Research Data Canada

# Stewardship of Research Data in Canada: A Gap Analysis

*October 2008*

Research Data Strategy Working Group
http://data-donnees.gc.ca/eng/index.html

## Introduction

The need for a coordinated approach to the stewardship of research data in Canada has been well documented in a series of reports published over the last decade. During this same period, billions of dollars have been invested in research in Canada, generating huge volumes of digital data. Collectively, these data represent a significant asset with virtually limitless opportunities to develop new knowledge through their re-use, if they are managed appropriately. To date, Canada has not taken action to institute, in a coordinated way, practices and services dedicated to the stewardship of research data. As a result, valuable data are under-utilized and at risk of being lost.

In January 2008, a working group comprised of representatives from a cross-sector of Canadian research organizations was established to provide recommendations and an action plan on a national approach to the stewardship of research data in Canada. The working group agreed that the best way to proceed was to form several task groups that would produce a succinct statement of the problem, develop a strategy for education and training of researchers, and undertake a gap analysis of data stewardship activities in Canada. This report provides the results of the gap analysis undertaken in the spring/summer of 2008.

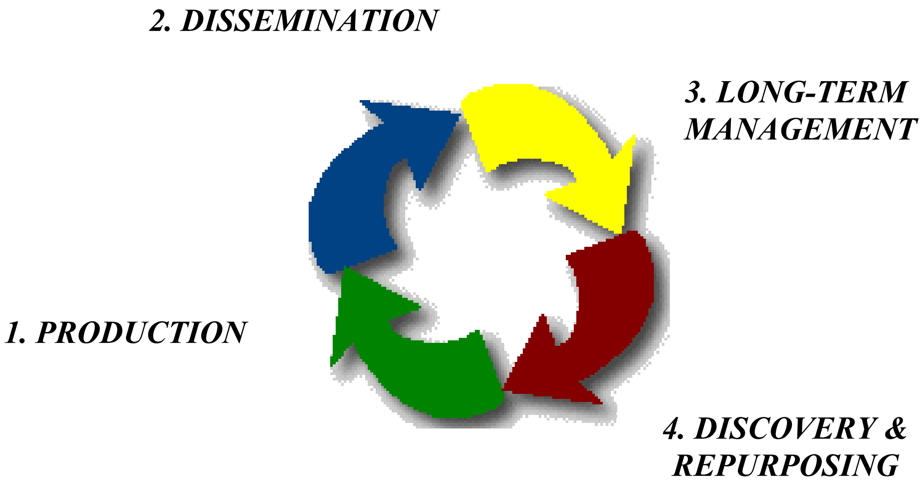
## Purpose and Methodology

The purpose of the gap analysis is to identify discrepancies between current and ideal states. The results will be used as evidence in a call for action and will contribute to the development of a practical strategy for improving data stewardship in Canada. The intent is not to reproduce what has already been done, but rather to pull together all existing information and evidence to describe the current state of research data stewardship in Canada. From there, a strategy will be formulated for moving forward. Thus, much of the information contained here comes from previous reports on the subject of research data. Several Canadian publications, all based on extensive consultations with stakeholders, undergird this investigation: *The National Consultation of Access to Scientific and Research Data (NCASRD) Final Report*, the *National Data Archiving Consultation Report* (NDAC), and *Canadian Digital Information Strategy* (CDIS). A number of other reports were also consulted, including the Association of Research Libraries report, *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering*, the U.S. National Science Foundation's, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*, the UKOLN's *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, the Canadian Association of Research Library draft report, *Survey of Canadian and International Data*

*Initiatives*. In addition, information has been gathered from other sources (e.g., websites) as well as input from working group members.

## The Data Lifecycle

A lifecycle model provides a useful framework for analyzing the state of data stewardship in Canada. The lifecycle illustrates the various stages through which data travel as they are collected, used, preserved and reused again, ensuring their value is maximized. For the purposes of this report, we have chosen a simple 4-stage model of the data lifecycle: data production, data dissemination, long-term data management, and data discovery and repurposing.



Each stage represents a group of related processes in the data lifecycle.  Within a stage, specific processes or activities, when view collectively, represent a significant component in conducting research. While some activities and products are intrinsic to each stage, others flow across stages. For example, the design of an experiment or survey will be integral to the Data Production stage, while data products emanating from this stage will flow throughout the model. It should be noted that the aim of this model is to facilitate the identification of gaps in data-related activities, rather than to provide a comprehensive and detailed account of data stewardship in all its complexity.

1. Data Production: This stage includes all activities involved in the planning, collecting, processing, analysis and maintenance of data in the original research project. Among these activities are selecting a study design, constructing instruments for data collection, conducting data collection/creation, performing data editing/verification/validation, analyzing data, backing up data versions and preparing and tagging metadata.

2. Data Dissemination: This stage involves the preparation of data for use by others and the establishment of procedures and methods for disseminating data. To be understandable, data must be accompanied with descriptive metadata that is accessible in widely used formats. The data must also be available in commonly used formats; and mechanisms are needed to ensure appropriate levels of access, depending on security, privacy, or intellectual property restrictions.

3. Long-term Data Management: This stage involves activities supportive of the preservation and long-term access to research data. Collection development work is key to this stage and integral to managing access to research data. Long-term access will only be possible if sound preservation practices are followed. Among these activities are appraising, selecting, depositing or ingesting data into a repository, ensuring authenticity, managing the collection of data and metadata, refreshing digital media, and migrating data to new digital media.

4. Data Discovery and Repurposing: Data are increasingly being recognized as a research asset with value beyond the purpose for which they were originally intended. This stage involves mechanisms and activities to enable the discovery and re-use of data. Whether replicating previous findings or addressing unexplored research questions, discovery tools are needed to locate and retrieve relevant data. The repurposing of data, which involves creating new data by combining data appropriately from a variety of existing files, is emerging as a new method. Repurposing data also generates new data products that did not previously exist. Among the activities in this stage are developing and supporting search tools that utilize standardized metadata, harmonizing the coding of data for specific variables, engineering new methods of combining data and generating and harvesting new data collections.

## Gap Analysis

The following list of indicators was used to conduct a gap analysis of data stewardship in Canada. For each indicator, the analysis provides a succinct statement of the ideal state of data stewardship; a description of the current state; and a discussion of the gaps.

I. Policies
II. Funding
III. Roles and responsibilities
IV. [Trusted digital] data repositories
V. Standards
VI. Skills and training
VII. Reward and recognition systems
VIII. Research and Development
IX. Accessibility
X. Preservation

The ideal states, which are deliberately expressed at a high level, are based on input from the working group. The analysis does not discuss the specific details of an ideal state; this work will be undertaken in the next phase of the project. The current states were determined from a review of existing literature as well as input from the working group. Gaps between these two states have been identified and a gap-level has been assigned to indicate the magnitude of the gap.

In a number of cases, we were unable to provide a complete account of the current state because of a lack of available information. Therefore, some variation exists in the completeness of the current states and gaps discussed here. This absence of information represents a significant gap in itself and has been noted within the relevant sections.

### I. Data Policies

**Ideal state:** *Canadian organizations have coherent and cohesive policies based on sound data management principles that govern the management of data across disciplines and throughout the data stewardship lifecycle and are reflective of relevant legislative initiatives.*

**Current state**
A list of the research data policies reviewed for this report is contained in the Appendix. In summary:

- No overarching policy guiding data management exists in Canada.
- Many organizations involved with research do not currently have policies governing the management of research data.
- Where policies exist, the principles on which they are based are often not explicitly stated.
- In some cases, ethics board and privacy policies requiring the destruction of data are in contradiction with university and funding agency policies requiring data retention and data sharing.

- Data archiving policies of some funding agencies (e.g., SSHRC) are not enforceable because the agencies have not implemented mechanisms to document compliance of policies.
- The majority of existing policies address how long researchers must maintain and be ready to share their data after a research project has been completed.

| Data Production | Data Dissemination | Long-term Data Management | Discovery and Repurposing |
|---|---|---|---|
| In limited cases- e.g., IPY* | Yes, but sometimes contradict each other- Universities, Ethics Boards, Privacy Act, PIPEDA* | In limited cases- CIHR*, SSHRC*, NSERC*, Federal and Provincial Agencies, some research projects (IPY*) | In limited cases- SSHRC*, CIHR*, Federal and Provincial Agencies |

*See Glossary

**Gaps**

There are large policy gaps throughout the data lifecycle. Few policies address the need for researchers to develop data management plans. In terms of data dissemination and long-term management, although the tri-councils have data sharing policies, they do not explicitly state that researchers should adhere to specific standards to ensure data can be accessed and preserved in the future. Only a few agencies have policies governing the deposit of data into repositories.

**Gap level** **Moderate**

### II. Funding

**Ideal state:** *Together, the range of funding mechanism cover costs throughout the data lifecycle, ensuring long-term support that benefits the many stakeholders of research.*

**Current state**

Funding for data stewardship in Canada comes from a variety of sources and is targeted at different stages of the lifecycle. A list of funding sources for data stewardship that were reviewed for this report is contained in the Appendix. In summary:

- Through discovery grants, the tri-councils provide funds to research teams for data production, data maintenance during the life of the project, and in some cases data dissemination.
- SSHRC offers funds for long-term management, but few grants have been approved for this purpose.

- CFI provides funds for the capital cost of databases and data acquisition, and some funds for maintenance during the life of a project, through the Infrastructure Operating Fund.
- CFI funds the Research Data Centres (RDC) for repurposing and discovery, and SSHRC and CIHR are providing the RDCs with operating funds for the Centres up to but not exceeding 50 percent of the operating expenses. Universities are covering the rest. CANARIE (through the Network Enabled Platform) provide some funds for discovery and repurposing tools.
- Funding for the long-term management of data is supported through data repository management by government departments, university data libraries and research charities, and from the tri-council funding agencies—but these repositories only accept selective data collections in certain disciplines.

| **Data Production** | **Data Dissemination** | **Long-term Data Management** | **Discovery and Repurposing** |
|---|---|---|---|
| Yes-CIHR, NSERC, SSHRC, CFI; and other research funding sources | In limited cases-CIHR, NSERC, SSHRC, CFI | In limited cases-Universities, Not-for profit research institutions, government agencies, private industry | In limited cases-, CFI, CANARIE, SSHRC, CIHR, some university support for infrastructure |

**Gaps**

In terms of funding for the various stages of the data lifecycle, there are large gaps in data dissemination, long-term management, and discovery and repurposing stages. In particular, the costs associated with preparing data for dissemination are not supported through existing funding mechanisms, nor are there many institutions that provide sustainable funding for data repositories.

**Gap level** 🟥 **Large**

### III. Roles and Responsibilities

**Ideal state:** *Roles and responsibilities are clearly defined and properly fulfilled. Each participant in the data lifecycle has a distinct set of responsibilities, and also, in partnership, must act with other participants collectively to pursue higher-level stewardship goals important to the entire community.*

**Current state**

The Appendix contains a list of roles and responsibilities identified for this report. In summary:

.

- Major responsibilities lie with principle investigators, who are often required (by funding agency and university policies) to ensure that research data are retained and are available for sharing for a given time period after a research project is complete (ranging from 2 to 5 years).
- Some government departments have responsibilities for collecting and preserving specific data types: Statistics Canada, Natural Resources Canada, Canadian Space Agency, Environment Canada, etc.
- In limited cases, research institutions (universities, other research centres) have taken on the responsibility for long-term management of research data.

| **Data Production** | **Data Dissemination** | **Long-term Data Management** | **Discovery and Repurposing** |
|---|---|---|---|
| Research team | Research team | Some government agencies, universities, research communities | Some government agencies, universities, research communities |

**Gaps**

There are no lines of custodial responsibilities along the data lifecycle and no mechanisms for bringing stakeholders together to determine roles and responsibilities. Data management plans, which outline the mechanisms and timelines through which the data generated by research will be made available, are rarely required by research funding agencies in Canada. With the exception of some government departments, there are no national institution(s) responsible for preserving, managing and making research data publicly accessible on the scale required to support the needs of stakeholder communities.

**Gap level** 🟥 **Large**

## IV. [Trusted Digital] Data Repositories

**Ideal state:** *Canada has a comprehensive network of trusted digital data repositories that provide reliable, long-term access to all research data deemed to be of enduring value.*

**Current state**
The Appendix contains a cursory list of data repositories in Canada identified for this report. In summary:

- Where repositories exist in Canada, they are managed by federal and provincial agencies, research communities, universities, and private industry.
- There is a very nascent network of institutional repositories being managed by research libraries, but few are capable of collecting research data in a manner that facilitates easy access and re-use.
- Where repositories do exist, few, if any, conform to the definition of a Trusted Digital Repository (TDR), which provides a policy, process, standards, and technology framework for digital preservation.
- There is also no certification process to establish that a data repository can be 'trusted' and there is a risk that data repositories may not be capable of preserving digital information even though their specifications may suggest that they can.

| **Data Production** | **Data Dissemination** | **Long-term Data Management** | **Discovery and Repurposing** |
|---|---|---|---|
| Short-term repositories designed to support data analysis. | In limited cases (e.g., DLI*) | Lack of coverage<br>Little implementation of sound data preservation activities. | In limited cases |

*See Glossary

**Gaps**
There are large gaps in both coverage and capacity of data repositories. Most research datasets are never deposited into a data repository. Repositories do not exist for all subject areas, and the vast majority of research data still rests on researchers' hard drives. Only a few active data repositories in Canada allow researchers to deposit their data.

**Gap level** 🟥 **Large**

## V. Standards

**Ideal state:** *There is widespread adherence to standards (object, process and instrumentation) throughout the data lifecycle and they are implemented independently of any one field.*

**Current state**

Specific details about the use of standards in various disciplines is not readily available; however, the literature indicates that:

- There are varying levels of adherence to data standards in Canada. Large, government, university, or community data centres tend to adhere to discipline-based standards.
- For smaller data collections there are reportedly low levels of compliance with standards.
- It is common practice for researchers to adhere to production criteria set out by software programs, rather than comply with international standards.
- In the social sciences, data centres regularly employ the Data Documentation Initiative (DDI) metadata standard, but no general model for the representation of scientific study metadata has emerged in Canada.

| Data Production | Data Dissemination | Long-term Data Management | Discovery and Repurposing |
|---|---|---|---|
| Discipline-based standards | Variable-DDI in social science; unknown for other research areas | Variable- regular use in large data centres (e.g., community, university or government data archives) and ad hoc implementation in other cases | XML is commonly used in the social sciences |

**Gaps**

Canada has no national agency that monitors, oversees, and sanctions specific standards for use by Canadian researchers. There is no coordinated effort to participate in the development of international research data standards, in metadata schemes such as Data Documentation Initiative, in tools for data access such as the Networked Social Science Tools and Resources (NESSTAR) project, and in collaborative international infrastructure projects such as the European Union Frameworks. Much work still needs to be done in the matter of interoperability of software and protocols and adoption of standards for metadata and for data exchange and data quality.

**Gap level** **Moderate**

## VI. Skills and Training

**Ideal state:** *Data stewardship activities are widespread and supported by specially trained data scientists and information professionals; and researchers are well educated on the principles of data stewardship and its importance, and aware of their own roles and responsibilities.*

**Current state**
No detailed survey has been done in terms of skills and training levels; however, the literature indicates:

- Many researchers are unfamiliar with data stewardship processes, including the importance of metadata.
- Few researchers have had specific training in database development and preservation.
- There is a reticence amongst many to assume responsibility for data stewardship beyond the researchers' immediate interests.

| Data Production | Data Dissemination | Long-term Data Management | Discovery and Repurposing |
|---|---|---|---|
| Yes | In limited cases: e.g., CNC/CODATA* workshops for researchers | Low skill level and few data scientists employed at institutions to support researchers | In limited cases: e.g., DLI* Training Program |

*See Glossary

**Gaps**
There are insufficient numbers of trained scientists and information professionals with knowledge of data cataloguing, metadata standards and processes, preservation management and assessing the value of data to support researchers. Data managers are not widely regarded as essential to the research enterprise and remain vulnerable to budget pressures, even more so when such "library overheads" require budget increases. There is also a general lack of awareness of the importance of data management in the research community and there are few opportunities for researchers to receive training on data management issues.

**Gap level** **Large**

## VII. Rewards and Recognition Systems

**Ideal state:** *Reward systems for researchers widely recognize contributions to research data and the development of tools for improved data management, use and preservation as significant performance indicators.*

**Current state**

- Researchers are measured primarily according to their publication record, not data management activities.
- SSHRC has a data archiving policy, but does not reward researchers who comply to this policy in future applications that they may submit.
- Funding agencies and universities have data sharing policies, but it is unclear to what degree compliance is recognized (or non-compliance punished).
- A few projects, such as the International Polar Year, require data management plans from researchers in order for projects to receive funds.
- In a few disciplines, there are journals that have special issues highlighting innovative databases developed in the community. There are also a few journals that publish reviews of databases (e.g., Journal of Bioinformatics).

| Data Production | Data Dissemination | Long-term Data Management | Discovery and Repurposing |
|---|---|---|---|
| Built-in incentives linked to the successful undertaking a research project. | Some disincentives in place at universities and funding agencies for researchers who do prepare data so that it can be shared. | In limited cases-certain research projects, funding agencies, and journals require deposit of data into data repositories. Presumably, non-compliance would result in loss of future funding or publication. | Unknown |

**Gaps**

The research cultures in many domains do not embrace data preservation or data sharing. Consequently, researchers do not understand the risk levels associated with their current practices. Reward systems for researchers do not recognize sound data management or data sharing. There is little recognition for leadership in the compilation of, or major contribution to, high value, open access databases and datasets, nor in the development of tools that enhance the value of data (for example, developing methods for mining or combining databases across disciplines). Researchers in academic institutions do not receive recognition in their tenure and promotion reviews for significant contributions to research data or its management.

**Gap level** **Large**

### VIII. Research & Development

**Ideal state:** *Canada has a coordinated approach to R&D activities in support of data stewardship needs, with well-articulated priorities and adequate funding.*

**Current state:**

- In Canada, there are a number of unconnected projects looking at different issues around data stewardship. For example, the InterPARES Project (International Research on Permanent Authentic Records in Electronic Systems: http://www.interpares.org/), centred at the University of British Columbia, has been contributing to international research in the area of preservation and authenticity of digital information, including data.
- Data repositories are developed by individual research projects, but these project-based solutions usually cannot be generalized and used by others.

| **Data Production** | **Data Dissemination** | **Long-term Data Management** | **Discovery and Repurposing** |
|---|---|---|---|
| Discipline-based repositories are being developed in the context of individual research projects to facilitate data use and analysis | Unknown | Some research being undertaken by InterPARES in Canada. | Unknown |

**Gaps**

Numerous research and development challenges remain in the field of data stewardship. Research into technologies, organizational models, standards and practices, and interoperability are needed on an ongoing basis as the volume of data grows and becomes more complex. R&D projects are not coordinated, nor guided by national priorities. There is no coordinated effort to engage in international research and development in this area.

**Gap level** **Moderate**

## IX. Access

**Ideal state:** *There is widespread access to publicly funded research data, with appropriate mechanisms in place for regulating access that takes into account security, ethical, legal, and economic interests where appropriate.*

**Current state**

- Much of the research data being produced today is hard to access by other Canadian research communities, and is often not ideally structured to be as useful or as open as possible, even within the discipline for which it is being constructed.
- There are large reservoirs of existing data not in current use and not available online.
- Researchers are reluctant to share data because they feel it is their intellectual property.
- Researchers lack the expertise to ensure that data are accessible by others in the future.
- While there is a growing international trend towards free access to data held in repositories (e.g., GeoConnections), many repositories still charge fees for access through mechanisms, such as licenses and pay per view, or restrict access to community members only (e.g., Statistics Canada and Natural Resources Canada).
- Few research organizations have policies requiring researchers to provide access to data. Most research institutions require that data be retained for a period of five years and shared with others upon request. Where policies exist, they do not cover all types of research data. Privacy and ethics policies require some types of data to be anonymized or destroyed.
- Priority is given to short-term use for analysis and publication of articles.

| Data Production | Data Dissemination | Long-term Data Management | Discovery and Repurposing |
|---|---|---|---|
| Priority is on immediate analysis not dissemination and reuse by others | Researchers often do not have the time or skills for preparing data. Researchers are reluctant to share data. Researchers are unclear about who owns data. | There are large gaps in both coverage and capacity of data repositories. | Pay per view and licensed access still commonplace.<br>Most data rests on the hard-drives of researchers and is inaccessible to others who wish to use it.<br>Many organizations that own the data are not equipped to provide dissemination services, such as good documentation, |

| | | | |
|---|---|---|---|
| | | | standardized formats, ensuring ethical clearance and confidentiality requirements, communication of conditions and terms of use, documenting volume of use. Without these things, online access is meaningless. |

**Gaps**

Researchers often do not have the time or skills to prepare data for dissemination, are reluctant to share data and are unclear about who owns data. Few reward or recognition mechanisms exist for sharing research data. Many organizations are not equipped for dissemination services, which includes the ability to deliver good documentation and data in standardized formats, to ensure ethical clearance and confidentiality requirements, to communicate conditions and terms of use and to monitor overall use.

**Gap level** **Moderate**

## X. Preservation

**Ideal state:** *Research data with enduring value are preserved using standards-based, active management practices throughout the data lifecycle; furthermore, data are integrated into an enduring institutional environment supporting trusted digital repositories.*

**Current state**

- Few research organizations have policies regarding preservation of research data.
- Where policies exist, they do not cover all types of data.
- The current SSHRC data archiving policy is unenforceable.
- Some large collections of data exist in Government of Canada databases, such as those maintained by Statistics Canada, Fisheries and Oceans Canada, Natural Resources Canada, and Environment Canada, and are being actively preserved
- Research funding agencies only rarely require data management plans.
- Many researchers do not have time or skills to prepare data for preservation
- Priority is given to short-term use for data analysis and publication of articles.
- Few reward or recognition mechanisms exist for preserving research data.
- Where preservation activities do exist, quality control, data storage and backup, and descriptive metadata are the most commonly cited practices.

| **Data Production** | **Data Dissemination** | **Long-term Data Management** | **Discovery and Repurposing** |
|---|---|---|---|
| Few researchers have time or skills to prepare data for preservation | Unknown | Limited coverage and capacity of data repositories in Canada. Limited funding for data repositories. | N/A |

**Gaps**

Data collections are supported by relatively small budgets, often through research grants funding a specific project, and therefore do not have preservation as a priority. Most of the data collected through research is not deposited into data repositories. Few if any repositories have full preservation capacity as defined by trusted digital repository status. There is also a lack of awareness in the research community about preservation standards. There is no national agency to provide guidance to researchers in terms of standards. Canada will continue to lose valuable research data without funding for both researchers to prepare data and institutions to collect and preserve data once a research project is complete.

**Gap level** 🟥 **Large**

## Conclusion and Next Steps

With today's growing volume of research data, numerous options exist to use and exploit these resources in ways to discover new knowledge and provide Canada with a competitive edge. However, without the proper management of these digital resources throughout the data lifecycle, Canada will have squandered this opportunity. We have assessed the status of several key indicators across the stages of this lifecycle model to determine the priorities and work needed to place Canada within a sound research data environment. The results show serious gaps between our current state and an ideal state for our research sector. These are summarized in Tables 1 and 2.

### Table 1: Gaps across the Data Lifecycle

**Data Production**

- Priority is on immediate use, rather than potential for long-term exploitation.
- Limited funding mechanisms to prepare data appropriately for later use.
- Few research institutions require data management plans.
- No national organization that can advise and assist with application of data standards

**Data Dissemination**

- Lack of policies governing the standards applied to ensure data dissemination.
- Researchers unwilling to share data, because of lack of time and expertise required.
- Some policies require certain types of data be destroyed after a research project is over.

**Long-term Management of Data**

- Lack of coverage and capacity of data repositories.
- Preservation activities in repositories are not comprehensive.
- Limited funding for data repositories in Canada.
- Few incentives for researchers to deposit data into archives.

**Discovery and Repurposing**

- Most data rests on the hard drives of researchers and is inaccessible by others.
- Per per view and licensed access mechanisms are common where data are available
- Many researchers are reluctant to enable access to their data because they feel it is their intellectual property.

**Table 2: Summary of Gap Analysis**

| Indicator | Gap level |
|---|---|
| Policies | Moderate |
| Funding | Large |
| Roles and responsibilities | Large |
| [Trusted digital] data repositories | Large |
| Standards | Moderate |
| Skills and training | Large |
| Reward and recognition systems | Large |
| Research and Development | Moderate |
| Access | Moderate |
| Preservation | Large |

These gaps exist for a number of reasons: because of a lack of policies, infrastructure, and funding mechanisms for data stewardship, as well as a research culture that does not recognize the value of data management. Data stewardship in Canada is suffering without a strategic, national vision. The existing piecemeal approach has resulted in serious gaps throughout the lifecycle. This is particularly apparent in the final three stages. Significant amounts of data are rendered inaccessible at the data dissemination stage because of the absence of services and procedures to deliver data to other researchers. The woefully inadequate number of trusted data repositories in Canada contributes to the gap identified in the long-term management stage. Consequently, most research data created in Canada are greatly underutilized and are at a high risk of being lost.

There are considerable variations across disciplines, which contributes to the complexity of these issues. The social sciences are ahead in some areas because of the international data documentation standard DDI, which has been embraced in Canada by the Data Liberation Initiative. In the sciences, a distinction can be made between "big science" projects, which collect huge amounts of data that are systematically preserved, versus "small science", which tends to generate very diverse types of data that generally are not readily accessible or actively preserved. Meanwhile, the practice of data archiving in the arts and humanities is relatively rare.

There are significant risks associated with doing nothing. Canadian researchers will not have the tools they need to remain at the leading edge and Canada's innovation capacity will ultimately decline. Countries such as the UK and US are already far ahead in terms of policy and infrastructure development. They recognize that data will be a critical driver of new discoveries in the future. For example, the National Science Foundation in the US has recently invested $100 million in developing a sustainable and interoperable network of data repositories to support scientific research.

Given the scope of this challenge, it is clear that these issues cannot be resolved in isolation. They must be addressed collectively with participation from all sectors and disciplines in the research community. The next steps for the working group will be to develop a multi-pronged strategy that will articulate the need for a systematic approach to data stewardship and also identify some achievable outcomes for the short and medium terms, so that progress can be made towards filling in the gaps identified above.

## Glossary

**CANARIE-** CANARIE Inc.: Facilitates the development and use of its network as well as the advanced products, applications and services that run on it.

**CFI-** The Canada Foundation for Innovation: an independent corporation created by the Government of Canada to fund research infrastructure.

**CIHR**- Canadian Institutes of Health Research: Government of Canada's health research funding agency

**CNC/CODATA**- Canadian National Committee for CODATA (the Committee on Data for Science and Technology)- The Canadian voice of CODATA, the Committee on Data for Science and Technology. CODATA is an interdisciplinary Scientific Committee of the International Council for Science (ICSU), which was established 40 years ago.

**DLI**- Data Liberation Initiative: A Statistics Canada program that enables participating institutions to pay an annual subscription fee that allows their faculty and students unlimited access to numerous Statistics Canada public use microdata files, databases and geographic files. Use of these files is limited to academic research and teaching purposes.

**IPY**- International Polar Year: a large scientific programme focused on the Arctic and the Antarctic from March 2007 to March 2009.

**NSERC**- The Natural Sciences and Engineering Research Council: Government of Canada's major natural sciences and engineering research funding agency

**PIPEDA**- The Personal Information Protection and Electronic Documents Act (PIPEDA): Canada’s private sector privacy law.

**SSHRC-** Social Sciences and Humanities Research Council: the federal agency that promotes and supports university-based research and training in the humanities and social sciences.

# Appendix

**Table 1: Ideal states**

***Policies:*** *Canadian organizations have coherent and cohesive policies based on sound data management principles that govern the management of data across disciplines and throughout the data stewardship lifecycle and are reflective of relevant legislative initiatives.*

***Funding:*** *Together, the range of funding mechanism cover costs throughout the data lifecycle, ensuring long-term support that benefits the many stakeholders of research.*

***Roles and responsibilities:*** *Roles and responsibilities are clearly defined and properly fulfilled. Each participant in the data lifecycle has a distinct set of responsibilities, and also, in partnership, must act with other participants collectively to pursue higher-level stewardship goals important to the entire community.*

***Trusted digital data repositories:*** *Canada has a comprehensive network of trusted digital data repositories that provide reliable, long-term access to all research data deemed to be of enduring value.*

***Standards:*** *There is widespread adherence to standards (object, process and instrumentation) throughout the data lifecycle and they are implemented independently of any one field.*

***Skills and training:*** *Data stewardship activities are widespread and supported by specially trained data scientists and information professionals; and researchers are well educated on the principles of data stewardship and its importance, and aware of their own roles and responsibilities.*

***Rewards and recognition systems****: Reward systems for researchers widely recognize contributions to research data, the development of tools for improved data management, use and preservation as significant performance indicators.*

***Research and development:*** *Canada has a coordinated approach to R&D activities in support of data stewardship needs, with well-articulated priorities and adequate funding.*

***Access:*** *There is widespread access to publicly funded research data, with appropriate mechanisms in place for regulating access that takes into account security, ethical, legal, and economic interests where appropriate.*

***Preservation:*** *Research data with enduring value are preserved using standards-based, active management practices throughout the data lifecycle; furthermore, data are integrated into an enduring institutional environment supporting trusted digital repositories.*

## Table 2: Data policy attributes

| Agency | Scope | Timing for data access | Data archive | Requirements |
|---|---|---|---|---|
| CIHR | Selective data types-bioinformatics, atomic, and molecular coordinate data | Immediately upon publication | Appropriate public database | In addition, all grant recipients are required to retain original data sets for a minimum of five years after the end of the grant. This applies to all data, whether published or not. |
| SSHRC | All research data collected with the use of SSHRC funds | Within 2 years of completion of research project | Institution's library or data service if it can archive the data. If not, then list of data archives in Canada | 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 |
| NSERC Strategic Networks Program | Large data sets funded through the NSERC Strategic Networks Program | Reasonable period of time | None specified | An agreement regarding responsibility for the maintenance and preservation of large data sets must be in place at the outset of network activities. |
| Tri-Council Policy Statement on Accessing Private Information | Data with identifiable personal information | N/A | None specified | Researchers should ensure that the data obtained are stored with all the precautions appropriate to the sensitivity of the data. Information that identifies individuals or groups should be kept in different databases with unique identifiers |
| Sample University Policy (McGill University) | All data | Data must be retained for five years | None specified | Data must be organized in a manner that allows ready verification. Subject to exceptions based on a duty of confidentiality and the laws respecting intellectual property and access to information, after data are published, they must be made available to any party presenting a reasonable request to examine them. |
| Sample Human Research Ethics Board (University of Victoria) | Human research involving human participants, remains, cadavers, tissues, biological fluids, embryos, foetuses and other biological materials including human DNA, RNA or DNA and RNA fragments requires either an approval or a waiver from the | N/A | None specified | Researchers' plans for preserving or destroying participants' data must be appropriate to the field of research and the wishes of participants. For example, in oral history the best practice may be to archive the information collected (with the participants' consent) for future generations. With research where the release of information could harm participants, it may be best to destroy the data collected as soon as possible. |

| Agency | Scope | Timing for data access | Data archive | Requirements |
|---|---|---|---|---|
| | HREB before the research is begun. | | | Explain your plans for preserving and protecting participants' data or for destroying data in light of the best practices in your field of research and the wishes of participants. Some funding agencies, professional organizations and publishers have established minimum requirements for data retention (e.g., five years), after which time the data are to be destroyed. You must disclose their plans for data destruction that includes a time frame and the methods that will be employed to destroy the data (e.g., shredding, electronic file deletion). |
| Sample Federal Government Department Policy (Fisheries and Oceans Canada) | All DFO scientific data | Within two years of being acquired | A managed archive- The Marine Environmental Data Service, Science Sector, (MEDS) will provide co-ordination among regional, zonal and national centres as appropriate, to ensure that all data are properly managed. Where no data management centre exists in a Region, Science and Oceans managers will be required to designate and support indeterminate A-base staff positions that include data management responsibilities. | To ensure proper management and archival of data, all scientific data collected by the Department must be migrated to a 'managed' archive immediately after the data have been processed.<br>To obtain maximum benefit to the Department and to the user community at large, scientific data must be made available in a timely manner with full and open access, consistent with Departmental, national and international obligations with respect to its data holdings. |
| Genome Canada | Not specified | No later than the acceptance for publication of the main findings from any datasets | Genome Canada provides examples of databases where various data types or unique resources | Data sharing should occur in a timely fashion. Genome Canada expects data to be released and shared no later than the |

| Agency | Scope | Timing for data access | Data archive | Requirements |
|---|---|---|---|---|
| | | generated by a project | produced by Genome Canada-funded projects may be deposited. | acceptance for publication of the main findings from any datasets generated by a project. For large datasets that are collected over several discrete time periods or phases, it is reasonable to expect that the data be released in phases as they become available or as main findings from a research phase are published. However, at the conclusion of a project, all data must be released without restriction. |
| Sample research project policy (International Polar Year-IPY) | IPY data are those data generated during the IPY timeframe (March 2007 - March 2009) by the specific projects endorsed by the ICSU/WMO Joint Committee as IPY projects. | Shortest feasible timescale | The IPY Data and Information Service should work with the relevant operational centers, data centers, and other organizations to ensure the preservation of relevant IPY related data not explicitly produced by IPY projects. | The IPY Joint Committee requires that IPY data, including operational data delivered in real time, are made available fully, freely, openly, and on the shortest feasible timescale.<br>Recognizing that the true value of scientific data is often realized long after they have been collected, and to ensure the lasting legacy of IPY, it is essential to ensure long-term preservation and sustained access to IPY data. All IPY data must be archived in their simplest, useful form and be accompanied by a complete metadata description. |
| Privacy Act | Personal Information | N/A | N/A | A government institution shall dispose of personal information under the control of the institution in accordance with the regulations and in accordance with any directives or guidelines issued by the designated minister in relation to the disposal of that information. |
| Personal Information Protection and Electronic Documents Act | Personal information-information about an identifiable individual, but does not include the name, title or business address or telephone number of an employee of an organization | Data must be destroyed once it is no longer required to fulfill the identified purpose | N/A | Personal information shall not be used or disclosed for purposes other than those for which it was collected, except with the consent of the individual or as required by law. Personal information shall be retained only as long as necessary for the fulfillment of those purposes.<br>Personal information that is no longer required to fulfill the identified purposes should be destroyed, erased, or made anonymous. Organizations shall |

| Agency | Scope | Timing for data access | Data archive | Requirements |
|---|---|---|---|---|
| | | | | develop guidelines and implement procedures to govern the destruction of personal information. |
| Legal Deposit Legislation | Not all online materials fall within the scope of the Legal Deposit legislation. LAC is focusing on collecting online material that is considered to be in "publication" form. The types of online publications that should be deposited are: books, magazines, annual reports, research papers, scholarly journals, etc. Types of online publications that do not need to be deposited are: forms, email correspondence, abstracts, press releases, portals, advertisements, schedules, timetables, databases, etc. | Upon publication | Library and Archives Canada | Library and Archives Canada's (LAC's) mandate is to preserve the documentary heritage of Canada for the benefit of present and future generations. The Library and Archives of Canada Act was assented to in April 2004 and through legal deposit regulations, LAC is able to build and preserve a comprehensive collection of Canada's published heritage. |

## Table 3: Sources for and recipients of funding for data stewardship activities throughout the life cycle

N.B.: This list is not comprehensive.

| Funding Source | Production life cycle stage | Dissemination life cycle stage | Management life cycle stage | Discovery and Repurposing life cycle stage |
|---|---|---|---|---|
| **Tri-councils** CIHR, NSERC, SSHRC Discovery Grants | Research Teams | | | |
| **SSHRC**- "Costs associated with preparing research data for deposit are considered eligible expenses in SSHRC research grant programs." | | Research Teams | Research Teams | |
| **CFI-** Funding guidelines state that CFI funding will focus on either the acquisition of a database, or the time-limited design and development of a database to the point that it is ready for exploitation by a designated research community. | Research Teams | | | |
| **CFI- Infrastructure Operating Fund**: "Costs of technical and other operational personnel where the costs are directly associated with the operation and maintenance of the infrastructure (e.g., cost for a technician to maintain or operate the infrastructure) | | | Research Teams | |
| **SSHRC, CIHR, CFI funding for Research Data Centres** | | Institutions | | Institutions |
| **Indirect Costs Program** | Institutions ("Information resources include databases, telecommunicatio | | Institutions- Operating costs of library or data centre | |

| **Funding Source** | **Production life cycle stage** | **Dissemination life cycle stage** | **Management life cycle stage** | **Discovery and Repurposing life cycle stage** |
|---|---|---|---|---|
| | ns, information technology and research tools.”) | | | |
| **CANARIE (Network-enabled platform)-** data acquisition, storage, manipulation, sharing and analysis tools that are used by the distributed community are of direct interest under this program. | Research Teams | Research Teams | Research Teams | Research Teams |
| **Provincial research funding agencies** (e.g., Ontario Research Fund) | Institutions- ORF- Research Infrastructure program funds the capital costs of acquiring, developing modernizing or leasing research infrastructure” | | | Institutions- ORF- Research Excellence program indirect costs: the overhead costs of doing research |

| Funding Source | Production life cycle stage | Dissemination life cycle stage | Management life cycle stage | Discovery and Repurposing life cycle stage |
|---|---|---|---|---|
| **Ontario Council of University Libraries and OntarioBuys** | | | | Library Consortia (OCUL)- Ontario Data Documentation, Extraction Service and Infrastructure Initiative (ODESI) |
| **Federal and Provincial Governments** | Government departments: Agriculture Canada; Canadian Space Agency/Herzberg Institute; Environment Canada; Fisheries and Oceans Canada; GeoConnections; Health Canada; National Research Council; Natural Resources Canada; Statistics Canada various provincial departments | Government departments: Agriculture Canada; Canadian Space Agency/Herzberg Institute; Environment Canada; Fisheries and Oceans Canada; GeoConnections; Health Canada; National Research Council; Natural Resources Canada; Statistics Canada; various provincial departments | Government departments: Agriculture Canada; Canadian Space Agency/ Herzberg Institute; Environment Canada; Fisheries and Oceans Canada; GeoConnections; Health Canada; National Research Council; Natural Resources Canada; Statistics Canada; various provincial departments | Government departments: Agriculture Canada; Canadian Space Agency/Herzberg Institute; Environment Canada; Fisheries and Oceans Canada; GeoConnections; Health Canada; National Research Council; Natural Resources Canada; Statistics Canada; various provincial departments |
| **Library and Archives Canada** | | | Preserves a limited number of research data sets that meet the definition of "publications". These are, however, a small sub-set of the research data sets requiring preservation in Canada. | |
| **Universities** | Labs | | Data centres, institutional repositories | Data centres, Data Liberation Initiative |
| **Research Charities** | Labs | | Data repositories | |

## Table 4: Current responsibilities for data stewardship in Canada

| | Production | Dissemination | Long-term Management | Discovery/Repurposing |
|---|---|---|---|---|
| **Researchers** | Meet standards for good practice. | Work up data for use by others. | Manage data for life of project.<br>Retain data for a given period of time | |
| **Institutions** | | | Manage data for life of project | |
| **Funding agencies** | Provide funds for the creation of data (CIHR, NSERC, SSHRC)<br>Develop research tools (CANARIE) | | Provide funds for management of data (CFI) | Provide funding for preservation (Indirect Costs) |
| **Research Libraries and Data Libraries** | | | Preserve some unique datasets. | Provide tools for re-use of data. |
| **Federal government departments** | Undertake research activities | Provide tools for re-use of data created by the organization. | Manage data for life of project.<br>Manage specific datasets for the long-term. | |
| **Library and Archives Canada** | | | Collects a limited number of research data sets that meet the definition of "publications". These are, however, a small sub-set of the research data sets requiring preservation in Canada. | |

## Table 5: List of data repositories organized by organizational model

N.B.: This list is not comprehensive.

| | |
|---|---|
| **Provincial or Federal Data Archive**- This type of repository collects data considered of provincial or national importance. In some cases, the mandate of the repository is also to facilitate science and research in Canada. Repositories are domain specific and usually have a fairly narrow scope in terms of data types collected. | Agriculture Canada<br>Canadian Space Agency/Herzberg Institute<br>Environment Canada<br>Fisheries and Oceans Canada<br>GeoConnections<br>Health Canada<br>National Research Council<br>Natural Resources Canada<br>Statistics Canada<br>Various Provincial Departments |
| **Virtual Organization**- Facilitate data sharing for e-Science, defined by Tony Hey as a “data-driven research methodology”. (note-some Federal data archives also aim to facilitate e-science) | Venus Data Management and Archive System<br>Project Neptune<br>The Atlas Project;<br>The Sudbury Neutrino Observatory<br>Genome Canada<br>The brain image database of the Montreal Neurological Institute<br>Eucalyptus<br>TRIUMF |
| **University Research Centres-** These repositories are managed by university department (there is some overlap here with Virtual Organizations) | The Androgen Receptor Mutations Database<br>Cadmium and Zinc Uptake by Grain Varieties Databank<br>Calcium Sensing Receptor Locus Mutation Database<br>Cambridge Structural Database<br>Canadian Lightsource<br>Canine Inherited Disorders Database<br>Canadian Institute for Advanced Research Program in Evolutionary Biology<br>CRYSYS (Cryospheric System in Canada)<br>Data for Evaluating Learning in Valid Experiments<br>Data on PAH (polyaromatic hydrocarbon) Aquatic Toxicity<br>David Dunlop Observatory Database of Galactic Classical Cepheids<br>Facility for the Analysis of Chemical Thermodynamics<br>Functional Group Electron density Databank for Carcinogenic Carbonyl Compounds<br>Fungal Mitochondrial Genome<br>Genomics and proteomics Advanced Applications Project<br>GOBASE - The Organelle Genome Database<br>Halogenated Organic Molecules Electron Density Databank<br>Hemoglobin Binding Affinity Constants Database<br>HEXdb - GM2 Gangliosidase Database Web Site<br>HumGen<br>Institute for Social Research Data Archive<br>International Infectious Disease Data Archive<br>Journal of Applied Econometrics Data Archive<br>The McGill Radar and Weather Data Archive<br>McMaster Cepheid Photometry and Radial Velocity Data Archive<br>The Northwest Atlantic Fisheries Centre<br>Phenylalanine Hydroxylase Locus Knowledgebase<br>PROMISE Software Engineering Repository<br>Organelle Genome Database (GOBASE)<br>The Organelle Genome Megasequencing Program (OGMP) |

| | |
|---|---|
| | Protist EST Program<br>Super Dual Auroral Radar Network (SuperDARN).<br>Viral Bioinformatics Resource Center<br>University of Waterloo Weather Station Data Archive<br>Wilson Disease Mutation Database |
| **Other Research Centres-** Repositories managed by other types of research centres | Arabic Genetic Disease Database - Toronto Sick Kids Hospital<br>Autism Chromosome Rearrangement Database<br>Biological General Repository for Interaction Datasets (BioGRID)- Mount Sinai Hospital<br>Canadian Barcode of Life Network<br>Chromosome 7 Annotation Project<br>Cystic Fibrosis Mutation Database- Toronto Sick Kids Hospital<br>Expression Profiles for C. elegans GFP:promoter Fusions<br>Genome Sequence Centre<br>Human Genome Segmental Duplication Database<br>Non-Human Segmental Duplication Database<br>Database of Genomic Variants<br>The Lafora Progressive Myoclonus Epilepsy Mutation and Polymorphism Database<br>National Database of FASD and Substance Use During Pregnancy Resources<br>Pseudomonas Genome Database V2 |
| **Data Libraries**- The primary mandate is to provide access to external collections, but some also collect selective research data sets, mainly in the social sciences. | University of Alberta Data Library<br>University of British Columbia Data Library<br>Carleton University Social Science Data Archive<br>University of Guelph Data Resource Centre<br>Queen's University Social Science Data Centre<br>Simon Fraser University Research Data Library<br>University of Toronto Data Library<br>University of Waterloo Leisure Study Bank<br>University of Waterloo Data Resource Centre<br>University of Western Ontario Data Resource Library<br>York University Institute for Social Research |
| **Institutional Repositories-** managed by research libraries, these are very nascent and focus mainly on collecting publications. However, a few IRs are collecting small data collections (e.g., images, and very rarely numerical data) | University of Alberta<br>University of British Columbia<br>University of Calgary<br>Carleton University<br>Canada Institute for Scientific and Technical Information<br>Dalhousie University<br>University of Guelph<br>International Development Research Centre<br>Université Laval<br>University of Lethbridge<br>University of Manitoba<br>McGill University<br>McMaster University<br>Université de Montréal<br>University of New Brunswick<br>University of Prince Edward Island<br>Université du Québec à Montréal<br>Queen's University<br>University of Regina<br>University of Saskatchewan<br>Simon Fraser University<br>University of Toronto<br>University of Victoria |

| | |
|---|---|
| | University of Waterloo<br>University of Windsor<br>University of Winnipeg<br>York University |

**Table 6: Barriers to access and preservation of research data in Canada**

| | **Barriers** |
|---|---|
| **1. Policies** | - Not all organizations have policies.<br>- Policies do not cover all research data.<br>- Some policies are in contradiction with others (privacy and ethics vs. access and retention).<br>- Policies are not being monitored.<br>- Few policies address issues of long-term preservation. |
| **2. Funding** | - Few funding mechanisms available for dissemination and long-term management of data.<br>- Pay per view and licensed access still commonplace. |
| **3. Roles and Responsibilities** | - Roles and responsibilities for ensuring access are not clearly defined or accepted.<br>- Responsibility for data retention lies with researchers only.<br>- Low levels of institutional commitment towards long-term preservation of data.<br>- Research organizations only rarely require data management plans. |
| **4. Data Repositories** | - Lack of repositories to support access and preservation.<br>- Repositories do not have preservation capacity. |
| **5. Standards** | - Lack of awareness of access and preservation standards.<br>- No national agency to provide guidance to researchers in terms of standards. |
| **6. Skills and Training** | - Lack of data scientists and information professionals to ensure long-term preservation.<br>- Researchers may not feel they have time or skills for preparing data |
| **7. Rewards and Recognition** | - Researchers are reluctant to share data.<br>- Researchers are unclear about who owns data.<br>- Priority is short-term use for analysis and publication of articles.<br>- Few reward or recognition mechanisms for sharing research data. |
| **8. Research & Development** | - R&D activities being undertaken in Canada are uncoordinated and not guided by national priorities. |