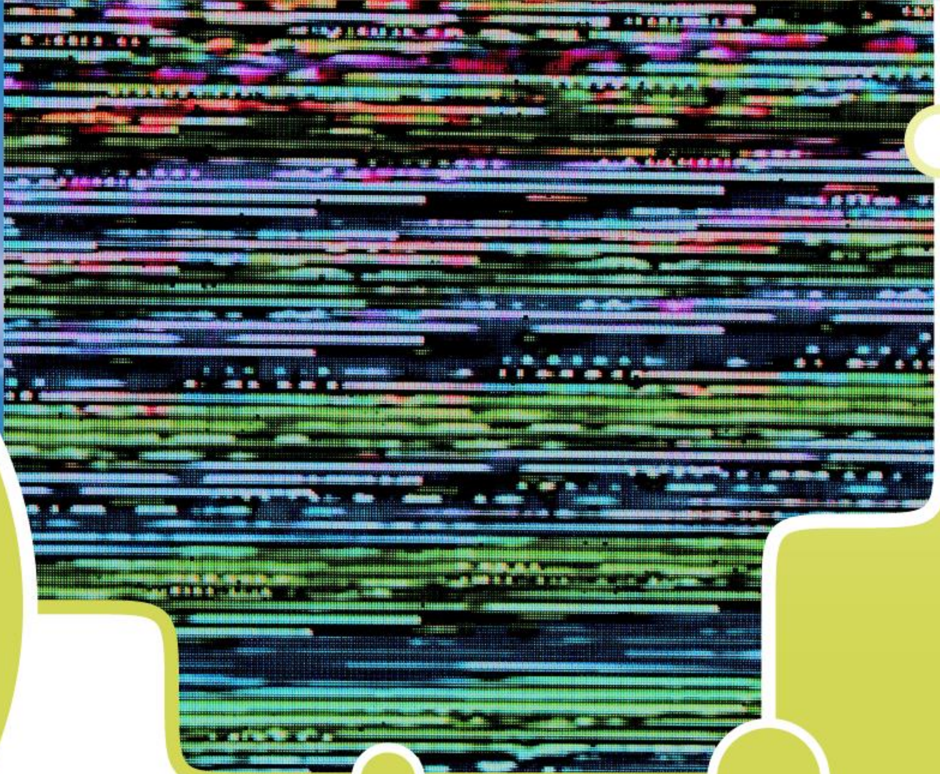
April 17, 2024 | 10:30am-12pm

Virtual Workshop + Sandbox Session

Sensitive Data Management

u.mcmaster.ca/scds-events

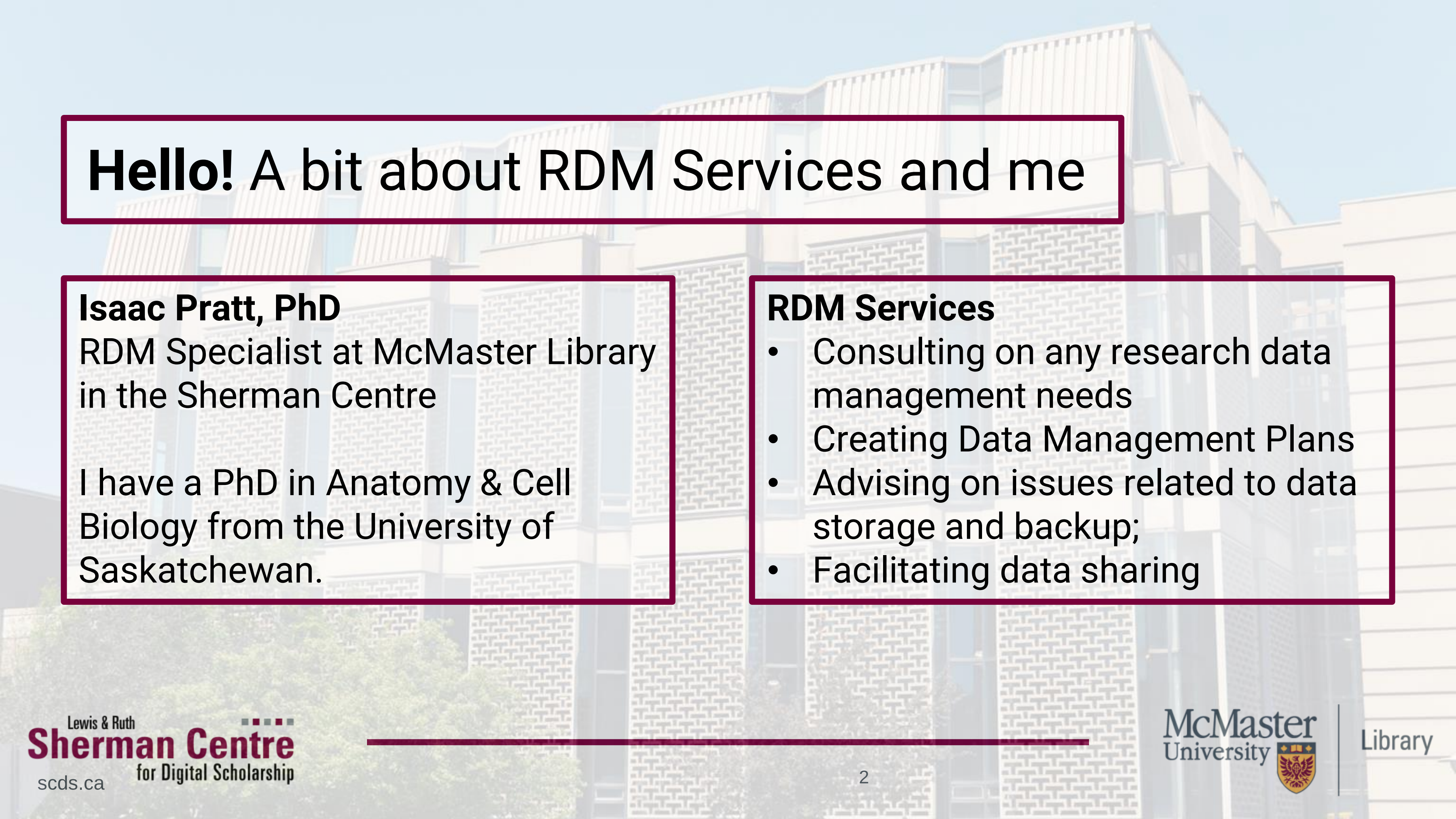


RDM

SCDS

Library

McMaster University



Hello! A bit about RDM Services and me

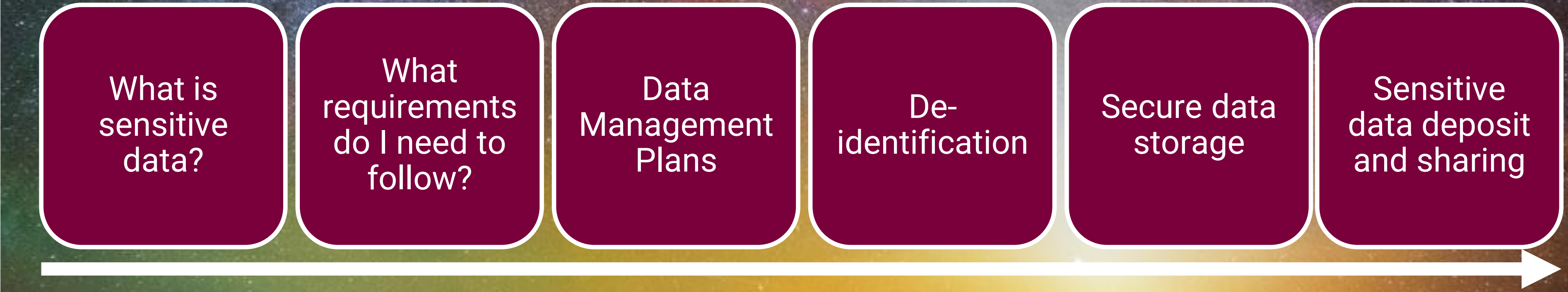
Isaac Pratt, PhD
RDM Specialist at McMaster Library in the Sherman Centre

I have a PhD in Anatomy & Cell Biology from the University of Saskatchewan.

RDM Services

- Consulting on any research data management needs
- Creating Data Management Plans
- Advising on issues related to data storage and backup;
- Facilitating data sharing

Outline



What is sensitive data?

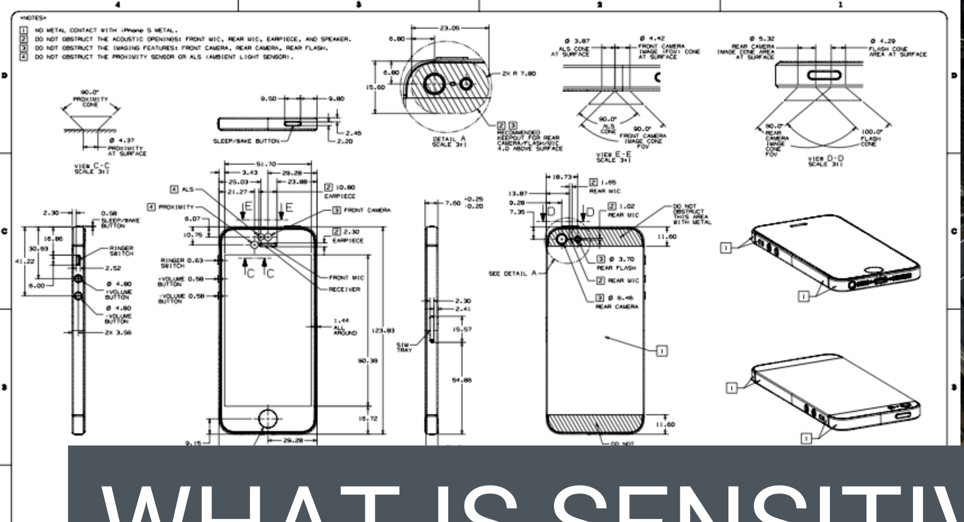
What requirements do I need to follow?

Data Management Plans

De-identification

Secure data storage

Sensitive data deposit and sharing


NOTES
1 NO METAL CONTACT WITH iPhone 5 METAL.
2 DO NOT OBSTRUCT THE ACOUSTIC OPENINGS: FRONT MIC, REAR MIC, EARPIECE, AND SPEAKER.
3 DO NOT OBSTRUCT THE IMAGING FEATURES: FRONT CAMERA, REAR CAMERA, REAR FLASH.
4 DO NOT OBSTRUCT THE PROXIMITY SENSOR OR ALS (AMBIENT LIGHT SENSOR).
DETAIL A SCALE 3:1
VIEW C-C SCALE 3:1
VIEW E-E SCALE 3:1
VIEW D-D SCALE 3:1
SLEEP/WAKE BUTTON
RINGER SWITCH
+VOLUME BUTTON
-VOLUME BUTTON
FRONT CAMERA
EARPIECE
FRONT MIC
RECEIVER
REAR MIC
REAR FLASH
REAR CAMERA
SIM TRAY
SEE DETAIL A
DO NOT OBSTRUCT THIS AREA WITH METAL

WHAT IS SENSITIVE DATA?

Reptile traffickers trawl scientific literature, target newly described species

“I knew: if I publish the exact location again, people will go look for it. I don’t want traders to use my information.” The paper where the species was described, published in the journal Zootaxa, mentions the location only as: “available on request, for fellow scientists.”

When mentioning a location, even listing the name of a village is risky, let alone publishing precise coordinates: “Geckos’ habitat is karst caves. You just go to that village and ask, ‘where is a cave around here?’”

McMaster University | Library

⚠ What is sensitive data:

Sensitive data is any data that would cause harm if released openly.

- Research data containing personal identifying information or personal health information
- Commercially sensitive data, including data generated under a commercial research funding agreement
- Data relating to rare or endangered species
- Data likely to harm an individual or community or have a significant negative public impact if released

What requirements do I need to be aware of?

Ethics

- Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2022)
- MREB, HiREB

Funders

- Tri-Agency Research Data Management Policy
- NIH Data Management & Sharing Policy

Data sources

- Data sharing/transfer agreements – MILO Sample Agreements

Journals

- Data availability statements and data sharing policies - Frontiers Materials and data policies

Legislative

- Health Canada Clinical Trial guidance
- Health information in Ontario is subject to the **Personal Health Information Protection Act** (PHIPA)

Privacy principles

There are 10 privacy principles that underpin the Canadian Privacy Legislation:

1) Accountability
2) Identifying purposes
3) Consent
4) Limiting collection
5) Limiting use, disclosure and retention
6) Accuracy
7) Safeguards
8) Openness
9) Individual access
10) Challenging compliance

Canadian Ethics principles - TCPS2

- **Respect for human dignity**
  - **Informed consent**, voluntary participation, limiting undue influence, assent process, deception and debrief, etc.
- **Concern for Welfare**
  - Holistic view of welfare, identify risks, minimize and mitigate risks, procedures in place to address likely harms, **data security**, etc.
- **Justice**:
  - Sharing burdens and benefits of research, just exclusion/inclusion, **disseminating results**, etc.
- **Ethical Duty of Confidentiality**
  - Researchers shall safeguard information entrusted to them and not misuse or wrongfully disclose it

Funder Requirements

Government of Canada Gouvernement du Canada

Search Canada.ca

MENU

Home › Interagency research funding › Policies and Guidelines › Research Data Management

Research Data Management

Tri-Agency Statement of

Tri-Agency Research Data Management Policy

Tri-Agency RDM Policy – CIHR:

- CIHR currently requires researchers to deposit bioinformatics, atomic, and molecular coordinate data into the appropriate public database.

Data Management Plans (in pilot phase)

- Grants will require data management plans (DMPs) to be submitted at the time of application.
- Use of DMP Assistant is encouraged; approx. 2-3 pages.

Data Deposit (launch tbd)

- Grant recipients will be required to deposit into a digital repository research data that supports journal publications and pre-prints from funded research
- Grant recipients are **not required to share their data**
- “researchers should only make data accessible if doing so is ethical, legal, and is in consonance with any commercial or other agreements the researcher has entered into”

Funder Requirements

National Institutes of Health (NIH) *Data Management and Sharing Policy*

Data Management Plans:

- NIH requires all applicants planning to generate scientific data to prepare a DMS Plan that describes how the scientific data will be managed and shared.

Data Sharing:

- "Scientific data should be made accessible as soon as possible."
- *"certain factors (i.e., ethical, legal, or technical) may necessitate limiting sharing to some extent. Foreseeable limitations should be described in DMS Plans"*
- *"Consider whether access to shared scientific data derived from humans should be controlled"*

NIH SCIENTIFIC DATA SHARING
Searc
DATA MANAGEMENT AND SHARING POLICY
GENOMIC DATA SHARING POLICY
PUBLIC ACCESS POLICY
VIEW MORE →

Expediting the Translation of Research Results to Improve Human Health.

Featured News & Events

Year of Open Science 2024 Conference

Journal Requirements: “Mandatory data and code sharing for research published by The BMJ”

https://doi.org/10.1136/bmj.q324 (Published 05 March 2024)

- The policy requires authors of all submitted trials to **post relevant trial data** in an enduring, publicly accessible repository before publication. A link to the trial data must be included in the data sharing statement in the article.
- The BMJ intends to broaden its data sharing policy to non-trial research in future
- The policy also requires submission of relevant analytical **code** in a supplementary file that will be permanently accessible alongside each paper
- A new **code availability statement** will be required in research papers

thebmj
EXCLUSIVE Cass gender care review p6-10
Doctors "censured" for Gaza concerns p4
Esketamine and prenatal depression p 17
Update on rheumatoid arthritis care p33
1.5 CPD hours in the education pages
NHS workforce
Built to survive?

Indigenous Data Sovereignty

When researchers are working with Indigenous communities, they should abide by principles of indigenous data sovereignty like the CARE principles. For Canadian First Nations, the OCAP principles are recommended:

- **Ownership**: a community or group owns information collectively in the same way that an individual owns his or her personal information.
- **Control**: First Nations, their communities, and representative bodies are within their rights to seek control over all aspects of research and information management processes that impact them.
- **Access**: First Nations must have access to information and data about themselves and their communities regardless of where it is held.
- **Possession** is the mechanism by which ownership can be asserted and protected, through direct physical control of data.

Create a Data Management Plan

- A **Data Management Plan (DMP)** is your plan for how you will interact with your research data both **during** the active phases of your research and **after** the completion of the research project.
- Engage research partners and collaborators in ongoing conversation about how to best manage research data.
- Establish and consistently lay out data practices for a lab.
- Set up storage and security systems, with timelines for backups, retention, and updates.
- Ensure contingency plans and responsibilities for unexpected events – illness, moving universities, ransomware attack.

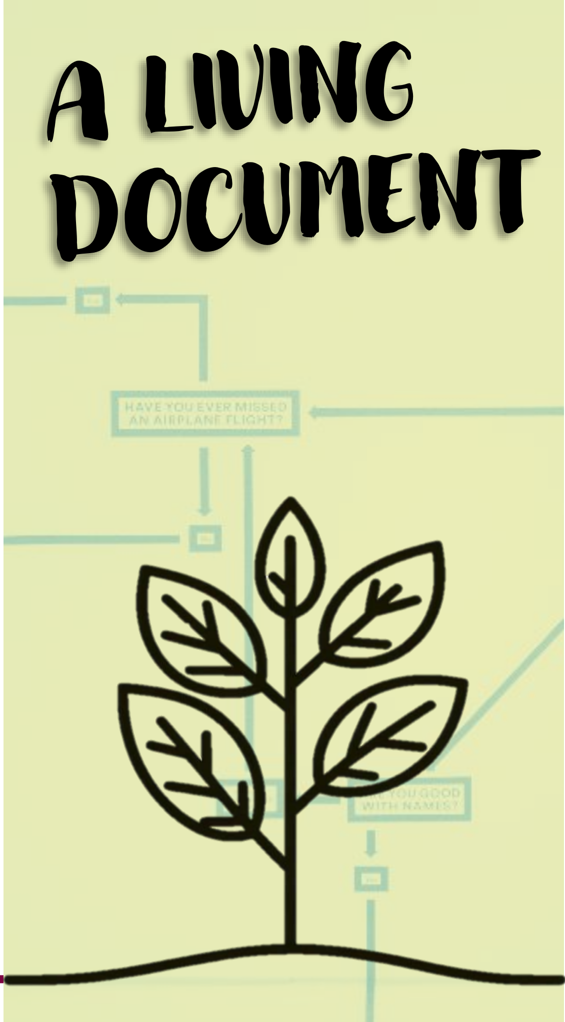
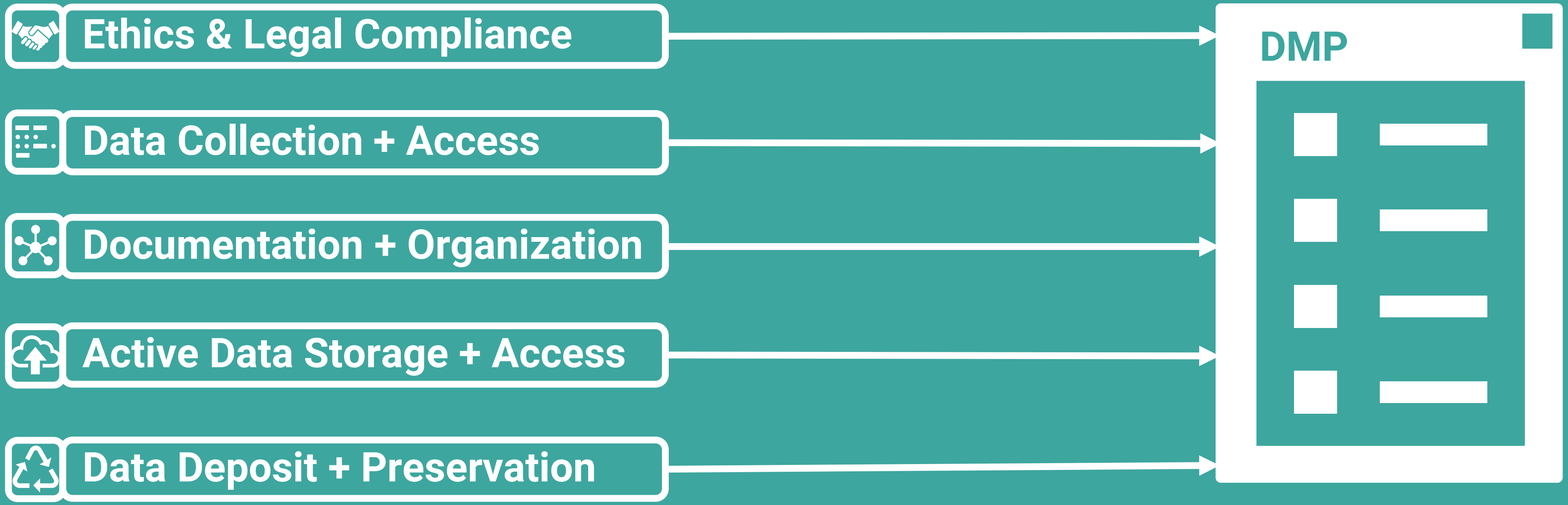


Photo by Hanna Morris on Unsplash.

What goes in a Data Management Plan?



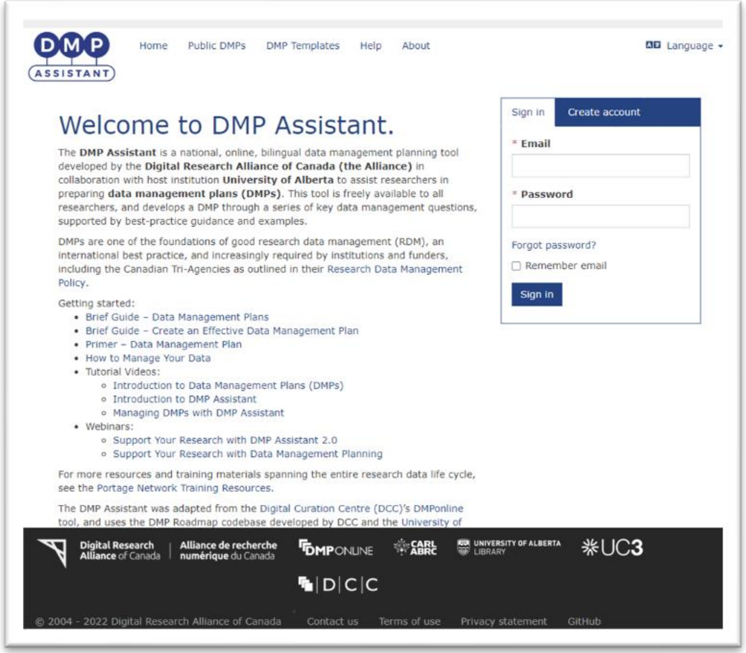
Ethics & Legal Compliance

Data Collection + Access

Documentation + Organization

Active Data Storage + Access

Data Deposit + Preservation



Welcome to DMP Assistant.

The **DMP Assistant** is a national, online, bilingual data management planning tool developed by the **Digital Research Alliance of Canada (the Alliance)** in collaboration with host institution **University of Alberta** to assist researchers in preparing **data management plans (DMPs)**. This tool is freely available to all researchers, and develops a DMP through a series of key data management questions, supported by best-practice guidance and examples.

DMPs are one of the foundations of good research data management (RDM), an international best practice, and increasingly required by institutions and funders, including the Canadian Tri-Agencies as outlined in their Research Data Management Policy.

Getting started:

- Brief Guide – Data Management Plans
- Brief Guide – Create an Effective Data Management Plan
- Primer – Data Management Plan
- How to Manage Your Data
- Tutorial Videos:
  - Introduction to Data Management Plans (DMPs)
  - Introduction to DMP Assistant
  - Managing DMPs with DMP Assistant
- Webinars:
  - Support Your Research with DMP Assistant 2.0
  - Support Your Research with Data Management Planning

For more resources and training materials spanning the entire research data life cycle, see the Portage Network Training Resources.

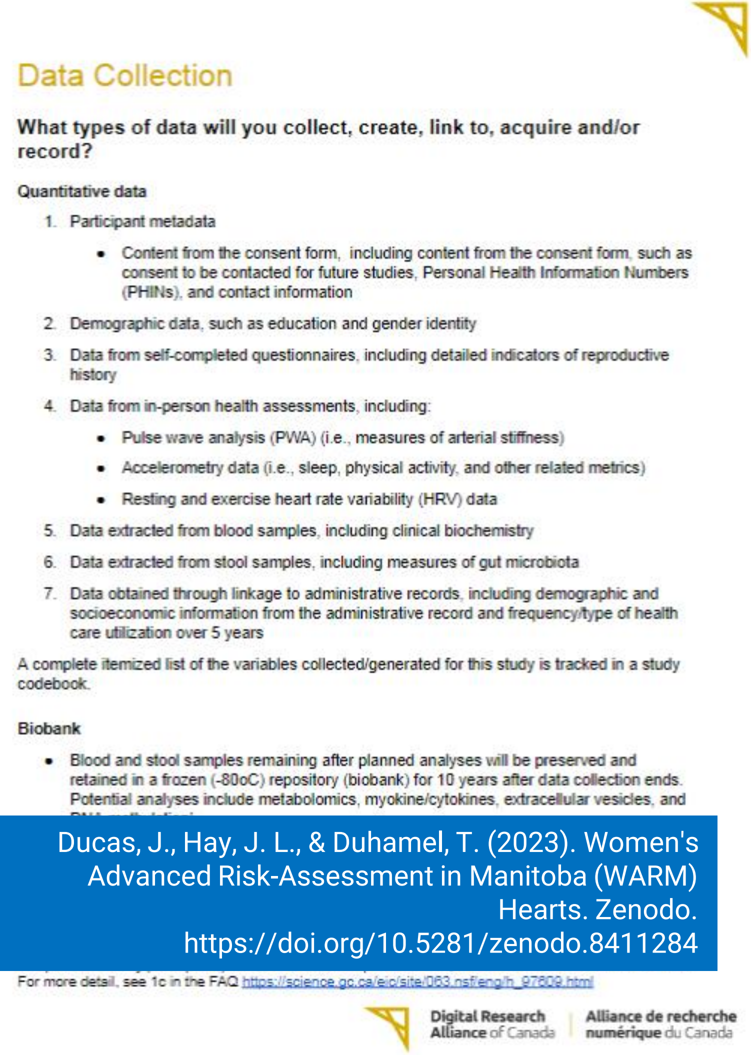
The DMP Assistant was adapted from the Digital Curation Centre (DCC)'s DMPonline tool, and uses the DMP Roadmap codebase developed by DCC and the University of

© 2004 - 2022 Digital Research Alliance of Canada   Contact us   Terms of use   Privacy statement   GitHub

- A web-based, bilingual data management planning tool
- Offers templates with built in McMaster specific guidance to help you create a DMP
- DMPs can be exported into .pdf, .docx
- We can support your DMP development through consultations and review
- https://dmp-pgd.ca/

McMaster DMP Database

- 200+ example DMPs from resources across the world.
- Search by field, **location**, funder (NIH)
- Submit your DMP for other researchers!
- rdm.mcmaster.ca/dmps



Data Collection

What types of data will you collect, create, link to, acquire and/or record?

Quantitative data

1. Participant metadata
   - Content from the consent form, including content from the consent form, such as consent to be contacted for future studies, Personal Health Information Numbers (PHINs), and contact information
2. Demographic data, such as education and gender identity
3. Data from self-completed questionnaires, including detailed indicators of reproductive history
4. Data from in-person health assessments, including:
   - Pulse wave analysis (PWA) (i.e., measures of arterial stiffness)
   - Accelerometry data (i.e., sleep, physical activity, and other related metrics)
   - Resting and exercise heart rate variability (HRV) data
5. Data extracted from blood samples, including clinical biochemistry
6. Data extracted from stool samples, including measures of gut microbiota
7. Data obtained through linkage to administrative records, including demographic and socioeconomic information from the administrative record and frequency/type of health care utilization over 5 years

A complete itemized list of the variables collected/generated for this study is tracked in a study codebook.

Biobank

- Blood and stool samples remaining after planned analyses will be preserved and retained in a frozen (-80oC) repository (biobank) for 10 years after data collection ends. Potential analyses include metabolomics, myokine/cytokines, extracellular vesicles, and

For more detail, see 1c in the FAQ https://science.gc.ca/eic/site/063.nsf/eng/h_97609.html

Digital Research Alliance of Canada
Alliance de recherche numérique du Canada

Ducas, J., Hay, J. L., & Duhamel, T. (2023). Women's Advanced Risk-Assessment in Manitoba (WARM) Hearts. Zenodo. https://doi.org/10.5281/zenodo.8411284

Data Management Practices

Good data management

Research project management

De-identification

Secure data storage

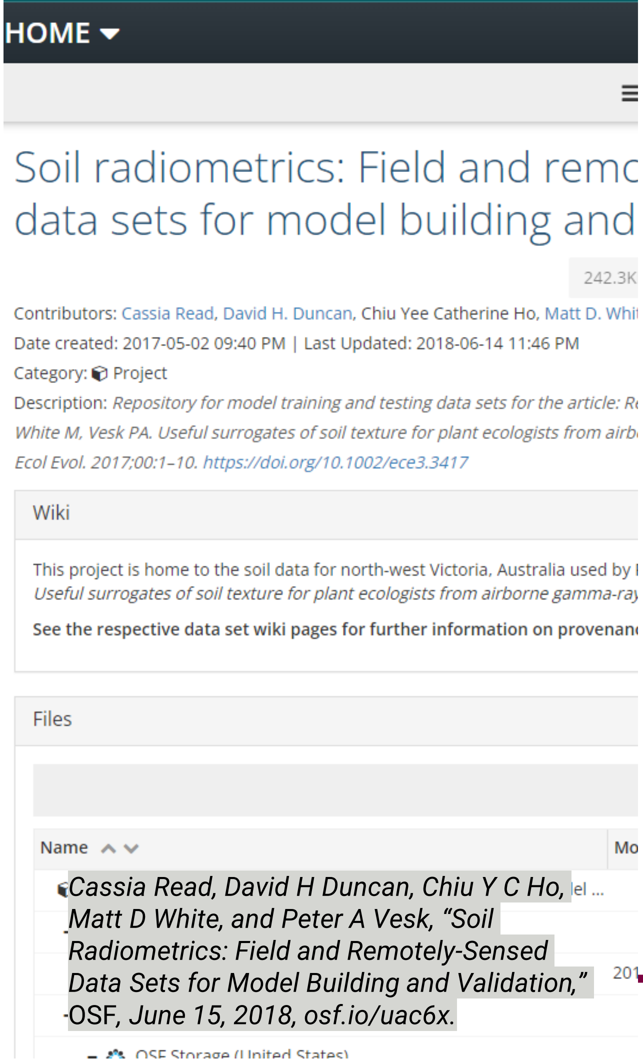
Data storage practices and controls

3rd party tools and services

Cybersecurity practices

Encryption

Research Project Management



Soil radiometrics: Field and remc
data sets for model building and

242.3K

Contributors: Cassia Read, David H. Duncan, Chiu Yee Catherine Ho, Matt D. Whi

Date created: 2017-05-02 09:40 PM | Last Updated: 2018-06-14 11:46 PM

Category: Project

Description: *Repository for model training and testing data sets for the article: R*
White M, Vesk PA. Useful surrogates of soil texture for plant ecologists from airb
Ecol Evol. 2017;00:1–10. https://doi.org/10.1002/ece3.3417

Wiki

This project is home to the soil data for north-west Victoria, Australia used by
Useful surrogates of soil texture for plant ecologists from airborne gamma-ray

See the respective data set wiki pages for further information on provenan

Files

Name | Mo

Cassia Read, David H Duncan, Chiu Y C Ho, Matt D White, and Peter A Vesk, “Soil Radiometrics: Field and Remotely-Sensed Data Sets for Model Building and Validation,” OSF, June 15, 2018, osf.io/uac6x.

el ...

201

OSF Storage (United States)

- **Collaboration**: Microsoft teams let you control your team and share and work on documents together in real-time, avoiding multiple versions and copies sent by email.
- **Reference Management**: Zotero or Endnote support collaboration through shared citation libraries.
- **Notetaking software**: Obsidian, OneNote, or an Electronic Lab Notebook allow you to create organized, linked notes that you can use to document your research practices
- **REDCap:** REDCap is a powerful web tool for collecting and organizing longitudinal data.

Learn more at **rdm.mcmaster.ca/organize**

McMaster University | Library

De-identification

De-identification is the general term for the process of removing personal information from a record or data set

Name:	Camila
Age:	23
Ethnicity:	Hispanic
Birthplace:	Hamilton
Address:	L8P 0C8
Gender:	Female
Employer:	McMaster
Unit:	Physics
Job Title:	Assistant Professor

Name:	-------
Age:	19-24
Ethnicity:	Hispanic
Birthplace:	Hamilton
Address:	--- ---
Gender:	Female
Employer:	[University]
Unit:	Physics
Job Title:	Assistant Professor

De-identification issues: small cell

Some people are more identifiable than others

Name:	Averi
Age:	36
Ethnicity:	Caucasian
Address:	Caledonia N3W 1C2
Gender:	Non-binary
Employer:	McMaster
Unit:	Biology
Job Title:	Education Technologist

Name:	-------
Age:	35-39
Ethnicity:	Caucasian
Address:	[Hamilton] --- ---
Gender:	[Other]
Employer:	[University]
Unit:	Biology
Job Title:	[Other]

K anonymity

k-anonymity is a mathematical approach to ensuring a dataset is anonymous.

A dataset has k-anonymity when a particular individual in the dataset cannot be distinguished from k other individuals in the dataset.

‘k’ is a number set by the researcher - most commonly set to 5. This means it should not possible to isolate a group of fewer than 5 identical individuals.

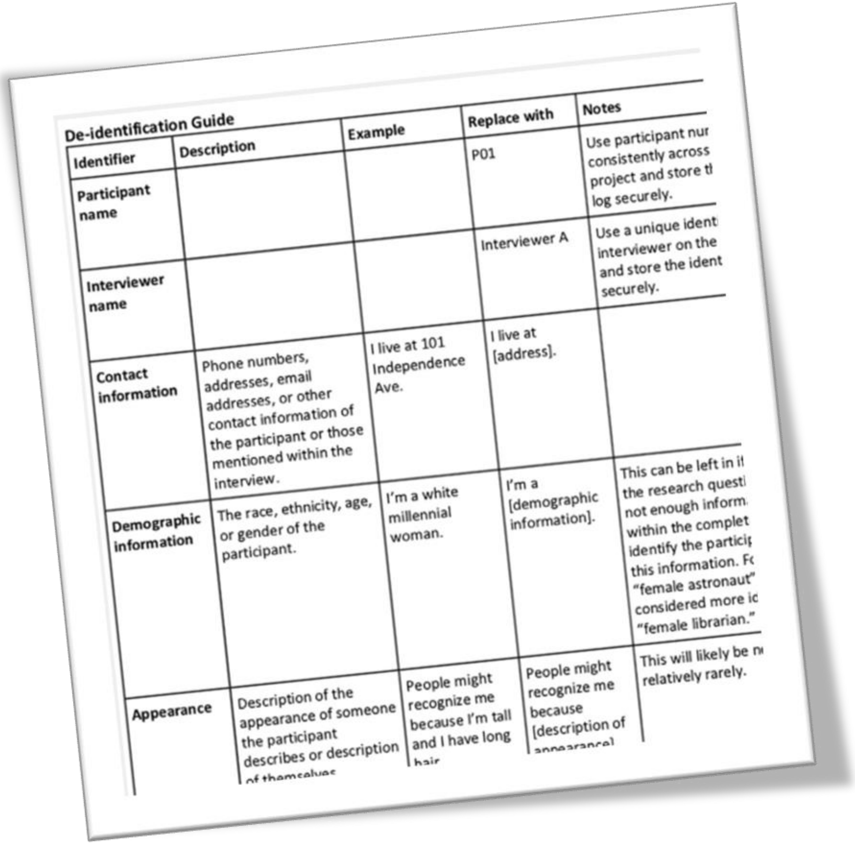
Amnesia https://amnesia.openaire.eu/

sdcMicro https://cran.r-project.org/web/packages/sdcMicro/index.html

For a more comprehensive overview see the Portage Network’s Reducing Risk Webinar and slides

De-identification protocol

- Note all the direct and quasi- identifiers you are collecting
- Write out how you will de-identify each variable
- How long will you keep a linking key?
- Attach this to your **Data Management Plan**
- This is particularly helpful if working as part of a team.



De-identification Guide

Identifier	Description	Example	Replace with	Notes
Participant name			P01	Use participant nur consistently across project and store tl log securely.
Interviewer name			Interviewer A	Use a unique ident interviewer on the and store the ident securely.
Contact information	Phone numbers, addresses, email addresses, or other contact information of the participant or those mentioned within the interview.	I live at 101 Independence Ave.	I live at [address].	
Demographic information	The race, ethnicity, age, or gender of the participant.	I'm a white millennial woman.	I'm a [demographic information].	This can be left in if the research questi not enough inform within the complet identify the particip this information. Fo "female astronaut" considered more ic "female librarian."
Appearance	Description of the appearance of someone the participant describes or description of themselves	People might recognize me because I'm tall and I have long hair	People might recognize me because [description of appearance]	This will likely be n relatively rarely.

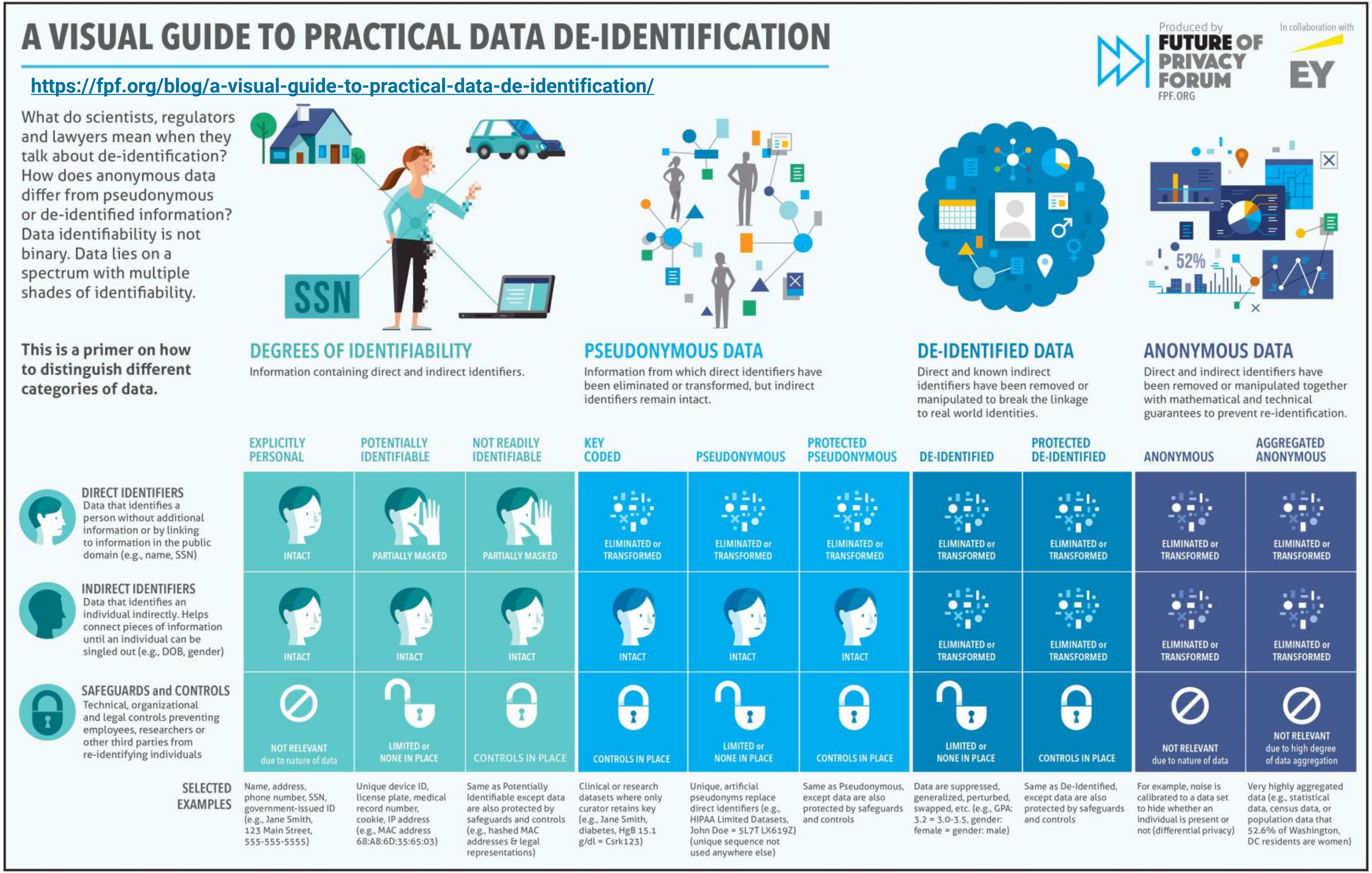
De-identification definitions

From the TCPS2:

- **Coded information** – direct identifiers are removed from the information and replaced with a code. With access to the code it is possible to re-identify specific participants.
- **Anonymized information** – the information is irrevocably stripped of direct identifiers, a code is not kept, and risk of re-identification of individuals from remaining indirect identifiers is low or very low.
- **Anonymous information** – the information never had identifiers associated with it (e.g., anonymous surveys) and risk of re-identification is low or very low.
    - De-identification is not a 'guarantee' of privacy and risks of re-identification can often remain.

Linking file/key

- A file linking the participant names and IDs or pseudonyms. Data is not considered anonymous if a linking file exists!
- Linking files should be encrypted and stored on separate devices or systems than the data.
- Linking files and the included personal information should be destroyed/deleted when no longer required to increase privacy.



A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

https://fpf.org/blog/a-visual-guide-to-practical-data-de-identification/

Produced by FUTURE OF PRIVACY FORUM FPF.ORG

In collaboration with EY

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.

This is a primer on how to distinguish different categories of data.

DEGREES OF IDENTIFIABILITY

Information containing direct and indirect identifiers.

PSEUDONYMOUS DATA

Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.

DE-IDENTIFIED DATA

Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.

ANONYMOUS DATA

Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
DIRECT IDENTIFIERS Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)	INTACT	PARTIALLY MASKED	PARTIALLY MASKED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
INDIRECT IDENTIFIERS Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)	INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
SAFEGUARDS and CONTROLS Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals	NOT RELEVANT due to nature of data	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	NOT RELEVANT due to nature of data	NOT RELEVANT due to high degree of data aggregation
SELECTED EXAMPLES	Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)	Same as De-Identified, except data are also protected by safeguards and controls	For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)	Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)

Evaluating 3rd party services

- High risk data should stay on campus and with the researchers
- Terms of use should be examined closely to see what platforms are doing with data
- Data storage location should be in Canada, PHI should be in Ontario
- Data should be shared to team-members and contractors in a de-identified form when possible
- Individual contractors should sign confidentiality agreements

Photo by Smit Patel on Unsplash.

Recordings and transcripts

- Looking for a transcription service? Use McMaster’s Guidance on Using Transcription with Data from Human Participants
- Video and audio recordings are inherently more identifiable than transcripts
- Researchers should limit access to the original recordings, and delete them or store them offline if they are no longer needed.
- Direct and quasi identifiers in transcripts should be pseudonymized or generalized where possible

Photo by Soundtrap on Unsplash.

Data storage

- RDM Webinar recording: Strategies for research data storage and backup

February 14, 2024 | 10:30am-11:30am
Virtual Workshop

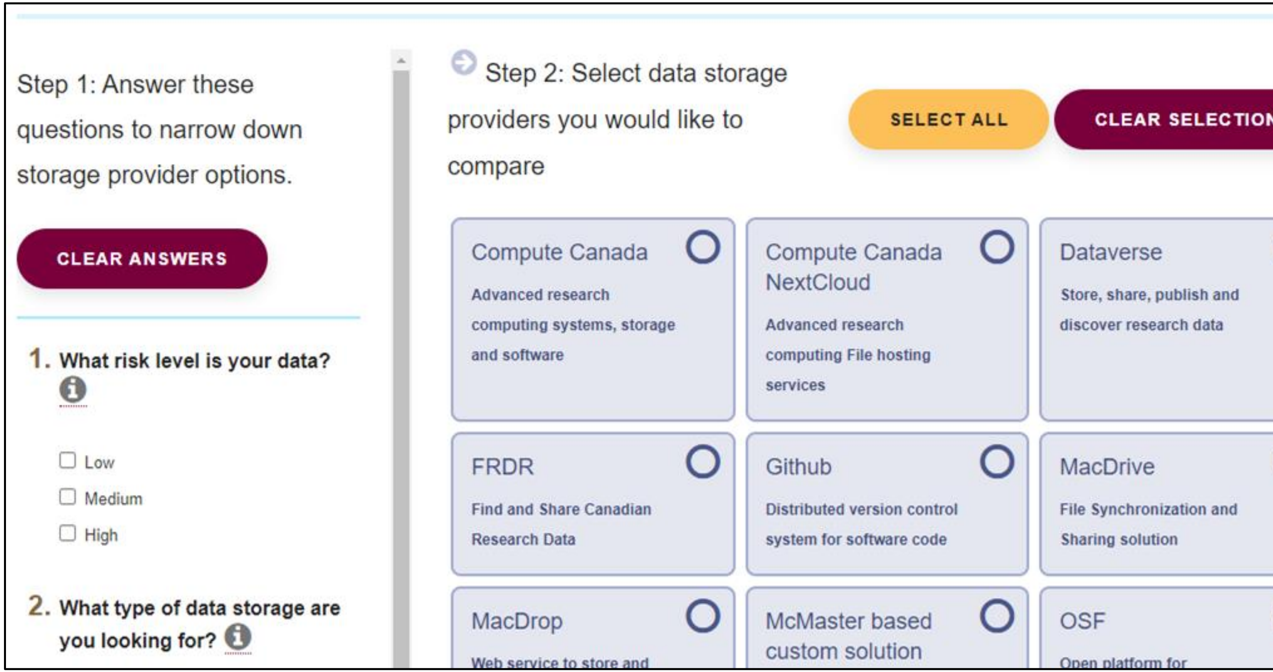
Storage Scores: Store & Back Up Data at McMaster

u.mcmaster.ca/scds-events

Photo by Eric Brazier

Research Data Storage Finder u.mcmaster.ca/storagefinder

- McMaster RDM Services has a **Data Storage Finder**, an interactive tool to help you find a vetted storage provider depending on risk, volume, and other needs.
- This tool also allows you to compare feature sets of selected options.



Step 1: Answer these questions to narrow down storage provider options.

CLEAR ANSWERS

1. What risk level is your data?

□ Low
□ Medium
□ High

2. What type of data storage are you looking for?

Step 2: Select data storage providers you would like to compare

SELECT ALL

CLEAR SELECTION

○ Compute Canada
Advanced research computing systems, storage and software

○ Compute Canada NextCloud
Advanced research computing File hosting services

Dataverse
Store, share, publish and discover research data

○ FRDR
Find and Share Canadian Research Data

○ Github
Distributed version control system for software code

MacDrive
File Synchronization and Sharing solution

○ MacDrop
Web service to store and

○ McMaster based custom solution

OSF
Open platform for

Backup Strategies (3-2-1)

A good data storage plan needs to balance accessibility and convenience against security and reliability.

Example:
1 copy stored locally on **hard drive** for analysis
1 copy stored on **cloud storage** platform
1 copy stored in a **secure campus drive**

3 Copies of your data (at least!)

2 Copies are on-hand (easily accessible) on different systems (internal hard drive, cloud storage, etc.)

- a “**production**” (working) copy
- a “**production backup**” copy

1 Copy is in another location (“off-site”) from the others with a ***trusted*** service provider

Data storage administrative controls

- Don’t collect identifiers that aren’t relevant to the research
- Data should be de-identified as soon as possible, with pseudonyms replacing names
- Researchers should work from de-identified data and not from identifiable data where possible
- Linking files/keys should be stored separately from de-identified data
- Identifiable data should only be made accessible to team members who **need** access

Data storage technological controls

- Store data on password protected devices
- Data stored on internet-connected devices needs to be encrypted
- Data should be stored in a secured environment or server rather than on individual computers or devices
- Back up devices need to follow the same requirements
- Data must be encrypted and password protected when shared – email should not be used for high risk data

Encryption

Encrypt **individual files**

- Microsoft Office or other applications can be used to password protect and encrypt documents on a file by file basis.

Encrypt your whole drive

- **Full disk encryption** is available on Windows, Mac, iOS, and Android. This protects every file on your device so you don't need to worry about missing a file. You can also encrypt entire external drives.

Create "**virtual encrypted disks**"

- VeraCrypt (3rd party software) can create encrypted virtual disks, where you can store sensitive data files

Recording: How to Implement Encryption to Protect your Research Data Online

Cloud storage

- Public cloud services (Google Drive, Dropbox) cannot be used for medium/high risk data but are fine for low risk data
- Institutional services such as MacDrive or OneDrive accessed through McMaster may be used in combination with encryption
- OneDrive is less flexible when working with outside collaborators but Teams
- MacDrive can create a shared folder that collaborators can access and can create encrypted folders
- Researchers using cloud storage should be careful about who they share files with and should enable security features like MFA

Long term/archival storage

- Consider how you will preserve data over the long term. Do you have any requirements to retain data for a set period? Ethics/Health Canada/Funder/etc.
- Does that data need to be identifiable or can it be anonymized?
- Storing data on a campus server (department/faculty/RHPCS/SEAL) is preferable to storing data on an external drive
- **Keeping data on your laptop isn't a good preservation plan**

Publishing Data

A good way to ensure data is preserved is to publish it by depositing it in a data repository or a data paper

Consider whether you can publish your data in an online repository for preservation and sharing.

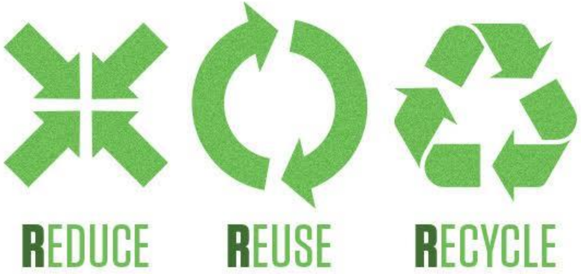


Photo by Lars Kienle on Unsplash.

Data Sharing

- Reproducible research **increases confidence in research results** and avoids article retractions.
- Leads to new **collaborations** – potential for **meta-analyses** over a wider topic area.
- Better **informed policymakers** in healthcare and science as well as hospital stakeholders, professional associations, patient advocates.
- Long term **preservation** and **archiving** of data by established repositories.

How does data sharing work?

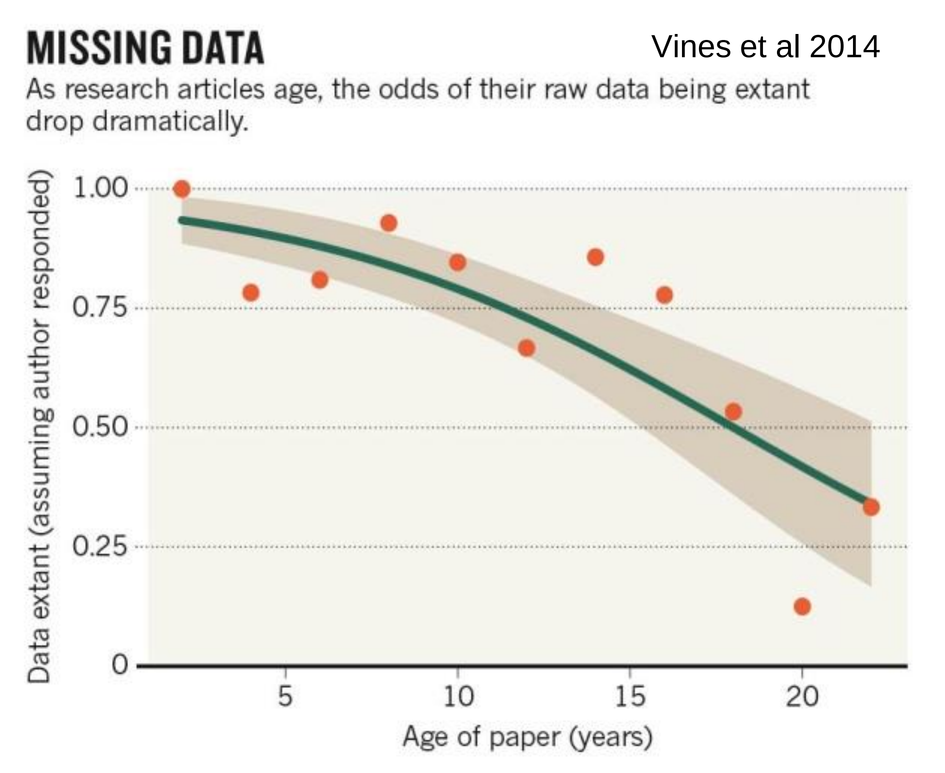
Researcher/
research team

Citation

Publica citing
your findings

F

U



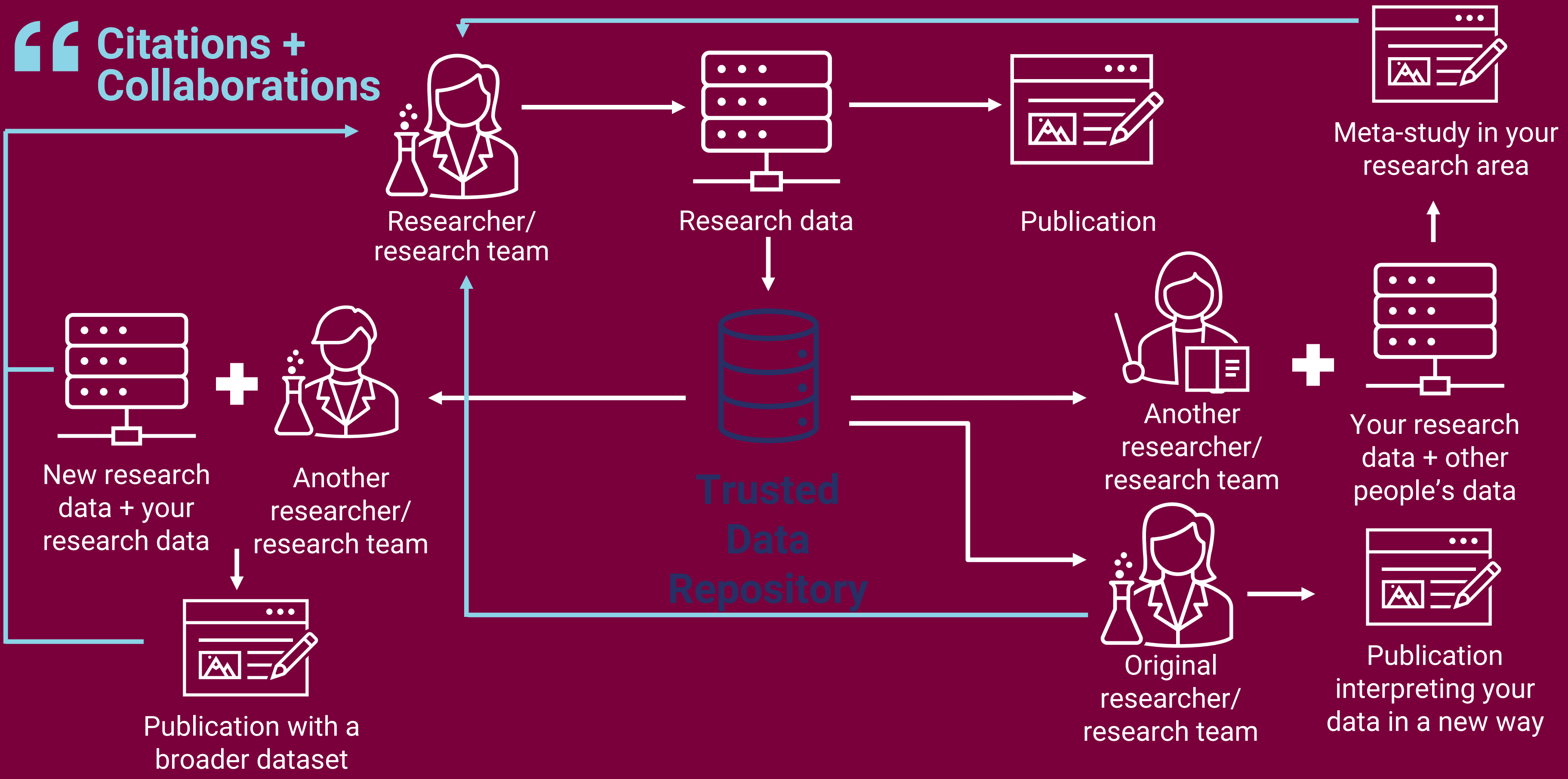
MISSING DATA

Vines et al 2014

As research articles age, the odds of their raw data being extant drop dramatically.

research team

What is data sharing?



Citations + Collaborations

Researcher/ research team

Research data

Publication

Meta-study in your research area

Trusted Data Repository

New research data + your research data

Another researcher/ research team

Publication with a broader dataset

Another researcher/ research team

Your research data + other people's data

Original researcher/ research team

Publication interpreting your data in a new way

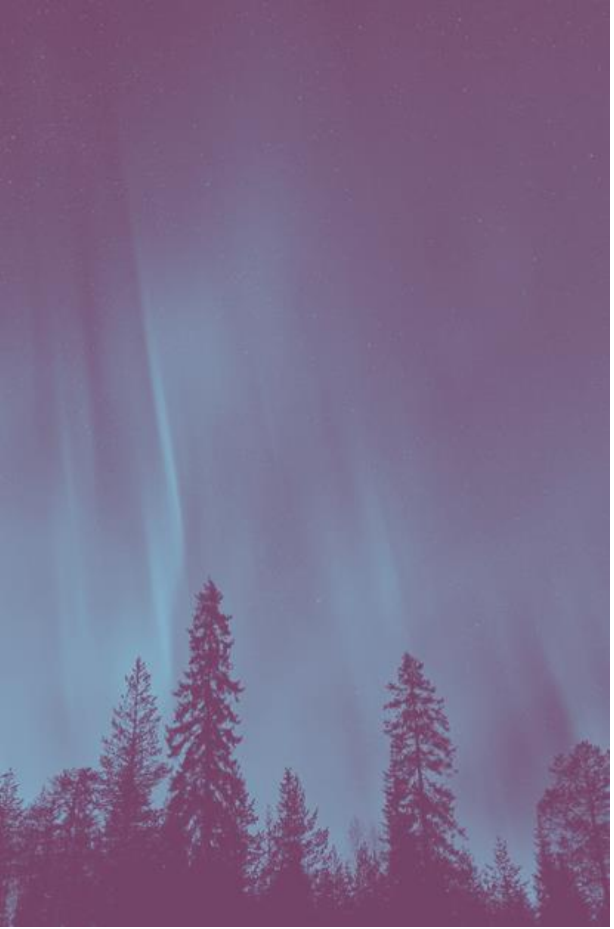
Sharing sensitive data

If you want to publish or share sensitive data, there a few main options:

- **Anonymize the dataset:** remove, replace, or redact all sensitive information from datasets prior to upload in an open repository.
- Data can be shared through closed access portals with restricted access mechanisms and **Data Sharing/Transfer Agreements**
  - Examples of this kind of web portal include ICES and CIHI

Remember you must have patient/participant consent to share data

- Portage's Research Data Management Language for Informed Consent



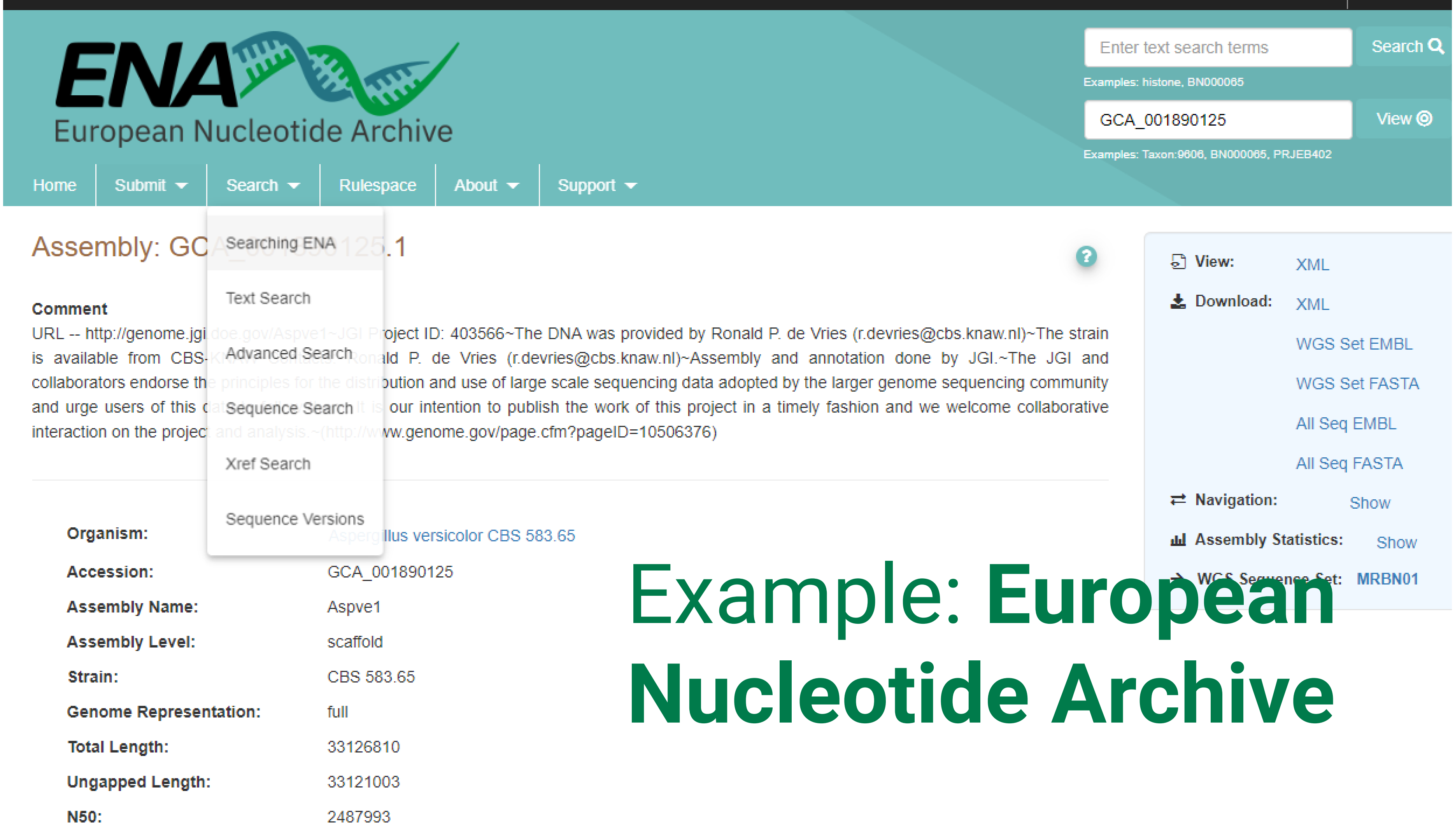
Ok, so where do I put everything?

A **data repository** is a web platform and storage space for researchers to deposit data sets associated with their research. Repositories provide:

- long-term storage and access to research data beyond the life of a grant, research project, or individual careers.
- Discoverability and findability for datasets through features like indexing and DOIs.
- Easy-to-use shared platforms made for research.

RDM Recording: Essentials of open data sharing.

Example: **European Nucleotide Archive**



Assembly: GCA_001890125.1

Comment

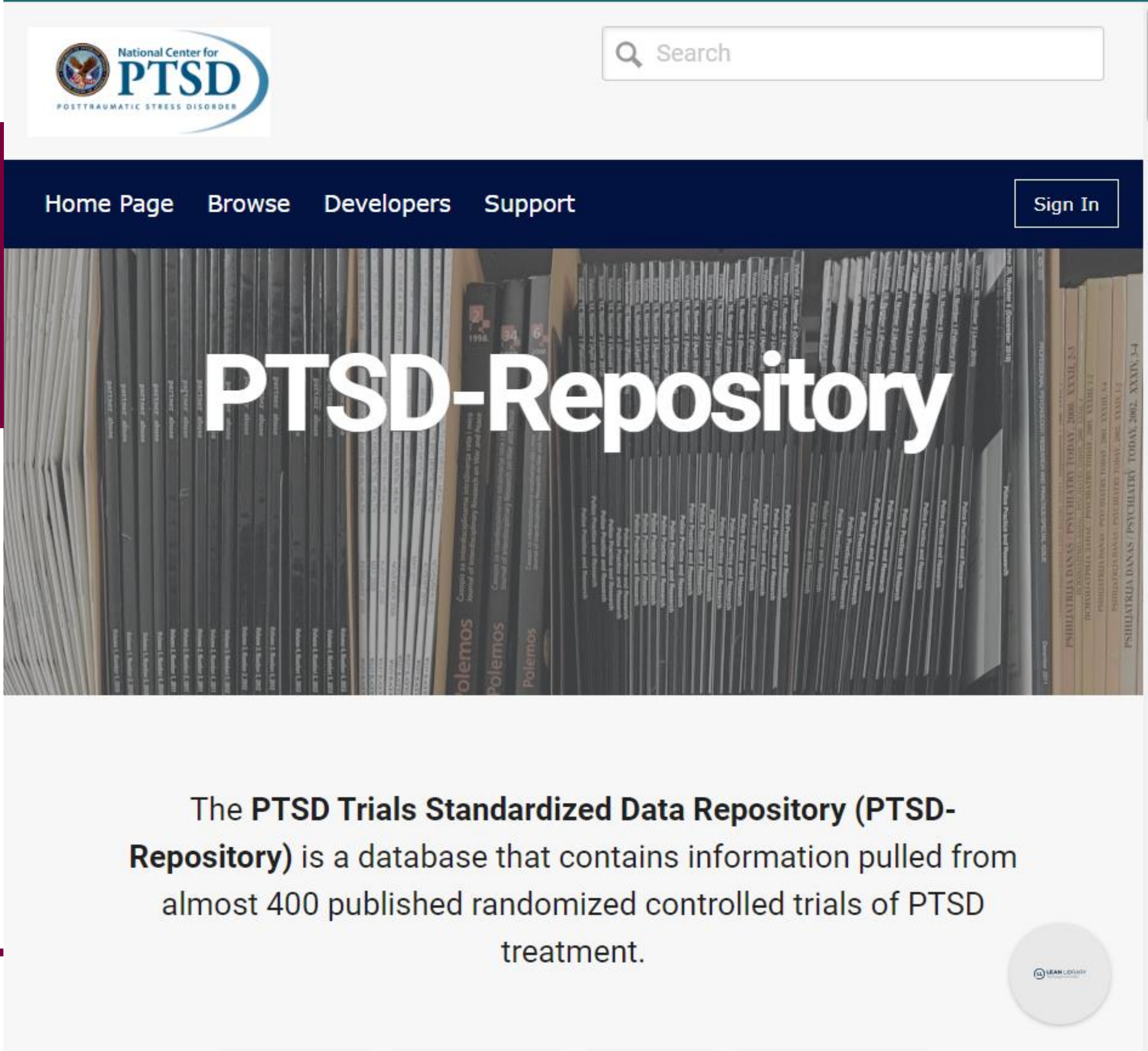
URL -- http://genome.jgi.doe.gov/Aspve1~JGI Project ID: 403566~The DNA was provided by Ronald P. de Vries (r.devries@cbs.knaw.nl)~The strain is available from CBS-KNAW. Contact: Ronald P. de Vries (r.devries@cbs.knaw.nl)~Assembly and annotation done by JGI.~The JGI and collaborators endorse the principles for the distribution and use of large scale sequencing data adopted by the larger genome sequencing community and urge users of this d It is our intention to publish the work of this project in a timely fashion and we welcome collaborative interaction on the project and analysis.~(http://www.genome.gov/page.cfm?pageID=10506376)

Organism:	Aspergillus versicolor CBS 583.65
Accession:	GCA_001890125
Assembly Name:	Aspve1
Assembly Level:	scaffold
Strain:	CBS 583.65
Genome Representation:	full
Total Length:	33126810
Ungapped Length:	33121003
N50:	2487993

View:	XML
Download:	XML WGS Set EMBL WGS Set FASTA All Seq EMBL All Seq FASTA
Navigation:	Show
Assembly Statistics:	Show
WGS Sequence Set:	MRBN01

Example:

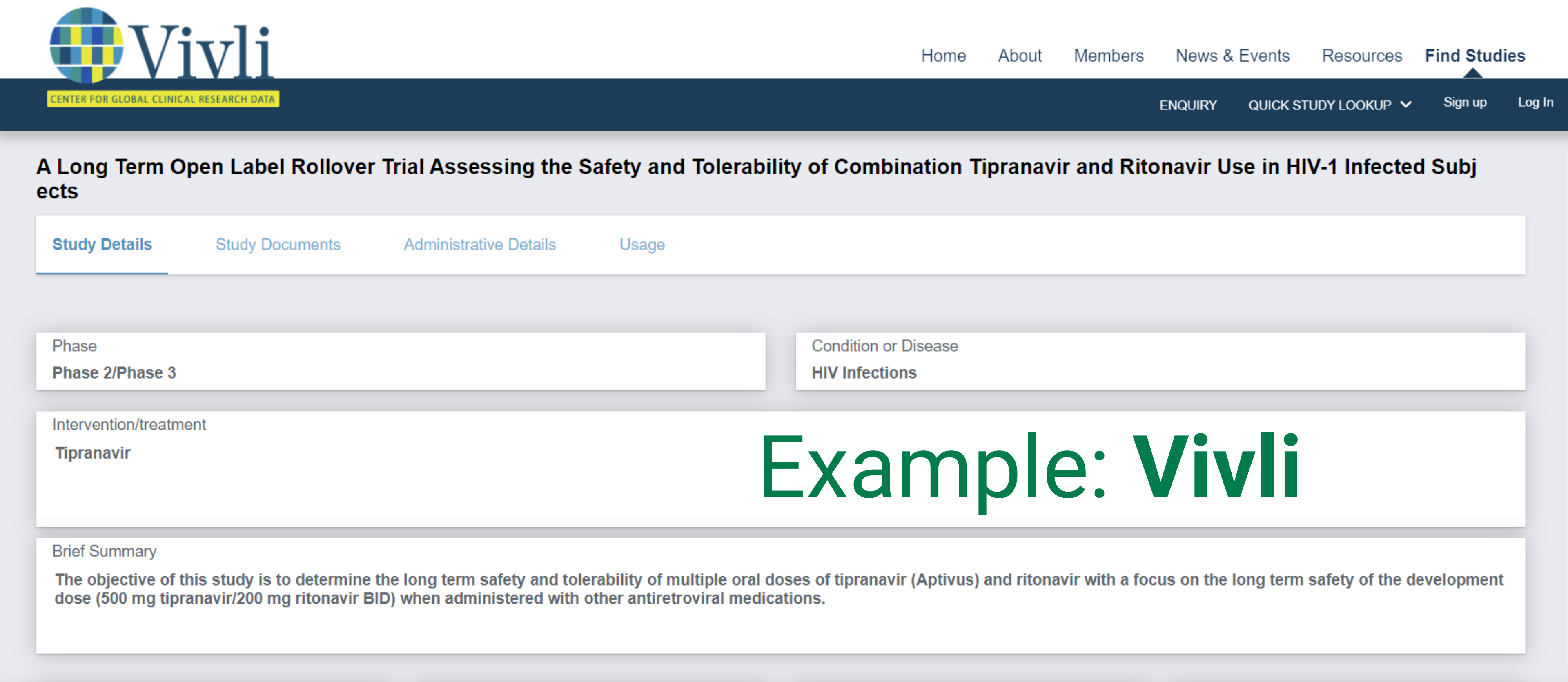
https://ptsd-va.data.socrata.com/



PTSD-Repository

The **PTSD Trials Standardized Data Repository (PTSD-Repository)** is a database that contains information pulled from almost 400 published randomized controlled trials of PTSD treatment.

Example: **Vivli**



A Long Term Open Label Rollover Trial Assessing the Safety and Tolerability of Combination Tipranavir and Ritonavir Use in HIV-1 Infected Subjects

Phase
Phase 2/Phase 3

Condition or Disease
HIV Infections

Intervention/treatment
Tipranavir

Brief Summary
The objective of this study is to determine the long term safety and tolerability of multiple oral doses of tipranavir (Aptivus) and ritonavir with a focus on the long term safety of the development dose (500 mg tipranavir/200 mg ritonavir BID) when administered with other antiretroviral medications.

Digital Research Alliance of Canada Controlled Access Management for Research Data

- Digital Research Alliance of Canada is a pilot project to enable research organizations and existing data repositories to meet researcher needs related to long-term storage, security, sharing and re-use of restricted-access research data.
- McMaster is participating along with the Borealis and FRDR repositories
- Interested in learning more? Email us: rdm@mcmaster.ca

Research Data Management Services

McMaster RDM webpage:	rdm.mcmaster.ca
Contact RDM services at:	rdm@mcmaster.ca
Upcoming RDM webinars:	rdm.mcmaster.ca/events
Recorded RDM webinars:	u.mcmaster.ca/learn-rdm

Make an appointment with a Research Data Management Specialist:
u.mcmaster.ca/rdm-appointments

Lewis & Ruth
Sherman Centre
for Digital Scholarship
scds.ca

McMaster
University | Library

RDM Community of Practice

- Monthly meetings of people interested in RDM at McMaster. Connect with other researchers practicing RDM across the university!

- Next Session on **Exit Protocols + Data Transfer** featuring Dr. Maureen MacDonald and Dr. Blaise Bourdin
  - **April 25th – 11 AM**

- https://u.mcmaster.ca/rdm-community

Photo by CHUTTERSNAP on U