Managing your research data
Resources, tools, strategies

Vivek Jadon (vivek@mcmaster.ca) & Jay Brodeur (brodeujj@mcmaster.ca)
McMaster University Library
FHS Graduate Student Workshop | 06-November, 2018
Overview

✧ A brief Research Data Management primer (10 mins)
✧ Data management planning (20 mins)
✧ Active data management
  ➢ Data storage & backup (10 mins)
  ➢ In-project data management & collaboration (15 mins)
✧ Data archiving & sharing (20 mins)
✧ Helpful data management resources (5 mins)
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 Management is the active organization & maintenance of data throughout the research cycle to ensure its security, accessibility, usability, and integrity.
The research data lifecycle
Research Data Management is the active organization & maintenance of data throughout the research data lifecycle to ensure its security, accessibility, usability, and integrity.

Applying RDM best practices will benefit...

<table>
<thead>
<tr>
<th>Researchers and their collaborators</th>
<th>Improves research efficiency and productivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Provides extra credit for research work</td>
</tr>
<tr>
<td></td>
<td>Increases research impact</td>
</tr>
<tr>
<td></td>
<td>May help to meet funding requirements</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Research Communities</th>
<th>Accelerates discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Enables validation and verification</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Funders, governments and the public</th>
<th>Improves return on investment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Increases research transparency</td>
</tr>
<tr>
<td></td>
<td>Data as a public good</td>
</tr>
</tbody>
</table>
Requirements and Expectations

Funding Agencies and Governments

Publisher and Research Communities

Institutions

- Office of Science and Technology Policy
- Australian Government, Australian Research Council
- RESEARCH COUNCILS UK
- European Commission
- Wellcome
- Genome Canada
- University of Prince Edward Island
- University of Oxford
- University of Edinburgh
- University of Sydney
- University of Alberta
Tri-Agency Statement of Principles on Digital Data Management

http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1
Tri-Agency Statement of Principles on Digital Data Management: Expectations

➢ Data management planning
➢ Constraints and obligations
➢ Adherence to standards
➢ Collection and storage
➢ Efficient and cost-effective

➢ Metadata
➢ Preservation, retention and sharing
➢ Timeliness
➢ Acknowledgement and citation

Tri-Agency Statement of Principles on Digital Data Management
http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1
Tri-Agency Statement of Principles on Digital Data Management: **Responsibilities**

**Researchers**
- incorporating best practices
- developing DMPs
- adhering to policies and standards

**Research Communities**
- developing & promoting standards
- fostering excellence
- selecting repositories

**Research Institutions**
- supporting best practices
- providing access to resources
- creating guidance and policies

**Research Funders**
- developing policies & guidance
- promoting data management
- providing peer reviewers
Draft Tri-Agency RDM Policy

Tri-Agency draft data management policy\(^1\) — June, 2018
✧ Applies to grant recipients and institutions administering tri-agency funds.
✧ June-Sep, 2018: consultation period; feedback to inform policy

3 pillars:
1. **Institutions**: Institutional Strategy
2. **Researchers**: Data Management Plans
3. **Researchers**: Data Deposit

Phased and incremental implementation

---

Planning for RDM

- Considerations
- DMP Assistant
A research Data Management Plan (DMP) should:
✧ Describe how you will manage data through all stages of your research
✧ Communicate a strategy for creating share-worthy data products

A good DMP will:
✧ Be completed at the time of study design
✧ Ensure compliance with policies / obligations
✧ Document and organize research activities
✧ Help identify support requirements
✧ *(Likely)* evolve with your study...
Portage DMP Assistant

- A web-based, bilingual data management planning tool.
- Available to all researchers in Canada.
- A guide for best practices in data stewardship.
- Exportable data management plans.

https://assistant.portagenetwork.ca/

https://www.youtube.com/watch?v=ZgLaJpJFehQ
Activity - Develop a DMP (~15 minutes)

Using the CARL Portage DMP Assistant:

✧ Navigate to assistant.portagenetwork.ca
✧ Create an account & login (check email for verification)
  ○ Or login with user: mcm.rdm.demo@gmail.com | pass: demopass
✧ Develop a DMP for a current, upcoming or prospective research project
✧ Browse through the DMP questions
✧ Complete the answerable questions in rough form (or internally)
✧ Note any confusing, problematic or difficult questions
Navigate to the DMP Assistant:
assistant.portagenetwork.ca

Sign in with your email / password

Click “Create plan”

Set organization to “McMaster University”
Select the “Portage” template
Enter relevant personal, organizational and study information

Expand boxes to complete questions for each section

Provide answers in the left-hand box
Use right-hand box for guidance/comments
Reflections and Discussion

What are your reflections on:

✧ The overall experience?

✧ The interface and its content?

What sections / questions were particularly challenging? Howso?

What kind of support or resources would help you complete this plan?
Planning

What are the stipulations in institutional, funder or publisher data policies to be followed?

What resources do you require to manage your data?

Who is responsible for data management and long-term stewardship?
In-project management

- Storage & backup solutions
- Managing files
- Managing collaboration
Things to consider...

What types of data (and how much) will you collect?

How will you organize, secure, and backup your data?

Are there ethical or commercial restrictions?

How will you describe your data so that others understand it?

How will you control access to your data?

How will you manage data versions?
Quiz time!

Go to menti.com and use the code: 45 20 33
Portage Network’s
“Good Enough RDM”

https://bit.ly/2zv54Eg
1. **Save your raw data in original format**

1.1 Don't overwrite your original data with a cleaned version.

1.2 Protect your original data by locking them or making them read-only.

1.3 Refer to this original data if things go wrong (as they often do).
2 Backup your data

2.1 Use the 3-2-1 rule: Save three copies of your data, on two different storage mediums, and one copy off site.

2.2 Do not backup or store sensitive data on a commercial cloud (Dropbox, Google Drive, etc.).
copies of your data

copies are on-hand (easily accessible)

- 1 “production” (working) copy
- 1 “production backup” copy

copy is in another location (“off-site”), with a trusted service provide
“Production” copy ➔ Where you work with the data

- PC, laptop, mobile device, etc.

“Production backup” copy ➔ Easily accessible (+ versioning?) backup

- External hard drive with backup software
- MacDrive (seafile): https://macdrive.mcmaster.ca
- MacDrop
- Dropbox, Google Drive, etc.
Considerations for “Archived” Backup

- Security (physical and electronic)
- Automation
- Availability (and time to recover)
- Versioning
- Integrity-checking and error correction
- Data storage (locational) requirements
- Cost
<table>
<thead>
<tr>
<th></th>
<th>RHPSCS - Backup Services</th>
<th>RHPSCS - Hosted Server Packages</th>
<th>MacDrive</th>
<th>Microsoft OneDrive / Teams</th>
<th>MacDrop</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Storage Quota</strong></td>
<td>1 TB; more available for a fee</td>
<td>1 TB; more available for fee</td>
<td>300 GB per account</td>
<td>1 TB per account; up to 5 TB by request</td>
<td>2 GB for free; more available for a fee.</td>
</tr>
<tr>
<td><strong>Rates / cost</strong></td>
<td>$500 / yr + one time set up fee ($125 / machine)</td>
<td>$500 - $4000 / yr</td>
<td>No cost to users</td>
<td>No cost to users</td>
<td>$10/month for 100GB</td>
</tr>
<tr>
<td>Additional space:</td>
<td>$300 / TB</td>
<td>Setup fee: $500 - $1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Restore services:</td>
<td>$125 / hour</td>
<td>Additional space: $450 / TB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Backups / versioning</strong></td>
<td>Nightly, 14-day rotating cycle; Restore services through RHPSCS</td>
<td>Nightly, 14-day rotating cycle; Restore services through RHPSCS</td>
<td>Ongoing real-time sync</td>
<td>Ongoing real-time sync</td>
<td>Ongoing real-time sync</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nextcloud sync service available.</td>
<td>4-month version history</td>
<td>Unlimited version history (?)</td>
<td>Versions are kept indefinitely so long as versions don’t exceed 50% of user’s quota</td>
</tr>
<tr>
<td><strong>Who can use this service?</strong></td>
<td>Any subscribing users or research group</td>
<td>Any subscribing users or research group</td>
<td>McMaster Faculty and Staff Graduate students can obtain zero-quota accounts</td>
<td>All McMaster faculty, staff and students</td>
<td>All FHS users</td>
</tr>
<tr>
<td><strong>Server location</strong></td>
<td>A.B. Bourns building</td>
<td>A.B. Bourns building</td>
<td>Replicated clusters in Gilmour Hall and JHE</td>
<td>OneDrive: Canadian servers Teams: Soon in Canadian servers only</td>
<td>Health Science Centre</td>
</tr>
<tr>
<td><strong>Other notes</strong></td>
<td></td>
<td></td>
<td>Supports encrypted libraries, file and directory sharing, Desktop client, web interface</td>
<td>Supports file and directory sharing, Desktop client, web interface</td>
<td></td>
</tr>
</tbody>
</table>

Access the matrix: [goo.gl/45iy38](https://goo.gl/45iy38)
3 Describe your data

3.1 **Machine Friendly:** Describe your dataset with a metadata standard for discovery.

3.2 **Human Friendly:** Describe your variables, so your colleagues will understand what you meant. Data without good metadata is useless. Give your variables clear names.
4

Process your data

4.1 Make each column a variable.
4.2 Make each row an observation.
4.3 Store units (e.g. kg or cm) as metadata (in their own column).
4.4 Document each step processing your data in a README file.
Open Science Framework

- Free, open source web application
- Connects and supports the research workflow
- Integrates with existing services (e.g. Google Drive, Dropbox, etc.)
- Supports collaborative workflows
- Facilitates data publishing to selected data repositories

Workshop demo: https://osf.io/mug5f
Archiving and Sharing

- FAIR principles
- Data repositories
- Scholars Portal Dataverse
5 Archive and preserve your data

5.1 Submit final data files to a repository assigning a persistent identifier (e.g. handles or DOIs).

5.2 Provide good metadata for your study so others could find it (use your discipline’s metadata standard, e.g. Darwin Core, DDI, etc.).
Things to consider...

How will your data products be stored in the long-term?
✧ How to ensure that it remains **integral** and **secure**?
✧ Who will assume long-term **responsibility** for your data?

How will others access your data products?
✧ What data (if any) can/should be shared? Who should have access?
✧ How will you manage legal, commercial & ethical constraints?

How to maximize credit for sharing your data?
✧ In which repository should you deposit your data?
✧ How to ensure that your data is **FAIR** *(findable, accessible, interoperable and reusable)*?
The FAIR Guiding Principles

**F1**: (meta)data have a globally unique and eternally persistent identifier
**F2**: data are described with rich metadata
**F3**: metadata clearly and explicitly includes the ID of the data it defines
**F4**: (meta)data are registered and indexed in a searchable resource

**A1**: (meta)data retrievable by their ID using a standardized protocol
**A1.1**: protocol is open, free and universally implementable
**A1.2**: protocol allows for AuthT/AuthZ where needed
**A2**: metadata is always accessible

**I1**: (meta)data use a formal, accessible, shared, broadly applicable language for knowledge rep.
**I2**: (meta)data use vocabularies that follow FAIR principles
**I3**: (meta)data include qualified references to other (meta)data

**R1**: (meta)data richly described with accurate and relevant attributes
**R2**: (meta)data released with a clear and accessible data usage license
**R3**: (meta)data associated with detailed provenance
**R4**: (meta)data meet domain-relevant community standards

---

Domain vs. General Repositories

**Domain-specific Repositories:** Archives of digital and/or digitized information related to an area of research. They can be broad in scope, such as ICPSR at University of Michigan, or more tightly focused, such as Milkulski Archive at Space Telescope (MAST) at the Space Telescope Science Institute.

**General Repositories:** The archives of digital and/or digitized information is not related to an area of research and hence it is general in nature. Scholars Portal Dataverse is an example of general repository.
re3data.org

Registry of research data repositories
Scholars Portal Dataverse

✧ A data repository for researchers at Ontario's universities -- **free** and **open** for all researchers in Canada

✧ An online platform to share, preserve, cite, explore and analyze research data.

✧ Allows researchers to control how they share their data.

✧ Supports data DOI registration through Datacite Canada.

[https://www.youtube.com/watch?v=UDFGqRY61fQ](https://www.youtube.com/watch?v=UDFGqRY61fQ)

[http://dataverse.scholarsportal.info](http://dataverse.scholarsportal.info)
Dataverses, Datasets, Files
Create your own (demo) Dataverse

- Scholars Portal Demo Dataverse site: https://demodv.scholarsportal.info
Sign-up & Institutional Affiliation
Institutional Dataverse

You can have as many Dataverses within your Institution Dataverse as you like!
Customize your Dataverse

- Customizable features:
  - Branding/logo
  - Featured Dataverses
  - Facets
  - Contact
Linking Datasets and Dataverses
Self-service Deposit Model

• Anyone with an account can publish to Dataverse

• SP Terms of Use covers removal of data if necessary

http://guides.scholarsportal.info/dataverse
Storage Size Limits

- Currently SP Dataverse supports
  - 2 GB file size upload
  - Will be increased to 10 GB in the near future
  - Unzipping, tabular ingest, processing intense
Dataset Management / File Versioning

The Growth of the Network Media Economy in Canada, 1984-2016


Version 2.0

Files (Added: 2); View Details

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Summary</th>
<th>Contributors</th>
<th>Published</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>Files (Added: 2); View Details</td>
<td>MacOdrum Library Cataloguing and Collections</td>
<td>December 13, 2017</td>
</tr>
<tr>
<td>1.0</td>
<td>This is the first published version.</td>
<td>MacOdrum Library Cataloguing and Collections, MacOdrum Library</td>
<td>December 7, 2017</td>
</tr>
</tbody>
</table>
Dataset Management / File Versioning

Upload → Draft Dataset → Published Dataset v1

Publish Version 1: Authors, Title, Year, DOI, Repository, V1

Note: A Published Dataset cannot be deleted (only deaccessioned, if legally needed).

Publish Version 1.1: small metadata change; citation doesn’t change.

Publish Version 2: File change (automatic); big metadata change; or citation changes.

Authors, Title, Year, DOI, Repository, UNF, V2
Dataset and File Permissions

• User/groups and roles
  • Assign permissions for collaborators, curators, file downloaders (access)

• Granular file-level permissions

• IP Group based permissions
Licensing

• Default to CC0 (Open Data)

• OR

• Custom “Data Usage Agreement”
Guestbooks

• Who is downloading my data?
  • User fills out guestbook form

• Owner downloads Guestbook report in Excel format

• Can be used to mediate access / approval of access to data
Metadata Standards

- DataCite, Dublin Core, DDI (Citation)
- DDI (Social Science)
- Virtual Observatory VOResource (Astronomy)
- ISA-Tab (Life Science)
Metadata Templates

• Setup templates available at the Dataverse level to reduce need to copy information
• Can be set to default template, or can be copied
Data Citation and Journal Cross-referencing

• Get a DOI / data citation before publishing data
  • DataCite Canada DOI minting and indexing
  • Cite data in a paper, access for review process

• ORCID ID field (coming soon... ORCID sign-in)

• Reference a publication that uses data
Widgets support for Research Promotion
Data Analysis and Visualization

• Data Explorer (cross tabulations)
Data Analysis and Visualization
Preservation – Archivematica Integration

- Free and open-source digital preservation system designed to maintain standards-based, long-term access to collection of digital objects
- Currently work is underway to support Archivematica integration with Dataverse
Helpful data management resources
Helpful resources

✧ CIHR’s RDM Learning Module (beta)
✧ Portage’s "Good Enough" Research Data Management Guide
✧ UBC Library DataGuide
✧ MANTRA Research Data Management Training
✧ Scholars Portal Dataverse User Guide
Thank You.

For more information:

RDM @McMaster

library.mcmaster.ca/services/rdm
rdmgmt@mcmaster.ca

McMaster Library’s RDM webpage and primary contact

portagenetwork.ca/

CARL Portage Network page: Access to RDM information and DMP Assistant