Research Data Management

Modified based on

Shanda Hunt, Libraries

Alicia Hofelich Mohr, CLA LATIS

What is Data Management?

Practices across the lifecycle of a project that:

- ensure integrity of the data & facilitate replication
- protect the security of data
- enhance efficiency & reliability of the research

No, but really, what is it?

Thinking about and planning for:

- File names, folder structures, and their management
- Documentation & metadata
- Storage, backups, and security
- Sharing and/or preserving data

Why should you care?

- Data Management Plans - NSF, NIH, & many others
- Journal requirements and professional society ethical standards
- Saving time
- Demonstrating your integrity as a scientific researcher - open methods/open data

Why should you care?

Imagine: Three years after completing a study, a researcher contacts you wanting to use some of the data or materials to replicate your work.

- Can you locate these files or materials?
- Are they stored someplace you can still access?
- Are any of the files corrupt? Or in a format that you/others no longer have the software to read?
- Do you have documentation about how you created, analyzed, and made use of the data/materials that still makes sense to you and others?

Data Management Plan (DMP)

Five main questions to consider:

1. What type of data or files will be produced or used?
2. What standards will be used for documentation and metadata?
3. What steps will be taken to protect privacy, security, confidentiality, intellectual property or other rights?
4. If you allow others to reuse your data, how will the data be accessed and shared?
5. How will the data be archived for preservation and long-term access?

What data will be produced/used?

- Consistency is key - all data & materials in grant application should appear in the DMP
- Why?
  - Evidence of feasibility
  - Later decisions on sharing/preservation hinge on types of data

What is considered “data”?

Types of data/files

- Physical collections, survey forms, artifacts
- Spreadsheets/Statistical files
- Interviews, focus group notes
- Field notes
- Audio/videos
- Code

You don’t have to share everything

- Preliminary drafts, personal notebooks, etc
- Identifying or proprietary information

Protecting Data: Store Data Securely

Protected Health Information/ Sensitive Identifiable Data

- ✓ AHC Managed Servers
- ✓ Box Secure Storage*
- ✓ Encrypted Drive/Container

- x Shared/personal drive
- x UMN Google Drive
- x Dropbox
- x Amazon (personal)
- x Email

De-identified Data

(direct identifiers removed)

- ✓ AHC Managed Servers
- ✓ Box Secure Storage*
- ✓ Shared/personal drive
- ✓ UMN Google Drive

- x Dropbox
- x Amazon (personal)
- x Email

***NEW!**

Protecting Data: Backup

To avoid losing data, Use the 3-2-1 rule:

3 copies of your work (1 working copy, 2 backups)

On **2** different kinds of storage

At least **1** copy off site

North Quad, University of Michigan
https://www.si.umich.edu/news/

How will you document your data?

Describe strategies for keeping track of all the data and materials created

Why?

- Promotes reproducibility, transparency, and efficiency of research
- Ensures research can be understood later

Documentation/Metadata

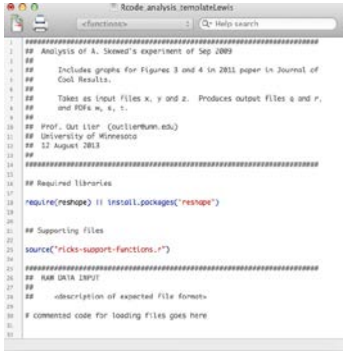
What should you document?

- Sources of data
  - When, where, & how data/materials were collected
- Study decisions (methods, coding, etc)
- Statistical analyses
- Software and instruments used
- Where data/documentation are stored
- Future research ideas and plans

How should you document it?

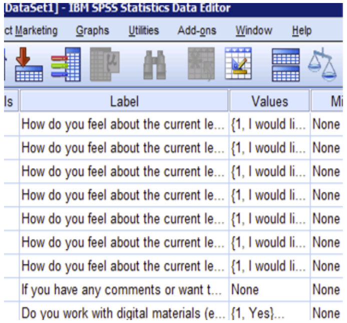
A research log

Statistical Syntax



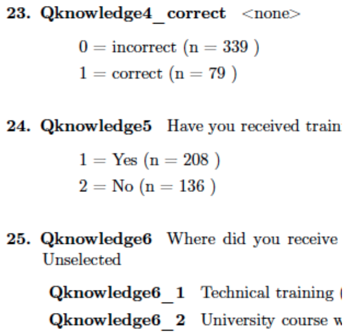
```
##################################################################################
##  Analysis of A. Skewed's experiment of Sep 2009
##
##      Includes graphs for Figures 3 and 4 in 2011 paper in Journal of
##      Cool Results.
##
##      Takes as input files x, y and z.  Produces output files q and r,
##      and PDFs w, s, t.
##
##  Prof. Out Lier  (outlier@umn.edu)
##  University of Minnesota
##  12 August 2013
##
##################################################################################

## Required libraries

require(reshape) || install.packages("reshape")


## Supporting files

source("ricks-support-functions.r")

##################################################################################
##  RAW DATA INPUT
##
##     <description of expected file format>

# commented code for loading files goes here
```

Dataset metadata



ls	Label	Values	Mi
	How do you feel about the current le...	{1, I would li...	None
	How do you feel about the current le...	{1, I would li...	None
	How do you feel about the current le...	{1, I would li...	None
	How do you feel about the current le...	{1, I would li...	None
	How do you feel about the current le...	{1, I would li...	None
	How do you feel about the current le...	{1, I would li...	None
	How do you feel about the current le...	{1, I would li...	None
	If you have any comments or want t...	None	None
	Do you work with digital materials (e...	{1, Yes}...	None

Codebooks



23. Qknowledge4_correct <none>

0 = incorrect (n = 339)
1 = correct (n = 79)

24. Qknowledge5 Have you received train

1 = Yes (n = 208)
2 = No (n = 136)

25. Qknowledge6 Where did you receive
Unselected

Qknowledge6_1 Technical training (
Qknowledge6_2 University course w

Dataset Metadata

Be sure to create:

- short, but descriptive variable names
- variable labels (especially if it's derived from other variables)
- assign variable values (1=Yes, 2=No)

If you're using Excel:

- row/column headings
- descriptions of abbreviations, variable names, range of variables
- note any formulas
- note any modifications (What is that column doing there?)
- what to follow up on (Oh yes, that's why those cells are yellow)

Codebooks

Qualitative

- code
- description
- guidelines for when to use
- guidelines for when not to use
- examples of both

Quantitative

- Variable name
- Variable values & labels
- Survey questions
- Recode/calculations
- Descriptive statistics
- If and how missing information is coded

Folder structure

Possible organizational strategies:

By data type: databases, text, images, models, etc.

By research activities: interviews, surveys, experiments, etc.

By materials: data, documentation, publications, etc.

File Naming

- **Be Descriptive:** interview.txt is not helpful. Instead: 20150814_interview_site01_respondent04.txt (up to 255 characters)
- **Don’t embed information solely in folder structures**: 2015/august/minneapolis/interviews/reactionmemo.txt
- **Use consistent structure:** create a useful order (for sorting) and decide on shared terminology
- **Use numerical dates:** YYYYMMDD rather than Dec09 or December 9

Here’s why you should use numerical dates

Sort, without numerical dates

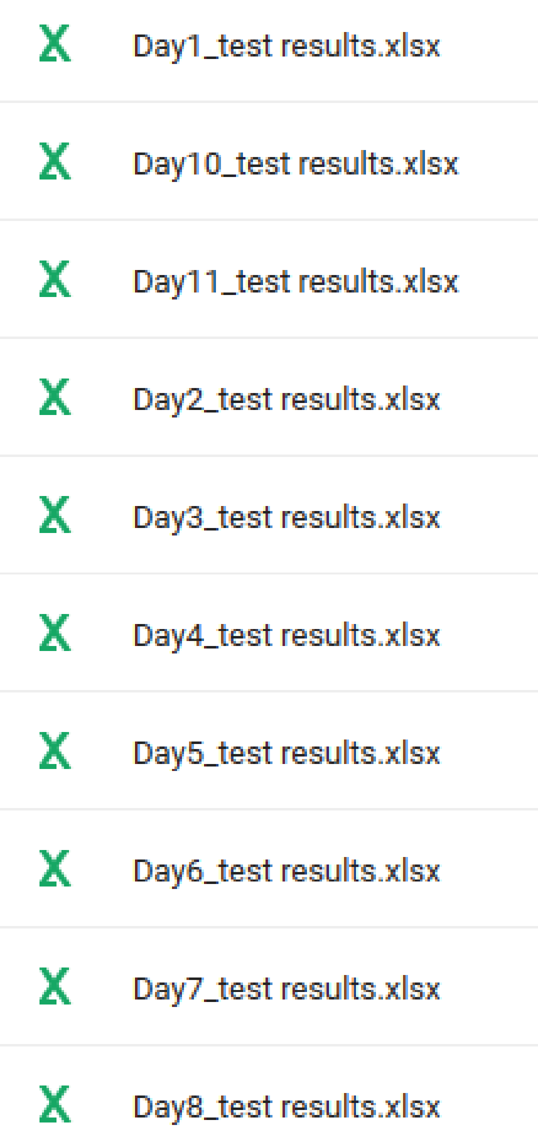
Code_descriptions_**12-8-15**.docx
Code_descriptions_**2-14-2015**.docx
Code_descriptions_**8-1-2015**.docx

Sort, with numerical dates

Code_descriptions_**20150214**.docx
Code_descriptions_**20150801**.docx
Code_descriptions_**20151208**.docx

Here’s why you should use leading zero

Sort, without leading zero



Day1_test results.xlsx

Day10_test results.xlsx

Day11_test results.xlsx

Day2_test results.xlsx

Day3_test results.xlsx

Day4_test results.xlsx

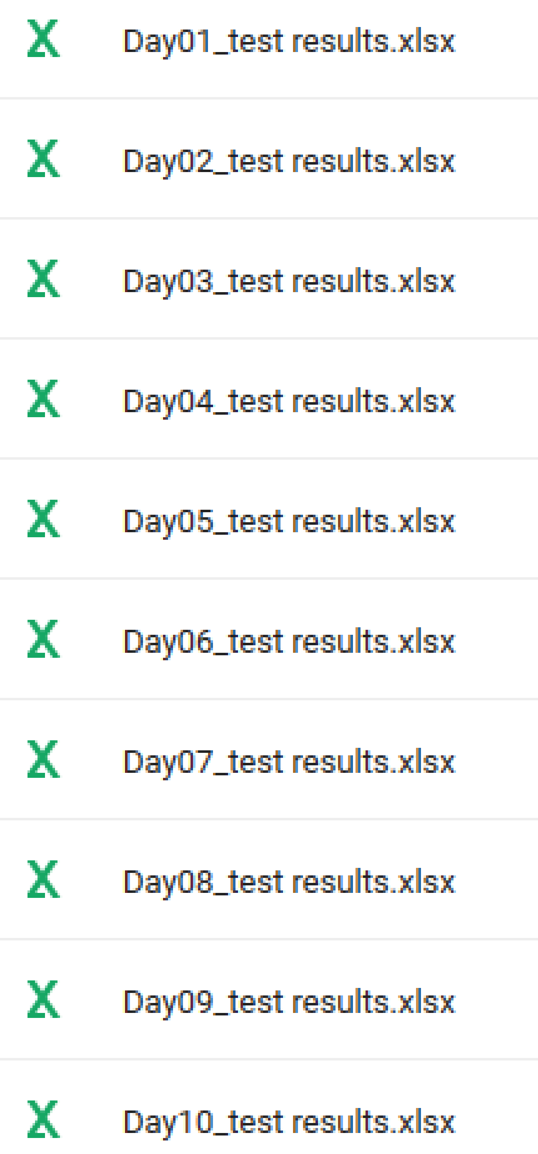
Day5_test results.xlsx

Day6_test results.xlsx

Day7_test results.xlsx

Day8_test results.xlsx

Sort, with leading zero



Day01_test results.xlsx

Day02_test results.xlsx

Day03_test results.xlsx

Day04_test results.xlsx

Day05_test results.xlsx

Day06_test results.xlsx

Day07_test results.xlsx

Day08_test results.xlsx

Day09_test results.xlsx

Day10_test results.xlsx

How will you share your data & handle privacy?

Describe plans for making data available

- Shared as widely as possible and appropriate
  - – How will data privacy/sensitivity and intellectual property rights be handled?
- Why?
  - – Reuse/dissemination of data
  - – Replication of results

Things to think about for sharing

1. IRB Considerations

2. Copyright/Intellectual Property

3. Data Privacy issues - locations, identifiers

4. Where & How to share

Make sure IRB aligns with your plans

Promise to eventually destroy the data?

- Destroy direct identifiers, linking information, identifying audio files
- Retain & preserve de-identified transcripts

Responses only seen by research team?

- Identifying data will be kept confidential
- Only de-identified data will be shared

Data will only be shared in aggregate?

- Individual responses will only be shared in ways that will not identify you

Understand Data Ownership Before Sharing

- Intellectual Property
- Copyright Policy
- UMN Research Data Management Policy

Sharing Human Subjects Data

- De-identify for broader use
  - Evaluate direct and indirect identifiers
  - Recode or collapse categories with small frequencies
- Choose appropriate access methods
  - Not all data should be shared publically
  - Share through restricted-access methods or use repositories with data use agreements

According to US Census, 87% of population can be identified with DOB, sex, and zip code.

Considerations

1. The sensitivity of the data

2. Promises made to research participants

3. With whom and how data are shared

"Data can either be useful or perfectly anonymous but never both." -Paul Ohm

How to share your data/materials

Consider depositing data in an archival repository

Data Repository for U of M

- Free for UMN-TC students, staff, faculty
- All data types & topics
- Curated by disciplinary experts
- Public data sharing

ICPSR

- Free for members; charge for OpenICPSR
- Social Science data
- Curated by social science experts
- Restricted data sharing options

How will you preserve your data?

How will you ensure the data will be around beyond the life of the project?

- How long? ("forever" is not a realistic plan)
- How will you do it? ("keep it indefinitely on hard drive/server" not enough)

Storage vs. Archiving

During Project

Storage

- Back-ups
- Active data

Actions

- Documentation

After Project

Archival Storage

- Final versions
- Off-line

Actions

- Preservation

Created by Rigo Peter
from Noun Project

Library Resources

LIBRARIES ONLINE WORKSHOPS

Series of short videos on different steps of data management

www.lib.umn.edu/datamanagement/workshops

DATA MANAGEMENT CAMPS

Intensive bootcamp for graduate students to introduce to data management topics & skills

Offered in January and August

www.lib.umn.edu/datamanagement/workshops/summer2016

Campus Resources (links)

- UMN’s data management page
- OIT’s Storage and Data Protection Services
- Box Self-Help Guide
- Training for Researchers and Students at UMN libraries
- List of resources on campus for managing data of all types

Full slides available: http://z.umn.edu/cehddatamanagement