Open Access, Data Sharing and Archiving: What It Means for Researchers

TABLE OF CONTENT

Introduction

As of May 1, 2015, all researchers receiving funding from either CIHR, NSERC, or SSHRC are required to comply with the Tri-Agency Open Access Policy on Publications. Grant recipients are reminded that by accepting Agency funds they have accepted the terms and conditions of the grant or award as set out in the Agencies' policies and guidelines. What this means to researchers is that you are required to ensure that any peer-reviewed journal publications arising from Agency-supported research are to be made freely accessible to the public within 12 months of publication. For CIHR funded research projects certain data are required to be made public as well:

> Deposit bioinformatics, atomic, and molecular coordinate data into the appropriate public database (e.g. gene sequences deposited in GenBank) immediately upon publication of research results.

There is growing support for sharing research data for the same ethical reasons as cited for supporting open access to research publications: it minimizes the burden on participants by encouraging reuse of research data thereby reducing duplication of research studies, it reinforces open scientific inquiry and accountability, and maximizes research benefits to society at large. When properly managed, responsibly shared

research data enables researchers to ask new questions, test alternative hypotheses, deploy innovative methodologies and collaborate with other researchers around the globe. However, the Tri-Agencies have not taken a clear position on it to date, even though there is mention of sharing data in a SSHRC Research Data Archiving Policy[1] that dates back to 1999 (and last modified in April 08, 2014) which states: All research data collected with the use of SSHRC funds "must be preserved and made available for use by others within a reasonable period of time." The document goes on to define a 'reasonable period of time" as within 2 years.

The SSHRC Policy provides this explanation:

> Sharing data strengthens our collective capacity to meet scholarly standards of openness by providing opportunities to further analyze, replicate, verify and refine research findings. Such opportunities enhance progress within fields of research, avoid duplication of primary collection of data, as well as support the expansion of inter-disciplinary research. In addition, greater availability of research data will contribute to improved training for graduate and undergraduate students, and, through the secondary analysis of existing data, make possible significant economies of scale. Finally, researchers whose work is publicly funded have a special obligation to openness and accountability.

In a recent NIH policy titled "Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research"[2] (February 2015) it states: "NIH intends to make public access to digital scientific data the standard for all NIH-funded research." The NIH policy defines digital scientific data as:

> "…the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens."

Furthermore, the policy cites similar reasons as the SSHRC policy above for its justification of its open access policy. However, making scientific digital data publicly available still has

[1] SSHRC Research Data Archiving Policy (1999): http://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-eng.aspx

[2] National Institutes of Health Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research (2015): https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf

a few kinks to work out as a 2013 document from the EU points out ("Open Data Access Policies and Strategies in the European Research Area and Beyond"[3]):

> The heterogeneous nature of scientific data certainly is a challenge for the development of OA data. The emergence of OA scientific data as a valid, citable form of reference is limited by the difficulties associated with the standardisation of data and metadata formats, poor indexation by internet browsers, as well as by the scarcity of directories or registries that could make data more visible. Initiatives from academia and from the non-profit and private sectors seek to address these limitations. A few fields of research use highly standardised formats that facilitate the aggregation and reuse of data; genomics, proteomics, chemical crystallography, geography, astronomy and archaeology are among those fields. The archiving of other, less standardized types of data needs to be carefully thought out in order to generate datasets that will be usable by other researchers. The proliferation of data archiving standards indicates that this issue is addressed by communities of researchers, librarians, and database administrators but probably will not be settled in the near future.

The last sentence captures the present view of the Tri-Agencies. Admittedly, while the world seems to be moving ahead on finding ways to make open access to publications and digital scientific data a requirement of publicly funded research, Canadian agencies seem content to take a 'watch and see' approach to open access to data for the time being.

If researchers are involved in data sharing or the conduct of secondary analysis of data, then the TCPS2 has following to say on the matter.

TCPS2 Chapter5, Section D: [top]

> "Reasons to conduct secondary analyses of data include: avoidance of duplication in primary collection and the associated reduction of burdens on participants; corroboration or criticism of the conclusions of the original project; comparison of change in a research sample over time; application of new tests of hypotheses that were not available at the time of original data collection; and confirmation that the data are authentic. Privacy concerns and questions about the need to seek consent arise, however, when information provided for secondary use in research can be linked to individuals, and when the possibility exists that individuals can be identified in published reports, or through data linkage (Article

[3] Open Data Access Policies and Strategies in the European Research Area and Beyond, Aurore Nicol, Julie Caruso, & Éric Archambault, 2013: http://www.science-metrix.com/pdf/SM_EC_OA_Data.pdf

5.7). Privacy legislation recognizes these concerns and permits secondary use of identifiable information under certain circumstances."

For Help with Publications, Data Sharing or Archiving

If you require more information about Open Access and online repositories, journals that offer open access, data sharing and archiving please go to the Queen's Library website or communicate with: For information about publications contact Rosarie Coughlan, Scholarly Publishing Librarian, at 613-533-6000 x 77529, or email: rosarie.coughlan@queensu.ca. For information regarding data sharing and archiving contact Jeffrey Moon, Data Librarian and Academic Director, Queen's Research Data Centre, 613-533-6000 x 77992, or email: moonj@queensu.ca.

[top]

The Tri-Agency Statement of Principles on Digital Data Management

The Tri-Agency Statement of Principles on Digital Data Management (2015) offers this advice to researchers:

1. It is advisable to have a data management plan.
2. Data should be managed in accordance with the most appropriate and relevant standards and best practices, recognizing that these are in a state of rapid evolution.
3. To determine whether data should be shared and preserved, researchers should consider the data needed to validate research findings and results, support replication and reuses, and consider the potential benefit to their own field of research, fields other than their own and society at large.
4. Data should be collected and stored throughout the research project using software and formats that ensure secure storage and enable preservation of and access to the data well beyond the duration of the research project.
5. All research data should be accompanied by data documentation also known as metadata in accordance with community best practice to enable future users to access, understand and reuse the data.
    - Metadata (also known as data documentation) generally includes statements about who created the data and when; information on how the data was created; information about data quality, accuracy and precision; information to facilitate understanding and reuse of data, such as code books and/or user guides that provide detailed descriptions of the data, variables, coding, etc..

6. Data release can be staged as research progresses, starting with metadata, but data should be shared no later than upon the publication of results. Where possible, data should be linked to relevant publications.
7. A defined period of exclusive use of data for primary research (i.e., an embargo period) is reasonable in some cases.

[top]

Researchers' Ethical Responsibilities

Privacy and Confidentiality

- The TCPS2 (2014), Article 2.1 states: "Where researchers seek to collect, use, share and access different types of information or data about participants, they are expected to determine whether the information or data proposed in research may reasonably be expected to identify an individual."

  It is not necessary that all research data be destroyed after completion of a study. The TCPS2 (2014) does not state that research data must be destroyed. Some laws and regulations might state a minimum retention period. For example, Health Canada requires clinical trial data to be retained for a minimum of 25 years, whereas the 2015-2019 Queen's-QUFA Collective Agreement states a minimum of 5 years.

  A researcher's main ethical consideration concerning their research data is to ensure participant privacy and confidentiality is maintained throughout the retention period of the data. For the purpose of this document "research data" will be defined as "recorded material that validates research findings and results, and enables reuse or replication."[4]

The Letter of Information and Consent Form

The TCPS2 (2014), Article 3.2 states: "Researchers shall provide to prospective participants, or authorized third parties, full disclosure of all information necessary for making an informed decision to participate in a research project."

[top]

TCPS2 (2014) Article 3.2 (i) further elaborates on essential information in the Letter of Information and Consent Form:

> an indication of what information will be collected about participants and for what purposes; an indication of who will have access to information collected

[4] Tri-Agency Statement of Principles on Digital Data Management

about the identity of participants, a description of how confidentiality will be protected (see Article 5.2), a description of the anticipated uses of data; and information indicating who may have a duty to disclose information collected, and to whom such disclosures could be made.

Paragraph (i) touches on issues of privacy and confidentiality, secondary use of data, and the possibility of compelled disclosure by the researcher to third parties for administrative and/or legal purposes. These issues are addressed in further detail in Chapter 5 and, in particular, Article 5.2.

What is important is that participants are informed in the Letter of Information and Consent Form about **how** their privacy will be protected and **what** steps the researcher will take to maintain their confidentiality throughout the full data retention period (see Appendix 1 for suggested Letter of Information and Consent Form language examples).

[top]

Required Information in Letter of Information and Consent Form

- Participants should be informed about what information will be retained and for how long (even if it is indefinitely).
- Participants should be informed that the data may be used for secondary purposes.
- If appropriate, participants should be asked if they want to be contacted in the future by other researchers regarding their data.
- Participants should be informed of whether they can withdraw their data or whether there are restrictions to withdrawing their data and when and if those restrictions apply.
- Researchers should not use words in the Letter of Information or Consent Form like anonymized, aggregate or de-identified but instead explain what information will be removed from the data in lay terms (see Appendix 1 for suggested LOI/CF language examples).

Data Security

TCPS2 (2014) Article 5.3 states: Researchers shall assess privacy risks and threats to the security of information for all stages of the research life cycle, and implement appropriate measures to protect information.

[top]

> Security measures should take into account the nature, type and state of data: the data's form (e.g., paper or electronic records); content (e.g., presence of direct or indirect identifiers); mobility (e.g., kept in one location or subject to physical or electronic transport); and vulnerability to unauthorized access (e.g., use of encryption or password protection). Measures for safeguarding information apply both to original documents and copies of information.

In order to assess the risks to participants' privacy and confidentiality over the duration of the data retention, REBs need to be informed about what safeguards are in place to protect against the possibility of participant re-identification if and when data are shared.

Factors relevant to the REB's assessment of the adequacy of the researchers' proposed measures for safeguarding information include:

a. the type of information to be collected;
b. the purpose for which the information will be used, and the purpose of any secondary use of identifiable information;
c. limits on the use, disclosure and retention of the information;
d. risks to participants should the security of the data be breached, including risks of re-identification of individuals;
e. appropriate security safeguards for the full life cycle of information;
f. any recording of observations (e.g., photographs, videos, sound recordings) in the research that may allow identification of particular participants;
g. any anticipated uses of personal information from the research; and
h. any anticipated linkage of data gathered in the research with other data about participants, whether those data are contained in public or personal records (see also Section E of Chapter 5 regarding Data Linkage).

[top]

Other things to consider:

- Will the data be de-identified or anonymized?
- Who will have access to the participant identification key? For how long will the participant identification key be retained?
- Where will data be archived?

The research data that are archived and shared are typically stripped of all identifiers. The researcher will retain the study participant identification key (i.e., if the data is de-identified) for the duration deemed necessary to satisfy retention requirements by journals, regulators, sponsors, etc... If a researcher is using secondary data and they

perceive a risk of re-identification of participants through data linkage or indirect identifiers, then this should be brought to the attention of the data custodian and the REB promptly.

Secondary Use of Data

[top]

TCPS2 (2014) Articles 5.5A/B:

Article 5.5A Researchers who have not obtained consent from participants for secondary use of identifiable information shall only use such information for these purposes if they have satisfied the REB that:

a. identifiable information is essential to the research;
b. the use of identifiable information without the participants' consent is unlikely to adversely affect the welfare of individuals to whom the information relates;
c. the researchers will take appropriate measures to protect the privacy of individuals, and to safeguard the identifiable information;
d. the researchers will comply with any known preferences previously expressed by individuals about any use of their information;
e. it is impossible or impracticable (see Glossary) to seek consent from individuals to whom the information relates; and
f. the researchers have obtained any other necessary permission for secondary use of information for research purposes.

If a researcher satisfies all the conditions in Article 5.5A (a) to (f), the REB may approve the research without requiring consent from the individuals to whom the information relates.

Article 5.5B Researchers shall seek REB review, but are not required to seek participant consent, for research that relies exclusively on the secondary use of non-identifiable information.

Application The onus will be on the researcher to establish to the satisfaction of the REB that, in the context of the proposed research, the information to be used can be considered non-identifiable for all practical purposes. For example, the secondary use of coded information may identify individuals in research projects where the researcher has access to the key that links the participants' codes with their names. Participant consent would be required in this situation. However, the same coded information may be assessed as non-identifiable in research projects **where the researcher does not have access to the key**. Participant consent would not be required in this situation.

The main concern for the secondary use of data is the risk of identifying participants if participants were given assurances in the LOI and Consent Form that their privacy and confidentiality will be protected throughout the full retention period. Many problems can

be avoided if the appropriate language is used in the Letter of Information and Consent Form, which will be discussed in Appendix 1 below.

[top]

The U.S. National Institutes of Health offers this helpful guidance to their researchers[5]:

1. **The informed consent form for my recently completed study states explicitly that only my research team will see the data provided and that we will not share the data. Am I now expected to share it?**

   No, but if you plan to collect additional data from those subjects under a grant with a data-sharing plan, you should revise the consent procedure to be consistent with the data-sharing plan. In preparing and submitting a data-sharing plan during the application process, **investigators should avoid developing or relying on consent processes that promise research participants not to share data with other researchers**. Such promises should not be made routinely or without adequate justification described in the data-sharing plan.

2. **How can I protect the privacy of my subjects?**

   It is the responsibility of the investigators, their IRB, and their institution to protect the rights of participants and the confidentiality of their data. Data should be redacted to strip all individual identifiers, and effective strategies should be adopted to minimize risk of disclosing a subject's identity. Options to protect privacy include:

   - withholding part of the data;
   - statistically altering the data in ways that will not compromise secondary analyses;
   - requiring researchers who seek data to commit to protect privacy and confidentiality; and
   - providing data access in a controlled site (sometimes referred to as a data enclave).

   Some investigators use hybrid methods, releasing a redacted dataset for general use but providing access to more sensitive data through a user contract or data enclave. In

[5] National Institutes of Health (NIH) Grants and Funding, Q&A: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm#923

most instances, sharing data is possible without compromising participant confidentiality and privacy.

[top]

Appendix 1:

[top]

Recommended Informed Consent Language for Data Sharing[6]

Language to Avoid

Promises in the informed consent can appear to limit an investigator's ability to share data with the research community. In reality, investigators can inform study participants that they are scientists with an obligation to protect confidentiality and still share the study data with the broad scientific community. Many effective means exist to create public-use data files or share restricted-use data files under controlled conditions. That is, data can be modified to reduce the risk of disclosure or shared with additional safeguards while preserving their value for science.

Model Language

Here are two model statements investigators may use in informed consents to describe protection of confidentiality that also allows data sharing:

> **Sample 1.** Study staff will protect your personal information closely so no one will be able to connect your responses and any other information that identifies you. Federal or provincial laws or regulations may require us to show information to university or government officials (or sponsors), who are responsible for monitoring the safety of this study. Any personal information that could identify you will be removed or changed before files are shared in any way, including with other researchers, or results are made public.
>
> **Sample 2.** The information in this study will be used only for research purposes and in ways that will not reveal who you are. Federal or provincial laws or regulations may require us to show information to university or government officials (or sponsors) who are responsible for monitoring the safety of this study. However, an assigned number will be used to designate your study record with your answers and not information that personally identifies you. You will not be identified in any publication from this study or in any data files shared with other researchers.

[top]

Known Concerns and Recommended Alternatives

Concern 1:

> Terms such as "anonymous" and "de-identified" are undefined and left open to interpretation. Some data are collected anonymously as directly identifying information is

[6] Slightly modified from the University of Michigan's Inter-University Consortium for Political and Social Research (ICPSR) Recommended Informed Consent language for Data Sharing, http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/conf-language.html

never obtained. De-identification may involve more than removing direct identifiers. Indirect identifiers in the file may still be used in combination to isolate a participant that is unique on certain characteristics. Even using the “safe harbor” method of de-identification by removing 18 specified elements still requires the covered entity to affirm it has no “actual knowledge that the remaining information alone or in combination with other information can be used to possibly identify the participant” (source [PDF 158KB]).

[top]

Recommendations:

Use descriptive sentences that state what information will not be shared:

- "Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public."
- "Your answers to the questions I ask will be anonymous. That is, I will not ask for your name, and we will not attach your name or jail number to your answers."
- "The Personally Unidentified Study Data does not include your name, address, telephone or social security number."

Use descriptive sentences that state what may be retained in data if shared with other researchers:

- "Personally Unidentified Study Data may include your date of birth, initials, and dates you received medical care. Personally Unidentified Study Data also may include the health information used, created, or collected in the research study."

Concern 2:

Confusion on whether language refers to identifiable participant information or research data that are separate from participant contact information or other direct identifiers.

Recommendations: [top]

Establish a term or phrase that identifies the identifiable information (i.e., contact information or other direct identifiers) that will not be shared.

Establish another term or phrase for the research data (i.e., the "coded" information or "your answers" that does not contain the contact information or other direct identifiers but still may include indirect identifiers).

Use these terms consistently throughout the form. Avoid indefinite language, such as "your data," "your study information," "all information collected about you," or "study results."

Concern 3:

Promises made that the data will be seen or accessed only by the research team.

Recommendations:

Explain the form that the identifying information will take and who has access:

- "Your identifying information will be replaced with codes. Only the research team will have access to information that identifies you to carry out this research study. Your identifying information will not be shared with others outside this research study."

[top]

If no "research data" will be released to persons in official or unofficial capacities, make sure that sharing data with other researchers is not left to be interpreted in this category but is allowed through a statement such as:

- "Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public."
- "During the project, information from this study will be kept in locked files that only the research staff can open. Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public."
- "Any answers that you give during the surveys will be kept confidential during this project. Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public."

State it explicitly when personally identifying information will be destroyed, removed, or changed:

- "All personally identifying information collected about you will be destroyed once it is no longer needed for the study. Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public."

[top]

Concern 4:

Descriptions of how the information will be stored during the project imply the data are stored only during the project and do not allow the storage of any data beyond the research project so the research portion of the data can be shared.

Directly identify elements that need to be stored separately from the "research data" (i.e., the data for analysis), such as the participant identification key, and must be destroyed within a specified period after the end of the research project. The research data can be shared if appropriately de-identified or as a limited dataset (aka restricted-use dataset).

Recommendation:

Explain the duration of the data storage and what happens to directly identifying information:

- "If you decide to be in this study, the study researchers will get information that identifies you and your personal health information. This may include information that might directly identify you, such as your name and address. This information will be kept for the length of the study [e.g., five years]. After that time it will be [(A) destroyed or (B) de-identified, meaning we will replace your identifying information with a code that does not directly identify you.] The principal investigator will keep a link that identifies you to your coded information, but this link will be kept secure and available only to the principal investigator or selected members of the research team [state expected retention period as per hospital, law or regulation requirement]. Any information that can identify you will remain confidential. Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public.

Concern 5: [top]

The phrase "shared anonymously" may prohibit sharing data using a limited-use (aka restricted-use) dataset if the data cannot be completely anonymized or de-identified.

Recommendation:

If the personally identifying elements were collected (i.e., the data were not collected anonymously), use a statement that acknowledges the directly identifiable information will be removed but does not promise all indirect identifiers will be removed:

- "Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public."

Concern 6:

The informed consent and, therefore, the authorization to use information (i.e., research data) expire, e.g., "Only valid while the study is being done."

Recommendations:

- Do not mention an expiration time period.
- State the authorization never ends unless the participant revokes it.
- State the retention of personally identifiable information expires but the data without the personally identifiable information may be used for future research.

[top]