Current trends in open access for research data

Presenters:
Erin MacPherson, Research & Instruction Librarian, Dalhousie University
Maggie Neilson, Geospatial and Numeric Data Librarian, Acadia University
Lee Wilson, Interim Service Manager, Portage/ACEnet
Open in Order to...

The theme of Open Access week this year is “Open in Order to…” and is an invitation to answer this question about the concrete benefits of making scholarly outputs openly available.

In this 1 hour, we will talk about:

• What is Research Data Management?
• Why should I care?
• What tools exist?
• Who can help me?
• Where should I start?
What is Research Data Management?

A broad term used to describe the structure, organization, maintenance, and overall stewardship of research data.

These elements are given consideration at each stage of research from project conception to its conclusion (and beyond!)
Research data are contents that are used as primary sources to support research, scholarship, artistic activity or research-creation, and that are used as evidence in the research process and commonly accepted in the research community as necessary to validate research findings and results.
Research Data Lifecycle

http://data.library.virginia.edu/data-management/lifecycle/
Why Manage Research Data?

To be ORGANIZED

For SAFE-KEEPING

Prepare for CHANGE

Grant, journal or Research Ethics Board REQUIREMENTS
Why Manage Research Data?

To help ensure DEPENDABILITY, ACCURACY AND VALIDITY of data

Enhance REPLICABILITY of research

Solidify commitment to a research culture of TRANSPARENCY AND ACCOUNTABILITY

SHARE it
Trends in Research Data Management

Journal submission & publication requirements

Institutional repositories as data repositories

FUNDING REQUIREMENTS
Journal Requirements

Some journals may require you to share your data as a condition of publication. Often, data sharing policies can be found in the "Instructions for Authors" or "Author Guidelines."

Some examples of data-sharing policies are below:

**Nature**: Availability of data, material and methods

**Wiley**: Wiley's Data Sharing Service
Tri-Agency Statement of Principles on Digital Data Management

Outlines the Agencies’ overarching expectations with regards to research data management, and the roles and responsibilities of researchers, research communities, research institutions and research funders.

Aims to establish a common set of principles that will serve as the basis for the development of data management requirements that the Agencies will establish collectively or individually.
Intended Impact of a Tri-Agency Digital Data Management Policy

• Strong data management as an accepted signifier of research excellence

• More Canadian datasets cited, and valued as a product of research in tenure, promotion and peer review processes;

• Canadian researchers equipped and ready to engage in international research collaboration

• Canadian research institutions ready to support the management of the data their researchers produce;

• Increased ability for research data to be archived, found and responsibly reused, to fuel new discovery and innovation.
Draft Tri-Agency Data Management Policy

For consultation

Feedback will inform final policy

Proposed policy includes 3 possible requirements

Phased, incremental implementation
1. Institutions: Institutional Strategy

Each institution administering Tri-Agency funds could be required to create an institutional research data management strategy.

The strategy could outline how the institution will provide its researchers with an environment that enables and supports world class research data management practices.
2. Researchers: Data Management Plans

Grant recipients could be required to create data management plans (DMPs) for research projects supported wholly or in part by Tri-Agency funds. Grant recipients could submit these plans to their institution’s research office as a condition of the release of grant funds.

For specific funding opportunities, the agencies could require DMPs to be submitted to the appropriate agency at time of application.
3. Researchers: Data Deposit

For all research data and code that support journal publications, pre-prints and other research outputs that arise from agency-supported research, grant recipients could be required to deposit these data and code in an appropriate public repository or other platform that will ensure safe storage, preservation, curation, and (if applicable) access to the data.
About Portage

Portage is a national RDM network launched by the Canadian Association of Research Libraries (CARL) in 2015 that coalesces initiatives to build RDM capacity and infrastructure with libraries.

Networks of Expertise:
- Pan-Canadian RDM expertise
- Provide access to resources, tools, and experts in the area of RDM

Infrastructure Platforms:
- Working with library consortia, institutions, and other infrastructure partners to assemble essential RDM infrastructure and service components
Portage Network of Experts

6 Expert Groups

Data Management Planning
Curation
Data Discovery
  - Data repository collection dev
  - Metadata for discovery
Preservation
Training
  - CIHR Training Modules
  - Online training for RDM & DMPs
Research Intelligence

4 Working Groups

Dataverse North
FRDR Service Model
Institutional RDM Strategies*
Ethics of Managing Sensitive Data*

* In the process of being formed
Dataverse North Working Group

What is it?

The Dataverse North Working Group is a Portage Working Group. It is a community of practice whereby Dataverse providers and libraries can come together to coordinate infrastructure development, training, support services, outreach strategies, and promotions.

What is our mandate?

Develop support materials and recommended best practices for adoption by hosts and users;
Aim to provide broad access to Dataverse across Canada, through common delivery models and services.
Portage Supported Platforms

Dataverse

- A free open source web application
- Developed by the Institute of Quantitative Social Science at Harvard University
- Designed specifically for research data
- Platform to contribute, share, preserve and find data.
- Dataverse is used all over the world. In Canada there are several installations available, with others coming on board (including Dalhousie, UNB in Atlantic Canada).

FRDR

- FRDR = Federated Research Data Repository (DFDR en français)
- A scalable, federated platform for digital research data management and the discovery of Canadian research data
- Discovery, Storage, Curation, Preservation
Share, publish, and archive your data. Find and cite data across all research fields.
FRDR ([www.frdr.ca](http://www.frdr.ca))
What is a DMP and why should I care?

A Data Management Plan, or DMP, is a formal document that outlines the steps taken before, during, and after your research project to ensure that the data will be useable, discoverable, and accessible beyond the purpose for which it was originally collected.

Benefits of a DMP:

- Start thinking about your data at the outset of a project
- Makes preservation of data easier
- Cuts down on the duplication of efforts
- Increases visibility of research
- Satisfies funder requirements
Components of a DMP

1. Data Summary
2. Data Quality
3. Metadata Standards
4. Preservation & Access
5. Ethical & Legal Compliance
6. Responsibilities & Resources

Note: The DMP Assistant or your institution may have a slightly different template, but these are the general components of a DMP.
Data Summary Questions

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

What file formats will your data be collected in? Do these formats allow for data re-use, sharing and long-term access? E.g., are file formats proprietary or open source?

What are the anticipated storage requirements of your project?

Avoid:

One-word answers. Funding agencies expect researchers to be able to write in complete sentences!

Generic responses lifted from project descriptions. The purpose of the exercise is to have researchers really thinking about their data; encourage them to be specific.
Data Summary

Data types
E.g., numeric, images, audio, video, text, tabular data, modeling data, spatial data, instrumentation data.

Formats
- Long-lasting; proprietary; “neutral” or industry standard?
- E.g., PDF, CSV, TIFF, JPEG, SVG, MPEG-4, HTML
- Need to justify the use of proprietary formats
- Transformations possible?

Quantity/Volume
- Important for costing DM
- Short and long-term storage considerations
Data Quality Questions

How will data quality be assured and controlled?

What provisions are in place for data security including data recovery, backup, secure storage, transfer of data, and version control?

What standards or applications (e.g., notebooks) are you using to help to structure, name and version control your files?
Data Quality

Internal Data Curation

Consistent file management practices

Codified processes and policies for data collection & analysis

File naming conventions
  - Keep file names short and relevant
  - Many tools available for bulk file-renaming

Version control for data files

Github, Gitlab, Jupyter Notebook, NVivo

Examples of data documentation:

- Experiment protocols
- Instruments & calibration
- Data collection conditions
- Database schema

Common elements of file names:

- Version number
- Name of creator
- Description of content
- Name of research team/department
- Publication date
- Project number

Example: 2017-10-26_oa-week-presentation_final
Tools for file renaming

Windows:
Ant Renamer (www.antp.be/software/renamer)
RenameIT (sourceforge.net/projects/renameit)
Bulk Rename Utility (www.bulkrenameutility.co.uk/)

Mac:
Renamer4Mac (renamer4mac.com/)
Name Changer (mrrsoftware.com/namechanger/)

Linux:
GNOME Commander (www.nongnu.org/gcmd/)
GPRename (http://gprename.sourceforge.net/)

Unix:
The use of the `grep` command to search for regular expressions
Data Quality

Storage & Security

Where will your data be stored?
- PC/laptop, departmental server, external hard drive, the cloud
- The cloud is useful; beware the cloud

Back up policy and procedures
- 3-2-1 rule
- Check storage media often

Managing access and permissions

Does your data require encryption?
- Consider potential harms arising from security breach
- If unsure consult information professional or privacy officer

http://blog.snapinspect.com/top-3-security-concerns-for-property-inspection-apps-cloud-storage/
What elements in the metadata and other supporting documentation are needed to ensure that data is read and interpreted correctly in the future?

What specific standard(s) are being used for metadata?
Metadata Standards

- Data is more than just data; metadata is data about data
- Standardized or controlled vocabularies to describe contextual information, variable names, and values
- Descriptions of your data in plain text (README file) allow an “informed outsider” to understand your data
  - What is it?
  - When and how was it collected?
  - What tools were used?
  - Were there any special conditions (e.g., machine calibration) that might affect re-use?
- Using a standardized method to describe these questions helps with machine-to-machine interoperability
- Helps information professionals make informed decisions around sharing and ultimately preserving your data

Example Metadata Standards:
- Dublin Core (general purpose)
- DataCite (research data)
- Darwin Core (biology)
- Ecological Metadata Language (ecology)
- Data Documentation Initiative (research data)
- Animal Acoustic Telemetry Standard
- ISO 19115
- SensorML

Preservation and Access Questions

What data will you be publicly sharing and in what form?

When will the data be made available for re-use? If an embargo period is requested specify why and how long this will apply, bearing in mind that data should be made available as soon as possible.

What type of end-user license to include with your data?

Where will you deposit data for long term preservation and access?

What steps will be taken to help the research community know that your data exists? E.g., what metadata catalog will you use to increase discoverability?
Preservation & Access

Consider what data you plan on sharing publicly and what form

How will you license your data?
• Creative Commons
• Open Data Commons

Can your data be shared? Is it proprietary?

When to begin sharing data?

Where will you deposit data for long term preservation and access?

How will you transfer your data?

What steps will be taken to help the research community know that your data exists?
Preservation and Access: The Good, the Bad, and...

Common issues:

The belief that others will not be interested in your data is not a reason for declining to share it.

“By request only” sharing is valid in some cases, but needs to be justified. I.e., the data is sensitive

Encourage open platforms and repositories. A researcher’s hard drive has a limited reach and shelf life

A journal publication is not a substitute for data sharing

Data published as supplementary materials is a good start; an open access data repository is better
Ethical and Legal Compliance Questions

Will approval from your University’s research ethics board be required?

Are there any legal, ethical or intellectual property issues with sharing data?

If applicable, how will sensitive data be securely managed and accessible only to project members? Does ethics alone address this, or are non-disclosure agreements required?

If applicable, what steps need to be taken before publicly releasing data? E.g., anonymization/de-personalization of data.
Ethical & Legal Compliance

Are there any sensitivities in your data that would affect sharing?
- E.g., nesting grounds of endangered species; Sacred sites

Ensure that Personally Identifiable Information (PII) is removed from data

Consult the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2, 2014)*
- Ch. 5: Privacy and Confidentiality

Adhere to federal and provincial privacy legislation
- E.G., PIPEDA and HIPA (SK)

Personally Identifiable Information (PII)

Directly identifying information
- Name
- SIN
- Personal health number

Indirectly identifying information
- DoB
- Place of residence
- Unique personal characteristics

Before sharing, data may be:
- Anonymized/De-identified
- Aggregated
- Perturbed (statistical methods)

In special circumstances, sensitive data may still be shared in a limited manner.
Responsibilities and Resources

Identify who will be responsible for managing the project’s data and the major data management tasks for which they will be responsible.

- Nominate a “data champion”
- Remember succession planning

What resources will you require to implement your data management plan? (E.g., training, storage space, and large data transfer capabilities). If applicable, try to estimate the costs associated with data management and build these into your project’s budget.

And prepare for the unexpected!
The DMP Assistant will help you with all of this!

Access the DMP Assistant tool via:

Portage website: https://assistant.portagenetwork.ca/

OR

Your institution's website: http://dal.ca.libguides.com/rdm/plan
Note: For Librarians and those who may have RDM responsibilities - you can customize the DMP Assistant for your institution. There is a guide available on the Portage website to help: https://portagenetwork.ca/working-with-portage/dmp-assistant-custom-guide/
Since I already have an account, we are brought directly to the “My Plans” section upon logging in. You can see some of the sample plans I’ve created. This section lists the name, template used, who owns the plan, if it has been shared, last edit date. You can also edit existing plans, share, export and delete them.
I’m going to choose Dalhousie’s branded version with the standard portage template. Institutions can customize template questions and add guidance. Researchers need to only answer questions that are applicable to them.
These fields can all be edited - you probably don’t want to call your plan “My plan”. You can also save each section as you go, and come back later.
These are the questions that you can fill out for your DMP. You only need to answer questions that are relevant to your project. Guidance is provided, and your institution can also add additional guidance.

<table>
<thead>
<tr>
<th>Sections</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Collection</td>
<td>- What types of data will you collect, create, link to, acquire and/or record?</td>
</tr>
<tr>
<td></td>
<td>- 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?</td>
</tr>
<tr>
<td></td>
<td>- 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?</td>
</tr>
<tr>
<td>Documentation and Metadata</td>
<td>- What documentation will be needed for the data to be read and interpreted correctly in the future?</td>
</tr>
<tr>
<td></td>
<td>- How will you make sure that documentation is created or captured consistently throughout your project?</td>
</tr>
<tr>
<td></td>
<td>- If you are using a metadata standard and/or tools to document and describe your data, please list here.</td>
</tr>
<tr>
<td>Storage and Backup</td>
<td>- 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?</td>
</tr>
<tr>
<td></td>
<td>- How and where will your data be stored and backed up during your research project?</td>
</tr>
<tr>
<td></td>
<td>- How will the research team and other collaborators access, modify, and contribute data throughout the project?</td>
</tr>
<tr>
<td>Preservation</td>
<td>- Where will you deposit your data for long-term preservation and access at the end of your research project?</td>
</tr>
<tr>
<td></td>
<td>- 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.</td>
</tr>
<tr>
<td>Sharing and Reuse</td>
<td>- What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final).</td>
</tr>
<tr>
<td></td>
<td>- Have you considered what type of end-user license to include with your data?</td>
</tr>
<tr>
<td></td>
<td>- What steps will be taken to help the research community know that your data exists?</td>
</tr>
<tr>
<td>Responsibilities and Resources</td>
<td>- 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.</td>
</tr>
<tr>
<td></td>
<td>- 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?</td>
</tr>
<tr>
<td></td>
<td>- What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?</td>
</tr>
<tr>
<td>Ethics and Legal Compliance</td>
<td>- 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?</td>
</tr>
<tr>
<td></td>
<td>- If applicable, what strategies will you undertake to address secondary uses of sensitive data?</td>
</tr>
<tr>
<td></td>
<td>- How will you manage legal, ethical, and intellectual property issues?</td>
</tr>
</tbody>
</table>
Once you fill out the basic plan details, you are brought to the Question section. Guidance to the questions is provided on the right. You can also add links, tables and format the text. On the right, you also have the option to provide notes for any collaborators. You can also save each question as you answer it and return to the plan later.
You can add collaborators and add permissions. There are three permission levels. Collaborators will have to register for the DMP Assistant if they don’t already have an account.
If you need to make changes, you can always go back and edit plan details or the questions anytime.
Click on the “Settings” option for more great features before you export your document. You can select which sections to export (including plan details and questions). For some formats, you can also choose your font type, size and margins.
### Plan title

**DMP title**

### Included Elements

#### Admin Details

- **Project Name**
- **Project Identifier**
- **Grant Title**
- **Principal Investigator / Researcher**
- **Project Data Contact**
- **Description**
- **Funder**
- **Institution**

#### Sections

**Data Collection**

- What types of data will you collect, create, link to, acquire and/or record?
- What will be your data be collected for? Will these formats allow for data re-use, sharing and long-term preservation?
- What conventions and procedures will you use to structure, name and version-control your files to help you and others find and use your data?

**Documentation and Metadata**

- What documentation will be needed for the data to be read and interpreted correctly in the future?
- How will you ensure that documentation is created and curated consistently throughout your project?
- If you are using a metadata standard and/or tools to document and describe your data, please list here.

**Storage and Backup**

- What are the anticipated storage requirements for your project in terms of storage space (in megabytes, gigabytes, etc.)?
- How and where will your data be stored and backed up during your research project?
- How will the research team and other collaborators access, modify, and contribute data throughout the project?

**Preservation**

- Where will you deposit your data for long-term preservation and access at the end of your research project?
- Indicate how you will ensure your data is preservation-ready. Consider preservation-friendly file formats, etc.

**Sharing and Reuse**

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

**Responsibilities and Resources**

- Identify who will be responsible for managing the project’s data during and after the project and the major data management responsibilities.
- How will responsibilities for managing data activities be handled if substantive changes happen in the person in charge?
- What resources will you require to implement your data management plan? What do you estimate the overall cost to be?

**Ethics and Legal Compliance**

- If your research project includes sensitive data, how will you ensure that it is securely managed and accessed?
- If applicable, what strategies will you undertake to address secondary uses of sensitive data?
- How will you manage legal, ethical, and intellectual property issues?
This is an example of a plan exported in pdf. This is just an example – in a real plan, I would put much more detail in the answers!
Any Questions?
Acknowledgements

Portage Network
The Community of Practice for Ocean Data Management
Angela Grant (CHONe II)
MEOPAR
ACENET/Compute Canada
NSERC/SSHRC/CIHR for their slides

References


