

Uncompressed (?)

DROID:
www.digital-preservation.github.io/droid/

Library of Congress:
www.loc.gov/preservation/resources/rfs
Sharing active data

Ensure easy access
Avoid duplication
Control versions
Keep a list!

https://www.flickr.com/photos/ryanr/142455033
Document, Document, Document!

Data only useful if understandable!

Metadata

Readme.txt (use a template)

Data Dictionaries

Electronic Lab Notebooks

Codebooks/lab books/field notes

Storage

University of Texas

• Departmental
• 2 TB in Box
• UTMail (Google Drive)
• 5 TB at TACC
• ITS: VMs

Other Cloud

• DropBox
• Google Drive
• iCloud
Security

- Passwords
- Encryption
- Updates
- Backup strategies
- Sensitive data
Managing Sensitive Data

Personally Identifiable Information (PII):
“Any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means” (US Department of Labor)

Protected Health Information (PHI):
“Any [individually identifiable] information, whether oral or recorded in any form or medium, that—
(A) is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse; and
(B) relates to the past, present, or future physical or mental health or condition of any individual, the provision of health care to an individual, or the past, present, or future payment for the provision of health care to an individual.” (US Department of Health and Human Services)
Restrictions

Embargos

Access Restrictions

Data Use Agreements

Ownership

https://www.flickr.com/photos/nycstreets/8358489746/in/photostream/
Discussion

• Paradigm shift
• What does it feel like/mean to you?
• What are you hearing by way of attitudes from faculty?
• How can we communicate, support, or change those attitudes?
• Where/how can we intervene?
• What approaches work? What don’t?
Publishing and Citing Data
What is publication?

- Uploading to a repository
- Submitting with an article
- Including as an appendix/supplement
- Making available on a public website

Why cite?

• For the same reasons you would cite a journal article – to get and give credit
• To help data stand on their own as scholarly output
Citation serves several purposes

- Provides appropriate credit to data producers & data publishers
- Enables other researchers to access the data
- Assists in measuring the impact of data
- Helps data producers know how their data is being utilized

Force 11 Data Citation Principles: [https://www.force11.org/group/joint-declaration-data-citation-principles-final](https://www.force11.org/group/joint-declaration-data-citation-principles-final)
How do you cite?

• Citing is easy – making data citable is harder
• Be deliberate about publication
• Request a persistent identifier
• Different versions of the same dataset should get different identifiers
• Data citation should go to a “catalog” page instead of directly to the data
• The data should link to any associated publications
Elements of a citation

Author/creator

Title

Version

Publication Date

Publisher/archive

Identifier/locator
Style

Data Cite:

• Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier

APA:

• Creator (PublicationYear). Title (Version Number) [Description of form]. Retrieved from http://

DOI Citation Formatter: http://crosscite.org/citeproc/
Questions about data citation

At what level of granularity should data be made citable?

What about regularly updated datasets?

Should ORCID and/or ISNI be included in a data citation?
Preservation and Archiving
Goals of this part of data lifecycle

• To have ongoing, consistent, citable access to data after a project is complete.
  • Allows review, re-use, interpretation, and re-creation of the data
• Ensure the integrity of the data
How long to keep it?

Are there any retention guidelines you need to follow?
• e.g. NIH requires 3 years from the close of grant

How long will it likely be useful to yourself and others?

Be realistic – this will probably cost you money
What to keep long-term?

- Data that can’t be replicated (e.g. weather data)
- Can be replicated but would be prohibitively expensive
- Major discovery
- High impact researcher
- Raw and final, processed files but not intermediate files
- Technical documentation is comprehensive and data is in a format that allows for ease of use and preservation
What is deposited?

• Raw data and processed data
• Description of project methodology
• Explanation of how data was handled post-collection
• Codebooks or other records
• Project generated software or code that was created to analyze the data
• Related records (e.g. human subjects protocols)
What about stuff you don’t need to keep?

Delete when finished or after the retention period

Keep basic records of the things you delete
  • date, format, info about project/grant, reason for destruction

https://www.flickr.com/photos/ladymixy-uk/4059154289
Types of repositories

- Institutional repositories (e.g. Texas ScholarWorks)
- Disciplinary repositories (e.g. ICPSR, Dryad)
- Open repositories (e.g. Dataverse)

http://www.re3data.org/
Possible limitations & requirements

- Size limitations
- Costs for deposit
- File format requirements
- Metadata requirements
Features to look for:

• OAI-PMH compliant
• Should assign a persistent identifier
• Preservation functions
  • Regular back-ups/replication (preferably with some geographic separation)
  • Check-sums or similar integrity checks
  • Migration plan
  • Succession plan if repository folds
• Automatic recording of provenance metadata
Why use a repository?

Digital content is fragile

Websites (especially personal ones) are ephemeral

• No integrity checking
• Likely not very visible to search engines
• Require upkeep and technological dependencies
• Links may not be persistent

Funding agencies expect it
Questions?

• References: