MREB Data Storage and Security Tools: Cover Page

Last Revised: 2020-11-27
Version: 0.2

This document is a collection of four complimentary data storage and security tools, which are intended to guide researchers in determining the data storage and security procedures for handling their research data.

MREB Data Storage & Security Guide: This guide provides an overview of the main issues in the storage and security of research data. It also lists data storage services available at McMaster.

Frequently Asked Questions: The FAQ answers some of the common questions about data storage and security – including questions on encryption, cloud storage, and sharing data with co-investigators at other institutions.

Research Data Risk Matrix: The matrix provides data storage and security guidance for different risk levels of data – low, medium, and high.

MREB Data and Information Storage Glossary: Defines some of the data storage and security terms used in the other documents (e.g. cloud service, portable storage device, anonymous data, etc.).

MREB Data Storage & Security Guide

Last Revised: 2020-11-27
Version: 0.2

The following Data Storage & Security Guide provides information on collecting and storing data for research involving human participants. This guide follows the CASRAI definition of research data (or 'data'), which is "data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results".

Why is data storage relevant?

Creating a data storage plan is essential to research projects at any scale. A thorough and purposeful plan ensures that data are easily retrievable in the short and long term, facilitates appropriate access, and prevents potential data loss.

Why is data security relevant for researchers?

Researchers often collect personal, confidential or sensitive data. It is important to ensure that appropriate measures are established to prevent unauthorized access to the information to protect participants, researchers, and the University.

Regulation of Data Storage & Security

In June 2016, the Tri-Agency released a Statement of Principles on Digital Data Management. It re-emphasized the importance for data to be stored using secure software and formats, and for data to be stored in a manner that enables preservation of and access to the data after the research project is completed. This was followed by a Draft Policy for Research Data Management in 2018. The Draft policy is structured around 3 pillars: Institutional research data management strategies for institutions, data management plans for researchers, and data deposit requirements. The Final policy has not yet been released.

In Ontario, the Personal Health Information Protection Act (PHIPA) 2004 regulates the collection, use, and disclosure of personal health information, stressing the importance for appropriate measures of data storage and security relative to the level of data sensitivity. All researchers must comply with this legislation. PHIPA was most recently updated in 2020.

Considerations for Data Storage

- Inform study participants of the data storage procedures during the consent process
- Always attempt to collect the least amount of specific information possible (e.g. initials instead of name, age instead of date of birth). Data should be anonymized as soon as possible given the context of the study (e.g. for a longitudinal study anonymization is not feasible).

- Consider the longevity of the file formats when storing data. Store data using more basic file formats (.csv, .txt) rather than Microsoft Office or Adobe Suites
- Provide instructions on how to interpret the data in case data are used in future projects (e.g., a code book, readme file, structured metadata etc.)
- When backing up your data, use the ***3-2-1 rule***
    - Save 3 copies of the data, stored in 2 different storage mediums, and have 1 copy off-site (using a trusted service provider)

Considerations for Data Security

Below are data security considerations for three types of security: physical, administrative, and technical.

Physical Security:

- Control access to buildings, rooms, cabinets where data, computers, media, or hardcopy materials are held
- Log the removal of, and access to, media or hard copy material in storage rooms
- Transport sensitive data only under exceptional circumstances, even for repair purposes. *For example, giving a failed hard drive containing sensitive data to a computer manufacturer may cause a breach of security*

Administrative Security:

- Develop organizational rules about who has access to research data
- Impose non-disclosure agreements for managers or users of confidential data
- Do not store sensitive data such as those containing personal information on servers or computers connected to an external network, particularly servers or computers that connect to the internet.

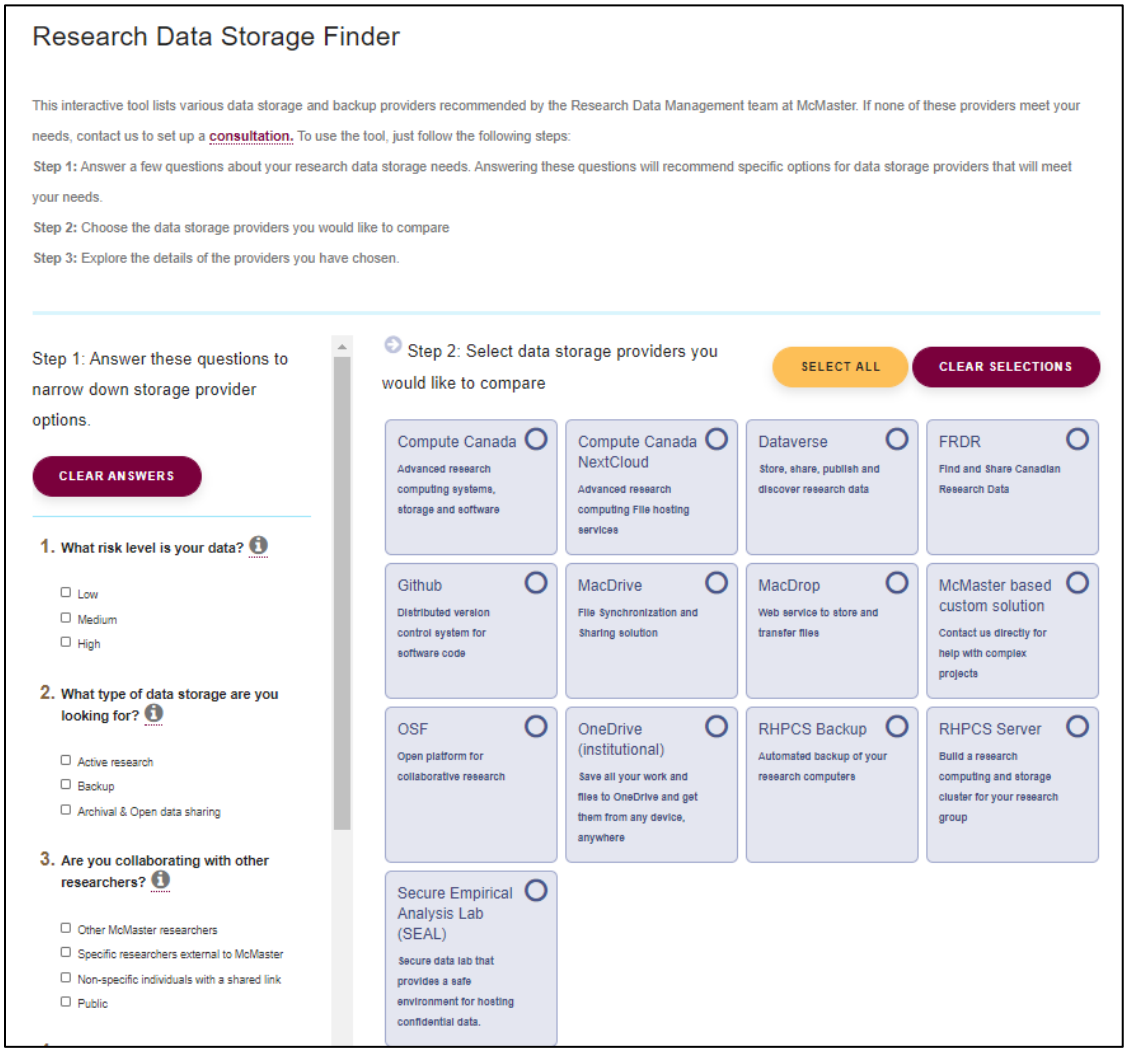
Technical Security:

- Ensure that all computer systems storing confidential data are password protected and that data are stored on encrypted drives wherever possible
- Apply firewall protection and security-related upgrades and patches to operating systems to avoid viruses and malicious code
- Ensure computer software is up-to-date, including anti-virus software
- Implement password protection and controlled access to data files. *For example, 'no access', 'read-only', 'read and write' or 'administrator-only' permissions*
- Control access to files, folders or entire hard drives
- Anonymize personal information that is collected
- De-identify personal information upon data collection
- Do not send personal or confidential data via email

Prior to selecting the data storage and security measure for a specific project, consider the following questions:

- What kind of storage will the research data require? *For example, will you need a physical space to store consent forms? How much digital storage will you need to hold all project files? Where will the data be stored?*
- What types of data are you collecting for the project? Will different levels of security will be necessary for this type of data? *For example, if you are collecting personal health information, how will you ensure that the data are kept secure? Will the data be anonymized? Will the data be encrypted, with identifying information removed?*
- How will the project be managed, if multiple collaborators are involved? *Will the data storage require sharing capabilities? How will you ensure that the project data are only authorized to those who have permission?*
- How will the data be backed up? *How long will the data be stored? Can the data be published to a repository for archival and sharing?*

The McMaster Research Data Storage Finder tool provides an interactive overview of the different data storage providers recommended by RDM@McMaster. Note that this is not an exhaustive list of storage providers - there are many not covered here. The tool focuses on storage providers either supported by McMaster or focused on research or academic needs.



NOTE: While Cloud-based storage (Dropbox) is easy to use, there are limitations to the data security. Some of these Cloud-based storages are based overseas or in another country (e.g. Dropbox server is housed in the US), meaning governments may be able to acquire access to the data upon request of the company. Best practice suggests that commercial storage systems, like Dropbox, should be avoided for confidential data. Alternatively, McMaster does provide locally hosted Cloud-based storage options which include, MacDrive, MacDrop, and Dataverse. Additionally, as part of McMaster's agreement with Microsoft, the use of the McMaster OneDrive is an option, provided that the data are encrypted both at rest and during transfer, and that the data are not synced to local machines in an unencrypted manner.

In addition to the above, there is another facility on campus that might be of interest to researchers. The Secure Empirical Analysis Laboratory (SEAL; formerly known as Public Economics Data Analysis Laboratory, PEDAL) is a medium security data laboratory located in the LR Wilson Building that is available to all researchers at McMaster (and beyond). SEAL allows for the storage and analysis of data in a medium security environment. Originating within the economics department with a focus on administrative data, SEAL now welcomes all researchers from across campus. While SEAL is primarily a BYOD (bring your own data) facility, it also houses a small number of specialized data sets that may be used by accredited researchers. SEAL is open to researchers contributing new datasets to this securely stored collection where there is sufficient demand. Depending on the level of data security required, SEAL's facilities allow for on-site, and/or remote encrypted, data analysis. SEAL operates on a cost recovery basis with prices depending on the level of utilization; please see their webpage for contact information.

Researchers with heavy computational needs are encouraged to look into Compute Canada. Compute Canada is a national governmental organization which provides researchers with free access to high performance computing and large amounts of data storage. Researchers can make a Compute Canada account by following the instructions linked here. They can then sponsor research staff and graduate students to make accounts.

Researchers requiring more secure facilities for data storage and/or analysis, or who are unsure about their security needs are strongly encouraged to contact the Research and High Performance Computing Support group before they start collecting data.

Note that other individual Departments and Faculties may also offer network storage solutions that are suitable for research data. Please consult with your local IT provider for more information.

Additional Resources

Regulatory Documents on Data Management

- Tri-Agency Statement of Principles on Digital Data Management
- Tri-Council Policy Statement (TCPS) II
- Personal Health Information Protection Act, 2004

Data Management Support at McMaster University

- Research Data Management (RDM)
- MacDATA Institute
- Research & High Performance Computing Support

External Resources

- Portage Network
- Portage DMP Assistant

For any questions regarding the contents of this document, please contact RDM@McMaster.

Frequently Asked Questions

Last Revised: 2020-11-27
Version: 0.2

The following is a list of common questions that MREB staff frequently receive about collecting and storing data for research that involves human participants:

NOTE: Many of the answers to these questions will depend on the level of risk associated with the data you have collected. To learn more about the risk level of your data see the Research Data Risk Matrix.

1. Is a password-protected laptop a secure place to store my data?

A password-protected laptop means that one must enter the necessary credentials (a username and password/pin/pattern) in order to use the laptop. Using a password-protected laptop is one of the most common ways to collect and store data. While this method might be secure enough for low-risk data, it is not secure enough for medium and high-risk data. If a laptop must be used for medium to high-risk data, then the files or the hard drive must be encrypted. It is important to note that a password-protected laptop does not necessarily mean that the hard drive is also encrypted. In fact, by default most password-protected laptops are *not* configured with drive encryption. For help with encryption, see the Secure page on the Library RDM website.

Even if you are collecting low-risk data, there are ways to make storage on a password-protected laptop safer. For example, encrypt your hard drive, scan your computer using anti-virus and anti-malware software regularly, update your computer as soon as security updates or patches are available, and use secure browser plugins. Perhaps most importantly, regularly back up and secure your data.

The use of physical security is also recommended for the storage of your data. Avoid common situations where your laptop may get stolen such as leaving it in a vehicle or public place unattended.

2. How long can/should I keep my data?

The short answer: As long as possible! But it depends on your data. Here are some things to consider:

- Researchers need to find a balance between the risks and benefits of retaining or deleting their data, paying special attention to how identifiable or risky the data are. To learn more about the risk level of your data see the *Research Data Management Matrix*. Depending on the project, it may be important to de-identify or anonymize data at the earliest possible step in the research process.
- Researchers need a plan to manage their data securely on an ongoing basis. Ask yourself: what is my data management plan if I move on from this institution, this field, or this career?
- Note that some funders, scholarly journals, organizations, and scholarly societies have guidelines about data retention for publication. For example, the Tri-Agency Open Access Policy on Publications notes that recipients of CIHR grants are required to retain original data sets for a minimum of five years after the end of the grant (or longer if other policies apply). This applies to all data, whether published or not.
- Pay attention to where you are in the world! Canada's TCPS-2 and the UK's *Data Protection Act* do not set out any specific minimum or maximum periods for retaining personal data. However, the U.S. *Office for Human Research Protections* requires research "records frequently held by investigators" such as consent forms or transcribed anonymized interviews to be held "for at least three years after completion of the research." There may also be cases where data sharing agreements impose stricter requirements.
- Researchers are encouraged to share or archive their data by uploading it to a data repository such as Dataverse, McMaster's institutional data repository.

3. What is encryption? When and how should I encrypt my data?

Encryption is a method of protecting your data so that only authorized parties can access content that is intelligible, whereas unauthorized parties will only see unintelligible content. The Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans states that "in general, identifiable data obtained through research that is kept on a computer and connected to the Internet should be encrypted." There are a couple of different methods of encrypting your data such as encrypting individual files, or encrypting your drive (a.k.a volume). Both of these methods have their pros and cons. See the McMaster RDM page on Data Security for more information on encryption.

Encrypting Individual Files

Encrypting only select files such as those that are research-related, or those that contain identifying information, keeps your data safe without any extra complications. Programs such as MS Office and Adobe, offer file-level encryption natively. These programs are recommended when there are few files to encrypt.

Encrypting Your Drive (or volume)

Encrypting your entire drive or the volume on which your data are stored protects against anyone accessing any of your data without your authorization. Encrypting your drive or volume is often more convenient and less prone to human error as all files are encrypted automatically.

For more information on **Encryption** and password protection, **Mobile Devices** (cell phones, laptops, USB keys), **Wi-Fi Security** (on public networks, and when traveling to other institutions), **Passwords** (best practices), **IT Security** (contacts, anti-virus software, spam/phishing emails, network drive security), and **Resources** (confidential waste bins, privacy impact assessments, other resources), please visit the McMaster Privacy Tools and Resource page.

4. What are Cloud services? Is it safe to store, transfer or share my data using the Cloud?

Cloud services store and share data by keeping it on remote servers accessed from the internet. For our purposes we can think of Cloud services as being either internal (i.e., provided by or endorsed by McMaster University) or external (i.e., provided by a third party). You may also see the terms public and private but these can be somewhat misleading. While any use of Cloud services comes with some inherent risk, the risks for internal and external Cloud services servers are somewhat different. The most important differences relate to where data are stored, and how users are authenticated. With external Cloud storage, data are stored on servers that could be located anywhere in the world, and thus subject to data security laws of that country (or countries). With internal Cloud services, your data are stored on locally controlled servers, in our case located at McMaster University, or on servers that are run by trusted partners. Which Cloud services you can use will depend on the risk level of your data. To learn more about the risk level of your data see the Research Data Risk Matrix.

External Cloud Services

- ***Example: Dropbox & iCloud***- McMaster has no agreement in place with these service providers. Users must use personal accounts, unaffiliated with the University, to store their data. MREB recommends against using these types of services for identifiable data.
- ***Example: OneDrive***- McMaster currently has agreements in place. These services are only recommended for low risk data. Data considered to have a medium-risk must be encrypted before uploading. The McMaster institutional OneDrive stores data within Canada.

Internal / McMaster Endorsed Cloud Services

McMaster University has 4 Cloud services that it endorses (MacDrive, Dataverse, OneDrive, and MacDrop). MacDrive and OneDrive are EFSS (Enterprise File Storage and Synchronization) solutions, similar in function to Dropbox, and house the data on-site. It is recommended that any data uploaded to these services are also encrypted. The following can help to distinguish between the internal solutions (MacDrive, OneDrive and MacDrop):

- MacDrive and OneDrive are both tied into the university's central authentication system and is a centrally supported service. MacDrive can provide encrypted web folders.

- MacDrop is only available to users within the FHS. MacDrop is a web service for file transfer of files up to 2GB, with additional space available if needed. It is not designed for file storage.

McMaster faculty, staff, and students have access to Scholars Portal Dataverse, a service provided by the Ontario Council of University Libraries (OCUL). Dataverse is an open-source research data repository system for archiving, describing, and publishing datasets to enable their long-term preservation and reuse.

5. Is it safe to store my data on portable storage devices such as cell phones, tablets, or USB keys?

Data should generally not be stored on any type of portable storage device (other than a laptop), no matter what the risk level is associated with the data. The biggest risk associated with all portable storage devices is that they can be easily lost or stolen. Using portable storage devices that have an internet connection (such as a cell phone or tablet) incurs additional risks that would otherwise not be applicable to a device that does not have an internet connection (such as a USB key). If data must be stored on a portable storage device, it must be stored in an encrypted format.

For internet-connected portable storage devices:

Pros: Collecting data on an internet connected portable storage device such as a cell phone or tablet can be a good choice because the technology is ubiquitous, familiar, and convenient. Additionally, it is often fast, accurate, and portable. These devices are extremely portable and relatively low cost to the researcher. Modern mobile devices including Android 10 phones and iPhones are encrypted by default if they are protected using a passcode.

Cons: When data is stored on portable storage devices it is more vulnerable to theft or loss. Additionally, when mobile devices can easily be connected to 3rd party and unsecured networks leading to a risk of data theft during data transmission. However, the use of encryption (both at the device level and during transmission) can greatly mitigate these risks.

For non-connected portable storage devices:

Pros: Non-connected portable storage devices are very convenient to use for storage and data transfer between computers.

Cons: Portable storage devices such as USB drives are not built for long-term storage and can become corrupted easily if they are not handled properly. In addition, these devices are easily lost or stolen.

6. What is the best way to share data with my co-investigators at other institutions?

Before you share any data collected from human participants in any way, the key is to render that data as low-risk as possible — for instance, by de-identifying or anonymizing it. Before

sharing data with research partners at other institutions; it is important to be mindful of legislation that may be applicable to your co-investigators (e.g. the US Patriot Act / Domestic Security Enhancement Act). If you intend to share data, you require a data sharing agreement and this must be described in your ethics application.

Select a method of sharing your data that is consistent with its risk level. To learn more about the risk level of your data see the Research Data Management Matrix.

- Low-Risk Data: All McMaster hosted Cloud services and McMaster email.
- Medium-Risk Data: Encrypted and password-protected files can be shared via McMaster approved Cloud services, encrypted files via OneDrive, and by McMaster email.
- High-Risk Data: Restricted data should be shared hand to hand on a password-protected and encrypted data storage device. Encrypted and password-protected files may be shared via McMaster approved Cloud services if approved by the MREB. Maintaining ethical high-risk data transfer between institutions may require individualized strategies. Contact the MREB more information.

7. What online survey software should I use?

McMaster University, through the Office of the VP Research, provides a survey service called LimeSurvey. LimeSurvey allows users to create online surveys that can work for large numbers of participants. The online survey software itself is self-guiding for the respondents. The McMaster LimeSurvey service stores data locally and provides templates for MREB and HiREB that were designed with the principles of the TCPS statement on ethics in mind. For this reason, we recommend all researchers to consider using this service rather than using commercial alternatives which may have higher inherent risk associated with them.

8. What is the difference between wireless and wired internet connections? Is one safer?

When it comes to connectivity, computers at McMaster University fall into one of three categories: computers that connect to the Internet wirelessly, computers that connect via wired networks, and computers with no internet connection at all.

When it comes to wired and wireless connections, the risks involved are relatively comparable. Risks on the device are much greater than those on the network (i.e., a user is more likely to contract malware). The difference in security between wired and wireless connections at McMaster is marginal. it is more important to maintain a clean computer.

Research Data Risk Matrix

Last Revised: 2020-11-27 **Version:** 0.2

The following Research Data Matrix is a guideline from the McMaster Research Ethics Board for collecting and storing data for research involving human participants.

	LOW RISK	MEDIUM RISK	HIGH RISK
TYPES OF DATA	Research data that does not contain any sensitive or identifiable information about individuals, organizations, or communities (e.g. data which have been de-identified). NOTE: If in doubt, assume that data are sensitive. Non-sensitive research documentation (e.g. non- confidential protocols and information sheets) Publicly facing information. While public facing information is often considered low-risk, there are cases where informed consent/risk of harm should be closely considered. For example, information regarding racial or ethnic origin could be found on public facing websites, but in certain study contexts could be considered medium or high risk data.	Research data that may or does contain confidential, sensitive, or identifiable information about individuals, organizations, or communities Some sensitive research-related documentation **Personally identifiable information** De-identified records of compensation Data and research protocols related to private or sensitive intellectual property	Research data that contains highly sensitive information about individuals, organizations, or communities (e.g. information about criminal activity) **Personal health information** Contact HiREB to determine the specific requirements for handling PHI Personal financial information such as banking information, income tax returns Data and research protocols related to highly sensitive intellectual property Identifiable data where disclosure, loss, or unauthorized modification of information may result in significant risk for the research participant including reputational damage, significant professional or personal disruption, financial consequences, physical or psychological harm, and legal liability.

EXAMPLES	Completely de-identified or anonymous data Blank consent forms and information sheets Information gathered from a public-facing website	De-identified financial information associated with research payments Identifiable demographic data and/or information about participants' beliefs, opinions, health, etc., that in the context of the study would be considered medium risk.	Depending on the study context, examples of high risk data could include information with regard to racial or ethnic origin; political opinions; religious beliefs or other beliefs of a similar nature; trade union membership; physical or mental health or condition; sexual life; the commission or alleged commission by the data subject of any offence. Video or audio recorded interviews depending on the content Identification keys and signed consent forms
DATA PROTECTION	Research data must always be stored according to protocols approved by the appropriate Research Ethics Board	Collect and store data on password- protected devices, preferably static devices in a secure location such as on a desktop computer in a locked office or on an appropriately protected server. Consider encryption where possible. All research data are subject to the TCPS2 which states "identifiable data obtained through research that are kept on a computer and connected to the Internet should be encrypted." See below for more information about secure data storage, access and transfer.	Collect and store data on password-protected and encrypted devices. Physical security of the data is required (i.e., stored in a locked office or on a protected server). All Research data are subject to the TCPS2 which states "identifiable data obtained through research that is kept on a computer and connected to the Internet should be encrypted."

DATA STORAGE	Local hard drive (e.g., C: drive, "My Documents") Removable storage media (e.g., USB drives, portable hard drives, etc.) University hosted file sharing and storage (e.g., UTS hosted shared network drives) Department hosted file sharing and storage (e.g., department shared network drives) University hosted Cloud based storage (e.g., MacDrive, MacDrop) University sanctioned Cloud based or third party storage (e.g., OneDrive) Department sanctioned Cloud based or third party storage (e.g., DropBox for Business, McMaster-based OneDrive) Personal Cloud storage (e.g., personal OneDrive, Dropbox)	A computer that meets the data protection requirements. Research data are subject to the TCPS2 which states "identifiable data obtained through research that is kept on a computer and connected to the Internet should be encrypted." Public Cloud services (DropBox, iCloud, etc.) for data storage or transfer are not recommended but might be suitable if specified in the REB protocol and <u>if identifiable data is encrypted</u>. Institutional Cloud services (e.g. MacDrive, Dataverse, and MacDrop) or Department sanctioned cloud services (DropBox for Business, McMaster-based OneDrive) are permitted <u>if identifiable data is encrypted</u>. Privacy and security risk are the reasons for preferring internal services over external, particularly those for which there is not an enterprise agreement. Central, departmental, and lab file shares that meet data protection requirements and have been identified in the REB protocol.	A computer or external electronic storage device that meets the data protection requirements. Central, departmental and lab file shares that meet data protection requirements and have been identified in the REB protocol. University hosted file sharing and storage (e.g., UTS/RHPCS hosted shared network drives) Must never be stored in any unsanctioned storage location. Must not be shared via email. Research data are subject to the TCPS2 which states "identifiable data obtained through research that is kept on a computer and connected to the Internet should be encrypted." All public Cloud services (DropBox, iCloud, etc.) for data storage or transfer are <u>strictly prohibited</u>. Use of University hosted or sanctioned Cloud services (McMaster-based OneDrive, MacDrive) to store high risk research data is allowed, however <u>such data must be encrypted</u> at-rest in the cloud. The use of these services is also subject to the restrictions listed below.

DATA ACCESS	No special handling required.	Access to confidential information must be restricted to authorized individuals who have been identified in the REB protocol only.	Access to confidential information must be restricted to authorized individuals only who have been identified in the REB protocol. Note for reviewers: Access should be restricted to the fewest number of individuals possible.
DATA TRANSFER	Can be shared via all Cloud services including public Cloud services (OneDrive etc.)	Encrypted and password-protected files can be shared via McMaster email and McMaster approved Cloud services.	Restricted data should be shared using direct system to system encrypted (TLS) transfer instead of portable devices. Although high-risk research data can be stored in MacDrive and McMaster-based OneDrive, transfer to local machines in an unencrypted format is prohibited. If you wish to sync high-risk data to a local machine, you need to ensure that the local machine complies with TCPS2 guidelines. Files may be shared if they are properly encrypted using password-protected, expiring links.

MREB Data and Information Storage Glossary

Last Revised: 2020-11-27
Version: 0.2

NOTE: For a generalized and comprehensive research data management glossary, visit the CASRAI glossary, the standard dictionary of research administration information. Many of the definitions below are drawn from the CASRAI dictionary.

Anonymized Data

Data which has had all identifying information irrevocably stripped out, with the risk of identification of individuals being very low. Note that this is not necessarily the same as de-identified data, which could include the use of a code that would allow re-identification of an individual with the aid of the coding key.

Anonymous Data

Data that never had identifiers associated with it (e.g. anonymous surveys), and risk of identification of individuals is very low.

Cloud Services

A method of storing and sharing data by keeping it on remote servers accessed from the Internet. Cloud services are maintained, operated and managed by a cloud service provider on storage servers. Cloud services can be public or private. Public cloud services include DropBox, iCloud and OneDrive. McMaster University endorsed private cloud services include Dataverse, MacDrive, and MacDrop. While any use of cloud services comes with some inherent risk, the risks for public and private servers are different. Some main differences include server location, server control, and attack surface.

With public cloud storage:

- data are stored in servers that could be anywhere in the world, and thus subject to that country's laws
- access both to data stored there and to the cloud services themselves are controlled by 3rd-party private companies
- might have sprawling infrastructure with many different points where an unauthorized user could attempt to extract data

With private cloud services:

- your data are stored on local servers on McMaster University premises
- access to stored data is controlled by McMaster University
- usually less open to attacks from unauthorized users

Cloud Computing

A large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms and services are delivered on demand to external customers over the Internet.

Data Lifecycle

The data lifecycle refers to all of the stages in the existence of digital information from collection to destruction. A lifecycle view is used to enable active management of the data over time, thus maintaining security, accessibility, and utility.

Data Management Plan (DMP)

A formal statement describing how research data will be managed and documented throughout a research project and the terms regarding the subsequent deposit of the data with a data repository for long-term management and preservation. Almost all DMPs contain the following core elements: metadata, policies for access and sharing, policies for re-use and redistribution, and plans for archiving preservation and destruction. McMaster encourages the use of the Portage DMP assistant, a bilingual tool for preparing DMPs that follows best practices in data stewardship and walks researchers step-by-step through key questions about data management.

Data Security

A description of security measures to protect the data - e.g. will data storage require additional security levels (such as residing on a non-internet connected private subnet, encryption states, etc.).

Data Sharing

The practice of making data available for reuse. This may be done, for example, by depositing the data in a repository or through data publication.

De-Identification

1. The act of minimally editing individual-level data to decrease the probability of discovering an individual's identity. It involves masking direct identifiers (e.g., name, phone number, address) as well as transforming indirect identifiers that could be used alone or in combination to identify an individual (e.g., birth dates, geographic details, dates of key events). If done correctly, de-identification is a defensible, repeatable, and auditable process that consistently provides assurance, based on generally accepted and repeatable statistical methodologies, so that there is a very small risk of re-identification of any data that are released.
2. The use of one or more techniques designed to make it impossible — or at least more difficult — to identify a particular individual from stored data related to them. The purpose of data anonymization is to protect the privacy of the individual and to make it legal for governments and businesses to share their data without obtaining permission. Such data have proven to be very valuable for researchers, particularly in health care. Data anonymization methods include removing personally identifiable information (e.g., names, addresses, social insurance numbers, Medicare numbers, etc.), or using obfuscation methods such as encryption, hashing, generalization, pseudonymization, and perturbation. As

governments move forward with open government initiatives, more data are becoming publicly available over the Internet. Much of these data have been scrubbed to create "limited datasets".

Deletion

The process of destroying data stored on hard disks, mobile devices and other forms of electronic media so that it is completely unreadable and cannot be accessed or used.

Encryption

Encryption is a method of encoding your data so that only you, or someone you authorize, can access it. The Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans states that "in general, identifiable data obtained through research that is kept on a computer and connected to the Internet should be encrypted." This can be done by either encrypting individual files, or by encrypting entire volumes (storage devices).

High-Risk Data

High-risk data requires very strong controls against unauthorized disclosure, loss, and modification. Disclosure, loss, or unauthorized modification of restricted information may result in significant risk for both the participant and the researcher including reputational damage, significant professional or personal disruption, financial consequences and legal liability. Depending on the study context, examples of high-risk data could include information with regard to racial or ethnic origin; political opinions; religious beliefs or other beliefs of a similar nature; trade union membership; physical or mental health or condition; sexual life; the commission or alleged commission by the data subject of any offence or criminal activity. Other examples of high-risk data may include Personally Identifiable Information (PII) (where a breach of confidentiality would carry a high risk for research participants), Personal Health Information (PHI) and credit card information (PCI). *See the Research Data Management Matrix for information handling guidance.*

Low-Risk Data

Low-risk data requires controls against unauthorized modification for the sake of data integrity rather than to prevent risk to researchers or research participants. Examples of unrestricted information may include completely de-identified or anonymous data, blank consent forms and information sheets, and information gathered from a public-facing website. *See the Research Data Management Matrix for information handling guidance.*

Medium-Risk Data

Medium-risk data requires strong controls against unauthorized disclosure, loss, and modification. Disclosure, loss, or unauthorized modification of confidential information may result in putting participants at risk. *See the Research Data Management Matrix for information handling guidance.*

Metadata

Literally, "data about data", metadata define and describe the characteristics of other data, used to improve both understanding of the data themselves as well as data-related processes. Business metadata includes the names and business definitions of subject areas, entities and attributes, attribute data types and other attribute properties, range descriptions, valid domain values and their definitions. Technical metadata includes physical database table and column names, column properties, and the properties of other database objects, including how data are stored. Process metadata is data that defines and describes the characteristics of other system elements (processes, business rules, programs, jobs, tools, etc.). Data stewardship metadata is data about data stewards, stewardship processes, and responsibility assignments.

Non-Identifiable Data

Data that could not lead to the identification of a specific individual, to distinguishing one person from another, or to personally identifiable information. These may be data that have been de-identified, or that could not lead to personally identifiable information in the first place.

Online Survey Software

Online survey software provide questionnaires that participants can complete over the Internet. They are usually Web forms along with a database to store the answers, and may also include statistical software to provide analytics.

Personal Health Information (PHI)

Personal health information (PHI), also referred to as protected health information, generally refers to demographic information, medical history, test and laboratory results, insurance information, and other data that a healthcare professional collects to identify an individual and determine appropriate care.

Personally Identifiable Information (PII)

1. Data which relate to a living individual who can be identified
   a. from those data, or
   b. from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller, and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual.
2. Any data that could potentially identify a specific individual. Any information that can be used to distinguish one person from another and can be used for de-anonymizing anonymous data can be considered personally identifiable data.
3. Data are identifiable if the information contains the name of an individual, or other identifying items such as birth date, address or geocoding. Data will be identifiable if the information contains a unique personal identifier and the holder of the information also has the master list linking the identifiers to individuals. Data may also be identifiable because of the number of different pieces of information known about a particular individual. It may also be possible to ascertain the identity of individuals from aggregated data where there are very few individuals in a particular category. Identifiability is dependent on the amount of

information held and also on the skills and technology of the holder.

Portable Device

A portable device is any small form factor computing device that is designed to be held and used in the hands. Portable devices are becoming an increasingly important part of personal computing as the capabilities of devices like laptops, tablets and smartphones continue to improve. A portable device may also be called a handheld device or mobile device.

Public Facing

A public facing resource accepts anonymous connection requests from any public internet protocol address. In other words, public facing resources are externally accessible resources that the public can access, without the requirement of being part of specific private subnets.

Raw Data

Data that have not been processed for meaningful use. Although raw data have the potential to become "information," they require selective extraction, organization, and sometimes analysis and formatting for presentation. As a result of processing, raw data sometimes end up in a database, which enables the data to become accessible for further processing and analysis in a number of different ways. Raw data have yet to be de-identified and thus, if there is any stage of the data lifecycle wherein your data will contain PII, it is this stage.

Research Data

Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. All other digital and non-digital content have the potential of becoming research data. Research data may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data, or repurposed data.

Research Data Management (RDM)

Data Management refers to the storage, access, and preservation of data produced from a given investigation. 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. Specific activities and issues that fall within the category of data management include: File naming (the proper way to name computer files); data quality control and quality assurance; data access; data documentation (including levels of uncertainty); metadata creation and controlled vocabularies; data storage; data archiving and preservation; data sharing and reuse; data integrity; data security; data privacy; data rights; notebook protocols (lab or field).