Research Data Management?

DATA 3101

Elizabeth Stregger & Doug Campbell

September 13, 2022

Setup / Reminders

- If you haven't already, go to Moodle and fill out the Student Information Survey. We need your Git username to set up your assignment repository.
- Accept invitation to join MtAData3101 Organization
- If you have any questions about installing and setting up Git, R, and RStudio, request an appointment with Elizabeth Stregger: https://mta.libcal.com/appointments/stregger (link also in syllabus)
- Recordings are available in Microsoft Teams: DATA 3101
- Presentations are uploaded in PDF format to Git

Plan for today

- Data & research data definitions
- Retraction Watch example
- Research data and research lifecycles
- Retraction Watch in-class activity, discussion

What are data?

Data are facts, measurements, recordings, records, or observations about the world collected by researchers and others, with a minimum of contextual interpretation. Data may be in any format or medium taking the form of writings, notes, numbers, symbols, text, images, films, video, sound recordings, pictorial reproductions, drawings, designs or other graphical representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing algorithms, or statistical records.

(Tri-Agency Research Data Management Policy FAQ)

What are research data?

Research data are data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, 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. All other digital and non-digital content have the potential of becoming research data. Research data may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data, or repurposed data.

(Tri-Agency Research Data Management Policy FAQ)

What is driving this conversation?

- Digital data is shareable but fragile
- Funder policies: accountability
  - Publicly funded research
  - Preventing data loss
  - Making data available for reuse
- Journal policies
- Reproducibility issues

Retraction Watch: One way students have contributed

- Making error detection easier – and more automated: A guest post from the co-developer of “statcheck”: https://retractionwatch.com/2015/11/17/making-it-easier-and-more-automated-to-find-errors-a-guest-post-from-the-co-developer-of-statcheck/
- R package written by graduate students
- See statcheck on Github: https://github.com/MicheleNuijten/statcheck

Retraction Watch: Example of problems identified using statcheck

- Prominent behavioral scientist's paper earns an expression of concern: https://retractionwatch.com/2021/07/29/prominent-behavioral-scientists-paper-earns-an-expression-of-concern/
- Expression of concern: Effort for Payment; A Tale of Two Markets https://journals.sagepub.com/doi/10.1177/0956797621103578 2

Retraction Watch: Criticism

- Psychological society wants end to posting error-finding algorithm results publicly: https://retractionwatch.com/2016/10/25/psychological-society-wants-end-to-posting-error-finding-algorithm-results-publicly/

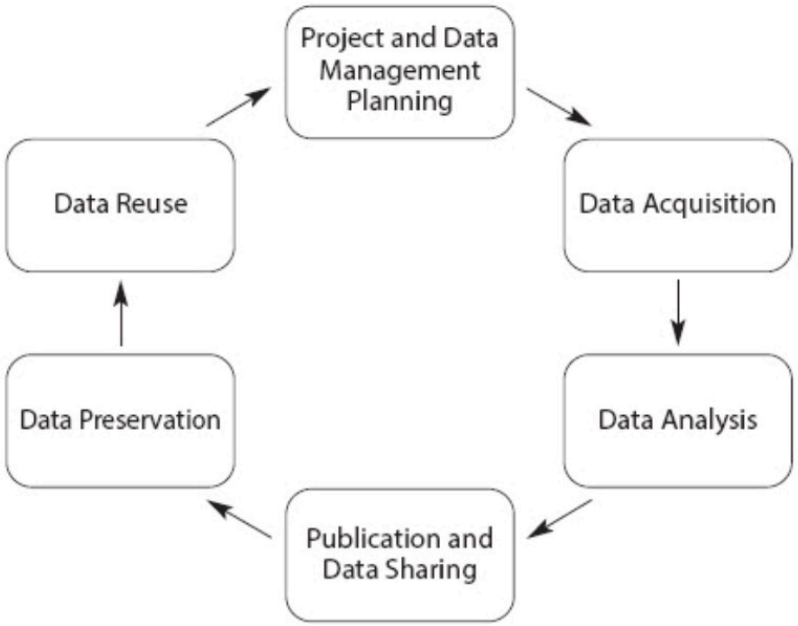
Tri-Agency Research Data Management Policy

Canadian Institutes of Health Research (CIHR), Natural Sciences and Engineering Research Council of Canada (NSERC),Social Sciences and Humanities Research Council of Canada (SSHRC) form the Tri-Agency

Tri-Agency Research Data Management (RDM) Policy (2021)

- Institutional Strategy
- Data Management Plans
- Data Deposit
- https://www.science.gc.ca/eic/site/063.nsf/eng/h_547652FB.html

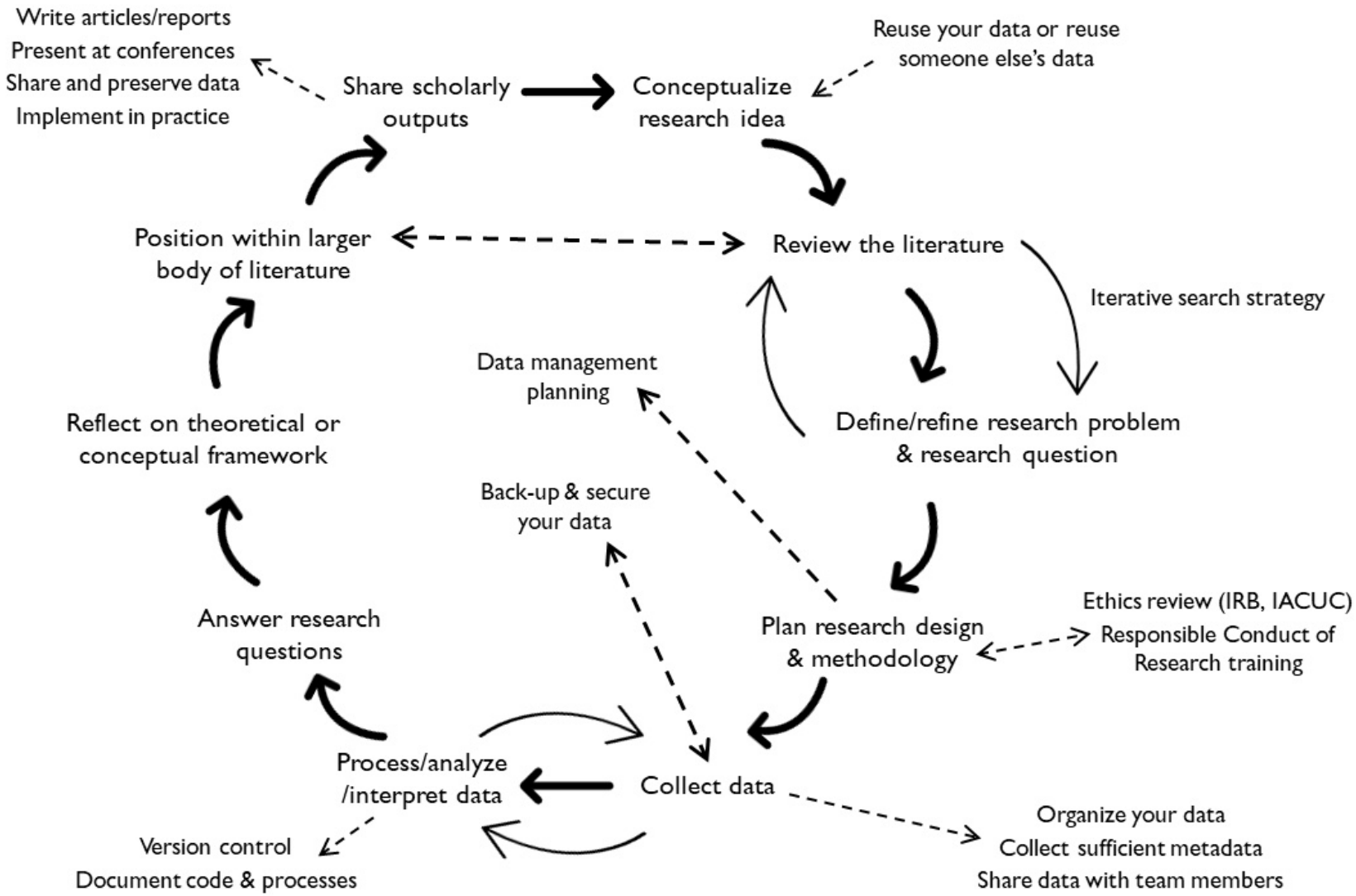
Research Data Lifecycle & Curation Goals



(Briney, 2015, p. 11)

"Data curation is the active management of research data throughout a project to produce datasets that are FAIR: Findable, Accessible, Interoperable, and Reusable."

(Portage)



Check-in Question

Retraction Watch example & research data lifecyle

- Where could Heyman & Ariely (2004) have improved their research process?
- How many articles cited Heyman & Ariely? Where could you find this information?

In-class activity: Part 1: Search & Describe

- Choose a retracted article from the Retraction Watch Database, using the filter “Concerns about data” or “Error in data” or “Unreliable data”
- Note: examples that don’t include falsified data are usually more interesting!
- Read the Retraction Watch blog article (if available), and the retraction notice (if available)
- In your own words, write a description of the problem with the data. Which step of the research data lifecycle would you identify as the place where these researchers had problems?

Introduce Mta Data3101 Organization on Git

https://github.com/MtADATA3101

- Includes course content
- Each of you will have a personal repository for your assignments. Over the term, this will build up to become your portfolio of work.
- We will give you feedback on your work and you can update it based on this feedback.

In-class activity: Part 2: Push to GitHub

- Create a new version control project in RStudio, linked to your personal repository
- Go to File – New File – R Markdown
- Give the file a human readable filename (can have spaces here)
- Add the text you created in Part 1 to the bottom of the bottom of the default text
- Save with a machine-readable filename (no spaces!)
- Go to Settings – Output Options – Advanced
- Check ”Keep markdown source file”
- Knit
- In the Git pane, stage the new folder, commit, and push