Research Data Management Workshop

http://libguides.uvic.ca/rdmp

Rebecca Raworth, MLIS
April 12, 2015
I have no conflicts of interest to disclose with regard to the subject matter of this presentation.
Learning objectives

By the end of this session you should:

• Understand why research data management plans are important

• Be able to identify the benefits and challenges of data management

• Be able to complete an initial plan for managing your own research data
Outline

• Why?
  – RDM today, in Canada
• What is a Research Data Management Plan?
• What you can do today
  – Portage – RDM
  – Dataverse – data storage & access
• Where to get help
Why manage your data?

• Satisfy grant & journal requirements
• Find your files
• Keep track of different versions of your data
• Organize and compile information at the end of a project
• Reproduce your work (if required for a journal or patent)
• Pass on your work to another researcher
• Share your work
• Satisfy research ethics board requirements
• Increase research efficiency
• Promote collaboration & maximize transparency
• Ensure data access & longevity
• Journal articles with published datasets get more citations
• Publishing data could prove you were first
Current RDM Requirements in Canada

- **SSHRC** *(SSHRC Research Data Archiving Policy)*
  - “All research data collected with the use of SSHRC funds must be preserved and made available for use by others within a reasonable period of time. SSHRC considers "a reasonable period" to be within two years of the completion of the research project for which the data was collected.”

- **NSERC**
  - None.

- **CIHR** *(OA Policy 5.1.2)*
  - CIHR now requires grant recipients to retain original data sets arising from CIHR-funded research for a minimum of five years after the end of the grant. This applies to all data, whether published or not. The grant recipient's institution and research ethics board may have additional policies and practices regarding the preservation, retention, and protection of research data that must be respected.

michael.steeleworthy@wlu.ca

13/11/2015
Data Management Plans will become a required section of future Tri-Council grant applications.

Draft Tri-Agency Statement of Principles on Digital Data Management

1. Preamble

The Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC) ("the Agencies") are federal granting agencies that promote and support research, research training, knowledge transfer and innovation within Canada.

As publicly-funded organizations, the Agencies are strong advocates for making the results of the research they fund as accessible as possible. In promoting access to research results, they aspire to advance knowledge, avoid research duplication and encourage reuse, maximize research benefits to Canadians, and showcase the accomplishments of Canadian researchers. These aspirations align with the Government of Canada's commitment to open science, as described in Seizing Canada's Moment: Moving Forward in Science, Technology and Innovation (2014).

Research data – recorded material that validates research findings and results, and enables reuse or replication[1] – are increasingly created or translated into digital formats. When properly managed and responsibly shared, these digital resources enable researchers to ask new questions, pursue novel research programmes, test alternative hypotheses, deploy innovative methodologies and collaborate across geographic and disciplinary boundaries. The ability to store, access, reuse and build upon digital research data has become critical to the advancement of science, supports innovative solutions to economic and social challenges, and holds tremendous potential for Canada's productivity, competitiveness and quality of life.

Governments and research funders across the globe are becoming increasingly aware of the value of digital research data, the importance of fostering reuse of digital research data, and the need for policies to facilitate excellence in data stewardship. Canada has joined other countries at the forefront of this movement, as exemplified by its support for the OECD Declaration on Access to Research Data from Public Funding (2004), commitment to the Open Government Declaration (2011), and approval of the G8 Science Ministers Statement (2013).

The Government of Canada's Action Plan on Open Government (2014) aims to maximize access to the results of federally-funded research to encourage greater collaboration and engagement with the scientific community, the private sector, and the public. The action plan includes a commitment to adopt policies to support effective data stewardship.

The Agencies believe that research data collected with the use of public funds belong to the fullest extent possible to the public domain, available for reuse by others, and they strongly support the creation of a robust and efficient environment for data stewardship in Canada and internationally. They have encouraged data stewardship through SSHRC's Research Data Archiving Policy (1990), the Tri-Agency Framework: Responsible Conduct of Research (2011), and data sharing provisions for CIHR grant holders in the Tri-Agency Open Access Policy on Publications (2015), and they will continue to promote excellence in data management practices within the Canadian research community.

http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1
Background

This statement of research data management (RDM) principles has been developed by a task group of senior university administrators and individuals from Research Data Canada (RDC) as follow-up to a meeting in Edmonton on November 17, 2015 co-sponsored by RDC and the University of Alberta with support from SSHRC on behalf of the federal granting agencies. The Edmonton meeting involved senior representatives of 16 institutions, as well as of the federal granting councils and other national organizations with an interest in research data. All who gathered in Edmonton were cognizant of RDM issues in part because of work underway to develop a set of guidelines on data management by the federal granting agencies. [http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1](http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1)

Those gathered in Edmonton delegated the development of a statement of principles to a task group. The group started its deliberations with a review of several comparable statements from around the world before selecting as a starting point the principles developed by the Research Councils of the United Kingdom [www.rcuk.ac.uk/research/datapolicy](http://www.rcuk.ac.uk/research/datapolicy). The authors of the current document acknowledge their debt to the RCUK.

Universities will develop RDM policies to suit institutional needs and circumstances, but they can learn from one another in a collaborative environment that avoids each one “reinventing the wheel”. If they all share the core principles articulated below, joint development of new tools, services, and
Research data means:

- 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.
Video

Data Sharing and Management Snafu in 3 Short Acts
http://youtu.be/N2zK3sAtr-4

https://www.youtube.com/watch?v=N2zK3sAtr-4, By Karen Hanson, Alisa Surkis & Karen Yacobucci, NYU Health Sciences Library
Take a moment to think about your research

- **Talk with your neighbour** and discuss the types of data you generate, where your data is stored and how it is organized.
- If you were asked to share your data with another researcher would they be able to make sense of your data?
- If you needed to locate your data files from 5 years ago, how easy would they be to find and use?
University of Toronto researcher Rachel Harding will be the first known biomedical researcher to welcome the world to review her lab notes in real time. The post-doctoral fellow with U of T's Structural Genomics Consortium (SGC) is also explaining her findings to the general public through her blog. She hopes her open approach will accelerate research into Huntington's disease.

Data management plans

• Describe how data is collected, formatted, preserved and shared, as well as how existing datasets will be used and what new data will be created

• FAIR Data Guiding Principles:
  – Data should be findable
  – Data should be accessible
  – Data should be interoperable
  – Data should be re-usable
The Data Management Plan: Common Misconceptions

• Does *not* require that *all* data must be shared
  – Sensitive information/patient privacy
  – Intellectual property rights and commercial value

• Sharing can take many forms

• Funders recognize that different disciplines have different “cultures” of data sharing

• Sharing “at no more than incremental costs and within a reasonable time.”
Best practices in data documentation

• At the very least you should document the following in a `readme.txt` file stored alongside your data:
  – context of data collection (the goal of your research)
  – data collection methods (protocols, sampling, instruments, coverage...)
  – structure of files
  – sources used
  – quality assurance (data validation, checking)
  – data modifications
  – confidentiality and permissions
  – names of labels and variables
  – explanations of codes and classifications
Metadata

**Descriptive:** title, author, keywords

**Administrative:** information needed to use the data, e.g. software requirements, copyright

**Structural:** how different data sets relate to one another, e.g. file formats, information about the relationship between data sets in a database
Metadata standards

• Many disciplines have created their own metadata standards to ensure that data records can be interpreted and compared across projects and fields

• List of standards in your field - http://www.dcc.ac.uk/resources/metadata-standards

• A few examples:

  – **CARMEN** A virtual laboratory for neurophysiology, enabling sharing and collaborative exploitation of data, analysis, code and expertise. Metadata must include the MIBBI-registered MINI recommendations.
  
  – **MIBBI** – Minimum Information for Biological and Biomedical Investigations
  
  – **Genome Metadata**

Also, remember to...

- Avoid jargon & symbols – use keywords
- Define terms and acronyms
- State limitations
- Use descriptive titles
- Be specific and quantify
- Use keywords
- Make it machine readable (avoid symbols)
File naming

- Keep file names under 32 characters
- Classify broad types of files (transcript, photo, etc.)
- Avoid spaces and special characters
- Use underscores instead of periods or spaces
- Make sure that file names are descriptive outside of their folders (in case they are misplaced or change locations); i.e., the file name should include all necessary descriptive information
- Include dates and format them consistently (international standard for date notation is YYYY_MM_DD or YYYYMMDD)
- Include a version number to track multiple versions of a document
- Use non-proprietary file formats (e.g. .docs, bmp/tif)
- Be consistent!
File types

- Chose a format that is:
  - Non-proprietary
  - Open, with documented standards
  - Used by your community
  - Encoded using standard character encoding
  - Uncompressed

http://www.data-archive.ac.uk/create-manage/format/formats-table
Data storage, security & backup

- Anticipated storage requirements

- Length of time for storage

- Where?
  - 3-2-1 rule:
    - 3 copies
    - 2 different media
    - 1 backup offsite

- Security
  - Passwords
  - Data encryption
  - Room secure
  - Network firewalls, etc…
How to share data

• Find a home for your data
  – Subject specific repository/archive (see re3data.org)
  – Institutional repository/archive
  – Journal website
  – Project website

• License your data (e.g. Creative Commons or Open Data Commons)

• Provide suggested data citation

Locating repositories

http://service.re3data.org/search?

http://oad.simmons.edu/oadwiki/Data_repositories

Schizophrenia

NOTE: The distribution refers to a collection of studies organized by disorder or theme (see FAQ). It is important to note that not all individuals in the distribution have DNA available.

Given the major public health implications of identifying genes responsible for severe neuro-psychiatric disorders, the National Institute of Mental Health (NIMH) has funded a Human Genetics Initiative. The goal of this Initiative is to establish a national resource of clinical data and biomaterials that are collected from individuals with Alzheimer disease, schizophrenia (SZ), or bipolar I disorder, in order to aid researchers in understanding the genetic bases of these disorders. The NIMH Schizophrenia Genetics Initiative is supported by the Office of Human Genetics & Genomic Resources in NIMH's Division of Neuroscience and Basic Behavioral Science (DNBBS). Since 1996 data and biomaterials (cell lines and DNA samples) have been available to qualified investigators who study the genetics of SZ, and may be accessed by following a set of instructions.

From 1991-98, SZ pedigrees were ascertained through three extramural sites (Columbia University, Harvard University, and Washington University). In 1998, NIMH issued a Request for Applications (MH-98-010) to solicit applications for collaborative research projects to collect large pedigree samples and conduct molecular genetic analyses on schizophrenia, bipolar disorder, and early-onset recurrent unipolar depression. Another solicitation (MH-99-005) was issued in 1998 that focused on schizophrenia and early-onset recurrent depression. Data and biomaterials collected and produced in these

http://oad.simmons.edu/oadwiki/Data_repositories
Ethics & legal compliance

- How will you ensure sensitive data is securely managed & accessible only to approved members of the project?
- What strategies are there for secondary use of sensitive data
Sharing and re-use

• What data will you be sharing and in what form?
  – Raw
  – Processed
  – Analyzed
  – Final

• What type of end-user license for your data?
  – Creative Commons
  – Open Data

Sharing and re-use

• Obtain a DOI (digital object identifier) for data
• Why?
• Provides a persistent link to facilitate data visibility
• Enable tracking of citations to provide proof of research impact


https://www.datacite.org/services/cite-your-data.html
Digital Object Identifiers (DOIs)

“Publication & citation are accepted by the scholarly community as the principal way of acknowledging the value and impact of a researcher’s contributions to the development of knowledge in their field.”


“Data are to be considered as first-class scholarly objects, and treated similarly in many ways to cited and archived scientific and scholarly literature.”

Portage, Canada’s DMP assistant

Has seven sections

1. Data collection
2. Documentation & metadata
3. Storage & backup
4. Preservation
5. Sharing & re-use
6. Responsibilities & resources
7. Ethics & legal compliance

https://portagenetwork.ca/
Dataverse – now available at UVic

A repository for research data that takes care of **long term preservation** and **good archival practices**, while researchers can **share, keep control of and get recognition** for their data.

Supports the sharing of research data with a persistent **data citation**, and enables **reproducible research**.
Resources to help you...

- UK Data Archive Checklist: http://data-archive.ac.uk/create-manage/planning-for-sharing/data-management-checklist
- Canada’s DMP Assistant: https://portagenetwork.ca/
- In USA (for NSF, NIH, etc.) – Data Management Planning Tool: https://dmptool.org/
- UK Digital Curation Centre Curation Checklists – http://www.dcc.ac.uk/sites/default/files/Curation%20Checklists.pdf
Uvic Library’s Research Data Subject Guide

http://libguides.uvic.ca/rdmp
Resources to help you

MANTRA is a free online course for those who manage digital data as part of their research project.
Acknowledgements


• *UBC Library DataGuide, (version 3.1, October 2015)* by Eugene Barskey, UBC Research Data Librarian
Thank you!