Research Data Management at McMaster

Offord Centre Lunch & Learn: May 18, 2016

V. Jadon, J. Brodeur
McMaster University Library

McMaster University LIBRARY

MAPS DATA GIS

Overview

1. Introduction to **R**esearch **D**ata **M**anagement (RDM)

2. Data management planning

3. Storage & backup

4. Preservation & sharing

If you shared your data another researcher or collaborator, would they be able to:

a. Interpret and understand it?
b. Use it in new analyses?

Would someone (including you) be able to find, interpret and use your data 20 years from now?

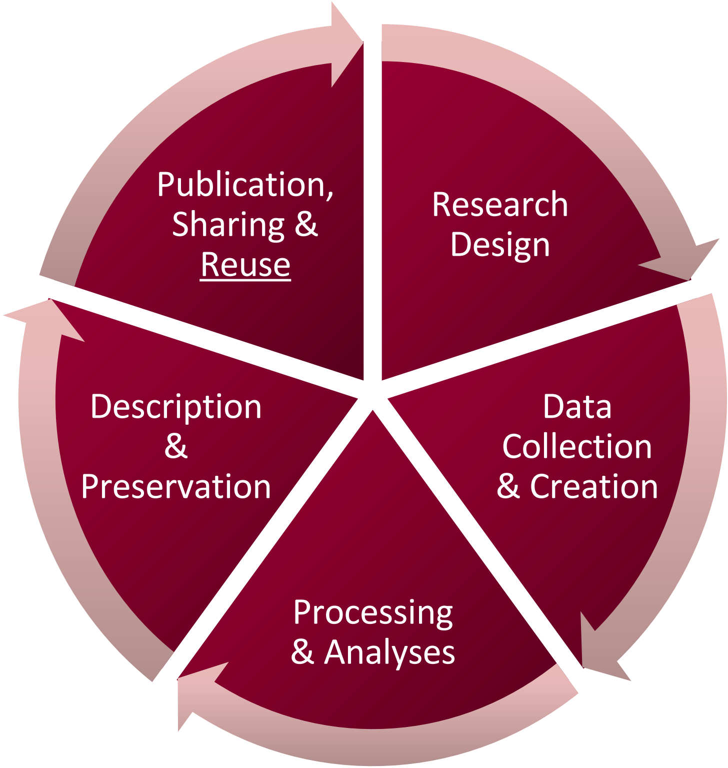
Research Data Management Primer: RDM in the Canadian Context

Research Data Management is…

… the active organization & maintenance of data

… the application of best practices to ensure data security, accessibility, usability, and integrity

… a set of activities resulting in self-describing data sets that can be discovered and reused.


Research Design
Data Collection & Creation
Processing & Analyses
Description & Preservation
Publication, Sharing & Reuse

Applying RDM best practices will benefit…

Researchers and their collaborators

- Improves research efficiency and productivity
- Provides extra credit for research work
- Increases research impact
- (May) help to meet funding requirements

Research Communities

- Accelerates discovery
- Enables validation and verification

Funders, governments and the public

- Improves return on investment
- Increases research transparency
- Data as a public good

Canadian Government & Funding Agencies

- Canada's Action Plan on Open Government (2014-2016)
- Tri-Agency Open Access Policy on Publications (Feb-2015)
- Tri-Agency Statement of Principles on Digital Data Management (Jul-2015)
- Comprehensive Brief on Research Data Management Policies (Aug-2015)

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

Canada Foundation for Innovation
Fondation canadienne pour l'innovation

SSHRC CRSH

NSERC
CRSNG

Tri-Agency Statement of Principles on Digital Data Management

Expectations:

- ➢ Data management planning
- ➢ Constraints and obligations
- ➢ Collection and storage
- ➢ Metadata
- ➢ Preservation, Retention and Sharing
- ➢ Timeliness
- ➢ Acknowledgement and Citation
- ➢ Efficient and cost-effective

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

Canada Foundation for Innovation
Fondation canadienne pour l'innovation

SSHRC CRSH

NSERC CRSNG

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

Publishers & Scientific Organizations

- Data sharing policies
- Recommended data repositories
- Publisher-supported data repositories

Science
AAAS

PLoS one
www.plosone.org

MENDELEY DATA

figshare
credit for all your research

nature

Support Organizations and Communities

RDA
RESEARCH DATA ALLIANCE

- Standards of practice
- Training, expertise and information
- Tools and resources

ARL®
Association of Research Libraries

SHARE

DCC

casrai
Connecting Research

DataCite

CARL ABRC

portage

McMaster University

LIBRARY

RDM

@McMaster

McMaster University LIBRARY | RDM @McMaster

What are researchers’ data management obligations?

What challenges do researchers face in managing their data?

How can the library help researchers address their data management needs?

MacDATA INSTITUTE

RDM @McMaster

RDA RESEARCH DATA ALLIANCE

ARL® ASSOCIATION OF RESEARCH LIBRARIES

portage

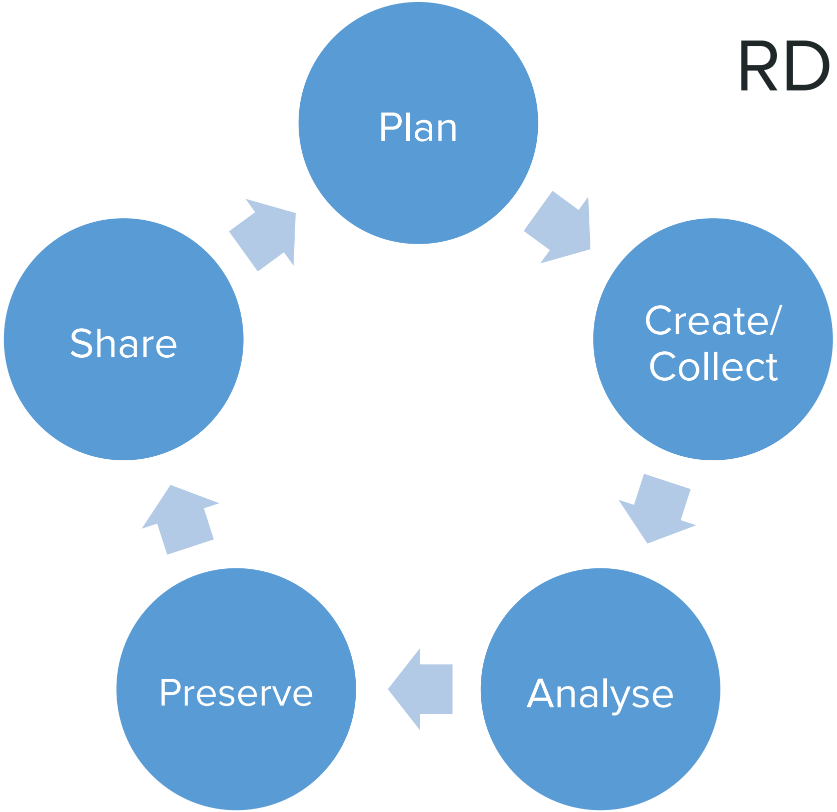
SHARE

DataCite

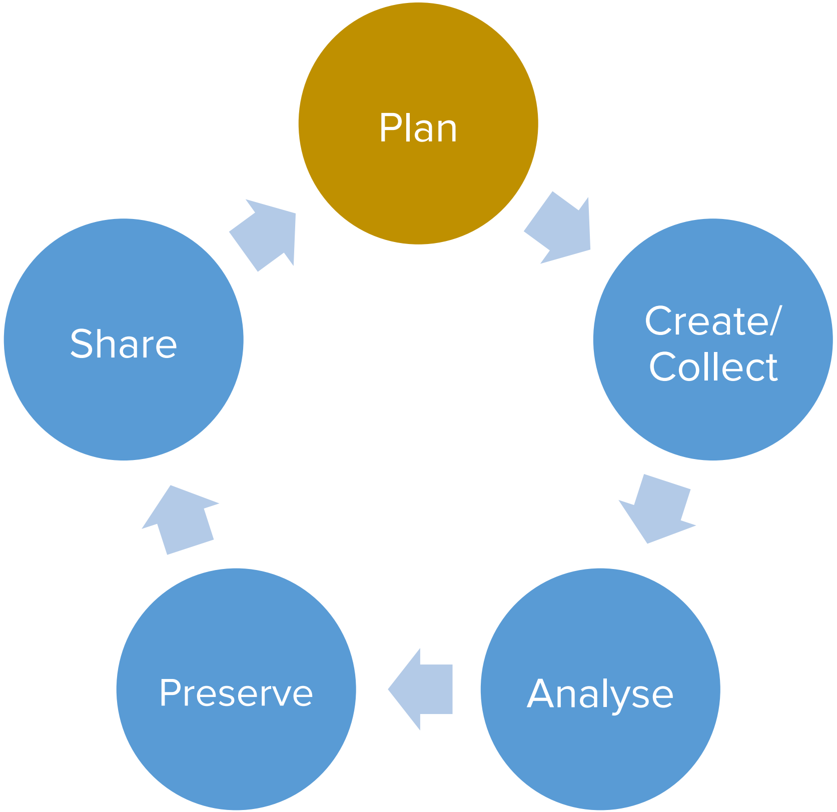
- ✧ Advocacy and communication
- ✧ Training, expertise and information
- ✧ Standards of practice
- ✧ Tools and resources

RDM in the Data Life Cycle: Common Challenges

RDM in the Data Lifecycle



Planning

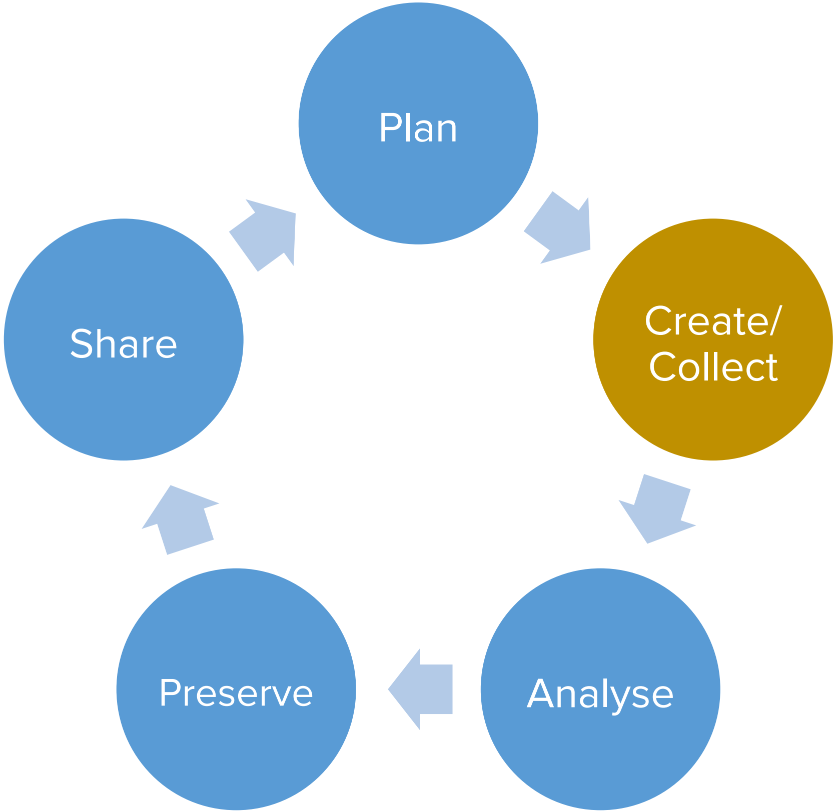


What are the stipulations in institutional, funder or publisher data policies to be followed?

What resources do you require to manage your data?

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

Creating and Collecting

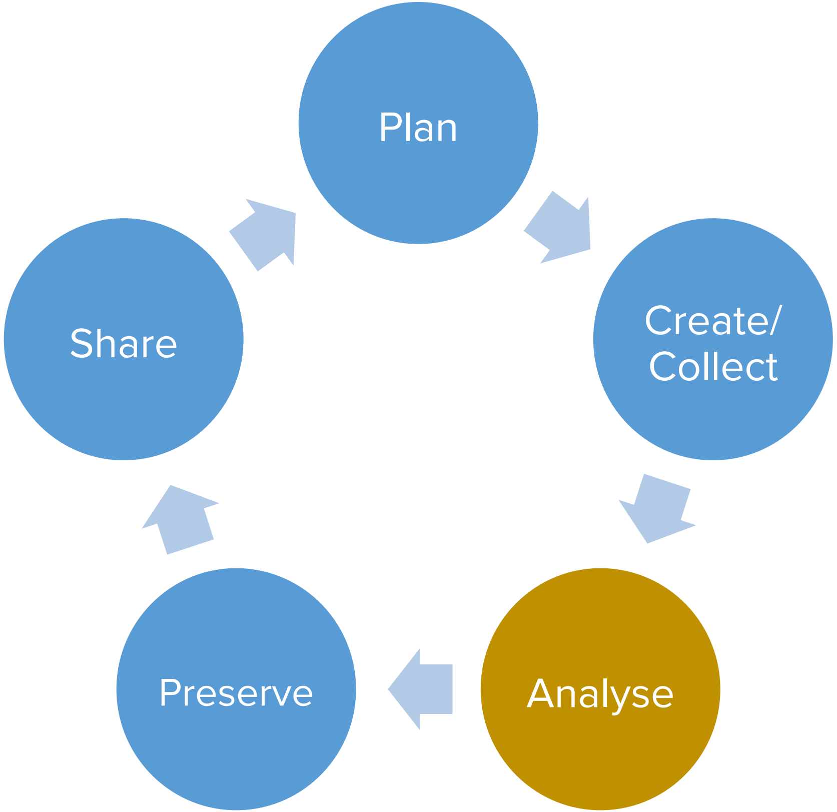


How much data will you collect or create? Where will you store it and back it up?

Is your created or collected data in a suitable format for sharing and long-term preservation?

What documentation and metadata should accompany the data?

Analyzing

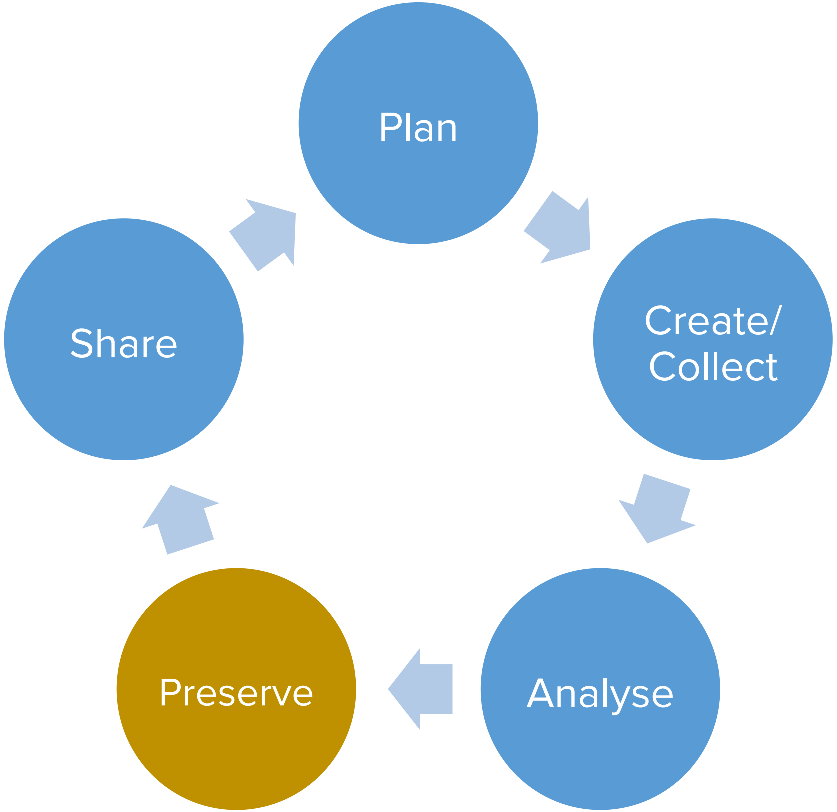


How will you manage any ethical or privacy issues before analysing the data?

How will you securely store (potentially large and cleaned) data pre- and post-analysis?

Who will have access to this data for analysis?

Preserving

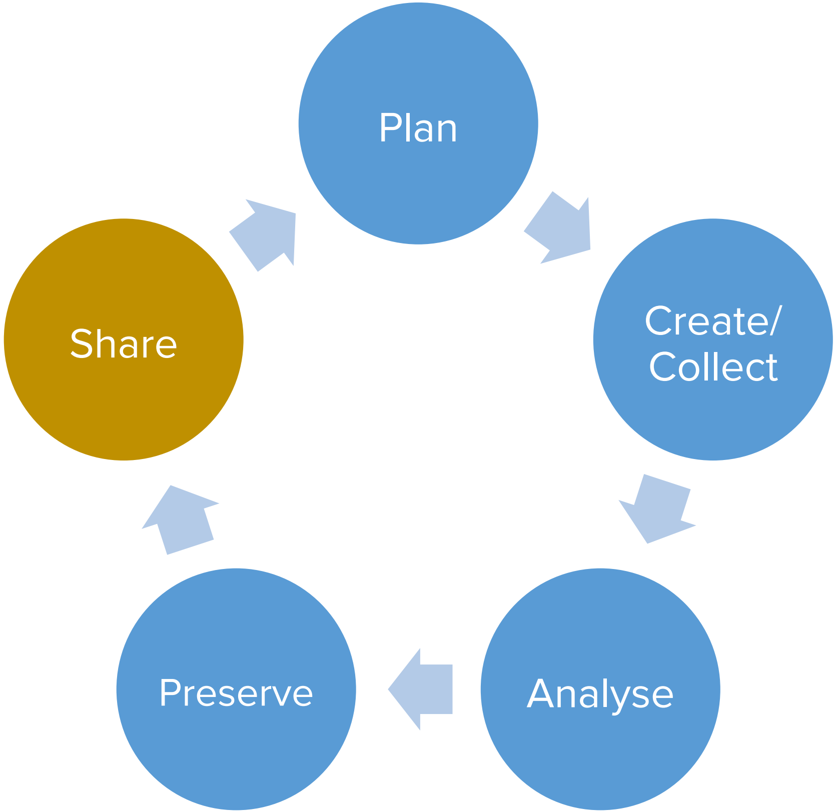


What data should be retained and preserved?

Where will you preserve your data? Who will have access to this preserved data?

For how long you are going to preserve your research data?

Sharing



What data will you make publicly available for research and reuse?

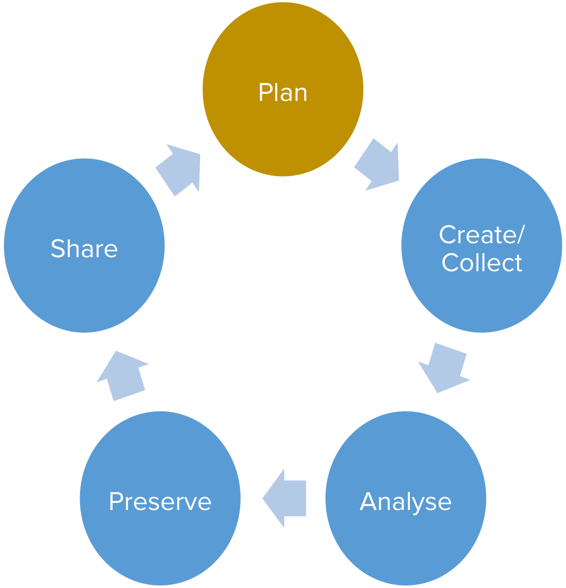
What resources are in place for sharing your data among multiple sites?

How will you manage restrictions e.g. license, privacy issues etc. associated with your data before sharing?

Research Data Management Planning

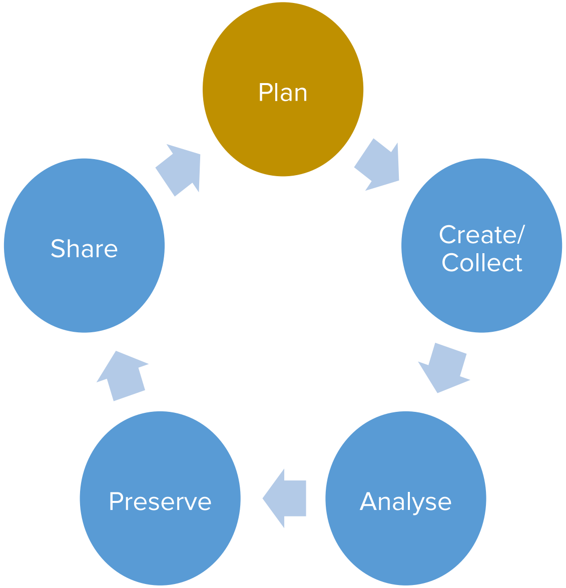
A Research Data Management Plan (DMP) *should:*

- ✧ Describe how you will manage data through all stages of your research
- ✧ Communicate a strategy for creating share-worthy and share-ready data products



An effective 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…



DMPs: The International Context

UK:

✧ AHRC, BBSRC, CRUK, ESRC, MRC, NERC, STFC, Wellcome Trust

US:

✧ NIH*, NSF,

EU

✧ Portugal, EC-Horizon 2020

Shearer, Kathleen. 2015. Comprehensive Brief on Research Data Management Policies. Science.gc.ca, April 2015. Available at: http://www.science.gc.ca/default.asp?lang=En&n=1E116DB8-1

DMPs: The Canadian Context

From the *Statement of Principles* (2015):

> "Data management planning is necessary at all stages of the research project lifecycle, from design and inception to completion."

Canada Foundation for Innovation
Fondation canadienne pour l'innovation

CARL ABRC

SSHRC CRSH

NSERC CRSNG

portage

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

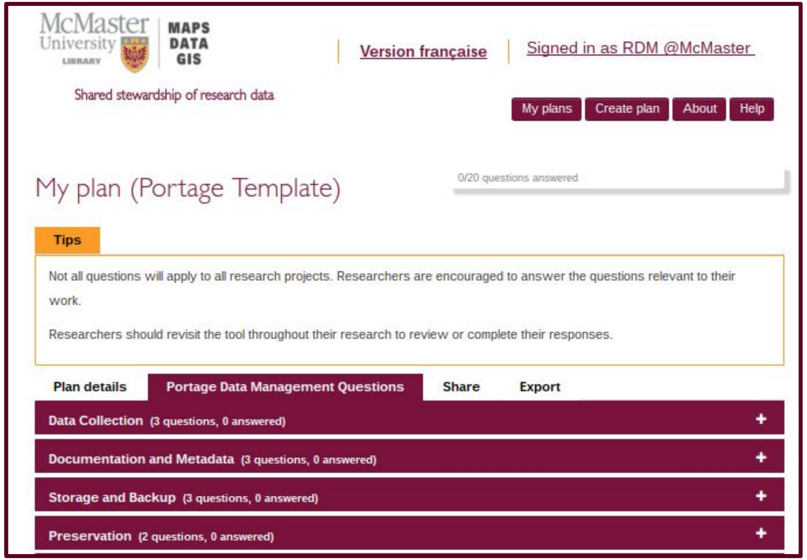
DMP ASSISTANT

DATA MANAGEMENT PLANS
PLANS DE GESTION DES DONNEÉS

PGD

A web-based, *bilingual* data management planning tool made available to all researchers in Canada through the Portage network

A guide for best practices in **data stewardship**



Shared stewardship of research data

My plan (Portage Template)

Not all questions will apply to all research projects. Researchers are encouraged to answer the questions relevant to their work.

Researchers should revisit the tool throughout their research to review or complete their responses.

Data Collection (3 questions, 0 answered)

Documentation and Metadata (3 questions, 0 answered)

Storage and Backup (3 questions, 0 answered)

Preservation (2 questions, 0 answered)

Available at: https://assistant.portagenetwork.ca

What comes next?

- Institutional customization of DMP Assistant
- Amalgamation of DMP tools DMPOnline/Assistant & DMPTool
- Integration with REBs?
- Funding agency guidelines / policies / mandates

DMP ASSISTANT
DATA MANAGEMENT PLANS
PLANS DE GESTION DES DONNEÉS
PGD

DMPonline
The DCC Data Management Planning Tool

DMPTool
Build your Data Management Plan

DMP guidance

Canada

- ✧ Portage Network website: https://portagenetwork.ca
- ✧ RDM@McMaster: http://library.mcmaster.ca/rdm/planning/dmp
- ✧ UBC Libraries: http://researchdata.library.ubc.ca/plan/

International

- ✧ CDL **DMP**Tool [US]: https://dmptool.org/community_resources
- ✧ DCC [UK]: http://www.dcc.ac.uk/resources/data-management-plans

Research Data: Storage & Backup

Tri-Agency Statement of Principles on Digital Data Management

Constraints and obligations

✧ “Research data must be managed in conformity with all commercial, legal and ethical obligations”.

Collection and storage

✧ “… using software and formats that ensure **secure** storage and enable **preservation** of and **access** to the data… “

CIHR IRSC Canadian Institutes of Health Research Instituts de recherche en santé du Canada

Canada Foundation for Innovation Fondation canadienne pour l’innovation

SSHRC CRSH

NSERC CRSNG

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

Challenges: Storage and Backup

- Volume
- Documentation
- Managing Access
- Ethical / Privacy Issues
- Security and Integrity
- Cost

Guiding Questions for Storage & Backup

What resources are available?

✧ Financial, infrastructure & services

How important is the data?

How much data is there?

Who needs access to data?

What level of data security is required?

✧ Ethical / legal requirements

Discussion

With your neighbours, take a few minutes to discuss the benefits and drawbacks of storing your data on:

- PCs / laptops / mobile devices
- External storage devices (hard drives, optical, USB)
- Networked drives
- Cloud services (Dropbox, Google Drive, etc.)

Digital Data is Fragile

...but it can be handled appropriately

3

2

1

3 copies of your data

2

1

3

copies of your data

2

copies are on-hand (easily accessible)

✧ 1 “**production**” (working) copy

✧ 1 “**production backup**” copy

1

3 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

$_3 2 _1$

“**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.

Off-site “Archived” Backup

Providers / Services:

- Campus / Consortium-hosted (RHPCS)
- Remote, Commercial
  - Backblaze, Iron Mountain, JustCloud, etc.

Considerations for “Archived” Backup

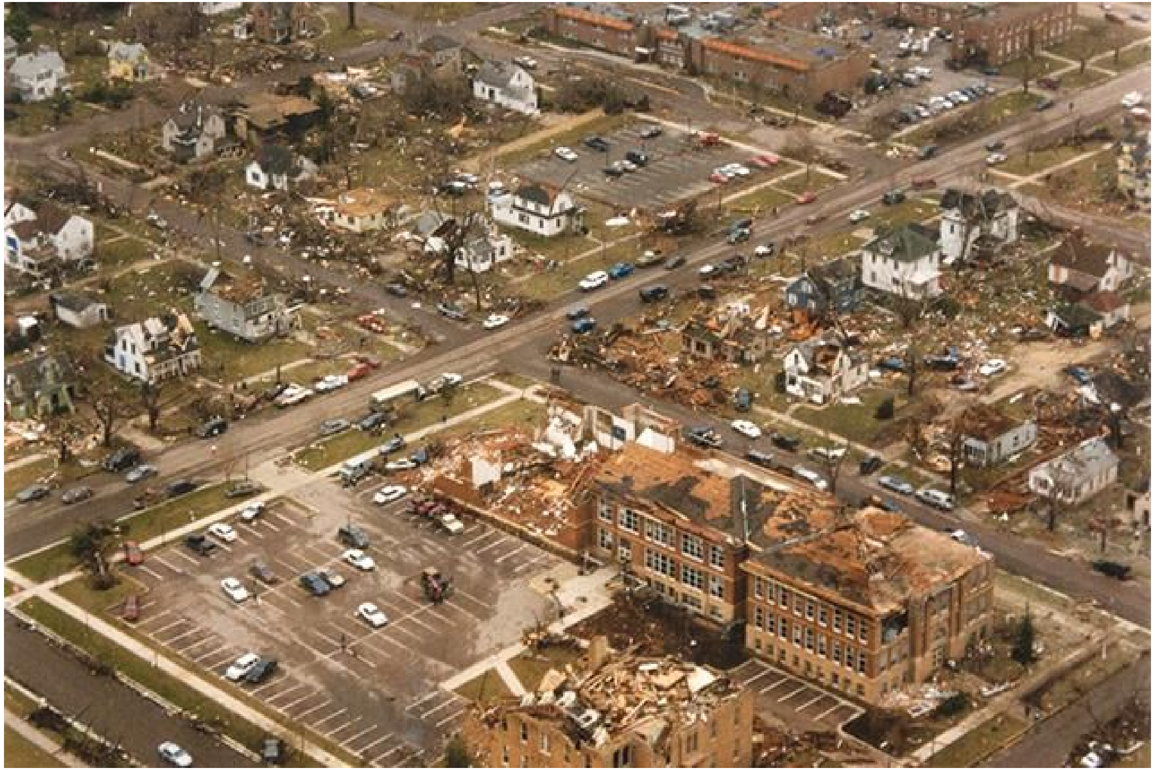
- Security (physical and electronic)
- Automation
- Availability (and time to recover)
- Versioning
- Integrity-checking and error correction
- Data storage (locational) requirements
- Cost

Why Distance is Important

Gustavus Adolphus College

March 29, 1998

F3 tornado



Resources

University of Edinburgh MANTRA RDM Training Kit

- ✧ http://datalib.edina.ac.uk/mantra/libtraining.html

McMaster Library’s Data Management Webpage

- ✧ http://library.mcmaster.ca/rdm/collecting

CARL Portage Network’s National RDM Information Page

- ✧ https://portagenetwork.ca/

Research Data: Preservation & Sharing

Tri-Agency Statement of Principles on Digital Data Management

Preservation, Retention and Sharing

- “All research data resulting from agency funding should normally be preserved in a **publicly accessible**, **secure** and **curated** repository…”

Timeliness

- Data should be shared **as early as possible** in the research process…”
- “…data should be shared **no later than upon the publication** of results”

CIHR IRSC
Canadian Institutes of Health Research Instituts de recherche en santé du Canada

Canada Foundation for Innovation
Fondation canadienne pour l’innovation

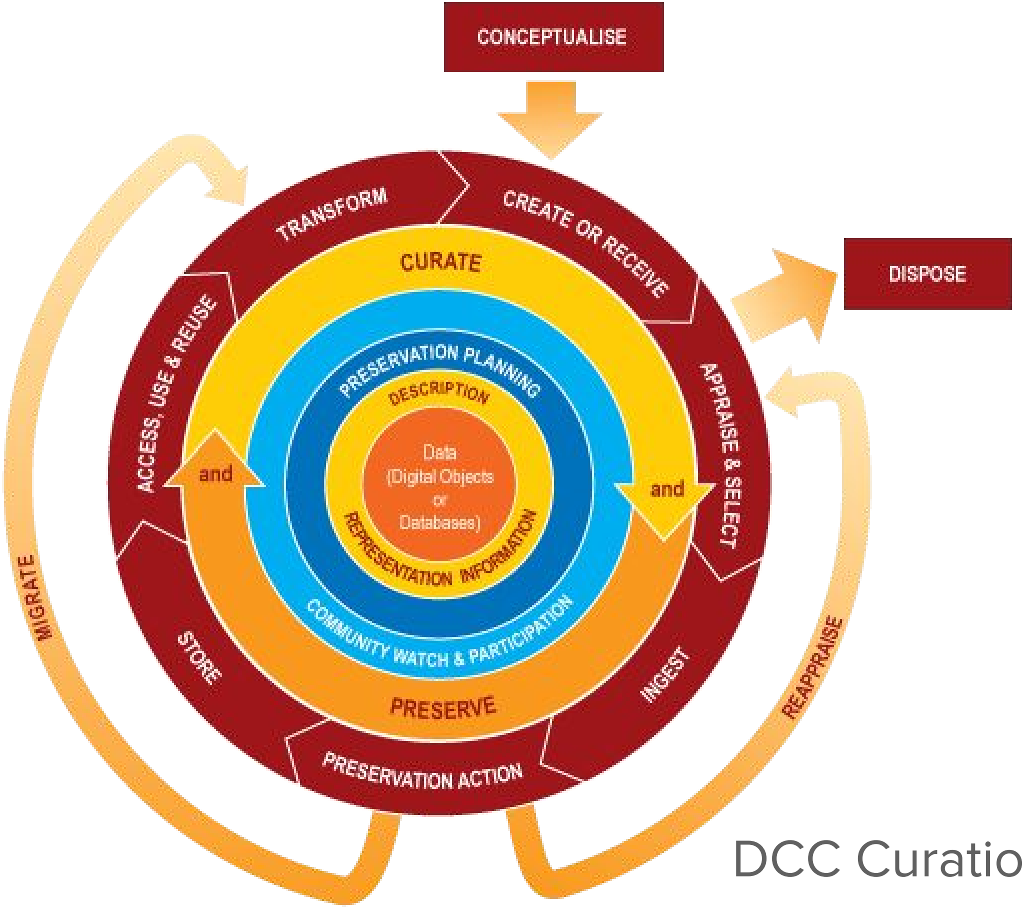
SSHRC CRSH

NSERC CRSNG

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

https://goo.gl/0jkkCP

Long-term storage isn’t necessarily preservation



DCC Curation Lifecycle Model

http://www.dcc.ac.uk/resources/curation-lifecycle-model

Challenges to Preserving Data

- Data requires software to interpret and render it
- Environments change rapidly ➜ formats/software become obsolete
- Storage media becomes obsolete
- Storage media decay over time (“bit rot”)
- Variety and complexity of formats
- Volume of data to preserve

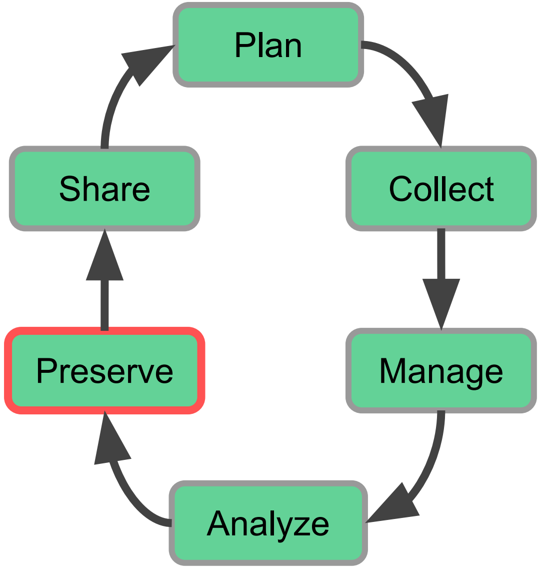
Preservation Considerations

- ✧ Backup schedules
- ✧ File formats & migration
- ✧ Bit integrity checking
- ✧ Version control
- ✧ Data security

Planning for Preservation

- Identify data with long-term value
  - Which will be most useful for future users?
  - What data would be most difficult to reproduce?
  - What documentation is required?

- Consider your obligations and restrictions
  - Funder, institutional policies
  - Journal requirements
  - Research group, discipline practices



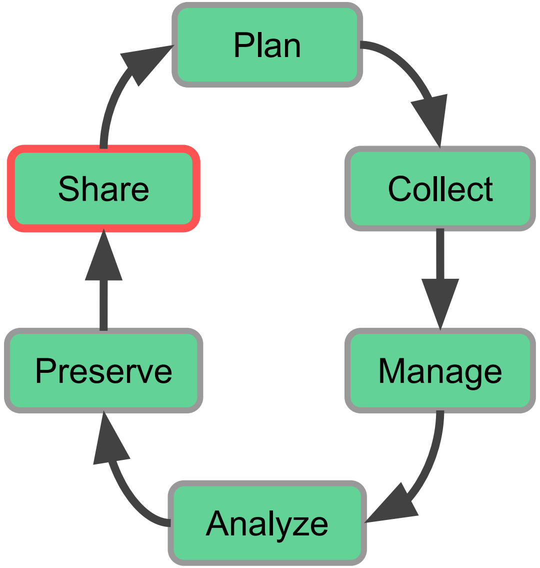
Documenting Data

Descriptive metadata for your ***data and study***:

- ✧ The digital context
- ✧ Personnel and stakeholders
- ✧ Scientific context
- ✧ Parameter information

Administrative metadata for management purposes

- ✧ Intellectual property rights
- ✧ Information for preservation purposes



Why Share Data?

Why Share Data? Internal and External Motivation

Internal

- Creating verifiable and reproducible results
- Increasing research efficiency
- Increasing citation rates and credit for your work

External

- Funders may soon (or already) require you to share your data
- Many publishers require that data be made accessible

Challenges for Sharing Data

- Time / Financial requirements
- Financial
- Confidentiality
- Ownership

Sharing Data Through Data Repositories

What is a Data Repository?

A data repository is a data centre / service with a mandate to:

- ✧ Archive and preserve data
- ✧ Enable discovery of data
- ✧ Manage sharing with others

Why Share in a Data Repository?

- ➢ Your data is secure and regularly backed up
- ➢ Your data is more discoverable ➔ visibility
- ➢ IP and licensing can be applied and administered
- ➢ Access can be managed and monitored

You don’t need to be the steward!

Structured Repositories

- Discipline-specific
- Strict expectations for standards, data formats, structure and metadata

re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

www.re3data.org → Registry of research data repositories

Unstructured Repositories

- Pan-disciplines (though discipline preferences may exist)
- More variation in data and metadata
- Enhanced functionalities and usability

MENDELEY DATA Beta

figshare credit for **all** your research

The Dataverse Project

zenodo

(GIGA)^n DB Revolutionizing data dissemination, organization, and use

DRYAD

GitHub

MacSphere McMaster University Libraries Institutional Repository

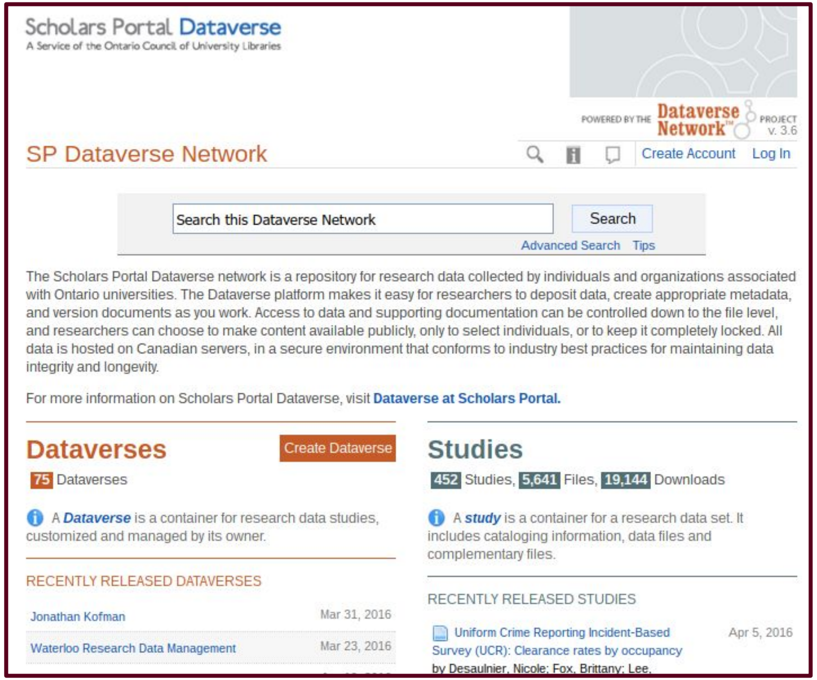
Scholars Portal Dataverse

A repository for research data collected at Ontario's universities.

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.


Scholars Portal Dataverse
A Service of the Ontario Council of University Libraries
POWERED BY THE Dataverse Network PROJECT v. 3.6
Create Account
Log In
Search this Dataverse Network
Search
Advanced Search
Tips
Create Dataverse

SP Dataverse Network

The Scholars Portal Dataverse network is a repository for research data collected by individuals and organizations associated with Ontario universities. The Dataverse platform makes it easy for researchers to deposit data, create appropriate metadata, and version documents as you work. Access to data and supporting documentation can be controlled down to the file level, and researchers can choose to make content available publicly, only to select individuals, or to keep it completely locked. All data is hosted on Canadian servers, in a secure environment that conforms to industry best practices for maintaining data integrity and longevity.

For more information on Scholars Portal Dataverse, visit **Dataverse at Scholars Portal.**

Dataverses

75 Dataverses

A ***Dataverse*** is a container for research data studies, customized and managed by its owner.

RECENTLY RELEASED DATAVERSES

Jonathan Kofman	Mar 31, 2016
Waterloo Research Data Management	Mar 23, 2016

Studies

452 Studies, 5,641 Files, 19,144 Downloads

A ***study*** is a container for a research data set. It includes cataloging information, data files and complementary files.

RECENTLY RELEASED STUDIES

Uniform Crime Reporting Incident-Based Survey (UCR): Clearance rates by occupancy — Apr 5, 2016
by Desaulnier, Nicole; Fox, Brittany; Lee,

Available at: http://dataverse.scholarsportal.info/dvn/

Not Long Term Preservation

“ARCHIVE YOUR DATA OR THEY DIE WITH YOU”

Life is a Transitory Illusion - Your data need not be

Staging for preservation

- Dataverse does not provide preservation services but can be used to organize files for ingest into a preservation service.

- A development project is underway in Canada to create an ingest link between Dataverse and Archivematica.

What comes next?

Compute Canada + CARL - National Research Data Platform

- Federated storage model
- Tools and services to support the curation, access, discoverability, and preservation of research data.
- Portage: Metadata, workflow, testing
- CC: Project management, software development, infrastructure

More Information: https://www.westgrid.ca/print/5498

Thank You.

RDM
@McMaster

Check out http://library.mcmaster.ca/rdm for more information

Contact us at rdmgmt@mcmaster.ca