Sensitive Data Toolkit for Researchers

Part 1: Glossary of Terms for Sensitive Data used for Research Purposes

Prepared by the Portage Network Sensitive Data Expert Group on behalf of the Canadian Association of Research Libraries (CARL)

SEPTEMBER 2020

Portage Network
Canadian Association of Research Libraries
portage@carl-abrc.ca

www.carl-abrc.ca

portage
SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA

CARL ABRC
CANADIAN ASSOCIATION OF RESEARCH LIBRARIES
ASSOCIATION DES BIBLIOTHÈQUES DE RECHERCHE DU CANADA

Introduction

The Sensitive Data Expert Group of the Portage Network has create a suite of tools for Canadian researchers. These tools have been created to help researchers understand how research data is involved in the research ethics process, and to address the evolution of research data management (RDM) practices such as data sharing and deposit in the context of existing research ethics frameworks.

This tool, a Glossary of Terms for Sensitive Data used for Research Purposes, is intended to provide definitions for a number of common terms used in the discussion of the management of sensitive data in the Canadian context. These definitions have been informed by an environmental scan of data policies, procedures, and forms at Brock University; Mount Saint Vincent University; Research Data Canada, CANARIE; St. John's, Newfoundland and Labrador Health Research Ethics Authority; University of Alberta; University of Ottawa; University of Toronto; and Queen's University. Additional resources which were consulted include documents from the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (TCPS2), the Consortia Advancing Standards in Research Administration Information (CASRAI), the General Data Protection Regulation of the European Union (GDPR), the Ontario Personal Health Information Protection Act (PHIPA), and the United Nations Educational, Scientific, and Cultural Organization (UNESCO).

The Portage Network's Responsible RDM Practices for Sensitive Data Expert Group is composed of a broad membership from research communities - including research ethics professionals, representatives of funding agencies, and members of Indigenous organizations - with direct interests in sensitive research data. The group works together to develop practical guidance and tools for managing sensitive data in the Canadian landscape.

Glossary of Terms for Sensitive Data used for Research Purposes

Note: Due to the evolving nature of research data management, this is a living tool. Where there are discrepancies in definitions between Canadian and international bodies, the Canadian definition will supersede. Where there are discrepancies between two Canadian definitions pertaining to human participant research, the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (TCPS2) definition will supersede. To add items or suggest revisions to this document, please contact the Portage Network.

Administrative data: Information collected primarily for administrative, and not research purposes. This includes profiles and curriculum vitae of researchers, the scope and impact of research projects, funding, citations, and research outcomes. This type of data is collected by government departments and other organizations for the purposes of registration, transaction and record keeping, usually during the delivery of a service.

Anonymized data (De-identified): Data that are irrevocably stripped of direct identifiers; a code is not kept to allow future re-linkage of data to direct identifiers, and the risk of re-identification of individuals from any remaining indirect identifiers is low or very low. According to the TCPS2 (2018), secondary use of anonymized (de-identified) human participant data for research purposes requires Research Ethics Board (REB) review and clearance.

Anonymous data: Data that have never had direct identifiers from the point of original collection; the risk of identification from indirect identifiers is low or very low. Note that according to Article 2.4, TCPS2 (2018), REB review is not required for research that relies exclusively on secondary use of anonymous human participant information, or anonymous human biological materials, so long as the process of data linkage, recording, or dissemination of results does not generate identifiable information.

Authorized data access: When personnel (as individuals or according to role) are given access to data by the researcher.

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. 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 potential vulnerabilities.

- **Public cloud services** are companies that provide free or for-fee storage to multiple customers using remote servers that are managed by a provider. Data is stored on the provider's server and the provider is responsible for the management and maintenance of the data center. While public cloud services are shared among different customers, each customer's data and applications running in the cloud remain hidden from other cloud customers. Examples of public cloud services include Google Drive, DropBox, iCloud and OneDrive (personal). Of particular note to Canadian academic institutions, public cloud storage is provided using servers that are outside of the institution's control and that could be anywhere in the world, and thus subject to the host country's laws. While public clouds employ privacy and security measures, they have sprawling infrastructure with many different points where an unauthorized user could attempt to extract data. In some cases private services may be less open to such attacks.

- **Private cloud services** store data on internal servers using local infrastructure. A private cloud is not shared with any other organization. Management of and access to data stored in private cloud services is controlled by the home institution or organization. Data resides on the organization's intranet or hosted data center where it is protected behind a firewall. Management, maintenance and updating of data centers is the responsibility of the institution. Private clouds may offer an increased level of security as they share very few, if any, resources with other organizations. However; not all institutions have the infrastructure and/or personnel needed to host private services.

Confidential data: information entrusted to a person, organization or entity with the intent that it be kept private and access to that information be controlled or restricted.

Data: facts, measurements, recordings, records, or observations about the world. Data may be in any format or medium and take many forms, including writings, notes, numbers, symbols, text, images, films, video, sound recordings, pictorial reproductions, drawings, designs or other graphical representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing algorithms, or statistical records.

Data access: the right or opportunity to use or look at information kept in a database or a repository/research environment/ virtual research environment. Data access is an important aspect of research data management.

Data custodian: An information technology (IT) individual or organization responsible for the IT infrastructure providing and protecting data in conformance with the policies and practices prescribed by data governance.

Data deposit: Placing of data into a recognized repository (institutional, national, open source, disciplinary, multi-disciplinary, or other) for retention and possible use by other researchers.

Data lifecycle: All of the stages in the existence of data from collection to destruction. A lifecycle view is used to enable active management of data over time, thus maintaining security, accessibility, and utility.

Data management: An active process involving application of various standards and best practices throughout the research cycle to increase efficiency and enable reusability of the data and its products. This includes, but is not limited to, standards and practices related to data management planning, metadata, storage, licensing, ethics, discovery, sharing, preservation, and reuse.

Data management plan (DMP): A formal statement describing how research data will be managed and documented throughout a research project. Most DMPs contain the following core elements: metadata; policies for data access, sharing, re-use and redistribution; and plans for archiving preservation and destruction.

Data publication: The release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way.

Data repository: A place where researchers can submit their data to be stored and managed. Data repositories may have specific requirements regarding subject, research domain, format or data re-use. Others are open to receiving a broad variety of data. Materials in data repositories are ideally curated and stewarded to ensure they are authentic, discoverable, and appropriately accessible over the medium term.

Data security: Ways of keeping data safe so that the researcher can appropriately access the data when needed (data availability); so that the data are not altered (data integrity); so that the confidentiality of the data is preserved (data confidentiality); and so that the data is carefully preserved and disposed of appropriately (retention/destruction).

Data sharing: The practice of making data available for discovery and reuse. This may be done by depositing the data in a repository for access or through other means of data publication. Data sharing may be subject to conditions and limitations, particularly when data are sensitive, subject to legal or regulatory requirements or when data are proprietary in nature.

Data steward: an individual responsible for data definition (i.e., defining the characteristics of the elements in a database) and access authorization, particularly

data access and disclosure to third parties. Data stewards, along with principal investigators, provide support for:

(a) Data collection, data integration, or reuse of existing data;

(b) Review of data quality;

(c) Description of scientific workflow/process;

(d) Provision of standards-compliant metadata; and,

(e) Submission of data and data productions.

Data stewards are responsible for, and principal investigators are consulted and informed on:

(a) Preservation of data and data products; and,

(b) Provision of formats (e.g., web services, NetCDF, etc.) for data discovery and integration.

De-identification: The act of changing individual-level data to decrease the probability of disclosing an individual's identity. This can involve masking direct identifiers (e.g., name, phone number, address) as well as transforming (e.g., recoding, combining) or suppressing 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 minimizes and therefore mitigates risk of re-identification of any data shared or released.

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 recovered, accessed or used.

Digital Preservation (or Archiving): The active process of managing digital content over time. This process involves activities such as selecting content for preservation, preparing and maintaining content in formats and environments that ensure ongoing usability, and having strategies in place to ensure content can be accessed over the long term.

Downstream risk/harm: Negative consequences or effects upon individuals, groups, or entities that are unanticipated and unintended at the time a decision or action is made. When collecting sensitive data and considering storage, maintenance, and possible secondary use, the potential for both immediate and downstream harms should be carefully considered.

Encryption: A security method that involves translating information into a different form or encoding data so that only people who have access to a decryption key or password can read it. The purpose of encryption is to protect the confidentiality of digital data stored on computer systems and transferred using the internet or on portable storage devices. Encrypted data appears scrambled and unreadable to anyone who does not have the secret code, password or decryption key.

Harm: negative consequences or effects to the welfare of individuals, groups, or entities. The nature of harm may include social, behavioural, psychological, physical, economic, ecological, or legal consequences. Harm also includes impact on privacy, security, reputation and status.

High risk data: are data which require strong controls against unauthorized disclosure, loss, and modification that could result in significant risk of harm to both researchers and research participants, be they individuals, communities, organizations, or entities. Examples of high risk data include, but are not limited to, information related 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; and the commission or alleged commission of any offence by the participant.

Identifying information: information that identifies an individual, organization or entity, or information for which it is reasonably foreseeable, under given circumstances, could be utilized, either alone or with other information, to make such an identification.

> **Directly identifying information:** information that identifies a specific individual, organization or entity through direct identifiers (e.g., name, social insurance number, personal health number).
>
> **Indirectly identifying information:** information that, while not directly identifying, when used or considered in combination with other information, could reasonably be expected to identify an individual, organization or entity (e.g., date of birth, education level, place of residence or other detailed geographic information, or unique personal characteristics).

Identifiable person: One who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to their physical, physiological, mental, economic, cultural or social identity. Although the word "person" is used, the same considerations could be relevant for a community, organization or entity.

Legally accessible data: data that is accessible to the public based on the presence of a legally designated custodian/steward (see **data steward**) who protects the privacy

and proprietary interests of the information (e.g., an access to information and privacy coordinator or a guardian of Canadian census data).

Linkage: The combining of two or more data sets with common elements to provide further information or new datasets. Note that growing numbers of databases and advancing technological capacity to link databases create new research opportunities, but also new privacy risks. In particular, linkage of de-identified or anonymized databases may permit re-identification of individuals. Article 5.7 of the TCPS2 requires that researchers seek REB review and approval prior to linking datasets obtained from human participant research.

Low risk data: Data which requires controls against unauthorized modification for the sake of data integrity rather than to prevent risk to researchers or research participants. Examples include but are not limited to, unrestricted information composed of completely de-identified or anonymous data, blank consent forms and information sheets, and information gathered from a public-facing website.

Medium risk data: Data which requires strong controls against unauthorized disclosure, loss, and modification and in, most cases is considered confidential. Disclosure, loss, or unauthorized modification of confidential information may result in putting research participants at risk. Examples of Confidential information may include Personally Identifiable Information (PII), Personal Health Information (PHI) and credit card information (PCI).

Metadata: Literally, "data about data"; data that defines and describes the characteristics of other data, used to improve both understanding of data and 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 is 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 upon initial collection can not lead to the identification of a specific individual, to distinguishing one person from another, or to personally identifiable information. Note that caution must be used if two such data sets are to be linked as the linkage could result in identifiable data.

Online survey software: Software used by researchers to collect data from participants through questionnaires completed over the internet, usually through web forms connected to a database to store the answers, and statistical software to

provide analytics. Researchers and participants should be aware of the privacy and security policies of specific software platforms and providers.

Open access (OA): a set of principles and range of practices through which research outputs are distributed online with limited or no barriers to access. Any kind of digital content can be OA, from texts and data, to software, audio, video, and multimedia. While most of these are related to text only, a growing number of OA sources are integrating text with images, data, and executable code. OA can also apply to non-scholarly content, like music, movies, and novels.

Open data: Structured data that are accessible, machine-readable, usable, intelligible, and freely shared. Open data can be freely used, re-used, built on, and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike.

Personal information: identifiable information pertaining to an individual, including, but not limited to, any of the following:

- name, address or telephone number,
- race, national or ethnic origin, colour, or religious or political beliefs or associations,
- age, gender, sexual orientation, marital status or family status,
- an identifying number, symbol or other particular assigned to the individual, or internet protocol identifier (IP address)
- the individual's fingerprints, blood type or inheritable characteristics,
- information about the individual's health care status or history, including physical or mental health conditions,
- information about the individual's educational, financial, criminal or employment status or history,
- the individual's personal views, opinions, evaluations, comments.

*Note that the assessment of whether information is identifiable is made in the context of the specific research.

Personal health information: identifying information about an individual in oral or recorded form, if the information:

(a) relates to the physical or mental health of the individual, including information that consists of the health history of the individual's family,

(b) relates to the providing of health care to the individual, including the identification of a person as a provider of health care to the individual,

(c) is a plan of service within the meaning of the Home Care and Community Services Act, 1994 for the individual,

(d) relates to payments or eligibility for health care, or eligibility for coverage for health care, in respect of the individual,

(e) relates to the donation by the individual of any body part or bodily substance of the individual or is derived from the testing or examination of any such body part or bodily substance,

(f) the individual's health number, or

(g) identifies an individual's substitute decision-maker.

Note that the assessment of whether information is identifiable is made in the context of the specific research.

Personally identifiable information (PII): Similar to direct identifiers, PII refers to information that can be used to uniquely identify, contact, or locate a person, organization, or entity (heretofore 'individual') or can be used with other sources to achieve the same end. PII includes but is not limited to the name of an individual, or other identifying items such as birth date, address or geocoding. Data coded with unique personal identifiers (UPI) are still identifiable if the holder of the information also has the master list or key linking the UPI to individuals. Data may also be identifiable on this level 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 unique characteristics of information, the amount of information held that may be combined resulting in identification, and on the skills and technology of the data holder that may allow such combinations.

Portable storage devices: Any device or media which is easily transportable, upon which information can be stored. This definition is not restricted to purpose built storage devices such as CD/DVDs, removable hard drives, and USB flash drives, but also may include laptop computers, tablets, smart phones, PDAs, and any other portable computing device. Portable storage devices can be either internet-connected or not-connected and different data security measures will apply in each case.

Primary data: information collected by a researcher from first-hand sources (e.g., directly from research participants), for the specific purpose of the research question at hand.

Pseudonymization: the process of assigning a fictional name or alias (a pseudonym) to a specific person, group or place in such a way that data can no longer be attributed to them without the use of additional information. In research, the process of pseudonymization is a type of de-identification used to protect the identity of research participants, and organizations or entities involved in research.

Public Facing: A resource which accepts anonymous connection requests from any public internet protocol address; in other words, externally accessible resources that the public can access.

Publicly available information: Any existing stored documentary material, record or publication, which may or may not include identifiable information, and that

(a) has no restrictions on its use or distribution and, in the case of human participant research data, holds no reasonable expectation of privacy; or

(b) may be released to the public under mechanisms set out by regulations based on the presence of a custodian/steward designated in accordance with access to information and privacy legislation who protects privacy and proprietary interests associated with the information (e.g., an access to information and privacy coordinator or a guardian of Canadian census data).

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. Unless raw data are anonymous, they have yet to be de-identified and may therefore be considered to be sensitive.

Research data: Any information that has been collected, observed, generated or created to conduct research, validate research findings and results, and enable reuse or replication.

Risk is the potential for harm to individuals, communities, organizations or entities . Risk is a function of the magnitude or seriousness of the harm, and the probability that it will occur, whether to research participants or to third parties. Consideration of risk should encompass immediate harm, delayed harm and downstream harm.

Secondary data: information, usually collected by someone other than the researcher, for purposes other than the research question at hand. Secondary data is pre-existing information re-used in a different context than that for which it was originally collected.

Sensitive data: information that must be safeguarded against unwarranted access or disclosure. May include:

- Personal information
- Personal health information
- Educational records
- Customer records
- Financial information
- Criminal information
- Geographic information (e.g., detailed locations of endangered species)
- Confidential personnel information
- Information that is deemed to be confidential; information entrusted to a person, organization or entity with the intent that it be kept private and access be controlled or restricted.
- Information that is protected by institutional policy from unauthorized access

Sensitive data includes any information relating to an identified or identifiable natural person, organization or entity.

Traditional knowledge/Indigenous sensitive data: The knowledge held by First Nations, Inuit and Métis peoples, the Aboriginal peoples of Canada. Traditional knowledge is specific to place, usually transmitted orally, and rooted in the experience of multiple generations. It is determined by an Indigenous community's land, environment, region, culture, and language. It may also be new knowledge transmitted to subsequent generations.

TCPS2: The *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (TCPS or the Policy) is a joint policy of Canada's three federal research agencies – the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council (SSHRC). This Policy provides the principles and guidelines that govern the ethical conduct of human participant research in Canadian institutions eligible to receive funding from the three federal agencies noted. This policy extends to all human participant research conducted under the auspices and/or jurisdiction of the institution.