Introduction to Research Data Management

Jane Fry, Carleton University
RDM Jumpstart Program
May 12, 2025

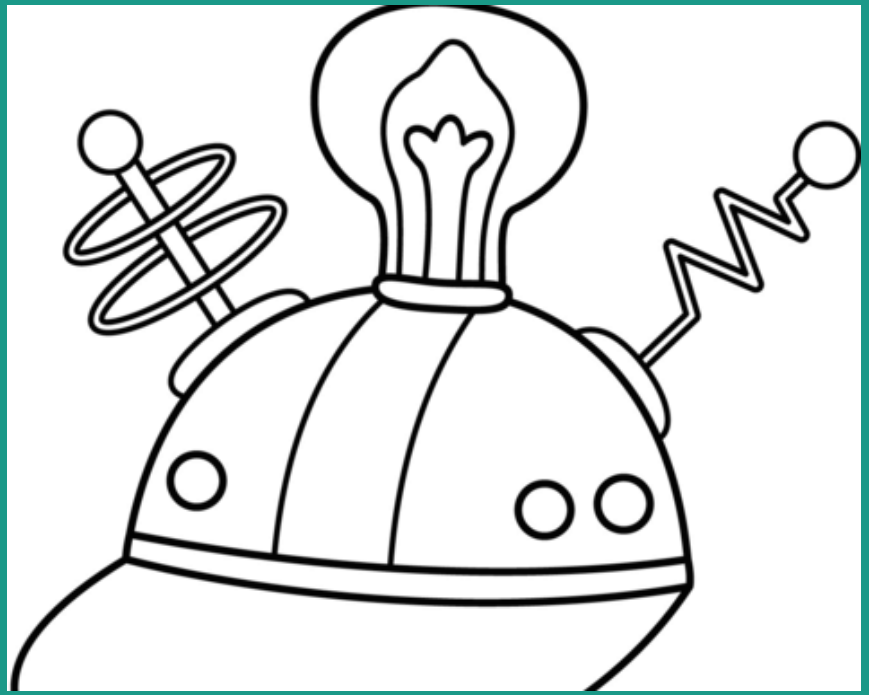
Outline of this session

- Research data
- Metadata
- The research data lifecycle
- Research data management (RDM)
- The importance of RDM

Acronyms

- Research Data Management
  - aka RDM
  - aka Data Management
  - aka DM
- Data Management Plan
  - aka DMP
  - aka Research DMP
  - aka RDMP

Research Data



- What are they?
- Need to know what they are before we start talking about them!

- Thinking caps on!

Source: https://rb.gy/jxyudl

Research Data (cont'd)

"Data are facts, observations or experiences on which an argument or theory is constructed or tested. Data may be numerical, descriptive, aural or visual. Data may be raw, abstracted or analysed, experimental or observational. Data include but are not limited to: laboratory notebooks; field notebooks; primary research data (including research data in hardcopy or in computer readable form); questionnaires; audiotapes; videotapes; models; photographs; films; and test responses. Research collections may include slides; artefacts; specimens; samples."

Source: https://blogs.ucl.ac.uk/rdm/2015/09/what-is-research-data/

Research Data (cont'd)

"Research data are the original sources or material that you have created or collated to conduct your research project. They can be digital or non-digital. The response to your research question is based on the analysis of these research data."

Source: https://blogs.ucl.ac.uk/rdm/2015/09/what-is-research-data/

Research Data (cont'd)

> *"Research data are data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or creative practice, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. Research data may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data, or repurposed data. What is considered relevant research data is often highly contextual, and determining what counts as such should be guided by disciplinary norms."*
>
> (Tri-Agency)

Source: https://tinyurl.com/yx6ehw2r

Research Data (cont'd)

- Why are they important?
  - Support original research
  - Deterrent to research fraud
  - When managed properly, they can be shared
    - Informally from one researcher to another
    - Deposited in a self-archiving system
    - Deposited in a specialist data repository dedicated to archiving, preserving and disseminating digital data
    - Deposited in an institutional repository
- Why is sharing important?
  - We don't conduct research in a bubble!
  - We will want to publish research about our data.

Sharing data considerations

- Meets data-management requirements
- Stability / security over time
- Makes your data more visible
- Requires documentation
- Offers curation expertise that adds value to data
- Facilitates data discovery and reuse through the development and standardization of metadata
- Achieves interoperability across scientific communities
- Authenticated online access to data / user access controls /licensing agreements
- Fully searchable
- Assigns unique persistent DOIs to files or collections

Reference: QDR & UKDS

Making Research Transparent

Requires:

- Data access
  - What data were used, where are they, are they available?
- Production transparency
  - Requires providing *documentation* describing how the data were generated / collected
- Analytic transparency
  - How were data analyzed to arrive at conclusions?
  - How are evidence and claims connected?

Reference: QDR & UKDS

Benefits of Transparency

- Makes research procedures clear
- Facilitates more and deeper collaboration
- Allows scholars to demonstrate the rigor and power of their work
- Acts as an incentive to do good work
- Facilitates learning of methodological lessons and teaching
- Makes it more likely that research will be more useful to others

SUMMING UP: more accessible, honest, rigorous, relevant, and useful research!

Reference: QDR & UKDS

Example …

- Political Persuasion and Attitude Change Study: The Los Angeles Longitudinal Field Experiments, 2013-2014
- Principal Investigator: Michael J. LaCour

Reference: https://www.openicpsr.org/openicpsr/project/100037/version/V8/view

Irregularities in LaCour (2014)

David Broockman, Assistant Professor, Stanford GSB (as of July 1),
dbroockman@stanford.edu
Joshua Kalla, Graduate Student, UC Berkeley, kalla@berkeley.edu
Peter Aronow, Assistant Professor, Yale University, peter.aronow@yale.edu

May 19, 2015

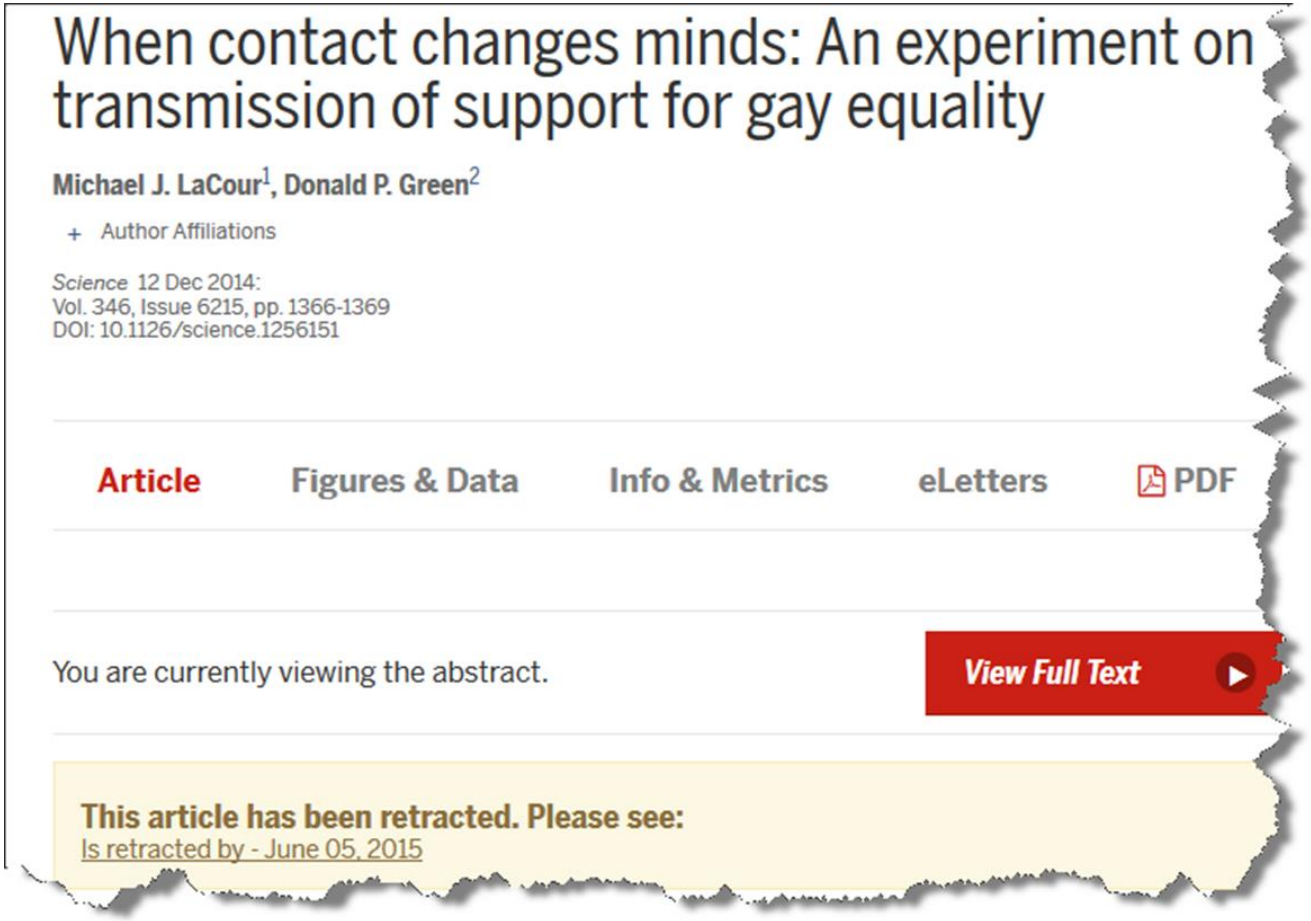
Summary

We report a number of irregularities in the replication dataset posted for LaCour and Green (*Science*, "When contact changes minds: An experiment on transmission of support for gay equality," 2014) that jointly suggest the dataset (LaCour 2014) was not collected as described. These irregularities include baseline outcome data that is statistically indistinguishable from a national survey and over-time changes that are unusually small and indistinguishable from perfectly normally distributed noise. Other elements of the dataset are inconsistent with patterns typical in randomized experiments and survey responses and/or inconsistent with the claimed design of the study. A straightforward procedure may generate these anomalies nearly exactly: for both studies reported in the paper, a random sample of the 2012 Cooperative Campaign Analysis Project (CCAP) form the baseline data and normally distributed noise are added to simulate follow-up waves.

Timeline of Disclosure

- January - April, 2015. Broockman and Kalla were impressed by LaCour and Green (2014) and wanted to extend the article's methodological and substantive discoveries. We began to plan an extension. We sought to form our priors about several design parameters based on the patterns in the original data on which the paper was based, LaCour (2014). As we examined the study's data in planning our own studies, two features surprised us: voters' survey responses exhibit much higher test-retest reliabilities than we have observed in any other panel survey data, and the response and reinterview rates of the panel survey were significantly higher than we expected. We set aside our doubts about the study and awaited the launch of our pilot extension to see if we could manage the same parameters. LaCour and Green were both responsive to requests for advice about design details when queried.
- May 6, 2015. Broockman and Kalla launch a pilot of the extension study.
- M[illegible]

Reference: http://stanford.io/2bzRWFo



When contact changes minds: An experiment on transmission of support for gay equality

Michael J. LaCour[1], Donald P. Green[2]

+ Author Affiliations

Science 12 Dec 2014:
Vol. 346, Issue 6215, pp. 1366-1369
DOI: 10.1126/science.1256151

Article | Figures & Data | Info & Metrics | eLetters | PDF

You are currently viewing the abstract.

View Full Text

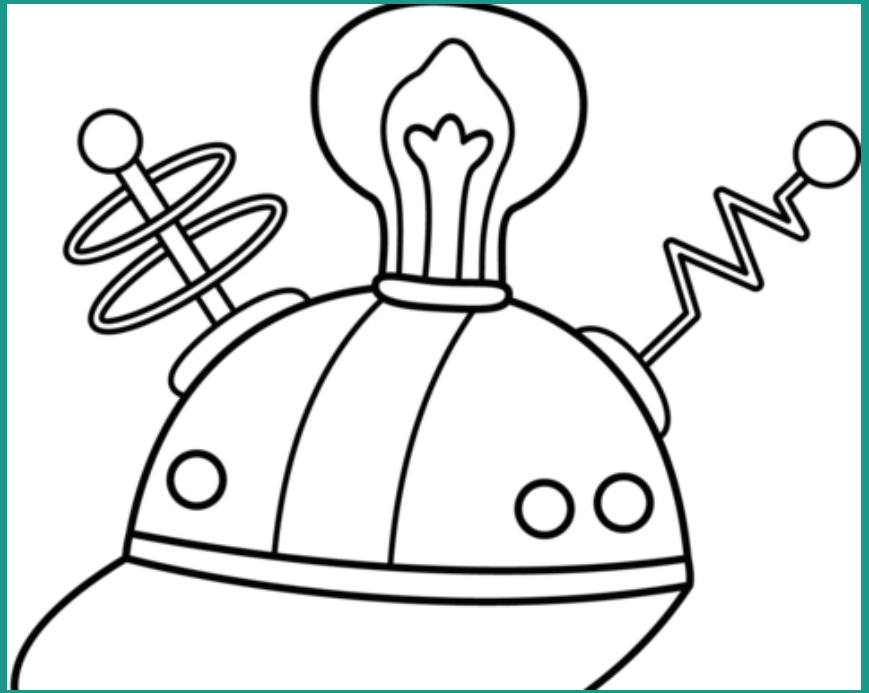
This article has been retracted. Please see:
Is retracted by - June 05, 2015

Reference: http://bit.ly/1NxWG5M

Something to think about …

- Check out this story when you get a chance.
- It will open your eyes about fake data and who is faking it.
- When you have time …
  - Retraction Watch

Metadata



- What are they?
- Need to know what they are before we start talking about them!

- Thinking caps on!

Source: https://rb.gy/jxyudl


Unico
Pâte de tomates
COLA
COLA
CLASSICO
pizza
sauce
à pizza
Lindt
HAZELNUT · NOISETTE
Flocons
de jambon
156 g

Metadata (cont'd)

- What is it?
- Explains …
- Why is it important?
- Who enters it?

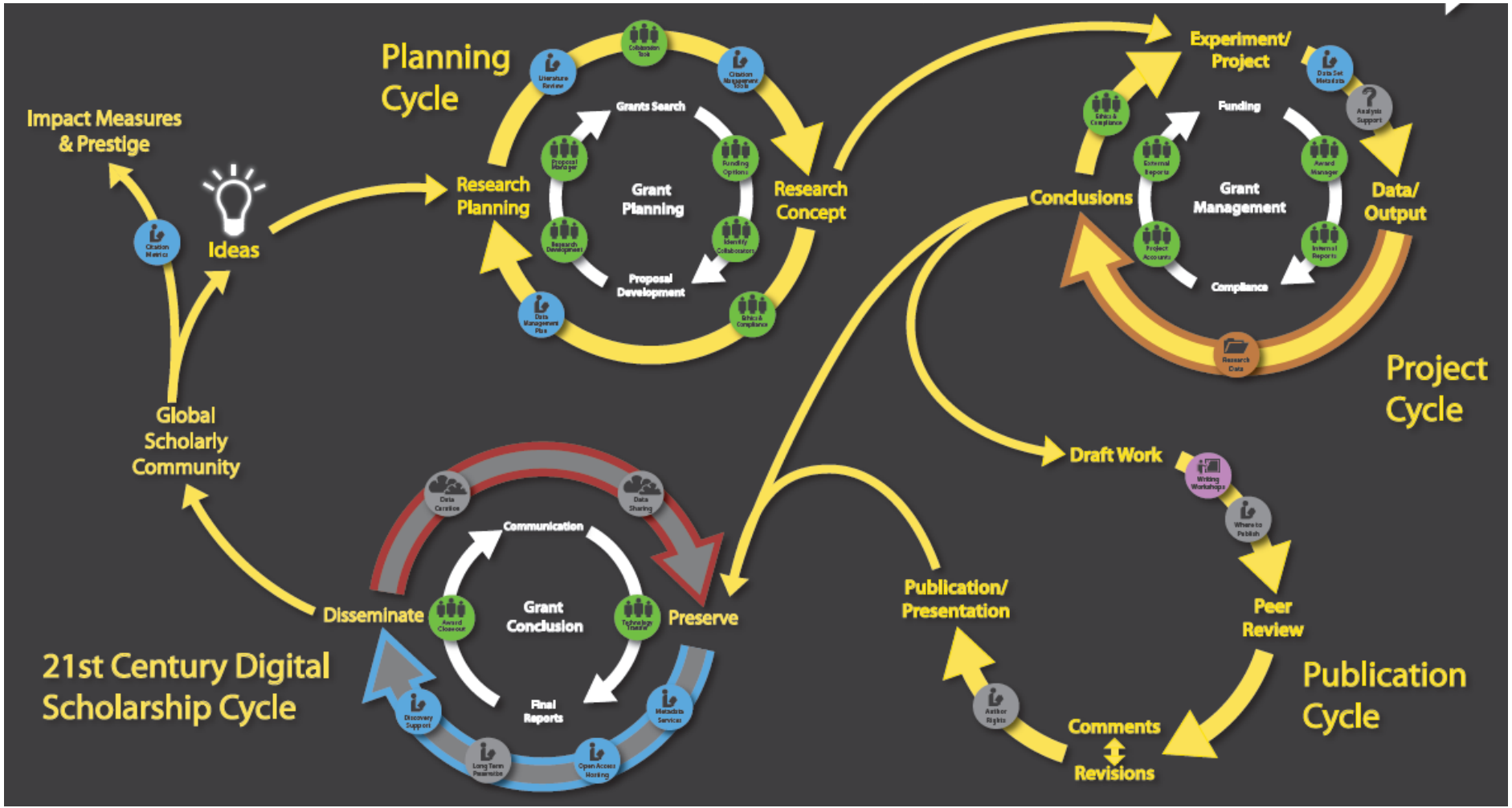
Metadata (cont'd)

- Why keep metadata?
- When to record it?
- What to keep?
- Using a readme file
- End goal …

Metadata (cont'd)

- Survey metadata
  - Questionnaire
  - Data collection
  - Interviewer instructions
  - …

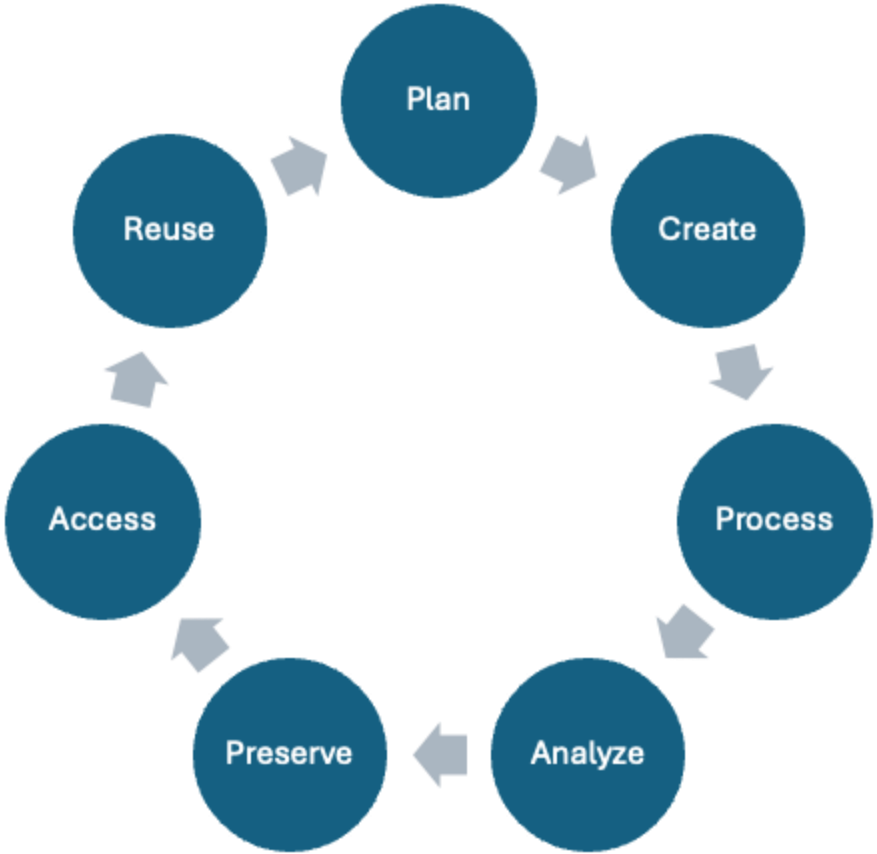
Questions?


Impact Measures & Prestige
Ideas
Global Scholarly Community
Planning Cycle
Research Planning
Grant Planning
Grants Search
Proposal Development
Research Concept
Experiment/ Project
Grant Management
Funding
Compliance
Conclusions
Data/ Output
Project Cycle
Draft Work
Peer Review
Comments
Revisions
Publication/ Presentation
Publication Cycle
Preserve
Grant Conclusion
Communication
Final Reports
Disseminate
21st Century Digital Scholarship Cycle

The Research Data Life Cycle

- Many out there!
- Many elements in common
- Can be discipline-specific

- The previous example is from the University of Centra Florida Libraries
- We will NOT be using that one!


Plan
Create
Process
Analyze
Preserve
Access
Reuse

Plan

- Data Management Plan
  - More on that later!

Create

- Design research
- Locate existing data
- Collect data (experiment, observe, measure, simulate)
- Capture and create metadata

Process

- Enter data, digitise, transcribe, translate
- Check, validate, clean data
- Anonymise data where necessary
- Describe data
- Manage and store data

Analyse

- Interpret data
- Derive data
- Produce research outputs
- Think about Author publications
- Prepare data for preservation

Preserve

- Migrate data to best format
  - Non-proprietary best
- Back-up and store data
  - 3-2-1
- Create metadata and documentation
- Archive data

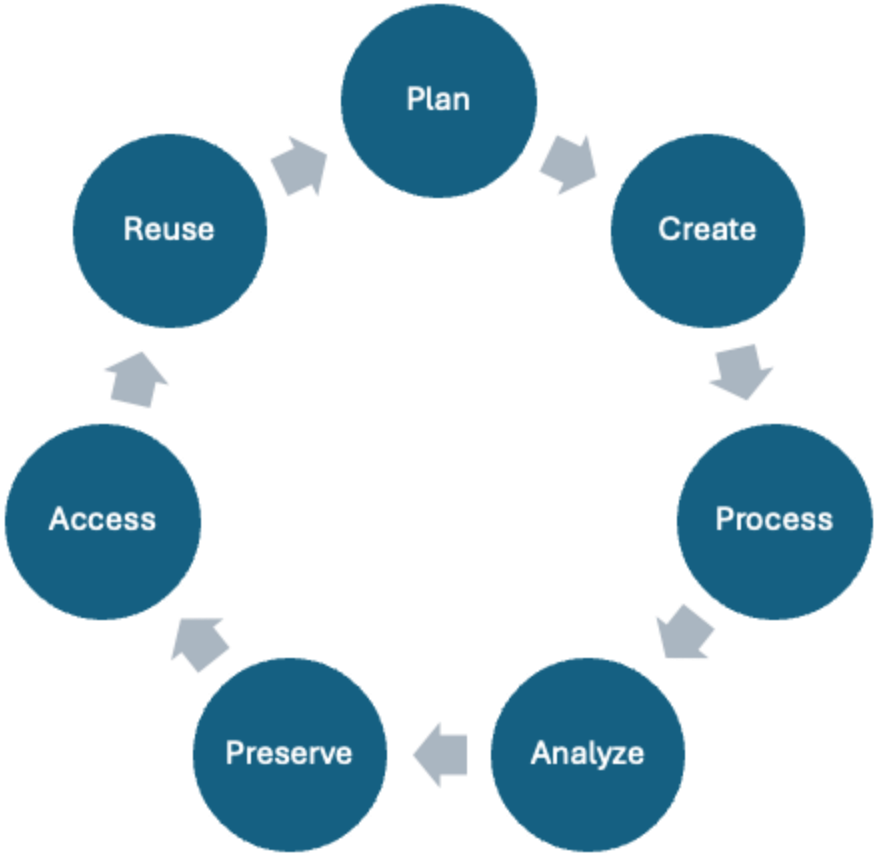
Source: https://www.datarecoverygroup.com/articles/types-data-loss

Access

- Determine what the access will be
- Distribute data
- Share data
- Promote data

Reuse

- Follow-up research
- New research
  - Scrutinize findings
- Teaching and learning


Plan
Create
Process
Analyze
Preserve
Access
Reuse

FAIR

An international guiding principle focused towards making data

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

Key Resource: go-fair.org
https://www.go-fair.org/fair-principles/

FAIR (cont'd)

- **Findable**
  - Data & supplementary materials have sufficiently rich metadata, and a unique and persistent identifier.
- **Accessible**
  - Metadata & data are understandable to humans and machines. Data is deposited in a trusted repository.
- **Interoperable**
  - Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- **Reusable**
  - Data & collections have clear usage licenses and provide accurate information on provenance.

Questions?

RDM

- We have been talking RDM since the beginning!
- What is RDM?
  - “It encompasses the processes applied throughout the lifecycle of a research project to guide the collection, documentation, storage, sharing, and preservation of research data, and allows researchers to find and access data.”
    - Source: https://alliancecan.ca/en/services/research-data-management
- Includes:
  - Sound practices; data curation; and data stewardship

Benefits of RDM

- Confirmation of original findings
- Further research
- Planning follow-up studies
- Bonus …

Tri-Council RDM Policy

- Finalized in 2021
  - After many years of consultation, revision ,,,
- Policy Objective
  - “to support Canadian research excellence by promoting sound data management and data stewardship practices. This policy is not an open data policy.”

Source: http://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html

Tri-Council RDM Policy (cont'd)

- to promote "the importance of data management to researchers, staff & students"
- to guide "researchers on how to properly manage data"
- to provide, or support "access to, repository services or other platforms that securely preserve, curate and provide appropriate access to research data"

Source: http://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html

Tri-Council RDM Policy (cont'd)

Created around 3 pillars

- Creating an institutional RDM strategy
  - by March 1, 2023
  - 220 research institutions have complied
- Data Management plans starting in spring 2022
  - Being rolled out slowly
  - Will eventually be a requirement for researchers for all funding received
- Data deposit requirement for researchers
  - Starting Fall 2025
  - Being rolled out slowly

Source: http://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html

And …

What could happen if you don’t practice good RDM?

https://www.youtube.com/watch?v=N2zK3sAtr-4#t=17

Summing it up …

- Hopefully you have learned
  - What is research data
  - What is metadata
  - What is RDM
  - Why is it important
  - RDM best practices

Questions?

Please log into OSF during the break!

References

Qualitative Data Repository & UK Data Services (2016). “Managing and Sharing Data and Achieving Research Transparency”. Syracuse University.
https://slideplayer.com/slide/12627589/

Retraction Watch
https://retractionwatch.com/

Tri-Agency Statement of Principles on Digital Data Management
https://science.gc.ca/site/science/en/interagency-research-funding/policies-and- guidelines/research-data-management/tri-agency-statement-principles-digital-data-management

Tri-Agency Statement of Principles on Digital Data Management
https://science.gc.ca/site/science/en/interagency-research-funding/policies-and-guidelines/research-data-management

More Resources

- Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide". August 2015. C. Tenopir et al. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0134826
- "ReadMe Best Practices" https://ubc-library-rc.github.io/rdm/content/03_create_readme.html
- "Store and back up" https://ubc-library-rc.github.io/rdm/content/06-3_Storage_Backup.html
- "The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data". A. Pienta, G. Alter, J. Lyle., November 2010. http://hdl.handle.net/2027.42/78307

More Resources

- Care Principles for Indigenous Data Governance
  https://www.gida-global.org/care
- OCAP®, The First Nations Principles
  https://fnigc.ca/ocap-training/