Introduction to Research Data Management

Alex Garnett
Adam McKenzie

Research Data

In sciences, social sciences, and humanities.

Records created in the course of a research project.

Why Manage Data?

Locate your files easily.

Keep track of versions.

Reproduce your work.

Collaborate.

Satisfy grant and journal requirements.

The RDM Cycle

Across the data lifecycle

1. Create/Discover
2. Process
3. Analyze
4. Preserve
5. Share
6. Re-use

Data Management Plans

A data management plan (DMP) helps you identify and mitigate future roadblocks.

Dear NSF,

I will store all data on at least one, and possibly up to 50, hard drives in my lab. The directory structure will be custom, not self-explanatory, and in no way documented or described. Students working with the data will be encouraged to make their own copies and modify them as they please, in order to ensure that no one can ever figure out what the actual raw data is…

http://ivory.idyll.org/blog/data-management.html

Demo: Portage DMP

Create Data

Design your research. Collect your data.

Capture and create metadata.

Metadata Describes

Who?

What?

When?

Where?

Why?

Metadata is Essential

You will not remember what the variable “multmemgp” represents.

Without metadata, your research is not reproducible.

It belongs in a unique file.

Documentation Levels

Project level

File or database level

Variable or item level

Demo: StatCan Metadata

2006 Census of Population [Canada] Public Use Microdata File (PUMF): Individuals File (version 2)

Process Data

Validate and normalize data.

Store your data.

Raw copy

File Naming Conventions

Keep file and folder names short, but meaningful.

Date format should be expressed as YYYYMMDD for easier sort and find.

Name it with more than your name or the document's title.

File names should be descriptive outside their folders.

Version	Create date	Creator	Description	Research team	Publication date	Project no.

by ULeicester, Research Data: Naming files and folders

Versioning

Ordinal numbers for major version changes (i.e. 1, 2, 3).

Decimals for minor changes (i.e. 1.1, 1.2) and fixes (1.1.1, 1.1.2)

Consider version control system (i.e. Git).

Smith_interview_July2010_V1_DRAFT

Lipid-analysis-rate-V2_definitive

2001_01_28_ILB_CS3_V6_AB_edited

[document name][version number][status: draft/final]

Storing Data

There are ethical implications to data storage.

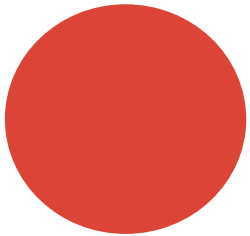
Cloud storage should be used judiciously. (Amazon, Google, Microsoft etc.)

Backing Up Data

Regularly backup data (both on and offsite)

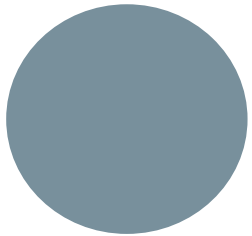
Three copies

original



external
local

external
remote



https://library.uoregon.edu/datamanagement/storage.html

Securing Data

Physical security.

Computer security.

Personal security.

Data must be destroyed and not merely deleted.

Sensitive Data

Anonymization.

Disk encryption, i.e. PGP Desktop, Bitlocker.

Schedule for retaining very confidential files.

Aggregation for sharing.

Preserve Data

To maintain files over time, they may need to be migrated to new formats.

Additionally, data needs to be fully described to achieve long-term access.

Bundled together for long term checking and completeness.

Best File Formats

Commonly used

Non-proprietary

Unencrypted

Share Data

Evidence that sharing may increase citation rate.

Null results are represented.

Ideologically good.

Raise your scholarly profile.

Increase research impact.

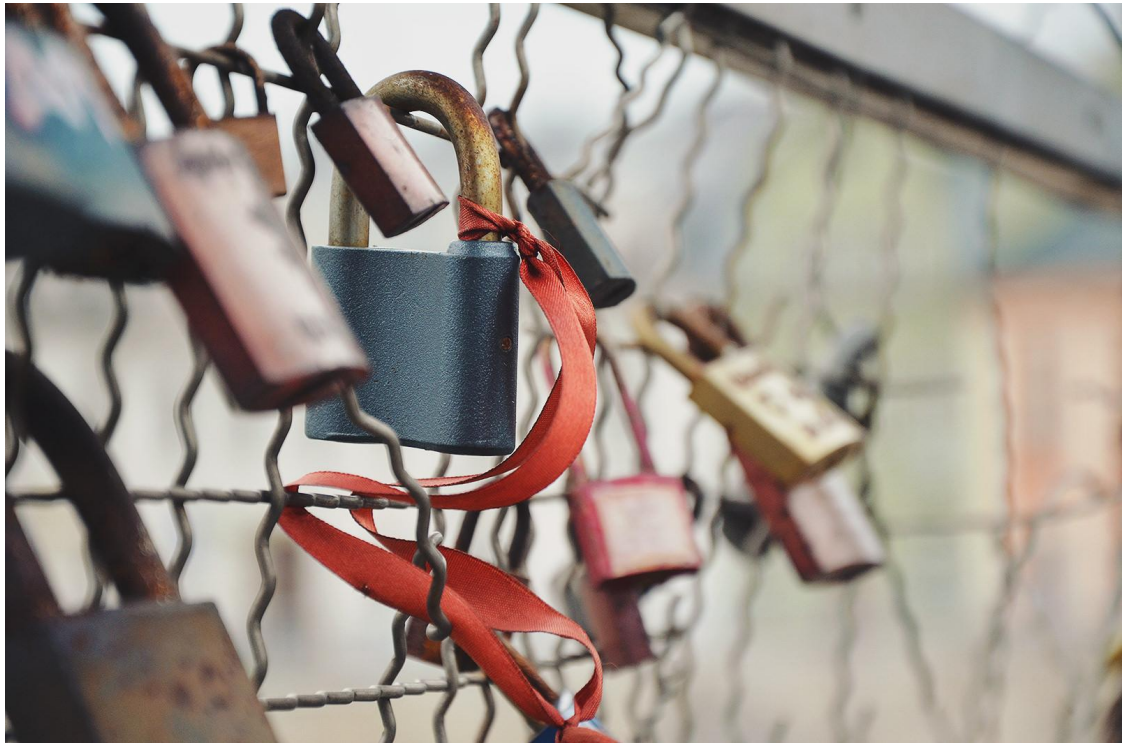
Kinds of Repositories

Institutional vs. subject

Restricted use vs. on demand

Curation level

Persistent identifiers




FRDR

Sharing in FRDR

Open licenses are encouraged.

Datasets can be temporarily embargoed or shared with a small group.

Tri-Agency Statement of Principles on Digital Data management

“The agencies believe that research data collected with the use of public funds belong, to the fullest extent possible, in the public domain and available for reuse by others."

Researchers are responsible for:

- incorporating data management best practices into their research;
- developing data management plans to guide the responsible collection, formatting, preservation and sharing of their data throughout the entire lifecycle of a research project and beyond;
- following the requirements of applicable institutional and/or funding agency policies and professional or disciplinary standards;
- acknowledging and citing datasets that contribute to their research; and
- staying abreast of standards and expectations of their disciplinary community

Challenges to Sharing

There may be legal and ethical limitations to sharing raw data files.

Evaluate your data.

Anonymization can be a solution.

Demo: DataCite & re3data.org

Re-Use Data

Citation is essential.

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

Challenges in sharing sensitive data

Sensitve data

- People or animals
- Generated or used under a commercial research funding agreement
- Potential to have significant negative public impact

Tri-Agency Statement of Principles on Digital Data management

“The agencies believe that research data collected with the use of public funds belong, to the fullest extent possible, in the public domain and available for reuse by others.”

AND

“Researchers should also consider whether any ethical, legal or commercial obligations prohibit sharing or preserving the data, and whether any of the data need to be de-identified or made available with restricted access.”

Ethnography

“The AAA supports the sharing of research data and encourages ethnographers to consider preserving field notes, tapes, videos, etc. as a resource accessible to others for future study. Ethnographers should inform participants of the intent to preserve the data and make it accessible as well as the precautions to be undertaken in the handling of the data.”

Permission for sharing **must** be obtained when participants are consenting to the research.

Managing data: Case study of re-use

- Clarence Gravlee re-used data to revisit landmark anti-eugenics study by Franz Boas (1912).

- While the original works were innovative and carefully done, there was doubt about their methodological soundness due to their age.
  - One method Gravlee used was analysis of variance.

Managing data: Case study of re-use

Current No.		Immi-gration	Age	LH	WH	WF	St	Ci	Wfi	Color	
Fam.	Ind.									Eyes	Hair
442	1496	1877 S	16	182	154	131	155	846	85.1		
443	2577	1877 S	15	177	149	130	156	842	87.3		
	4187	1877 S	13	176	146	127	141.5	830	87.0		
477	496	1906 S	11½	173	149	128	138	861	86.9		
S.	178	1906 S	11½	180	139	114	121	772	80.0		
	179	1906 S	9	168	148	121	116	881	88.7		

Detail of a page of Boas's data in *Materials for the Study of Inheritance in Man.*

In Gravlee et al. 2003. Used with permission of the American Anthropological Association.

- Gravlee and his co-authors (2003, 2005), using modern statistical methods, both substantiated and refined the original findings.
- Boas's reanalyzed data were in raw form; they are now digitized and available online and have been used by other researchers.

Managing Data: Basic Steps

- Think about ways to make data legible and meaningful to others beyond yourself and/or your research group.

- De-identify the data
- Anonymization isn't always enough -- if your surveyed groups are small, re-identifying participants may be possible
- Consider restricted use derivatives

Acknowledgements

Oths, Kathryn. “Cultural Anthropology: Principles and Practices of Digital Data Management.” In *Bringing Digital Data Management Training into Methods Courses for Anthropology*, edited by Blenda Femenías. Arlington, VA: American Anthropological Association, 2016. http://www.americananthro.org/methods

Rice, Robin. “Overcoming obstacles to sharing data about human subjects” Edinburgh, 10 June 2015

Thanks!!!

Questions?

garnett@sfu.ca

adam.mckenzie@usask.ca