RESEARCH DATA MANAGEMENT (RDM)

Michelle Edwards & Carol Perry

Starting your Research on the Right Foot Part 3

Objectives

- Understand the components of research data stewardship
- Apply concepts to your research project

WHAT IS RESEARCH DATA MANAGEMENT?

- Management of research data?
  - Management of data?

- We collect data, save it on our computers, analyze it using some software, write-up our results, and hopefully publish

- What do we need to manage????

Research Life Cycle – Interpret & Publish

DMP CHECKLIST

√ Organizing the data you've collected

√ Documenting your work

√ Managing your files – processing and analyzing your data

√ Storing, backing up, and securing your files

Preserving your data

Accessing, sharing, and reusing your data

ARCHIVING vs PRESERVING YOUR DATA

- Data archiving: moving data that is no longer needed to a separate storage area for long-term storage. This data is still important and may be needed in the future.

- Data preservation: ensuring that the data is still usable and accessible in the future. This requires a number of managed activities to ensure its use.

DATA PRESERVATION CHALLENGES - MEDIA

Data stored on:

- 1970s – magnetic tapes
- 1980s – 5.25” diskettes
- 1990s – 3.5” diskettes
- 2000s – CD-ROMs

- Can we access this data today?
- If not, is this data irrelevant?

Michael Leddy. Used under CC-NC-ND 4.0.
Retrieved from: http://mleddy.blogspot.com/2007/09/utnapishtims-word-processor.html

DATA PRESERVATION CHALLENGES - SOFTWARE

- Data collected in a variety of software…
  - VP Planner
  - Lotus 1-2-3
  - Quattro-Pro
  - first Mac version of Excel in 1985
  - first Windows version of Excel in 1987
  - Old excel version .xls

DATA PRESERVATION CHALLENGES

- New data recovery pilot project
  - exploring ways to move data from older media and software to platforms and non-proprietary software available today – where possible
- contact lib.research@uoguelph.ca for more information

DATA PRESERVATION – Format

- At the end of your project save all data, statistical coding procedures, and outputs in a non-proprietary format.
- NO Excel, NOWord, NO SAS, NO SPSS, etc…
- Non-proprietary formats include:
  - Numerical data
    - Ascii
    - Text
    - Csv
    - Tab
  - Tabular data
    - Transform to csv

DATA PRESERVATION – Format

- Data visualizations
  - Charts, save as image file – jpg, pdf
  - Mapping – save all files (shp, prj,sbx, sbn, dbf)
- Image files
  - Tiff 6.0 uncompressed
- Video
  - Mpeg 4
  - Motion jpeg 2000
- Text
  - Txt, pdf/A, pdf, xhtml, html

- Check for errors after converting files

DATA PRESERVATION – Documentation

- When preparing your data and project for preservation you will need all of your files and

Documentation!

- Without documentation
  - your data is not meaningful and cannot be used in the appropriate manner
  - your study and analysis cannot be replicated!

DATA PRESERVATION – Documentation

- Documentation to include:
  - Codebook
    - explains variables – remember those labels?
    - shows what your values represent
  - Syntax files
    - all of your SAS codes
    - document what action each set of code is performing
  - Readme files
    - provide user with notes and description of directory structure
  - Any additional documentation necessary to understand your data

SHARING, ACCESSING, and REUSING YOUR DATA

- Why would you want to share your data? What value does it add?
  - Increases impact of your research
  - Helps others replicate your research
  - Encourages further scientific enquiry
  - Reduces research costs by reducing duplication
  - Encourages transparency and accountability

SHARING, ACCESSING, and REUSING YOUR DATA

- What data do you want to share?
  - Raw data files?
  - Data supporting publication?
- When do you want to share it?
  - Immediately upon publication?
  - After an embargo period to have time to expand on research?

SHARING, ACCESSING, and REUSING YOUR DATA

- Before openly sharing your data you need to consider:
  1. Ethical and legal obligations and/or requirements
  2. Data anonymization
  3. Intellectual property rights

1. ETHICALAND LEGAL OBLIGATIONS

- Ethical and legal obligations
  - Research ethics board
  - Funding agencies –Tri-Agency (NSERC, SSHRC, and CIHR); Polar Data Canada
  - Journal requirements
  - Partnership requirements

- Before sharing any data, review your ethical and legal obligations. In some instances you may not be ALLOWED to share or you may be REQUIRED to share

2. DATA ANONYMIZATION

- Any shared data should be anonymized
  - no personal identifiers remaining in the dataset
  - users cannot recreate and identify individuals, units, etc..
- Methods of anonymization:
  - Aggregation
  - Pseudo anonymization

2. DATA ANONYMIZATION

- How can you identify individuals in a dataset:
  - Direct identifiers
    - names
    - addresses
    - identification numbers – student ID, OHIP number….
  - Indirect identifiers
    - birth date,
    - detailed geographic areas
    - detailed information on income, place of birth, etc.
  - A combination of indirect identifiers could lead to the identification of an individual

2. DATA ANONYMIZATION - METHODS

- Aggregation
  - Example:
  - Income of the producer was collected
    - to anonymize this piece of information, aggregate income data (create larger grouping)

Producer X original data = $45,660/year anonymized data = $40,000 - $49,999

Age of Observer original data = 34 yrs anonymized data = 30-39 years

2. DATA ANONYMIZATION - METHODS

- Masking – pseudo anonymization
  - Data masking is a form of pseudo anonymization
  - Maintaining the integrity of the variable or information that was collected but replacing the real data with fictitious and masked information
    - **cannot** identify the individual unit
    - **can** still use the information to describe the sample or population

2. DATA ANONYMIZATION - PITFALLS

- Tendency to replace all direct identifiers with pseudonyms or aggregate variables
- Avoid blanking out information
- Avoid over-anonymization – can lead to misleading conclusions
- Keep LOG of anonymization techniques!!
  - Secure data -> useable data

3. INTELLECTUAL PROPERTY RIGHTS

- Who owns the data?
- Manage your intellectual rights to the data that you've collected through a license
- License will determine who can use the data, and what they can do with it.
- For more information please see Data licensing at the University of Guelph.

SHARING, ACCESSING, and REUSING YOUR DATA

- We've finished our project
- Our data is clean and organized – and well documented!
- We CAN share our data
- So now what? Where and how do we share our data?
  - Pass out USB keys?
  - Provide the world with access to our Departmental Servers?

PRESERVATION AND ACCESS OPTIONS

- Institutional repository
- Publish data with results - journal
- Deposit in major data repository
- Deposit in discipline-specific data repository

UNIVERSITY OF GUELPH-BASED OPTIONS

- Atrium (institutional repository) - e-theses, articles, reports, videos, etc. https://atrium.lib.uoguelph.ca/

- Agri-environmental Data Repository- research data - OAC data

https://dataverse.scholarsportal.info/dataverse/ugardr

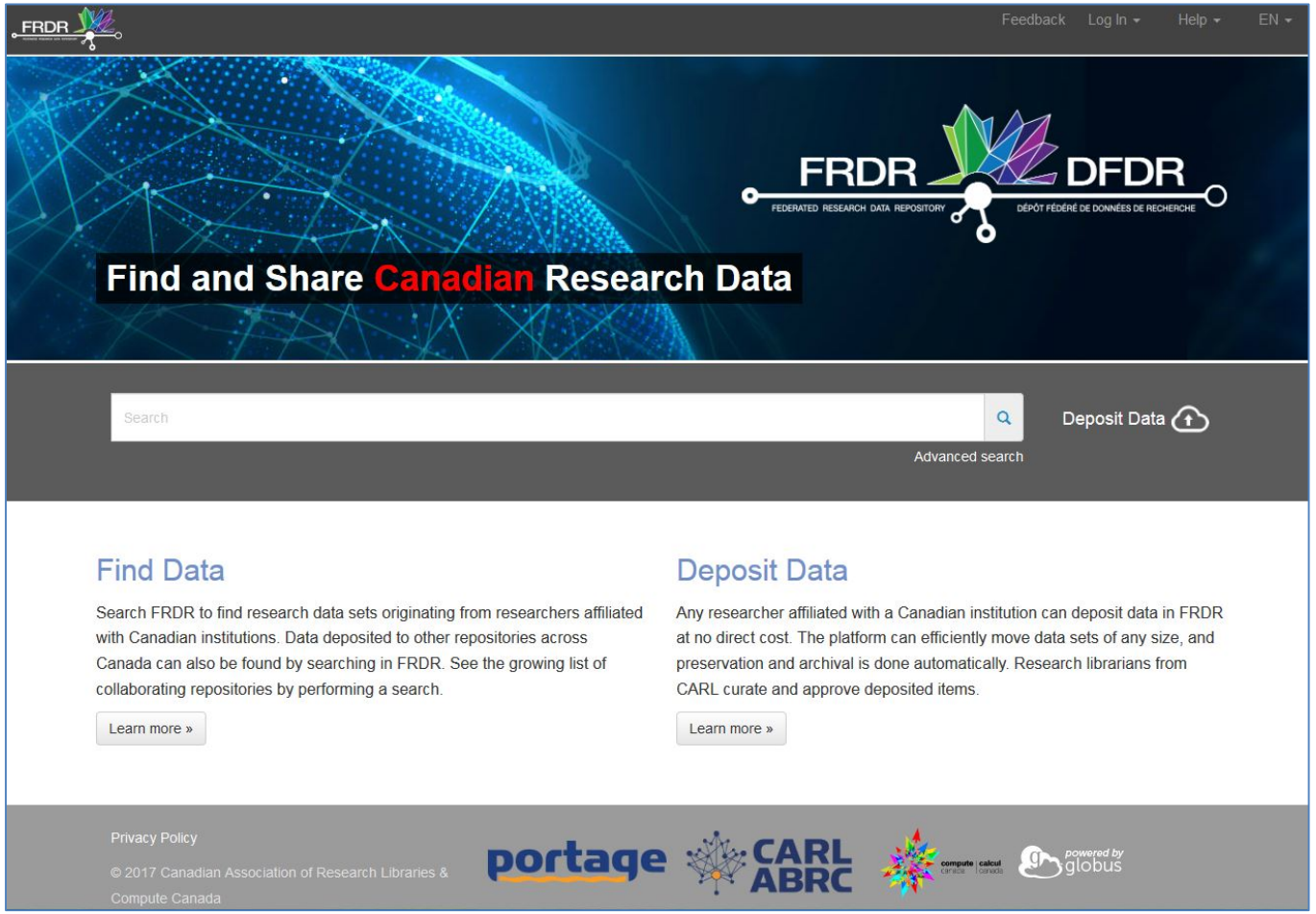
- University of Guelph Data Repository – research data – all disciplines

https://dataverse.scholarsportal.info/dataverse/ugrdr

Federated Research Data Repository (FRDR)

- https://www.frdr.ca

- Canadian
- Data portal
- Big data repository



DISCIPLINE SPECIFIC REPOSITORIES

- Re3data.org – global registry of research data repositories
  www.re3data.org/

- Stanford University Libraries – Guide to Domain-specific Data Repositories

  https://library.stanford.edu/research/data-management-services/share-and-preserve-research-data/domain-specific-data-repositories

CHOOSING A REPOSITORY

- Is it reputable?
- Will it accept your data? Is it a good fit?
- Will your data be safe?
- Does it assign a persistent unique identifier?
- Does it provide analytics on data usage?
- Are there fees?
- What are your obligations under the service?
- What are the obligations of the service provider?

DMP CHECKLIST

√ Organizing the data you've collected

√ Documenting your work

√ Managing your files – processing and analyzing your data

√ Storing, backing up, and securing your files

√ Preserving your data

√ Accessing, sharing, and reusing your data

DATA MANAGEMENT PLAN (DMP)

- Document everything you do in your project
- Use README file – for project level information
- Add comments in your SAS syntax file to describe what your analysis is doing
- All of this information is also called METADATA

METADATA = Documentation about DATA

DATA MANAGEMENT PLAN (DMP)

- KEY to:
  - keeping you organized
  - helping you manage your files
  - getting ready to share your project data
- New National online tool available to you to help you create a DMP

DMP Assistant available through Portage Network

https://portagenetwork.ca/

DMP ASSISTANT

7 Sections:

- Data Collection
- Documentation and Metadata
- Storage and Backup
- Preservation
- Sharing and Reuse
- Responsibilities and Resources
- Ethical and Legal Compliance

DMP ASSISTANT

- Create your own account
  - Will be accessible to you when you leave UG
  - Repositories for YOUR DMPs
- Work through each section and answer questions
  - 2 or 3 questions in each section
  - You already have the answers!!

DMP ASSISTANT

- Great tool to add to your current workflow
- As you start a project – work through the materials we covered here and create your DMP
- Update as you progress through your project
- Can save and create a PDF and pass onto your supervisor!
- Canadian funders are moving to make DMPs a requirement!

AGRI-ENVIRONMENTAL REPOSITORY

Let's see what we have in the Data Repository today

- https://dataverse.scholarsportal.info/dataverse/ugardr

RESEARCH DATA MANAGEMENT

- Managing our research data resources
- Enabling us to keep our project data organized and well-documented
- Recommending best practices for file and variable naming conventions
- 3-2-1 Storage and Backup
- Preserving data for future research use
- Creating a Data Management Plan to guide us through RDM

Contact

- Michelle Edwards
- edwardsm@uoguelph.ca

- Carol Perry
- carolp@uoguelph.ca
- lib.research@uoguelph.ca