(Some) Research Data Management Best Practices

James Doiron
Research Data Management Services Coordinator, University of Alberta Library
Academic Director, UofA Research Data Centre
Co-Chair, Portage DMP Expert Group

Jane Fry
Data Services Librarian, Carleton University
Co-Chair, Portage RDM Training Expert Group

September 16, 2020
RDM ‘Best Practices’: A bird’s eye view

1. Overarching principles: FAIR; CARE; OCAP

2. Pre-Research:
- Funder requirements
- Including RDM into funding applications
- Data Management Plans
- Participant consent & Information Letters
- Ethics applications

3. During Research:
- Primary data collection
- Data storage
- Data transferring
- Data access
- Data documentation/metadata
- File naming conventions

4. Post-Research:
- Publication requirements
- Data deposit

1. Overarching Principles

**FAIR**

**FAIR** is a set of guiding principles focused towards making data **Findable, Accessible, Interoperable and Reusable:**

- **Findable** - Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

- **Accessible** - Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

- **Interoperable** - Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

- **Reusable** - Data and collections have clear usage licenses and provide accurate information on provenance.

FAIR Principles: Key Resource

go-fair.org (https://www.go-fair.org/fair-principles/)

- Detailed information across the FAIR principles
- Implementation Networks
- News
- Event
- Resources!
1. Overarching Principles

CARE

CARE is a set of guiding principles for Indigenous data governance:

- **Collective benefit** for inclusive development and innovation, improved governance and citizen engagement, and equitable outcomes

- **Authority to control** - Recognizing rights and interests, data for governance, and governance of data

- **Responsibility** for positive relationships, expanding capability and capacity, and Indigenous languages and worldviews

- **Ethics** for minimizing harm and maximizing benefit, justice, and future use of data

*Key Readings available at: https://www.gida-global.org/resources*
CARE Principles: Key Resource

* gida-global.org/care*

- Detailed information across the CARE principles
- Foundational readings & publications
- News
- Events
- Resources!

---

The CARE Principles for Indigenous Data Governance can be downloaded here in summary or full.

The current movement towards open data and open science does not fully engage with Indigenous Peoples' rights and interests. Existing principles within the open data movement (e.g., FAIR: findable, accessible, interoperable, reusable) primarily focus on characteristics of data that will facilitate increased data sharing among entities while ignoring power differentials and historical contexts. The emphasis on greater data sharing alone creates a tension for Indigenous Peoples who are also asserting greater control over the application and use of Indigenous data and Indigenous Knowledge for collective benefit.

This includes the right to create value from Indigenous data in ways that are grounded in Indigenous worldviews and real opportunities within the knowledge economy. The CARE Principles for Indigenous Data Governance are people and purpose-oriented, reflecting the crucial role of data in advancing Indigenous innovation and self-determination. These principles complement the existing FAIR principles encouraging open and other data movements to consider both people and purpose in their advocacy and pursuits.
The First Nation Principles of OCAP are a set of standards that establish how First Nations data should be collected, protected, used or shared:

- **Ownership:** refers to the relationship of First Nations to their cultural knowledge, data & information - a **community/group collectively owns information** in the same way that an individual owns his/her personal information.

- **Control:** affirms that **First Nations communities have rights in seeking control over all aspects of research** - from start to finish - that impact them. This extends to control of resources and review processes and management of information.

- **Access:** First Nations must have access to information and data about themselves and their communities regardless of where it is held, and have the right to manage and make decisions regarding access to their collective information.

- **Possession:** Refers to the physical control of data - the mechanism by which ownership can be asserted and protected.

“OCAP® is a registered trademark of the First Nations Information Governance Centre (FNIGC)”

www.FNIGC.ca/OCAP
OCAP Principles: Key Resource

First Nations Information Governance Centre
(https://www.FNIGC.ca/)

- Fundamentals of OCAP online training program
- FNIGC data online
- First Nations Data Centre (data by request)
- First Nations surveys (i.e., regional health, early childhood, education, labour, oral health)
- FNIGC online library
2. Pre-Research

RDM Funder requirements

Awareness of funder requirements helps to identify:

- specific supports needed;
- collaborative opportunities;
- RDM supports to leverage.

When involved in funded research projects, know who the funder is and:

1) What general RDM related policies they may have; and

2) If there are any RDM related requirements pertaining to the specific call for funding.
2. Pre-Research

General RDM Funder requirements

Example:
‘Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans’ (TCPS-2)
2. Pre-Research

Specific RDM Funder requirements

Example: CRAft Digital Research Archive grants

- Research data management plan with a focus on data accessibility and stewardship. ***Please note that the RDM plan is not considered part of the 5-page proposal and should be included as an attachment. We recommend using the Portage DMP Assistant to generate an RDM plan.***
2. Pre-Research

Including RDM into funding applications

Including RDM within funding applications can help to strengthen funding applications by identifying:

- areas where essential RDM support may be needed
- collaborative opportunities before the research begins
- specific RDM supports that research projects may leverage

2. Pre-Research

Including RDM into funding applications

Example RDM statement:

**Research Data Management:** The University of Alberta Libraries system will provide research data management training and support for project researchers on a one-on-one and group basis (including HQP), host project research data in Dataverse, UAlberta’s data repository, and host project papers and publications, learning objects, digital images, etc. in its open-access Education & Research Archive.[i] Data management will extend beyond the project itself to ensure sustainability of the data for future researchers. We will form a Research Data Management Committee (reporting to the EC; see Governance) and use the Portage DMP Assistant[ii], a web-based open source application, to develop a Research Partnership Data Sharing Agreement at the beginning of the formal partnership. One of the major outcomes of this project will be well-documented, well-preserved **data sets** which can be used by future researchers and are themselves a form of scholarship.

*Special thanks to Dr. Carla Peck from the Faculty of Education, UofA, for permission to use this text from her recently successful application for SSHRC Partnership funding - “Thinking Historically for Canada’s Future”*
2. Pre-Research

What *is* a data management plan (DMP)?

A DMP:

- Is a formal document which clearly articulates the strategies and tools you will implement to effectively manage your data.

- Speaks to the management of data both *during* the active phases of your research and *after* the completion of the research project.

The objective of a DMP is to address issues related to data management prior to starting your research project!
A DMP provides information across key research lifecycle categories:

- Data Collection
- Documentation & Metadata
- Storage & Backup
- Preservation
- Responsibilities & Resources
- Sharing & Reuse
- Ethics & Legal Compliance

Diagram:

- DMP
- Data Collection
- Preservation
- Responsibilities & Resources
- Sharing & Reuse
- Ethics & Legal Compliance
- Storage & Backup
- Documentation & Metadata
Creating an effective DMP (English/French)

Exemplar DMPs:
- #1: Digital Humanities (English/French)
- #2: Digital Humanities and Secondary Data (English/French)
- #3: Mixed Methods (English/French)
Short tutorial video

https://libcasts.library.dal.ca/Portage/DMP_Assistant/
2. Pre-Research

Participant consent & Information letters

Research involving human participants requires informed consent.

Information letters must describe *how data are handled during active phases and beyond*.

What can/can’t be done will in part be determined by what is said in:

- Participant information letters/consent forms
- Approved ethics applications

**NOTE:** It can be very difficult, or even impossible, to go back to participants to revise their consent, so getting things right at the start of your research project if important!
2. Pre-Research

Participant consent & Information letters

Outline such things as:

- project background
- purpose of the study
- study procedures
- benefits/risks
- data preservation/destruction
- security/confidentiality
- voluntary participation
- freedom to withdraw

*FUTURE USE OF DATA:

Participant consent is required in order for data to be used beyond the scope of the immediate project.

Example statement:

“By participating in this research I hereby give consent for my *de-identified* information to be used for research purposes beyond this immediate project.”
2. Pre-Research

Ethics applications

An ethics application addresses such things as:

- Research design/methodology, risks/benefits, security/confidentiality, participant information, informed consent, data sensitivity, data collection, & data storage, retention and disposal.

Most research projects involving human participants require ethics approval.

Multi-institutional/regional projects require multiple ethics approvals.

Your institutional ethics office offers essential supports & services to help guide and support the ethical management of your research data!
2. Pre-Research

Ethics applications

*FUTURE USE OF DATA:

If there is potential for future use of data outside of the immediate project then this should be clearly stated within the ethics application.

Example statement:

“There are no plans to destroy these data. Data will be securely stored *enter details of storage methods’ *for *i.e., the minimum 5 years*."

*De-identified* data may be deposited into an institutional repository for discovery and possible repurposing. Any future use of these data outside of the immediate research project will occur only with all ethical and contractual obligations met.”
3. During Research

Data collection: immediate storage

A safe definition of ‘storage’ = 48+ hours

Storing data on mobile devices is **not** considered best practice.

Any electronic devices used for collecting/storing data should always be encrypted (*i.e.*, laptops, digital voice recorders, tablets, etc).

**Best Practice:**

- Clear & succinct data collection policies and protocols *define when and how data are transferred off of data collection devices*.
- Be aware of any data storage policies imposed by institutions, funders, data providers, etc.
3. During Research

Data collection: immediate storage

Definition of ‘storage’?
- Safe: 48+ hours

Storing data on mobile devices is not considered best practice

Any electronic devices used for collecting/storing data should always be encrypted (i.e., laptops, digital voice recorders, tablets, etc)

Best Practice:

- Clear & succinct data collection policies and protocols that define when and how data are transferred off of data collection devices
- Be aware of any data storage policies imposed by institutions, funders, data providers, etc.

$1M study at U of M paused after personal health info of 420 participants is breached

Study was led by Peter Jones, who was suspended by the university last year. It informed them their health information was “not handled, stored or secured properly.” According to the letter from the university’s access and privacy office, an audit of the program found data was being stored off-site, without encryption, allowing the potential for it to be accessed by a third-party company.

The announcement of the PHI breach means all the tests and data from the 420 participants will be destroyed and will no longer be used in the study.
3. During Research

Data collection: ‘longer term’ immediate storage

Sometimes it is necessary to store data on electronic devices - i.e., laptops, portable hard drives - for longer periods (collecting data in rural/remote areas).

Beyond security risks, these also introduce risk of data loss and/or corruption.

Best Practice:

*If* the use of laptops/desktops/hard drives is deemed necessary for longer term data storage, use the 3-2-1 rule:

*At least 3 independent copies of your data:*
*store copies on 2 different types of media; and keep 1 backup copy offsite.*
3. During Research

**Data transferring**

Transferring of data is a critical stage of the data collection process.

Regardless of whether data are collected from primary or secondary sources, transferring of data is a necessity.

**Some Risks:**

Data transfers may occur:

- from field (real world settings)
- from data providers
- between researchers
- between researchers & stakeholders
3. During Research

Data transferring

Transferring of data is a critical stage of the data collection process. Regardless of whether data are collected from primary or secondary sources, transferring of data is a necessity.

Some Risks:

- Loss of data
- Unintended copies
- Unintended recipients

WRONG ONE
3. During Research

Data transferring: Risks

Definite don’ts: e-mail, unencrypted devices

Typical do’s: Secure FTP; Securely supported MS Sharepoint; secure extranets

Best practice:
- Identify data transfer methods that you will use before the research begins.
- Talk to your local IT support to identify secure methods available.
3. During Research

Data storage: Cloud services

What is ‘Cloud Storage’?

- Physical storage typically spans multiple servers (sometimes in different locations).
- Data are easily available and remotely accessible → typically 24/7.
- Best practice → built in physical, technical, & administrative safeguards.
3. During Research: **Cloud Services**

*Do you know where your cloud is?*

‘Clouds’ are clusters of servers → servers need to live somewhere.

**Find out:**
- Where servers are physically located before using a cloud service → local, provincial, Canada?
- What security policies and procedures are in place → disaster recovery, back-ups, etc.
● **Rapid Access Service**
  → Freely available cloud space for Canadian researchers

● **Resource allocation competitions**
  → for greater needs

● **High performance computing**

● **Big data transfers** → i.e., TBs of data

● **Portals for specialized software & tools**

● **Data storage & back-up**
3. During Research: **Data documentation & metadata**

‘Metadata’ essentially refers to ‘data about your data’.

Descriptive information that describes your data, as well as to help others (and machines) to locate it and make it readable and useable.

**Some key examples of metadata documentation can include:**

- Data codebooks/dictionaries
- Data management and processing protocols
- Readme files
- Analytic plans
- Code
3. During Research: **File naming**

Having clear and standardized file naming helps to support:

- Organization
- Quality assurance
- File versioning
- Collaborative use
- Data analysis
- Dissemination
- Preservation & archiving
- Staff/Student training

**Elements of file names can include:**

- Project name/acronym
- File version
- Data type
- Participant codes/pseudonyms
- Geographic location
- Context information
- Date information
- Interviewer codes/initials
File versioning is an important component of research - it supports such things as participant confidentiality, organization, work efficiency, quality control, analysis, ...

As data are processed (cleaned), new ‘versions’ are created - from raw data, to the versions which will be used for analysis, and beyond.

Qualitative research example:

‘Raw’ audio data = the original digital audio recording

‘Raw’ transcript data = the original and unaltered transcript (text)

‘Master’ transcript data = the processed transcript → e.g., further de-identified, interviewer comments, typos fixed, etc.

‘Analytic’ transcript = working copy used for analysis/importing into analytic software
### Transcript File Naming Convention Table

<table>
<thead>
<tr>
<th>File Version</th>
<th>Illness</th>
<th>Data Type</th>
<th>Dyad Number</th>
<th>Dyad</th>
<th>Interview / FN/ Photo TYPE</th>
<th>Interviewer (first and last initial)</th>
<th>Locatio n (first three letters of city)</th>
<th>Day</th>
<th>Month</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>HF=Heart Failure</td>
<td>IV = Audio Interview</td>
<td>01-20=HF</td>
<td>PA=Patient</td>
<td>1= 1st in-person interview / 1st photo</td>
<td>Joanna=JC</td>
<td>EDM=E dmonton</td>
<td>Two digits</td>
<td>First three letters</td>
<td>Use four Digits</td>
</tr>
<tr>
<td>Master</td>
<td>LD = COPD/Lung Disease</td>
<td>IVT = Interview Transcribed</td>
<td>21-40=LD</td>
<td>FA=Family</td>
<td>2=2nd in-person interview / second photo</td>
<td>LW=Lade</td>
<td>VIC= Victoria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Draft (use when cleaning transcript)</td>
<td>RD=End Stage Renal Disease</td>
<td>FN=Field Notes</td>
<td>41-60=RD</td>
<td></td>
<td>3=Phone interview / third photo</td>
<td>LD=Lindsay</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Draft (use when cleaning transcript)</td>
<td>CA = Advanced Cancer</td>
<td>PH=Photo</td>
<td>61-80=CA</td>
<td></td>
<td>OTH#= Other (e.g. field note during intake call)</td>
<td>LC=Lynn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**EXAMPLE 1:** Raw_CA_FN_01_FA_OTH1_LD_VIC_15Apr2016

- **Raw Transcript**
- **Advanced Cancer**
- **Field note**
- **Dyad 01**
- **Family Member**
- **Field note made at a time not specific to an interview (e.g. dropping off the camera)**
- **Lindsay**
- **Victoria**
- **April 15, 2016**

**EXAMPLE 2:** MASTER_LD_IVT_39_PA_01_MA_VIC_15DEC2015

- **Master Transcript**
- **COPD**
- **Interview**
- **Dyad 01**
- **Participant**
- **First interview**
- **Marcy**
- **Victoria**
- **December 15, 2015**

4. Post-Research: Publication requirements

- Increasingly, journals are requiring research data to be made openly accessible.

- Talk to researchers early about journals to which they may be interested in submitting their articles.

- Find out what the journals’ data policies are.

- Most journals requiring data to be made openly available will have exceptions for data with legal and/or ethical considerations.

- Refusal to share data are grounds for rejection.

EXAMPLE: **PLOS ONE**

**PLOS One** is a peer-reviewed open access scientific journal published by the Public Library of Science (**PLOS**) since 2006. The journal covers primary research from any discipline within science and medicine.
What is a ‘research data repository’?

A research data repository is a technology-based platform that allows for research data to be:

- Deposited & described
- Stored & archived
- Shared & published
- Discovered & reused

There are different types of repositories including:

- Proprietary (paid for services)
- Open source (free to use)
- Discipline specific

RESOURCE! → Re3data.org is an online registry of data repositories, which can be searched according to subject, content type and country. Find a list of Canadian research data repositories.
**Dataverse Features**

- Digital Object Identifier (DOI) assigned
- User controlled data access
  - Many types of data
  - Built in data citations
  - Usage metrics
- File Versioning
- Customized Terms of Use
- **Portage Training Materials in development**

**Persistent identifier assigned when upload data**

- Automatically registered with DataCite
  - [https://www.datacite.org/](https://www.datacite.org/)

**From fully open to restricted access**

- Tabular (CSV, SPSS, R, etc)
- Documentation (pdf, doc, text)
- Geospatial data
- Multimedia (audio-visual)

**Madueke, Ijeoma Sylvia, 2018, "NIGERIAN LITERATURE IN FRENCH TRANSLATION (NILIFT)", [https://doi.org/10.7939/DVN/CHNOHA](https://doi.org/10.7939/DVN/CHNOHA)**

- Including a customizable guestbook
- Retain previous versions and metadata
- Create Terms of Use for data

**High level as well as in-depth web based training modules**
Why should I consider depositing my data?

- Helps you to meet both funding and journal requirements.
- Increases the impact and visibility of your research.
- Digital Object Identifier (DOI) = your data are discoverable and citable.
- Your data are made available, as appropriate, to others → supporting science.
- Obtain metrics on how your data are being used.
What is a Digital Object Identifier?

A digital object identifier (DOI) is a unique persistent identifier assigned by a registration agency to identify digital content and provide a persistent link to its location.

Anatomy of a DOI:

https://doi.org/10.7939/DVN/10591

- registration agency
- prefix (assigning body)
- suffix (resource)
4. Post research: DOI

**What does a DOI do?**

**DOIs** help data producers take control of the management of their research in that they:

- Support the discovery & appropriate sharing of data;
- Support data producers in receiving credit for their data through data citations;
- Make research data easier to access, reuse and verify; and
- Help to meet funder requirements → data deposit and assignment of DOIs are becoming increasingly expected, or even required, in academia.
4. Post Research: **DOIs**

**DOIs - supporting data usage metrics**

Without DOIs our understanding of data usage and impact is extremely limited:

- # of downloads → does not tell us how data are used
- Possible citations, but not easily tracked

A DOI supports data usage metrics by providing:

- Persistent identifier to the data
- Ready-made citation which includes the DOI
- Bi-directional discovery → data to outputs and back to data
- Ability to ‘scrape’ the web using the DOI
RDM ‘Best Practices’: A bird’s eye view

1. Overarching principles: FAIR; CARE; OCAP

2. Pre-Research:
   - Funder requirements
   - Including RDM into funding applications
   - Data Management Plans
   - Participant consent & Information Letters
   - Ethics applications

3. During Research:
   - Primary data collection
   - Data storage
   - Data transferring
   - Data access
   - Data documentation/metadata
   - File naming conventions

4. Post-Research:
   - Publication requirements
   - Data deposit
Questions & Discussion

James Doiron
Research Data Management Services Coordinator, University of Alberta Library
Academic Director, UofA Research Data Centre
Co-Chair, Portage DMP Expert Group
jdoiron@ualberta.ca

Jane Fry
Data Services Librarian, Carleton University
Co-Chair, Portage RDM Training Expert Group
Jane.Fry@carleton.ca