RESEARCH DATA MANAGEMENT

JOHN BROSZ, Visualization
jdlbrosz@ucalgary.ca

HEATHER GANSHORN, Librarian
Heather.Ganshorn@ucalgary.ca
Data Sharing and Management Snafu, [https://youtu.be/N2zK3sAtr-4](https://youtu.be/N2zK3sAtr-4)
OUTLINE

• Introductions
• What is research data management?
• What do I need to do to comply with funding requirements?
• Data Management Planning with the DMP Assistant
WHAT BRINGS YOU HERE TODAY?

• Discipline
• What are you hoping to get out of this session?
• What are your current data management practices?
WHAT ARE RESEARCH DATA?

- Pieces of (digital) information, structured through methodology for the purpose of producing new knowledge
- AKA research material
- Can include:
  - Images
  - Email, chat logs
  - Sound or video recordings
  - Presentations
  - Document data
  - Oral history & interviews
  - News stories
  - Bibliography
  - Measurements & statistics
  - 3D models, virtual environments, simulations
  - Software, source code
  - Transcripts
  - Procedures & workflows
RESEARCH DATA MANAGEMENT

• Application of standard practices towards the creation and analysis of data for current and future purposes
Why Research Data Management?

• Save your time
• Preserve your data
• Maintain data integrity (versions)
• Meet requirements
• Easier collaboration
• Promote new discoveries (get more citations)
• Provide evidence
  • Increase reproducibility & verifiability
Research Data Management

Before
- Using other’s data
  - Finding data
  - Copyright
  - Data licensing
- Funding guidelines
- Data management plan
- Ethics

During
- Documentation & metadata
- Storage & data protection
- Managing sensitive data
- Collaboration

After
- Sharing data
- Data repositories
- Preservation
- Data citation
DATA MANAGEMENT PLANS (DMPs)

• Short, living document
• Describe how you will organize, store, and share your research data through the lifetime of the research
• Does not contain data
DMP POLICIES IN CANADA:

- Publication-related Research Data
- For recipients of CIHR funding:
  - Deposit bioinformatics, atomic, and molecular coordinate data into the appropriate public database immediately upon publication of research results.
  - Retain original data sets for a minimum of five years after the end of the grant (published or not).
- Policy-mandated requirement
2016: TRI-AGENCY STATEMENT OF PRINCIPLES ON DIGITAL DATA MANAGEMENT

Statement objective:

to *promote excellence* in digital data management practices and data stewardship in agency-funded research.

It complements and builds upon existing agency policies, and *serves as a guide* to assist researchers, research communities and research institutions in adhering to the agencies’ current and future research data management requirements.
TRI-AGENCY STATEMENT OF PRINCIPLES ON DIGITAL DATA MANAGEMENT

Includes statements of responsibilities for:

• Researchers
• Research communities
• Research institutions
• Research funders
2018: DRAFT TRI-AGENCY RESEARCH DATA MANAGEMENT POLICY FOR CONSULTATION

• Released this summer for feedback
• Includes requirements for:
  • Institutions: Create an institutional RDM strategy
  • Researchers: Create a DMP for inclusion with grant submissions
  • Researchers: Deposit research data
• Implementation to begin incrementally
WORLD WIDE

● In the US, Data management policies and requirements for DMP have been introduced by NIH (2001, 2003), NSF (2011), and more to come for federally funded agencies.
● In the UK, “Common Principles on Data Policy” issued by Research Council UK (2011), and most UK funders now require a DMP
● In the EU, a pilot action on data management requirements within the Horizon 2020 framework. The approach is “as open as possible, as closed as necessary.”

Source: Comprehensive Brief on Research Data Management Policies, Kathleen Shearer, 2015
EXAMPLE SCENARIO

Louise is a new PhD student in the School of Public Health and Preventive Medicine. Her PhD topic relates to policy interventions to prevent the outbreak of infectious diseases like bird flu. She is interested in this topic because of her work as a policy analyst with AHS and her background in volunteering in developing countries and sees completing the PhD as a good way to further her policy career as well as her interests in social development.

Louise's research will involve several field interviews with health workers and policy makers in Canada, Vietnam, Indonesia and China. She has an iPhone and thought that she would use this to make audio recordings of the interviews, which she will later analyze (possibly using NVivo).

Louise also wants to access the policy documents of government agencies and health service providers (including hospitals) in Calgary and other jurisdictions in Canada and overseas. She thinks she will do content analysis on these, probably also using NVivo. Some agencies freely provide these documents on their websites, while other agencies have internal documents that are not readily available to the general public, which she may have to approach the organizations for directly. Louise wants to test her hypothesis that a speedy response from policymakers can reduce the spread of infectious diseases. This will require doing some cross-analysis of her findings from the policy documentation and interviews along with the World Health Organization’s Cumulative number of confirmed human cases of avian influenza A(H5N1) dataset, which is available for download from the WHO website as a series of PDFs published monthly.

In doing her literature review there are several industry publications and academic journals that Louise has identified as potential places in which she might try to publish later. There are also some big international conferences coming up, and her supervisor has encouraged her to consider presenting her results at these.
CREATING A DMP

DMP Assistant (Canadian, bilingual), https://assistant.portagenetwork.ca
DATA COLLECTION

1. What types of data will you collect, create, link to, acquire and/or record?
2. 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?
3. 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?
TIP: FILE TYPES

If possible, DO NOT use proprietary file formats
If you must use a proprietary format, also include an open version (ex. rtf with docx)
TIP: FILE ORGANIZATION

We strongly encourage you to use a reference manager for storing your articles/literature, such as:
Mendeley
Zotero
Papers
Endnote
"FINAL."doc

FINAL.doc!

FINAL_rev.2.doc

FINAL_rev.6.COMMENTS.doc

FINAL_rev.8.comments&5.corrections.doc

Track changes

FINAL_rev.18.comments&7.corrections&9.MORE.30.doc

FINAL_rev.22.comments&49.corrections.10.#@$%WHY%DID
ICOMETOGRADESchool????.doc

TIP: FILE NAMING

Be Descriptive:

• Unique identifier (ie. Project Name or Grant # in folder name)
• Project or research data name
• Conditions (Lab instrument, Solvent, Temperature, etc.)
• Run of experiment (sequential)
• Date (in file properties too)
• Version #
TIP: FILE NAMING CONVENTIONS

Be Consistent

YYYYMMDD  TimeDate  Sample00001234
MMDDYYYY  DateProjectID  Sample01234
YYMMDD  TimeProjectID  Sample1234
MMDDYY
DDMM

MAINTAIN ORDER

INCLUDE THE SAME INFORMATION

INCLUDE THE SAME INFORMATION
TIP: FILE NAMING CONVENTIONS

Files without a naming convention:

• Test_data_2013
• Project_Data
• Design for project.doc
• Lab_work_Eric
• Second_test
• Meeting Notes Oct 23

Files with a naming convention:

• 20130503_DOEProject_DesignDocument_Smith_v2-01.docx
• 20130709_DOEProject_MasterData_Jones_v1-00.xlsx
• 20130825_DOEProject_Ex1Test1_Data_Gonzalez_v3-03.xlsx
• 20130825_DOEProject_Ex1Test1_Documentation_Gonzalez_v3-03.xlsx
• 20131002_DOEProject_Ex1Test2_Data_Gonzalez_v1-01.xlsx
• 20141023_DOEProject_ProjectMeetingNotes_Kramer_v1-00.docx
1. What documentation will be needed for the data to be read and interpreted correctly in the future?
2. How will you make sure that documentation is created or captured consistently throughout your project?
3. If you are using a metadata standard and/or tools to document and describe your data, please list here.
METADATA

• Data about your data
• The main purpose of metadata is to facilitate the discovery or search of your data through relevant criteria
• Everything you’d need to recreate your data
TIP: METADATA

- Describe your data as thoroughly as possible
- Use standards as appropriate, ex.
  - DDI - Archiving and Social Science
  - Darwin Core – Biology
  - DIF - Scientific data sets
  - CSDGM - Geographic data
  - TEI - Humanities, social sciences and linguistics
- For data deposit, most data repositories will walk you through minimum metadata, but the more you can add, the better.
DATA DOCUMENTATION TOOLS

- README template
- Data documentation initiative: https://www.ddialliance.org
- DDI + Excel integration: https://www.colectica.com/software/colecticaforexcel/
1. What are the anticipated storage requirements for your project, in terms of storage space (in megabytes, gigabytes, terabytes, etc.) and the length of time you will be storing it?
2. How and where will your data be stored and backed up during your research project?
3. How will the research team and other collaborators access, modify, and contribute data throughout the project?
Working USB stick found in leopard seal's year-old frozen faeces

New Zealand puts out call to find owner of memory stick spotted in frozen poo sample

A functioning USB stick has been found in the scat of a rare Antarctic leopard seal, prompting New Zealand’s national science body to launch a hunt for the owner.

Volunteers at the National Institute of Water and Atmospheric Research (Niwa) recovered the device while examining the animal’s frozen faeces - which had been sitting in a freezer for over a year.

It revealed a video shot by a kayaker paddling around Porpoise Bay in the Catlins at the bottom of the South Island. The kayaker is not visible, but the tip of their blue kayak is, as well as footage of leopard seals playing in the cold coastal waters, including a mother and its pup.

Leopard seal scat is valuable to scientists because it holds a wealth of information about what the Antarctic predators eat, their general health and how long they have been in New Zealand waters.

Around the country, volunteers comb the coastline collecting it for marine biologists, in a bid to learn more about the rare animals, which usually live and hunt in packs in Antarctic waters, but are increasingly heading north to New Zealand, puzzling experts.

https://www.theguardian.com/world/2019/feb/06/working-usb-stick-found-in-leopard-seals-year-old-frozen-faeces
TIP: BACKUP

- 3-2-1 Backup strategy:
  - 3 total copies of your data, on at least 2 different mediums, and at least 1 copy offsite
- Keep your anti-virus software up-to-date
- Keep your passwords secure
- Consider using encryption
- Do not store master data copies on personal computers or laptops
- Be mindful of ethics constraints
STORAGE RESOURCES AT UCALGARY

- Secure Computing Program for sensitive, level 3 or 4, data
- Research Storage Service for level 1 or 2 data
- UCalgary Information Security Classification standard
Data Security Classification Levels

1. **Public data**
   - public announcements, telephone directory, published research data, open data

2. **Internal Use**
   - Info where disclosure or loss would be inconvenient but it is not likely to result in financial loss or reputational damage
   - Unpublished research data, anonymized or de-identified human subject data

3. **Confidential**
   - Restricted use information where disclosure or loss would cause harm or have to be reported to a regulating organization
   - Employment applications, personnel files, intellectual property, passwords

4. **Restricted**
   - Information that is confidential and subject to specific privacy and security safeguards. Problems would have to be disclosed to all affected individuals
   - Credit card info., health information linked to an identifiable individual, identifiable human subject research data

https://www.ucalgary.ca/policies/files/policies/im010-03-security-standard_0.pdf
PRESERVATION

1. Where will you deposit your data for long-term preservation and access at the end of your research project?
2. Indicate how you will ensure your data is preservation ready. Consider preservation-friendly file formats, ensuring file integrity, anonymization and de-identification, inclusion of supporting documentation.
STUDY: The Availability of Research Data Declines Rapidly with Article Age

• Study on biology data, looked at 516 articles published 1991-2011
• Dataset availability: decreases by 17% per year
• Working email address: decreases by 7% per year

**TIP: FILE FORMATS FOR PRESERVATION**

<table>
<thead>
<tr>
<th>IMAGES</th>
<th>TEXT</th>
<th>SPREADSHEETS</th>
<th>EBOOKS</th>
<th>DATABASES</th>
</tr>
</thead>
<tbody>
<tr>
<td>• TIFF, PNG, JPG</td>
<td>• TXT, CSV, PDF/A, ASCII, UTF-8</td>
<td>• CSV, TSV</td>
<td>• EPUB, PDF/A</td>
<td>• XML, CSV, SQL</td>
</tr>
<tr>
<td>VIDEO</td>
<td></td>
<td>• MPG, MOV, AVI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• MP3, FLAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOUND</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
RETENTION RULES FOR RESEARCH DATA

- UCalgary Master Records Retention Schedule
  - Research Ethics Process and Review
  - Research Involving Human Subjects
  - Research Data – Not Involving Human Subjects

Often research data/records must be retained for five years after end of project. Check with the Principal Investigator.
SHARING AND REUSE

1. What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final).
2. Have you considered what type of end-user license to include with your data?
3. What steps will be taken to help the research community know that your data exists?
TIP: SHARING CONCERNS

You must ensure the confidentiality of your subjects. Watch for:

- Personally Identifiable Information
- Protected Health Information
- Sensitive Information
- Information that seems generic, but could still be used to narrow information down to an individual
TIP: SHARING CONCERNS

If you can’t scrub your data?

• Embargoes
• Technological Access Restrictions
• Data Use Agreements
SHARING: DATA REPOSITORIES

Designed for data
- Secured-stored
- Authoritative copy of your data
- Discoverable through Google, etc.
- Easily citable
- Open access
- Some are free to use, some are not
TIP: DATA REPOSITORIES

- UCalgary PRISM Dataverse: [https://dataverse.scholarsportal.info/dataverse/calgary](https://dataverse.scholarsportal.info/dataverse/calgary)
  - One page [Quick Submit guide](#)
- Other disciplinary and general data repositories available, see here for good list: [https://www.lib.umn.edu/datamanagement/datacenters](https://www.lib.umn.edu/datamanagement/datacenters)
RESPONSIBILITIES AND RESOURCES

1. Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible.

2. How will responsibilities for managing data activities be handled if substantive changes happen in the personnel overseeing the project's data, including a change of Principal Investigator?

3. What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?
ETHICS AND LEGAL COMPLIANCE

1. If your research project includes sensitive data, how will you ensure that it is securely managed and accessible only to approved members of the project?
2. If applicable, what strategies will you undertake to address secondary uses of sensitive data?
3. How will you manage legal, ethical, and intellectual property issues?
DATA MANAGEMENT PLANNING RESOURCES

University of Calgary Guide:
https://library.ucalgary.ca/guides/researchdatamanagement

Portage Network:
https://portagenetwork.ca/
QUESTIONS?

Email: heather.ganshorn@ucalgary.ca