Foundations of Research Data Management

(Les bases de la Gestion des Données de Recherche)

Marc-Olivier Croteau et Hélène Tardif
Bibliothécaires
Direction des bibliothèques
Outline

1- Get organized like a librarian!

2- Introduction to Research Data Management (RDM)

3- Data Management Plan (DMP)

4- Organize research data

5- Conclusion
Get organized like a librarian!

Excellence is mundane!

Superlative performance is really a confluence of dozens of small skills or activities, each one learned or stumbled upon, which have been carefully drilled into habit and then are fitted together in a synthesized whole (Chambliss, 1989).

Reference:
Get organized like a librarian!

1. Organize your literature review (with EndNote or Zotero or...)

2. Organize your data (from the beginning...)
   - Organize the contents of your files (good practices)
   - Organize your files and team collaboration
   - Organize data for publication and reuse

3. Organize your writing (using a thesis model + good practices)

4. Organize your future in research (manage your digital reputation as a researcher)
Get organized like a librarian!

Tool box to find everything in one place:

Organiser sa recherche

Dès le début de son parcours de chercheuse ou de chercheur, l’approche gagnante est une bonne organisation de son travail :

Préparer sa revue de littérature
- Faire de la veille informationnelle et utiliser EndNote ou Zotero
- Réaliser une synthèse des connaissances

Organiser ses données de recherche
- Rédiger son plan de gestion des données
  - Exemples de PGD
  - Modèle de plan en format MS Word
- Gérer ses fichiers
  - Bonnes pratiques : How to name files et File naming best practices
  - Outils de versionnage : GIT, SVN

Documenter ses données
- Dico de données : Data dictionaries, How to make a Data Dictionary
- Fichier README en français ou en anglais
- Schémas de métadonnées : Darwin Core / DwC-A, EML-XMD

Optimiser sa rédaction
- Modèle de thèse DOCX ou LaTeX
- Rédiger un article scientifique

Planifier son avenir en recherche
- Identifiants et profils de chercheurs (ORCID, etc.)

In this Libguide: bib.umontreal.ca/informatique-mathematique-sciences-nature/sciences-biologiques?tab=5237710
Outline

1- Get organized like a librarian!

2- Introduction to Research Data Management (RDM)

3- Data Management Plan (DMP)

4- Organize research data

5- Conclusion
Introduction to RDM

• What is research data?
  Data to produce or validate research results

• Several possible sources
  ▪ observation : e.g. field data
  ▪ experimentation : e.g. chemical analysis
  ▪ simulation : e.g. climate change modeling
  ▪ derived data : data calculated from other data
  ▪ metadata : data about data
Introduction to RDM

What is Research Data Management?

A set of actions taken during a research project that better organizes the storage, sharing, use, documentation and preservation of research data.
Introduction to RDM: The data lifecycle

Data management practices cover the entire lifecycle of the data:

❖ from planning the investigation to conducting it...

❖ ...and from backing up data as it is created and used to long term preservation of data deliverables after the research investigation has concluded.

Introduction to RDM

Why organizing your research data is important?

- A good organization facilitates research work,
- Ensures continuity in the course of research,
- Avoids the loss or compromise of data,
- Facilitates the validation of results,
- Facilitates the sharing and reuse of data (and collaboration with a view to open science).
Introduction to RDM

The granting agencies / Les organismes subventionnaires

They increasingly recognize the importance of data as a research output. Many organizations require submission of a Data Management Plan (DMP) and data sharing.

*The DMP is a management tool.* [...] Its objective is to summarize the description and the evolution of the datasets of your research project. It prepares the sharing, reuse and sustainability of data.
Introduction to RDM

The policy includes suggested requirements related to three primary areas:
Introduction to RDM

Institutional strategy

• Published in February 2023
• Establishes units responsible for specific issues

https://recherche.umontreal.ca/vrrdci/gestion-des-donnees-de-recherche-la-strategie-institutionnelle/
Outline

1- Get organized like a librarian!

2- Introduction to Research Data Management (RDM)

3- Data Management Plan (DMP)

4- Organize research data

5- Conclusion
EDITORIAL  •  13 MARCH 2018

Everyone needs a data-management plan

They sound dull, but data-management plans are essential, and funders must explain why.
Data Management Plan (DMP)

Organize data at each stage of the data lifecycle!

- Tool for research project management
- Start writing at the beginning of the research cycle
- Update during research
- Goals = reflection, collaboration, transparency...
- Risk management
- Grant application
Data Management Plan (DMP)

Components of a DMP:

1. Data collection > data types, file formats, naming and version control
2. Documentation > ensure data can be read and interpreted
3. Data storage and backup throughout the research
4. Data preservation strategy for long-term access
5. Provisions for sharing and reuse
6. Sharing of responsibilities and necessary resources
7. Ethical and legal compliance

Tool available: UdeM MS Word form
Outline

1- Get organized like a librarian!

2- Introduction to Research Data Management (RDM)

3- Data Management Plan (DMP)

4- Organize research data

5- Conclusion
Organize research data

- Organize data for publication and reuse
- Organize the contents of your files
- Organize your files and team collaboration
Organize research data: Data collection

Organize the contents of your files:

- The goal of data entry is to create data that is valid or has gone through a quality assurance process...
- ...and which are organized in such a way as to facilitate their use or archiving.
- Commonly used tools: Excel, R, Google Docs forms, Open Office.
Organize from the start of taking or obtaining data

### Organize research data: Data collection

<table>
<thead>
<tr>
<th>Site_ID</th>
<th>Year</th>
<th>Month</th>
<th>Site_type</th>
<th>Reservoir_number</th>
<th>System_type</th>
<th>location_sampled</th>
<th>Latitude_GPS</th>
<th>Longitude_GPS</th>
<th>MeHg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW33</td>
<td>2016</td>
<td>Aug</td>
<td>natural</td>
<td>natural</td>
<td>GW</td>
<td>groundwater</td>
<td>51.36618</td>
<td>-63.87583</td>
<td>0.021072679</td>
</tr>
<tr>
<td>SW35</td>
<td>2016</td>
<td>Aug</td>
<td>natural</td>
<td>natural</td>
<td>GW</td>
<td>groundwater</td>
<td>51.21284</td>
<td>-63.92261</td>
<td>0.080348182</td>
</tr>
<tr>
<td>SW40</td>
<td>2016</td>
<td>Aug</td>
<td>natural</td>
<td>natural</td>
<td>GW</td>
<td>groundwater</td>
<td>51.36618</td>
<td>-63.87583</td>
<td>0.021072679</td>
</tr>
<tr>
<td>L001</td>
<td>2016</td>
<td>Aug</td>
<td>natural</td>
<td>natural</td>
<td>lake</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L001-K-epi</td>
<td>2016</td>
<td>Aug</td>
<td>natural</td>
<td>lake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L001-K-hypo</td>
<td>2016</td>
<td>Aug</td>
<td>natural</td>
<td>lake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L002</td>
<td>2016</td>
<td>Aug</td>
<td>natural</td>
<td>natural</td>
<td>lake</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L002-K-hypo</td>
<td>2016</td>
<td>Aug</td>
<td>natural</td>
<td>lake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR1</td>
<td>2016</td>
<td>Aug</td>
<td>natural</td>
<td>natural</td>
<td>LR_DS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR2</td>
<td>2016</td>
<td>Aug</td>
<td>natural</td>
<td>natural</td>
<td>LR_DS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR3</td>
<td>2016</td>
<td>Aug</td>
<td>natural</td>
<td>natural</td>
<td>LR_DS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R01-07</td>
<td>2016</td>
<td>Aug</td>
<td>reservoir</td>
<td>reservoir</td>
<td>lake</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R01-07-HK</td>
<td>2016</td>
<td>Aug</td>
<td>reservoir</td>
<td>lake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R01-08</td>
<td>2016</td>
<td>Aug</td>
<td>reservoir</td>
<td>reservoir</td>
<td>main_channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R01-08-A</td>
<td>2016</td>
<td>Aug</td>
<td>reservoir</td>
<td>main_channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R01-09</td>
<td>2016</td>
<td>Aug</td>
<td>reservoir</td>
<td>reservoir</td>
<td>main_channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R02-09-HK</td>
<td>2016</td>
<td>Aug</td>
<td>reservoir</td>
<td>main_channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R02-46</td>
<td>2016</td>
<td>Aug</td>
<td>reservoir</td>
<td>reservoir</td>
<td>main_channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R02-46-3</td>
<td>2016</td>
<td>Aug</td>
<td>reservoir</td>
<td>main_channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R02-46-4</td>
<td>2016</td>
<td>Aug</td>
<td>reservoir</td>
<td>main_channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R02-48</td>
<td>2016</td>
<td>Aug</td>
<td>reservoir</td>
<td>reservoir</td>
<td>main_channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Metadata information

**Title of dataset**: Mercury concentrations and water chemical variables measured in La Romaine hydroelectric reservoir samples in the summers of 2016, 2017 and 2018.

**URL of dataset**: DOI link is inserted here (once on Zenodo).

**Abstract**: This dataset includes chemical and physical data measured in situ in water in natural and recently dammed portions of the La Romaine River watershed in Northern Quebec, Canada. Samples were collected from 2016 to 2018 in August, as well as in June 2017. For most sites, the samples were collected close to the water surface (if it was deep) with a peristaltic pump following the clean hands-dirty hands sampling protocol to avoid any contamination by trace metals. Filtered samples were collected using an in-line Whatman 0.45 μm filtration capsule attached to the tubing. For June 2017, physicochemical variables collected with a YSI multiprobe, greenhouse gas partial pressure (methane and carbon dioxide), total nutrient concentrations and dissolved metals (iron and manganese) were also included. We used this dataset to explore the distribution in time and space of various forms of mercury and pools of carbon across the dammed watershed of La Romaine and in different aquatic systems (e.g. tributaries, lakes, groundwater, river and reservoir sites).

**Keywords**: Mercury, methylmercury, dissolved organic carbon, PARAFAC, metals, hydroelectric reservoir

**Dataset lead author**: Parent student

**Position of data author**: Parent student

**Address of data author**: Département de sciences biologiques, Université de Montréal, Complexe des sciences, 1315 Avenue Thérèse-Lavoie-Rose, Montréal, QC, H3V 0B3, Canada

**Email address of data author**: [Email address]

**Primary contact person for dataset**: [Contact person name]
Organize research data: Documentation

Organize for publication and reuse:

- Build documentation throughout the project.
  - Codebook
  - README File
  - Source codes (scripts that will read and process the data)
  - Metadata (which standard to use?)
Organize research data: Documentation

Metadata

*Data that provides information on the nature of certain other data in order to facilitate their understanding and management.*

In other words: Data about data

For humans and for machines!
Roselin familier – Génétique

Il s’agit d’une séquence nucléotidique de Carpodacus mexicanus (clone 6b).
(A = Adénine, G = Guanine, C = Cytosine, T = Thymine : bases d’acide nucléique).

> Seq1 [organism=Carpodacus mexicanus] C.mexicanus clone 6b actin (act) mRNA, partial cds

CCTTTATCTTAATCTTTGGACATGAGCTGGCATAGTTGGAACCAGCCCTCAGCCTCTCATCCGTGCAGA
CTTGGGAACCTGGAATCTCTCTCTGAGACGGACGACAAATAATTAATGGAATCTACTGACACTGGCCCACGCCTCTCG
TAATATTTTCTTATAGTAATACCAATCTAGATCGTGTTTCGGGAAACTGAGACTAGTCCCCACTGATAT
CGGGCCCGCCCGACATACATCTTCCCCGATATAAAAAACATAGCCTCTGACTACTTCCCCCATATTCTT
TTACTTCTAGATCCCTCCACAGTGAAGCTGGAGCGAGAAGGTTGAAAGCAGTATAGCCCTCTCTCGTG
GTAACCTAGCCCATGCGGTGTCTGACTAGACCTAGCATTTTCCTCCCTACCTTAGGAGGTGTTTCTCTC
TATCTAGGGTGCTATTACCTTTTATACAAACCAGGCAACATACAAAAACCCCACACCCCTCTCCCAATAACAA
ACCCCCCTATCTGATGATTACGCTTTATTACGCGGTCCCTTTCTCTACTCTCTCTTTCCAGTCTCCGTG
CTGGCATATTACCTATAACTACAAAGACCGGAAACCTAAGACACTAGCTTTGACCCAGCTGGAGGAGAGA
CCCAGTCCCCGATCAAACACCTTCTGATTCTGCCTGCTGCTGGCATATTATACTACTAACAGACCGAAACCTAAGACACTAGCTTTGACCCAGCTGGAGGAGAGA
Organize research data: Data storage and backup

Data horror stories ...

Have you ever lost or not been able to access data for any of these reasons?

- My lab was destroyed by fire/flood/earthquake
- My laptop was lost/stolen
- There was a hardware failure
- The file was corrupt
- I couldn’t find software to open the file
- I couldn’t find a computer to open the storage device

Organize research data: Data storage and backup

Organize files and team collaboration during research:

- Evaluate how much space is needed and who needs access
- Agree on a naming convention and consider using a versioning tool (like GitHub)
- Regular backup and if possible automated
- 3-2-1 backup rule (3 copies of your data, on 2 different media and 1 backup copy offsite)
- Use an electronic Lab Notebook (ELN) to promote the traceability, publication and sharing of research data.
Organize research data: Sharing and reuse

• Sharing data

• From a FAIR perspective

• As mentioned in the Policy

“to provide appropriate access to the data where ethical, cultural, legal and commercial requirements allow, and in accordance with the FAIR principles and the standards of their disciplines.”

Tri-Agency Research Data Management Policy
From data collection to publication

**FAIR** principles (FAIR Data)
Organize research data: Sharing and reuse

Data repository / Dépôt de données

• Space for publishing research results.

• Often provides a unique and permanent identifier, such as the DOI.

• Rarely for storage and sharing between collaborators during research.

• Borealis UdeM (for deposit and distribution with or without restriction).
Extra

Some attractions of the Bibliothèque des sciences on the Campus MIL

Salle d’enregistrement sonore

Atelier de fabrication numérique

Studio d’édition numérique
Conclusion

Librarians are here to help!

• Consulting service with the Data Management Plan.
• Assistance in the choice of a data repository.
• UdeM Dataverse (Borealis):
  • Assistance in using
  • Enhancement of descriptive metadata;
  • Support for data curation;
  • Support for standardization of data during research