Introduction to Research Data Management

Jane Fry
MacOdrum Library
March 22, 2022
Agenda

- Why the Library
- Define key components of RDM
- Why does RDM matter to me
- What is an RDMP
How to find Data Services

Source: https://library.carleton.ca/
Data

Need Canadian data?

Need international data? Start here

Search ICPSR or access ICPSR resources.

Research Data Management: Manage your data

Dataverse: Store your data

Source: https://library.carleton.ca/find/data
Acronyms

- **Research data management**
  - aka RDM
  - aka data management
    - *aka DM*

- **Research data management plan**
  - aka RDMP
  - aka data management plan
    - *aka DMP*
Why the Library?

- Research partner
- Support the research endeavor
- RDM expert
- Partner with CU Research Office
- The scholarly life-cycle
- Discipline-agnostic
Our role
- Information
- Consultation

Challenge
- Determine how we can help researchers advance their research

References: Rambo Neil; Shorish, Yasmeen
What are they?

Need to know what they are before we can start talking about RDM
“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/
“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/
Why are research data important?

Sharing research data

Check out the following examples …
Political Persuasion and Attitude Change Study: The Los Angeles Longitudinal Field Experiments, 2013-2014

Principal Investigator(s):
  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 19, 2015. Our work is in progress as follows: we completed a grand total of the

Reference: http://stanford.io/2bzRWFo
Reference: http://bit.ly/1NxWG5M
“New Study Links Vaccines To Autism. There's Just One Tiny Problem With It”

“... one of its own co-authors claimed that figures in the paper were deliberately altered before publication. The data had been tampered with. ...”

Source: http://bit.ly/2zSwAxo
“Researchers from the University of British Columbia are retracting their scientific paper linking aluminum in vaccines to autism in mice, because one of the co-authors claims figures published in the study were deliberately altered before publication — an issue he says he realized after allegations of data manipulation surfaced online.”

“…original data cited in the study is inaccessible, which would be a contravention of the university's policy around scientific research.”

“…the original data is in China, with an analyst who worked on the paper.”

(October 16, 2017)

Source: https://bit.ly/2kSjMRJ
RETRACTED: Subcutaneous injections of aluminum at vaccine adjuvant levels activate innate immune genes in mouse brain that are homologous with biomarkers of autism

Dan Li a, Lucija Tomljenovic a, Yongling Li a, Christopher A. Shaw a, b, c, d, e

Show more

https://doi.org/10.1016/j.jinorgbio.2017.08.035

Under a Creative Commons license
“A top Cornell food researcher has had 13 studies retracted. That’s a lot.”

- September 21, 2018
- Brian Wansink
- "committed academic misconduct,"
- "he would retire from the university on June 30, 2019"
- "has been removed from all teaching and research,"
- "will spend his remaining time … in an “ongoing review of his prior research.”"

Source: https://bit.ly/2xocljs
Wansink refuted these findings. “There was no fraud, no intentional misreporting, no plagiarism, or no misappropriation,” he wrote. “I believe all of my findings will be either supported, extended, or modified by other research groups.”

“In a press release, JAMA said Cornell couldn’t “provide assurances regarding the scientific validity of the 6 studies” because they didn’t have access to Wansink’s original data. So, Wansink’s ideas aren’t necessarily wrong, but he didn’t provide credible evidence for them.”

(JAMA = Journal of American Medical Assoc.)

JAMA Network Retracts 6 Articles That Included Dr. Brian Wansink as Author

FOR IMMEDIATE RELEASE: SEPTEMBER 19, 2018

Media advisory: To contact JAMA Network Media Relations email mediarelations@jamanetwork.org.

CHICAGO – JAMA, JAMA Internal Medicine and JAMA Pediatrics have retracted six articles that included Brian Wansink, Ph.D., of Cornell University, Ithaca, New York, as author. Below is the notice of retraction published online today by JAMA, which references the retracted articles (see references 4-9). Similar notices were published online today in JAMA Internal Medicine and JAMA Pediatrics.


Howard Bauchner, MD

Source: https://bit.ly/2OyCH1N
Missing and Murdered Indigenous Women and Girls Commission, 2019

CBC, Sept 3, 2019

“…the commission left out a word that had appeared in the original Statistics Canada figure.”

"Indigenous women and girls now make up almost 25 per cent of homicide victims" should have referred to their percentage share of female homicide victims …”

“In some cases, the inquiry report's footnotes cite government reports that do not support the footnoted statements.”

“MMIWG final report quietly altered after CBC inquired about errors”

“Report contained errors that were fixed online, but allowed to remain on the official record.”

“Tracking retractions as a window into the scientific process”

https://retractionwatch.com/

COVID-19

- Retracted Papers - 214
- Expressions of concern - 10
Metadata

- What is it?
▪ What is it

▪ Explains …

▪ Why is it important

▪ Who enters it
Why keep metadata
- Researchers re-use data
- Good research practice

When to record it

What to keep

End goal
Metadata (cont’d)

- Survey metadata
  - Questionnaire
  - Data collection
  - Interviewer instructions
  - ???
ICPSR 101: What are Metadata (and why are they so important)?
UKDA RDM Lifecycle

- Creating Data
- Processing Data
- Analysing Data
- Preserving Data
- Giving Access to Data
- Re-Using Data

Source
Creating data

- Design research
- Plan data management (formats, storage, etc.)
- Plan consent for sharing
- Locate existing data
- Collect data (experiment, observe, measure, simulate)
- Capture and create metadata

http://www.data-archive.ac.uk/create-manage/life-cycle
Processing data

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

http://www.data-archive.ac.uk/create-manage/life-cycle
Analysing data

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

http://www.data-archive.ac.uk/create-manage/life-cycle
Preserving data

- Migrate data to best format
- Migrate data to suitable medium
- Back-up and store data
- Create metadata and documentation
- Archive data

http://www.data-archive.ac.uk/create-manage/life-cycle
Giving access to data

- Distribute data
- Share data
- Control access
- Establish copyright
- Promote data

http://www.data-archive.ac.uk/create-manage/life-cycle
Re-using data

- Follow-up research
- New research
- Undertake research reviews
- Scrutinise findings
- Teach and learn

http://www.data-archive.ac.uk/create-manage/life-cycle
A set of guiding principles focused towards making data
- Findable
- Accessible
- Interoperable
- Reusable

Key Resource:
go-fair.org
https://www.go-fair.org/fair-principles/
▪ **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.
A set of guiding principles for Indigenous data governance

- Collective benefit
- Authority to control
- Responsibility
- Ethics

Resource:
https://www.gida-global.org/care
A set of standards that establish how First Nations data should be collected, protected, used or shared.

- **Ownership**
- **Control**
- **Access**
- **Possession**

**Resource:**

https://fnigc.ca/ocap-training/
Questions are guaranteed in life; Answers aren't.
What is RDM?

“...describes the activities researchers perform as they create and save their research data.”

- Source: http://researchdata.library.ubc.ca/learn/

We have been talking RDM since the beginning!

Includes

- Sound practices
- Data curation
- Data stewardship
Benefits of RDM

- Confirmation of original findings
- Further research
- Planning follow-up studies
- Bonus …
What’s next?

- Need an DMP
- Why an DMP
  - Safety
  - Efficiency
  - Quality
- If no DMP?
  - Potential problems
Why should I care?

- Requirement by funders
  - Tri-Council (SSHRC, CIHR and NSERC)
  - CFI
  - Genome Canada

- Tri-Agency Statement of Principles on Digital Data Management
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
▪ 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
Data Management Plans

“All grant proposals submitted to the agencies should include methodologies that reflect best practices in research data management. The agencies encourage grant applicants to complete data management plans (DMPs) as an essential step in research project design. For specific funding opportunities, the agencies may require DMPs to be submitted to the appropriate agency at time of application; in these cases, the DMPs may be considered in the adjudication process.”

Source: http://www.science.gc.ca/eic/site/063.nsf/eng/h__97610.html
“The content and length of DMPs depend on the research project. Generally, DMPs describe:

- how data will be collected, documented, formatted, protected and preserved;
- how existing datasets will be used and what new data will be created over the course of the research project;
- whether and how data will be shared; and
- where data will be deposited.”

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

Source: https://assistant.portagenetwork.ca/
Key features

- Bilingual
- Web-based, free
- Supports cross-institutional collaboration
- Endorsed by Tri-Agencies and other funding agencies
- Templates
- Keeps a list of all of your DMPs
- You determine DMP access
- Ability to edit/update whenever necessary
Portage DMP Assistant

- Data Collection
- Documentation and Metadata
- Storage and Backup
- Preservation
- Sharing and Re-use
- Responsibilities and Resources
- Ethics and Legal Compliance
Data collection

- Types of data
- File formats
- Conventions and procedures
Documentation and metadata

- Documentation
- Consistency
- Metadata standard and tools
Storage and Backup

- Storage requirements
- Storage and backup
- Access to data during the project
Preservation

- What data
- Where will you deposit your data
- Is it preservation ready
Sharing and reuse

- What data
- How
- End-user license
- Promotion
Responsibilities and resources

- Who
- How to handle change
- Resources
Ethics and legal compliance

- Sensitive data
  - Primary use
  - Secondary use

- Legal, ethical and IP issues
- Anyone
- Step-by-step
- The length
- Different agencies
- Remember …
- **Downloadable in Word or pdf**
  - [https://assistant.portagenetwork.ca/public_templates](https://assistant.portagenetwork.ca/public_templates)
  - Same questions as the DMP Assistant
  - Living document
  - Filling it in easy
    - *Request a consultation*
    - *Fill it in and then email for feedback*
PORTAGE NETWORK: PORTAGE TEMPLATE

DATA COLLECTION

What types of data will you collect, create, link to, acquire and/or record?

Guidance:
Examples: numeric, images, audio, video, text, tabular data, modeling data, spatial data, instrumentation data.

What file formats will your data be collected in? Will these formats allow for data re-use, sharing and long-term access to the data?

Guidance:
Proprietary file formats requiring specialized software or hardware to use are not recommended, but may be necessary for certain data collection or analysis methods. Using open file formats or industry-standard formats (e.g. those widely used by a given community) is preferred whenever possible.
Read more about file formats: UBC Library or UK Data Service.

What conventions and procedures will you use to structure, name and version-control your files to help you and others better understand how your data are organized?

Guidance:
It is important to keep track of different copies or versions of files, files held in different formats or locations, and information cross-referenced between files. This process is called 'version control'. Logical file structures, informative naming conventions, and clear indications of file versions, all contribute to better use of your data during and after your research project. These practices will help ensure that you and your research team are using the appropriate version of your data, and minimize confusion regarding copies on different computers and/or on different media.
Read more about file naming and version control: UBC Library or UK Data Service.
Where to deposit my data?

- Carleton U Data Repository
  - [https://library.carleton.ca/services/dataverse](https://library.carleton.ca/services/dataverse)
  - Easy to use
  - Provides a DOI
  - You choose who to share with
  - We will help you get started

The Dataverse Project
Some final tips

- Mark it down!
- No blank questions
- Use plain English
- It is not written in stone!

- Easy!
Still don’t believe me?

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

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

- **Help with RDM**
  - [https://library.carleton.ca/services/research-data-management](https://library.carleton.ca/services/research-data-management)
  - Consultations

- **Help with RDMPs**
  - Portage:
    - [https://assistant.portagenetwork.ca/](https://assistant.portagenetwork.ca/)
  - DMP Templates
    - [https://assistant.portagenetwork.ca/public_templates](https://assistant.portagenetwork.ca/public_templates)
Need help with data literacy?
- Check out this data literacy learning catalogue
- Short videos, …
  - *What you need to know for the data journey*
  - *Data terminology and concepts*
  - *Understanding and exploring data*
  - *Gathering data*
  - *Analyzing your data*
  - *Telling the story*
  - …

Research Computing Services

- [https://carleton.ca/rcs/](https://carleton.ca/rcs/)
- “to provide researchers with the support, expertise and resources necessary to conduct research computing at Carleton”
- Separate from Carleton’s ITS department
- Access to cloud computing resources
  - [https://carleton.ca/rcs/rcdc/](https://carleton.ca/rcs/rcdc/)
- Access to filesharing platform
  - [https://carleton.ca/rcs/sharefile/](https://carleton.ca/rcs/sharefile/)
- Check out their other services!
In summary

- You are now able to:
  - Define the key components of RDM
  - Know why it matters to you
  - Define an DMP
  - Create an DMP
- **RDM at Carleton**
  [https://library.carleton.ca/services/research-data-management](https://library.carleton.ca/services/research-data-management)

- **Portage DMP Assistant**
  [https://portagenetwork.ca/](https://portagenetwork.ca/)

- **Research Data Lifecycle (UK Data Archive)**
  [http://www.data-archive.ac.uk/create-manage/life-cycle](http://www.data-archive.ac.uk/create-manage/life-cycle)

- **Tri-Agency Statement of Principles on Digital Data Management**
References, plus


  http://acrl.ala.org/techconnect/post/the-library-as-research-partner

- Retraction Watch  
  https://retractionwatch.com/
Sometimes the questions are complicated and the answers are simple.
Thank you!

Contact Information:

Jane Fry
Data Services Librarian
MacOdrum Library
jane.fry@Carleton.ca

http://www.library.carleton.ca/find/data