Managing your research data

Resources, tools, strategies

Vivek Jadon (vivek@mcmaster.ca) & Jay Brodeur (brodeujj@mcmaster.ca)

McMaster University Library
CRESS Research in Focus | 06-March, 2018
Overview

✧ A (very) brief Research Data Management primer (5 mins)
✧ Data management planning (20 mins)
✧ Data storage & backup (10 mins)
✧ In-project data management (10 mins)
✧ Data archiving & sharing (30 mins)
✧ Helpful data management resources (5 mins)
In our previous episode...
Research Data Management is the active organization & maintenance of data throughout the research data lifecycle to ensure its security, accessibility, usability, and integrity.
Tri-Agency Statement of Principles on Digital Data Management

http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1
Draft Tri-Agency RDM Policy

✧ Tri-Agency draft data management policy — expected April, 2018
✧ 6-month consultation period; feedback will inform policy

Proposed policy — 3 possible requirements:
1. **Institutions**: Institutional Strategy
2. **Researchers**: Data Management Plans
3. **Researchers**: Data Deposit $^{a,b,c,d,e,f,g}$

✧ Phased and incremental implementation

[Link to M. Lucas’ slides — Portage Network Event](#) (30-Jan, 2018)
Planning for RDM

- Considerations
- DMP Assistant
...it all starts with a (good) plan

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...
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?
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://www.youtube.com/watch?v=ziGlNpJf6hQ

https://assistant.portagenetwork.ca/
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
assistant.portagenetwork.ca

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 you 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?
In-project management

- Storage & backup solutions
- 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: 10 42 71
copies of your data

2 copies are on-hand (easily accessible)
- 1 “production” (working) copy
- 1 “production backup” copy

1 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
- 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>RHPCS - Backup Services</th>
<th>RHPCS - Hosted Server Packages</th>
<th>MacDrive</th>
<th>Microsoft OneDrive / Teams</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Storage Quota</strong></td>
<td>1 TB; more available for 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>
</tr>
<tr>
<td><strong>Rates / cost</strong></td>
<td>$500 / yr + one time set up fee (&lt;$125 / machine)</td>
<td>$500 - $4000 / yr Setup fee: $500 - $1000 Additional space: $450 / TB</td>
<td>No cost to users</td>
<td>No cost to users</td>
</tr>
<tr>
<td><strong>Backups / versioning</strong></td>
<td>Nightly, 14-day rotating cycle; Restore services through RHPCS</td>
<td>Nightly, 14-day rotating cycle; Restore services through RHPCS Nextcloud sync service available.</td>
<td>Ongoing real-time sync 4-month version history</td>
<td>Ongoing real-time sync Unlimited version history (?)</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>
</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><strong>OneDrive</strong>: Canadian servers <strong>Teams</strong>: Soon in Canadian servers only</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>
</tr>
<tr>
<td><strong>More info</strong></td>
<td><a href="rhpcs.mcmaster.ca/current-rates">rhpcs.mcmaster.ca/current-rates</a></td>
<td><a href="rhpcs.mcmaster.ca/current-rates">rhpcs.mcmaster.ca/current-rates</a></td>
<td><a href="https://goo.gl/AvRGwx">macdrive.mcmaster.ca</a></td>
<td><a href="portal.office.com/">portal.office.com/</a> <a href="mcmaster.ca/uts/licensing">Documentation: mcmaster.ca/uts/licensing</a></td>
</tr>
</tbody>
</table>

Access the matrix: [goo.gl/45iy38](goo.gl/45iy38)
Open Science Framework

https://osf.io/

✧ 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
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

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

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

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
Federated Research Data Repository (FRDR)

- Compute Canada + Canadian Association of Research Libraries (CARL)
- A scalable, federated platform for digital research data management and the discovery of Canadian research data
- National data repository
- Framework for federating existing and future data repositories

https://www.frdr.ca
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

http://dataverse.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
Dataverses, Datasets, Files
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
Dataset Management / File Versioning

1. **Upload** → **Draft Dataset**
   - **Publish Version 1**: Authors, Title, Year, DOI, Repository, V1
     - Small metadata change; citation doesn’t change.

2. **Published Dataset v1**
   - **Publish Version 1.1**: File change (automatic); big metadata change; or citation changes.
     - Authors, Title, Year, DOI, Repository, UNF, V2

3. **Published Dataset v1.1**
   - **Published Dataset v2**

**Note:** A Published Dataset cannot be deleted (only deaccessioned, if legally needed).
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
Create your own (demo) Dataverse

- Scholars Portal Demo Dataverse site: http://demodv.scholarsportal.info
Helpful data management resources

- Plan
- Create/Collect
- Share
- Preserve
- Analyse
Helpful resources

✧ 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/rdm
rdmgmt@mcmaster.ca

portagenetwork.ca/

McMaster Library’s RDM webpage and primary contact

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