



Social Sciences and Humanities  
Research Council of Canada

Conseil de recherches en  
sciences humaines du Canada

Canada

National Archives  
of Canada

Archives nationales  
du Canada



## **National Research Data Archive Consultation**

Phase One: Needs Assessment Report

May 2001

# Table of Contents

<b>Executive Summary</b>	<b>1</b>
<b>1.0 Introduction</b>	
1.1 Background	4
1.2 Consultation Methodology	6
1.3 Responses to the Questionnaires	8
1.4 Stakeholder Submissions	10
<b>2.0 Gaps in Existing Research Infrastructure</b>	
2.1 Mandates of the National Archives of Canada and the National Library of Canada	11
2.2 SSHRC Data Archiving Policy	12
2.3 University Data Services	12
2.4 Authentication of Research Data	13
2.5 Canadian Presence on the International Scene	13
<b>3.0 Needs of the Research Community for Data Archiving Function</b>	
3.1 Lost Data	14
3.2 Lost Opportunities	14
3.3 Researcher Knowledge and Attitudes	14
3.4 Security of Research Data	15
<b>4.0 Benefits of a National Data Archiving Function</b>	
4.1 Beneficiaries	16
4.2 Building Resources for Multidisciplinary Analysis	16
4.3 Contributing to Education and Research Training	17
4.4 Building Capacity for Data Archiving	17
4.5 Preservation and Access for Electronic Text Resources	18
4.6 Non-Reproducible Data	18
4.7 Building Models for the Public and Private Sectors	18
4.8 Accessibility for Researchers to Expensive Resources	19
<b>5.0 Conclusion</b>	<b>19</b>
<b>6.0 Recommendation</b>	<b>20</b>
<b>Appendices</b>	
1) A Survey of SSHRC-Funded Researchers	21
2) List of Working Group and Resource Group Members	36
3) List of Submissions from Stakeholders	38
4) Research Data Archiving Reports and Other Related Documents	40

## Executive Summary

This document reports on the Phase One Needs Assessment of the National Data Archive Consultation. The report is submitted to SSHRC Council Board of Directors and the National Archivist of Canada.

The report presents evidence that there is a substantial gap in Canada's research infrastructure—a national research data archiving service or function. This evidence was gathered from the university-based social science and humanities research community; data archivists and librarians from both university and other public institutions, and a variety of stakeholders concerned about the preservation and management of digital research materials.

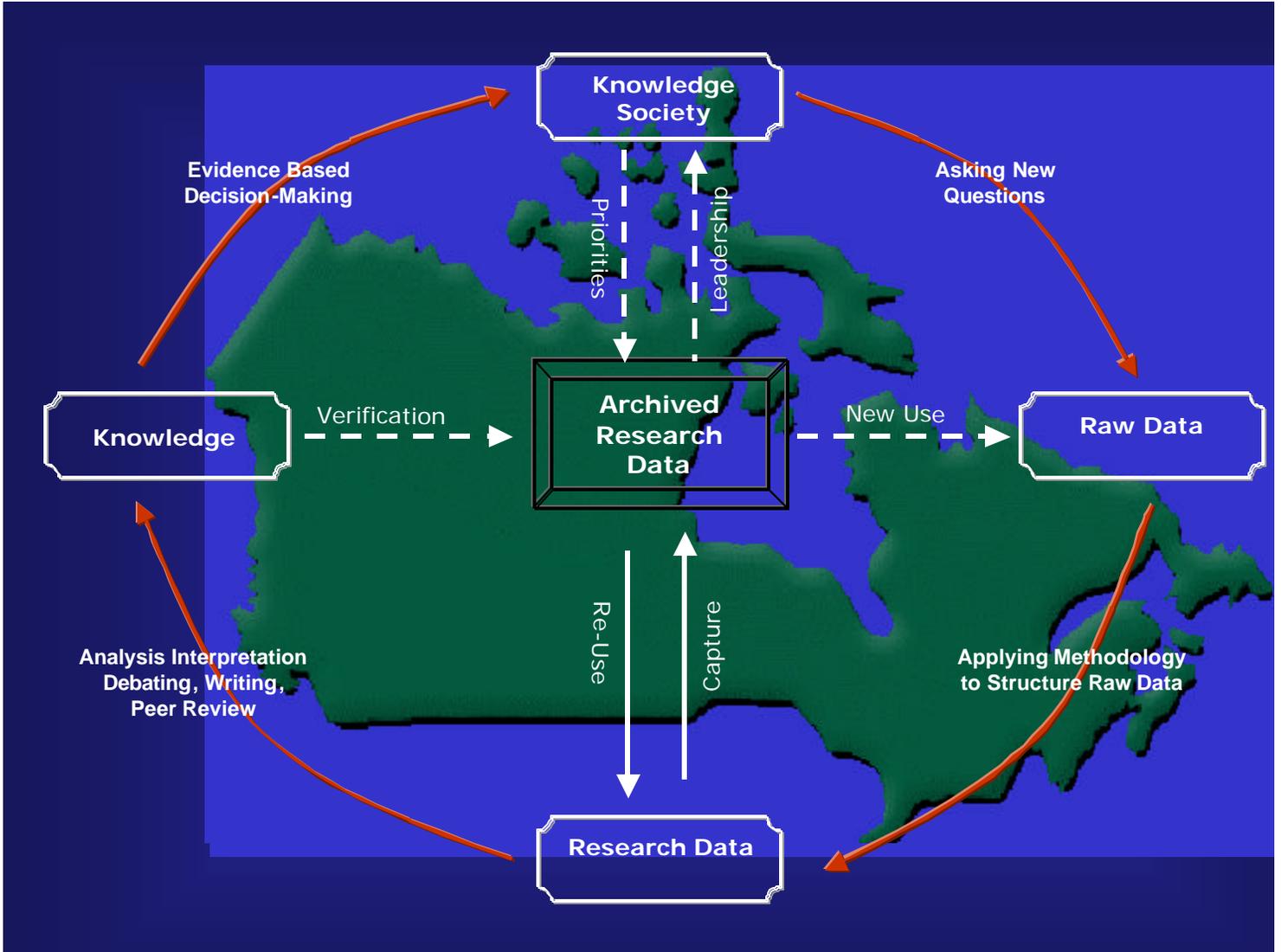
SSHRC and the National Archives of Canada asked a group of experts to investigate whether the structures and mandates of existing institutions meet the data archiving needs of the research community, what those needs are, and who will benefit from improved research data archiving services.

From the evidence gathered it is clear that there is no national institution mandated to preserve, manage and make accessible research data, that there are significant needs within the research community that are not being met, and that substantial benefits would result from the creation of a national research data archive. What form such an archive should take, however, has not yet been investigated.

It is also clear that, while a considerably large volume of research data is gathered each year by Canadian researchers, opinions about whether these data should be shared, and if so, under what conditions, are not unanimous. While the archival and library communities are convinced of the need for a national data archiving service, a significant number of researchers are not convinced or have a limited understanding of the value of a research data archive.

At this point, the Working Group that conducted the needs assessment is recommending that they be authorized to proceed to Phase Two of the consultation and investigate the most appropriate form of a national research data archive for the Canadian context.

# Archived Research Data in the Knowledge Society



**A knowledge economy is based on the ability of its participants to transform data, information and ideas into new knowledge. A data archive, like a new piece of diagnostic equipment, is a tool of innovation.**

## **1.0 Introduction**

In May 2000, a researcher asked the Data Archivist at the University of Toronto if she could get for her a statistical data file entitled "Access to Justice in Ontario, 1985-1988". A simple request. The file, however, is not to be found in Ontario, or anywhere else in Canada, but at the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan. The researcher no doubt received the requested file, uncorrupted, compatible with available software, complete with metadata, and delivered over a secure Internet channel. The situation would not be a problem, except that if the researcher had been from some other university, one that is not a dues-paying member of the ICPSR, her request could have been given a lower priority, levied a substantial fee, or been denied altogether.

This report provides the evidence that there is a substantial gap in Canada's research infrastructure. Increasingly, modern society depends on knowledge to function. Knowledge is the product of good information, which in turn, is based on reliable data derived from a variety of reliable sources. In Canada we have no national agency, institutional body or network in place to ensure that research data are preserved and made accessible for re-use, verification and replication of findings, for benchmarking, or for representing Canadian interests in the international arena. As a result, valuable research data are being lost or stored elsewhere and often fail to become part of the on-going knowledge-building process. As one of the stakeholders expressed this,

*Canadians, one way or the other, have paid a price to create these data files, because it was felt that there was a need to gather certain information. However, the potential of these data files is such that researchers can come back to them many years later and look for new and different relationships. If the data files are lost, then the investment made in creating them is not getting the return that it could with preservation.*

Today, it is technically possible for a researcher in Saskatoon or St. John's to find out not only which libraries have a particular book or journal, but also which archives have the related data sets used to produce the publication. It is also possible to have both efficiently delivered to his or her desktop computer. The delivery system, CA\*Net3, is in place, and various electronic journal projects are in development. The missing component is a national research data archiving function. A modern, comprehensive depository of research data sets has the potential to create tremendous synergies between libraries and archives, enabling researchers from across the country to access published materials, archival materials, and related data sets at the same time, from the same place.

There are a number of university-based research and research infrastructure projects currently in development that will require the services of a research data archiving function. A few examples include:

- The Visual Resource Centre at McGill University, that will develop a digital depository of visual images, commercial artefacts, photographs and cultural materials.
- The Text Analysis Portal for Research, a joint project involving the University of Victoria, University of Alberta, University of Toronto, McMaster University, Université de Montréal, and University of New Brunswick, that will provide Internet access to a wide variety of textual databases and software tools.

- The Supply Chain Research Institute at Athabasca University, that will develop a series of databases to support research on organizational management practices.
- The Video Archive at Brock University, that will provide access to digital recordings of television and film productions.

## Definition of a National Research Data Archiving Function

For the purpose of this report, a research data archiving function is defined as preserving, managing and making publicly accessible digital information structured through methodology for the purpose of producing new knowledge. Such an archiving function provides stewardship for those outputs of the research process that exist between raw research materials and published results. Acquisitions could include digital information produced by researchers and of interest to researchers. This statement emphasizes six points:

- Three important aspects of archiving are preservation, management and access.
- The materials to be archived are research data in digital format. Although several stakeholders felt that a function of a national data archive should be to provide access to paper and electronic publications, there are existing institutions designed to meet this need.
- The focus is on digital material or information has been gathered through a defined method, and with a specific purpose in mind. It is not unstructured information.
- The objective is to produce new knowledge. Not all digital information meets the requirement for preservation as research data.
- The function would bridge the gap between raw, unprocessed data and published results of investigative research.
- The function encompasses both research data compiled by researchers and data of interest to researchers, subject to the limitations of financial resources.

Throughout this report, we refer to a national data archiving “function” rather than an agency or institution. What form the function might take is the subject planned for Phase Two of the consultation.

By a “data archiving function” we mean the broad range of preservation, management and access services offered by many data archives in other countries. These services include off-site preservation, Internet based access to data sets, retention protocols, metadata creation, transportation of data across software systems and hardware platforms, providing input for the establishment of international standards, and many more. By “national” we mean a function that would provide these services to all Canadians.

Archiving data so that it is preserved and accessible to future users also involves advising researchers on best practices, active intervention, migration, refreshing, testing, and many other activities. The countless varieties of physical storage formats, media encoding, file formats, hardware and operating systems means that a great deal of work must be done when data are first acquired, so that the data are not affected by the obsolescence and disappearance of the hardware and software with which it was first created.

### 1.1 Background

A research data archive preserves, manages and makes accessible, in digital form, the scientific and cultural materials used by researchers to build our knowledge-based society and economy. These functions lie at the heart of information and knowledge management, and yet, in Canada, no institution, agency or network exists for developing a national research data management strategy or for accepting stewardship of digital research materials. The OECD/SSHRC Ottawa Workshop on Social Science Research Infrastructure held in October 1999 came to this

conclusion. So did organizations such as the Canadian Association of Public Data Users, the Canadian Global Change Program, and investigators such as Dr. John English, reporting on the National Archives and National Library. All have called for action to address data archiving and management issues in Canada.

In mid-2000, the Social Sciences and Humanities Research Council (SSHRC) formed a partnership with the National Archives of Canada to conduct a two-phase consultation on research data archiving. The first phase assessed the need among the research and archival communities for a national data archiving function. If the SSHRC Board and the National Archivist agree that there is sufficient need, the consultation will move to the second phase. The second phase will investigate the most appropriate form of a national data archiving function and how it might be established.

The collection of research data in Canada is a huge enterprise. The federal government alone spends over \$1 billion annually. At least this much again is spent by provincial governments, non-profit organizations and academic researchers. What the private sector spends is unknown. By the most conservative calculation, the total financial resources devoted to data collection are in the billions of dollars each year. Ensuring maximum return on this investment is a responsibility shared by all publicly funded researchers and research organizations.

The rapid development and adoption of computer technology in research and the enormous communication and co-ordination possibilities offered by the Internet have heightened both the need for a co-ordinated data management system, and the potential impact such a system could have on all areas of research. Just as important, the range of disciplines and research areas using computer databases as research tools has expanded dramatically. Traditional statistical data sets are no longer predominant and are now joined by geospatial data, digital maps, digitized texts, electronic journals, oral records, digital images, photographs and digitally recorded performances.

In fact, we are only beginning to realise the rich research potential of digital information in a computerised and Internet connected environment. Data sets can now be created from quantitative and qualitative information in a wide variety of formats. Through the Internet, these data sets can be transported over any distance in seconds. With the proper hardware and software, they can be searched, manipulated, transformed and integrated with ease. Stored digitally, such data requires only a fraction of the physical space necessary for paper or film records. This dramatically increases both the efficiency and possible scope of research activities.

Since the mid-1960's, many countries have recognized the utility of unified and co-ordinated data archives and have built national facilities in a wide variety of institutional models. Some are highly centralized state funded operations. Some blend user fees with public funds. Most are based at universities to better serve the research community. Data collections range from survey and polling data to literary works, television programs, and medical information. Services offered include on-site computer labs, automatic ordering of data sets from other depositories, software development and evaluation, training programs, guides and protocols for the creation of data sets, metadata development, transportation of data across technologies, custom designed teaching materials, full service Internet networks, and branch offices.

Today no one knows how many social science and humanities research data sets exist in this country. No one knows what surviving data sets contain or whether they are properly stored. There is no master index for any discipline, let alone general fields of study. We do know that major irreplaceable data sets have been lost—all Gallup polls before 1947 and most before 1952, a large portion of the public opinion polls taken before the 1995 Quebec referendum—but so poor is the documentation that we cannot estimate the losses. Evidence, however, suggests they are great and affect researchers significantly.

Data management, preservation and access are of increasing concern to researchers across many disciplines. This is particularly true in interdisciplinary studies where researchers gather information from a wide variety of often-unfamiliar sources. Tracking down data sets outside one's discipline is frustrating and time-consuming. And, as more social science and humanities research takes an issue orientation, rather than a disciplinary focus, the need for the co-ordination and accessibility of research materials increases.

*Data archiving is more than making a backup copy of a file... Data archiving involves the long-term commitment to the resources, expertise, and public service required to ensure perpetual access to data files, to describe and document the files, and to provide access to and intellectual control of those files. One of the reasons why researchers may not be excited about this issue is that it is difficult to find out what data have been collected. It only makes sense to use economies of scale and centralize the resources required for an enterprise of this magnitude.*

Questionnaire Respondent

SSHRC, the National Archives, and indeed Canada, are not alone in grappling with the issues of research data archiving. A number of countries have recently established a national research data archive, among them Finland, Japan, Estonia, Slovenia, and South Africa. Even existing national data archives in a number of countries are conducting consultations. The Economic and Social Research Council of Great Britain is reviewing its policy on data archiving. Since the mid-1990s, the US National Archives and Records Administration and the National Research Council have been studying how to manage and preserve the vast amount of government-sponsored scientific and technical data produced in the natural sciences. A number of years ago, the International Council of Scientific Unions established CODATA, an interdisciplinary committee to deal with data management, quality control and dissemination from all scientific and technical disciplines. For the past few years, the International Council for Scientific and Technical Information (ICSTI) has been commissioning investigative reports and holding international workshops in an attempt to build a co-ordinated approach to archiving that spans national and disciplinary boundaries.

As a key report commissioned by ICSTI points out, the challenges of managing, preserving and providing access to research data are shared by all scientific sectors and disciplines, including the humanities.<sup>1</sup> Computer technologies facilitate digital storage of a growing diversity of research data. Maps, photographs, sound recordings, hypertext, dynamic pages, geographic information systems, multi-media, and interactive video—all these are in active use today and all require effective stewardship and preservation if they are to remain usable building blocks in the knowledge creation process.

## 1.2 Consultation Methodology

SSHRC and the National Archives of Canada asked a Working Group of experts in various fields to conduct a thorough investigation of the need for a national data archiving function or service. The Working Group addressed the following questions:

- To what extent is there a need for a unified and co-ordinated data archiving function? Are modest changes to existing institutional policies and mechanisms adequate to meet current and future requirements?

---

<sup>1</sup> *Digital Electronic Archiving: The State of the Art and the State of the Practice*, ICSTI.

- What gaps exist in the mandates and structures of existing institutions in relation to management of research data?
- Who will benefit from the improved management of research data and to what degree?
- How will effective research data management, preservation and access contribute to Canadian research capacity?

Two groups of accomplished researchers and archival and library science experts participated in conducting the consultation. A nine-person Working Group, working in conjunction with a larger Resource Group, conducted the investigations and produced the Phase One Report. Chaired by Dr. John ApSimon, the former Vice-President of Research at Carleton University, the Working and Resource Groups brought together academics from both the social sciences and humanities, the director of a leading communications research centre, a university data archivist with extensive experience in promoting national systems, and a federal Treasury Board official who works to develop national infrastructure in digital environments. Both groups met face-to-face.

The consultation attempted to seek input from a variety of well-informed sources. In general, the Working Group concentrated on two areas; university-based researchers, and archivists and librarians both within universities and government. These are the principal producers, users and managers of research data, and it was felt they would have the best grasp of the data archiving needs in Canada.

The Phase One Needs Assessment involved the following activities:

- Letters were sent to 246 directors of university-based research institutes or groups announcing the consultation and requesting input and opinions on data archiving needs.
- In October 2000, an open Stakeholders Meeting was held at the National Archives. Fifty-five persons from a wide variety of professional backgrounds attended and participated in the discussions.
- Letters were sent to the Deputy Ministers of major federal government departments that have research as part of their mandates.
- SSHRC staff made several presentations to particular groups, such as the CANARIE E-Learning Projects Group, and to researchers the University of Calgary, University of Alberta, McGill University, and Brock University.
- In partnership with Natural Resources Canada, a publicly accessible Web site was created, offering background information and an open discussion forum.
- Individual investigations by Working and Resource Group members in areas such as social statistics, geospatial data, history, text analysis, security and authentication of electronic records, etc.

Separate needs assessment questionnaires were sent to four groups of data producers, data users and data managers and archivists:

- 1) A 20 per cent stratified sample of all researchers who received a research support grant from SSHRC between 1998 and 2000.
- 2) A group of stakeholders identified during the consultation process, including federal and provincial government departments, private sector organizations, archival associations, academic associations, research grant agencies, university research offices, academic library associations and individual researchers.
- 3) Directors of the 10 Data Repositories listed by SSHRC in its *Grant Holder's Guide*.
- 4) Data Liberation Initiative (DLI) contact persons at 66 Canadian universities. These are professional library staff responsible for managing publicly-accessible Statistics Canada data files as well as a variety of other research data.

Response rates for the questionnaires ranged from 20 per cent for the stakeholders group to 70 per cent for the directors of data repositories. The response rate for SSHRC-funded researchers was 26 per cent. Respondents covered all regions of Canada, and included both francophones and anglophones. Not surprisingly, the questionnaires completed by the identified stakeholders, the directors, and DLI contact, showed a great deal more familiarity with data archiving issues, particularly those of a complex technical nature.

*Publicly-funded research should require that the data generated, research instruments employed, design used and sampling frameworks etc. be archived and made available for other researchers. This would be very important to activities such as fostering collaborations, longitudinal studies, replication studies, comparative studies, creation of 'normative' question designs in certain areas of inquiry, and secondary analyses. Transparency, accountability and responsibility would be encouraged by requiring the archiving and access to data. Further, consideration of such data should become a more central attribute of planning 'new' primary research—less re-inventing the wheel and more imaginative and creative work might result. Thoughts—for what they are worth.*

Questionnaire Respondent

### 1.3 Responses to the Questionnaires

#### SSHRC-Funded Researchers

A twenty-percent stratified sample of all researchers receiving a grant from the SSHRC between 1998 and 2000 was drawn. A total of 116 responses were received. Each researcher received a questionnaire asking about her or his experiences in creating and archiving research data or using research data created by others. They were also asked about their attitudes toward fundamental issues that underlie the principles of archiving data and were asked to assess the importance of establishing national services for the preservation of research data.

In any given year, as many as one-half of SSHRC-funded researchers produce research data. For those who responded to this consultation, the figure is 55 per cent. This extrapolates to approximately 1200 data sets created by SSHRC-funded researchers between 1998 and 2000, or an average of 400 each year. As of January 2001, only 7 per cent of those researchers surveyed had archived their data, and only a further 18 per cent reported that they intended to do so. Of the 18 per cent that intend to archive their data, less than one half were able to identify an actual data archiving service or agency. Even assuming the best scenario, this means that, over a three-year period, we will fail to preserve and make widely accessible close to 950 publicly-funded data sets.

What are researchers attitudes towards data collection and archiving issues?

- 81 per cent stated that data, whether created by themselves or by others, is a valuable by-product of research.
- 79 per cent believe that research data belongs to the principal investigator as his or hers intellectual property.
- 78 per cent stated that secondary analysis of existing data is a valid research methodology.
- 73 per cent think that data should only be shared if the principal investigator decides to share it.
- 71 per cent said that the research councils should cover the costs of preparing data for sharing.
- 68 per cent agreed that researchers have a responsibility to act as trustees of data that cannot be easily reproduced.

- 50 per cent agreed that spending resources to prepare data for archiving would not be a waste.
- 48 per cent agreed that archiving should be an integral part of conducting research.

See Appendix 1 for a complete analysis of the SSHRC-Funded Researchers Questionnaire.

### **SSHRC Listed Data Depositories**

In 1990 SSHRC published a new appendix to its guide for applicants containing a list of ten university data libraries willing to serve as repositories for data files created from research funded by the SSHRC. Since this appendix was first published, over 8,500 SSHRC grants have been awarded yet the data for less than ten of these projects have been deposited. A survey was conducted of these ten data libraries to gain an understanding of the efficacy of the existing repository method and of the level of support that these data libraries are able to commit to archiving research data. Seven of the ten repositories completed the questionnaire.

When asked if they had received any files deposited directly from an SSHRC-funded project since the repository listing was first published, three of the seven data libraries had not. Of the four receiving a data deposit, one had received files from three projects, two had received files from two projects, and one had files from one project. When asked how successful they considered the current repository list in the *SSHRC Grants Guide* in directing researchers to deposit their data, five said very unsuccessful while two said unsuccessful. Following this question, they were asked how effective they thought the SSHRC repository list was in fulfilling the function of a data archive in Canada. Six responded that the repository list is very ineffective, while one said ineffective.

When asked what factors they thought determined whether a researcher will deposit her or his data, the replies tended to focus on the current data practices of researchers and the responsibilities of grant councils to support data preservation. First, the respondents felt that researchers lack knowledge both about archiving data and the requirements of the SSHRC to deposit data.

Second, they focused on the responsibilities of research councils. The existing SSHRC policy about depositing data is seen as neither strict enough nor enforceable. One respondent mentioned the need for follow-up by the Council to determine compliance with the deposit policy. Two respondents suggested sanctions by the Council in the event that data are not deposited. Three commented that the costs of preparing data for deposit currently are not covered by grant councils. One respondent summarized this concern in the following statement.

*[Researchers] will need instructions and assistance with the creation of proper documentation of their data in order that others may easily use their data. In the case of the two data sets that were deposited with us, the PI's never did produce proper documentation for their studies.*

### **Stakeholders**

The stakeholders questionnaire concentrated on three issues in particular: whether it is important to establish a national data archiving function or service, who would use such a service, and whether they knew of existing data that is at risk of being lost or destroyed. A total of 43 respondents completed questionnaires, which is roughly a 20 per cent response rate. Eighty-eight per cent of these respondents reported that they work in an academic setting while only 12 per cent work in the federal or provincial government.

The questionnaire administered in this survey began by asking stakeholders about the importance of national support for the preservation of research data. The vast majority (95 per cent) of the respondents agreed that it is “very important” (70 per cent) or “important” (25 per cent) for Canada to establish such national support. One hundred percent of the respondents from the federal and provincial sector felt that national support for preserving research data is “very important”, while 68 per cent from the university sector reported the same level of importance (27 per cent of those remaining said that it is important).

Sixty-four per cent of the respondents indicated that they know about data that are at risk of being lost. A follow-up question, which was answered by twenty-seven respondents, asked for a description of data that they know to be vulnerable. Most of these descriptions consisted of generalizations about types of data rather than titles of specific data collections. Two general categories emerged from the answers provided. First, data from a variety of producers were identified. This included data created from research funded by the SSHRC, producers of geo-spatial data, the data from graduate student thesis or dissertation research, data created by research units in federal and provincial government departments, community-based and regional historical databases, data from small-scaled projects, and private and government-sponsored polling data. The second category consisted of technical reasons that data are at risk. Examples of this category included data stored on legacy media and data that have not been properly documented.

Respondents named several specific projects at risk, including the Canadian Illness Survey, the Nutrition Canada National Survey, the Health Survey of Ontario, the Atlantic Canada Shipping Project, and the Canadian Families Project.

### **Data Liberation Initiative Contacts**

The DLI Contacts are a special sub-population of university professionals in charge of providing access to Statistics Canada materials and other research data at their institutions. Consequently, this group is knowledgeable of the importance of data to research, works to create access to research data on behalf of researchers and students, and is aware of the challenges in securing access to such data. Thirty-one responses were received from this group to a brief questionnaire that focused primarily on the importance of national data archive services to various sectors and research activities in Canada. In addition, they were asked specific questions about the efficacy of existing practices of preserving data and whose responsibility it is to preserve research data.

The librarians and staff providing access to data in their institutions recognize the importance of preserving research data from all levels of government and from the academy. They are less in agreement about the importance of preserving research data from the private sector. It may be that they perceive this as an unrealistic goal. Most agree that libraries are not in the position to provide data archiving services. However, the overwhelming consensus is that national data archiving services are needed.

### **1.4 Stakeholder Submissions**

Twenty submissions, equally divided between archivist-librarians and academic researchers and sampling views from coast to coast, unanimously and strongly recommend the establishment of a National Data Archive (see Appendix 3). University faculty, whose academic research SSHRC and other provincial and federal agencies fund, represented the humanities, information science, and the social sciences. All are clients for and donors of research data. Their briefs emphatically state that only a national data archiving function can ensure that credible, creative data-based research will continue.

Their observations and comments, along with those of the respondents to the questionnaires, and the members of the Working and Resource Groups, are presented below in the context of the questions posed in the Terms of Reference.

## **2.0 Gaps in Existing Research Infrastructure**

For the preservation, management and access of digital research materials, significant gaps exist in the mandates and structures of both federal institutions and agencies and university-based libraries and archives. Archivists from many sectors—federal, provincial, post-secondary, and public—testify that they cannot separately or together solve the massive, historic problem of data management without an overarching national data archiving function or agency. Many archives handle research data now, but none considers itself capable of providing the essential central infrastructure.

### **2.1 Mandates of the National Archives of Canada and the National Library of Canada**

Until 1986, when it was closed, the Machine Readable Records Division of the National Archives of Canada accepted digital research materials. Today, the National Archives focuses its preservation efforts on the records of the federal government and on private records of national significance, which it defines as documents that “record the efforts and experiences of individuals, groups, institutions, corporate bodies, and other organizations which have become nationally or internationally recognized. They also document the physical environment of Canada, as well as events and trends (cultural, political, economic, social, demographic, scientific, and religious) having a broad national scope.”<sup>2</sup>

The 1994 National Archives acquisition strategy lays out ten broad theme areas as a guide to selecting acquisitions. The objective of applying this thematic approach is to build a comprehensive representative sample that contains significant or unique information that will substantially enrich our understanding of Canada’s history, society, culture and people, such as the National Census conducted by Statistics Canada. Although this is an excellent approach for the National Archives, given its overall mandate and limited resources, it only partially meets the archival needs of the academic and non-academic research communities. Researchers need access to more than a representative sample.

In the end, the National Archives is an institution of the federal government charged with preserving the national memory. The Canadian research community, on the other hand, operates within a much broader geographic, social, cultural and economic analytic framework. Researchers create data as a means to assist in the examination of social, economic and cultural realities, often beyond of the borders of Canada. The mandate of the National Archives simply does not encompass the varied research data archival needs of Canada’s academic researchers.

This is also true of the National Library of Canada. Its mandate is to preserve the published heritage of Canada. In recent years, the mandate has expanded to include electronic publications, electronic journals and Web sites, but excludes virtually all research data. Data sets are not published, and thus outside the National Library’s mandate. Furthermore, neither the National Library or the National Archives currently have the statistical and technical expertise needed to close this gap in Canada’s research infrastructure.

---

<sup>2</sup> National Archives of Canada, *An Overview of the Acquisition Policy of the National Archives of Canada*.

## **2.2 SSHRC Data Archiving Policy**

In 1990, the Social Sciences and Humanities Research Council adopted a specific policy regarding data sets created by researchers using public funds. As stated in the guide for applicants,

SSHRC requires that data collected with its assistance, including machine-readable files and computer databases, become public property and be made available for use by others within a reasonable period of time, on condition that confidentiality of information and right to privacy are protected. Consequently, it also requires that the institution of the principal investigator or any other institution which becomes the repository of the data, take the necessary steps to preserve the data and facilitate its accessibility to researchers.

The guide provides a list of university data services (one of which is defunct) where researchers could deposit their data sets. It also states that costs associated with the archival preparation of data are eligible expenses.

The intention of this policy is to advance knowledge creation in the social sciences and humanities by encouraging research data sharing among researchers. Data sharing strengthens our collective capacity to meet academic standards of openness by providing opportunities for further analysis, replication, verification and refinement of research findings. These opportunities enhance the development of fields of research and the potential for inter-disciplinary work. In addition, greater availability of research data can contribute to improved training for graduate and undergraduate students, and make possible significant economies of scale through the secondary analysis of existing data. Finally, researchers whose work is publicly funded have a special obligation to maintain openness and accountability.

Unfortunately, this policy has not achieved its objectives. When SSHRC-funded researchers were asked if they had created data sets in past research projects, 80 per cent of those who said yes, had not archived their data. In fact, over an eleven-year period, only ten data sets have been deposited with the repositories listed in the SSHRC guide.

Among those researchers surveyed who did, or intended to, archive data from SSHRC-funded projects, several did not know where to send data sets. Others relied on university libraries. Still others intended to use US archives or Canadian institutions with no archival capacity. And, one researcher stated, "we will maintain our own archive, as we are not willing to archive it in a non-Canadian archive".

## **2.3 University Data Services**

University data archivists and librarians stated that universities do not have the resources necessary to preserve, maintain and update increasingly large and complex data bases, nor are they able to provide the broad range of services that are needed by academic researchers. Their main preoccupation is providing local patrons with access to readily-available data files. What these institutions have in common is that they directly confront the problems associated with the management, access and preservation of research data while they struggle, on a daily basis, to deal with the gap left by the absence of a national agency with sufficient resources, expertise and authority to offer the necessary services.

Existing Canadian university data archives are sparsely staffed and equipped. Of the seven repositories that responded to the Data Repositories questionnaire, none has more than 2.5 full-time staff members and one has less than 1 full-time staff member. This highlights their incapacity to handle the broad range of services necessary for the effective archiving of research data.

## **2.4 Authentication of Research Data**

Authentication is one of the means used to verify the identity and integrity of data. It refers to validating data files and/or the person accessing the data at a specific moment in time. Authentication ensures that those accessing the data files are valid users and that they do not alter the data in any way. The presence of this authentication implies that, at the time of its use, the data in question have been verified to be what they purport to be and have been created by the person or persons who purport to have created them. However, this is not sufficient to ensure ongoing authenticity.

Authentication procedures embedded in the processes of creation, transmission, receipt, use, maintenance and preservation of data files are the most effective way to ensure the authenticity of the data over time, especially considering the need for ongoing reproduction of the data. Given the rapid obsolescence of electronic systems, we cannot preserve the data themselves: we can only preserve reproductions of the data. Thus, we need national standards both for the authentication of the reproductions and for the protection of identity and integrity of these reproductions. Preservation of data is a proactive endeavour that requires ongoing measures applied to the data. This can only be undertaken by a permanent, active service organization.

## **2.5 Canadian Presence on the International Scene**

Canada has limited capacity to negotiate international data exchange agreements because, unlike many other countries, especially those in Europe, we have no national agency. Lacking a national research data function, Canada has no co-ordinated voice in the creation and establishment of international research data standards, in metadata schemes such as Data Documentation Initiative (DDI), in tools for data access such as the Networked Social Science Tools and Resources (NESSTAR) project and the Language Independent Metadata Browsing of European Resources (LIMBER) project, and in collaborative international infrastructure projects such as the European Union Frameworks. As well, Canada lacks national representation on such important bodies as the International Federation of Data Organizations and the Council of European Social Science Data Archives.

Data-sharing has become a necessary part of participation in the global economy, for it is the sharing of data that provides crucial information for countries developing social and economic policies to suit today's global context. While Canada does take part in some projects and does belong to some international organizations, our collaboration is, at best, ad hoc. We do not even know what Canadian data are available, where they are located or whether they are useable.

### **3.0 Needs of the Research Community for a Data Archiving Function**

Stakeholders noted that a national data archiving function could address three archiving needs of the academic community: (1) the archiving of data that researchers gather but lack the expertise to maintain; (2) the archiving of computerised records of government, quasi-public and private institutions that are of direct interest to researchers; and (3) the archiving of computerised records of individuals who either contribute to the knowledge building process or whose data sets are of interest to researchers.

#### **3.1 Lost Data**

Stakeholders repeatedly expressed dismay that, while we cannot and should not preserve all research data, many data collections, judged essential by any criteria, are being destroyed. Currently, there is a need to improve the archiving of entire classes of research data, such as:

- geospatial data
- social science and humanities data sets created with SSHRC funding
- social science and humanities data sets funded from other sources
- unpublished data created by the federal government
- unpublished data created by provincial, municipal or private-sector sources
- statistical survey data
- government or private polling data.

One specific example cited is the expensive data developed for Michael B. Katz's book *The People of Hamilton* (Cambridge: Harvard University Press, 1975), which is held by the Institute for Social Research at York University, where the equipment can no longer read the magnetic tapes on which the data are stored.

#### **3.2 Lost Opportunities**

One of the paramount problems researchers face today is difficulty in locating data relevant to their research. There is no 'union list' of data sets held by data producers, distributors or other researchers. This means that researchers may have to needlessly replicate costly studies or rely on anecdotal evidence rather than objective data. A national data archiving function could potentially place information on data sources right on the researchers' desktops, thereby saving time, money and other precious resources.

Both researchers and data archivists repeatedly made the point that preservation of spatial data sets is inadequate in Canada and that we are losing much data as a result. Few organizations producing spatial data have an archiving capacity and old data sets are often over-written or discarded. In addition, there is the problem of continuously updated databases which require provisions for "point in time" archiving.

#### **3.3 Researcher Knowledge and Attitudes**

Within the Canadian research community we find substantial differences in attitudes towards data archiving, as well as a great deal of confusion and lack of understanding of the importance and nature of data archiving. Only 60 per cent of the SSHRC-funded researchers surveyed clearly understand and appreciate the importance of establishing a national data archiving service or function in Canada. A substantial minority of researchers are unsure of, or undecided on, the

issues that underlie the preservation of research data. Indeed, one function of a data archive could be to inform the research community about the value of research data archival services.

*The general awareness of data archiving among researchers seems to be quite low. It isn't that they don't understand the concept, but rather that data archiving isn't part of their normal practices in conducting research. I see this fault lying with the training of researchers in their graduate school years and with senior researchers who should be mentoring junior researchers about data archiving. If the importance of the practice isn't taught as part of the research method, data archiving won't generally be discussed or perceived as an important research activity. This needs to be addressed by professional associations and deans of research responsible for the training of the next generation of researchers.*

Questionnaire Respondent

Several university data archivists pointed to self-interest among researchers and a research culture that does not emphasise data sharing. One archivist mentioned “concerns about losing control over the research potential of their data—their fear of getting scooped by other researchers who find and use deposited data.” Part of the problem is that those researchers who only collect and analyse their own data feel a strong sense of proprietary ownership and are reluctant to share. On the other hand, those researchers who use data collected by other agencies or other researchers are both much less proprietorial and much more aware of the need for a data archiving service. In fact, 76 per cent of the SSHRC-funded researchers surveyed reported that they had used data sets in previous research. Of the 45 respondents who reported that they had used or tried to use data produced by other researchers, three had been denied access to the data. Nine respondents reported that they had denied other researchers access to their data, despite the fact that this is a direct violation of SSHRC policy.

### **3.4 Security of Research Data**

The security of data a crucial need for the research community. Security standards need to be formulated and articulated at the national level in order to ensure both adequacy and consistency in the management of data archives. These standards should address: (1) methods for clearly identifying data assets and risk management procedures for assessing the vulnerabilities of data sets; (2) identification of the legal, statutory, regulatory and contractual requirements that must be taken into account, including ethics guidelines and intellectual property rights; and (3) a set of principles, methods and procedures that organisations must follow to ensure the reliable creation, secure maintenance, confidential use and authentic preservation of their data.

Security policies and procedures need to be developed at a national level, and organizations must be actively encouraged to apply them, in a consistent fashion, to all of the data files considered to be of national value. This is vital not only to ensure that the data are only accessed by those authorised to do so, and that the data are identifiable and genuine over the long-term, but above all to ensure that researchers and any other users are able to verify that the data are what they purport to be and have been maintained intact over time. If researchers cannot rely on the integrity of the data on which they base their research, and if the users of that research cannot in turn verify that data, the research results are useless.

## **4.0 Benefits of a National Data Archiving Function**

Ultimately, the Canadian public will benefit most from improved returns on investments in research and from the increased production of knowledge made possible by the repeated use of existing data sets. Canadian capacity to study how social, cultural, economic and medical phenomena change over time will benefit from the recovery of key historic data, as well as from improved and co-ordinated access to contemporary data of scattered origins.

### **4.1 Beneficiaries**

Among the respondents to the Stakeholders Survey, fully 88 per cent saw university research as the greatest immediate beneficiary of improved research data management. Benefits would include a significant reduction in research costs due to the prevention of duplication of data collection, improved access to a wider variety of data, which in turn will facilitate more thorough analysis and greater inter-disciplinarity, the ability to verify research results without having to duplicate data collection, and improved access to non-replicable time-series data sets.

In order of importance, the stakeholders feel that the most frequent users of a national data archive system or service would be (1) university researchers, (2) secondary and postsecondary teachers and students, (3) policy analysts and decision-makers in various levels of government, (4) private researchers and non-governmental organizations, and (5) the general public. Several respondents noted, however, that this scale is based on current capacity to undertake research using research data, and not on possible patterns of use in the future.

Successful data preservation and access will not be a remedial process (as happens now with paper resources) but rather a proactive process where preservation concerns will be a feature of resource creation. Academic libraries and archives, because they work so closely with the research community that creates these resources, can do much to inculcate a culture of effective archiving and preservation that facilitates continued access. As new research modes/resources evolve (e.g., massive genetic databases, dynamic 3D simulations, multimedia environments, virtual worlds) libraries and archives are in a position to work with users on how to ensure preservation and access from the time of project conception.

Stakeholders consistently stated that we can vastly enhance Canadian capacity for analysis of research data by creating a cohesive collection of research data with which we can then develop new research and statistical techniques tailored to the unique conditions in this country.

### **4.2 Building Resources for Multidisciplinary Analysis**

Preserving data permits opportunities across disciplines for innovative, and often unforeseen, uses of data. Let us look at a specific example; the task of establishing the genealogy of French-Canadian families from the New France period to the twentieth century. The project has benefited from considerable public-sector financial support and has generated a number of historical analyses that could not possibly have seen the light of day without the help of computers. But today, the new historical information created by demographers and historians serves other purposes precisely because it is available in electronic format and has been documented in accordance with explicit rules. Because it is available on a server, researchers can access it from wherever they work. We can see the most striking example of the use of this new historical data in the genetics and epidemiology work that has been carried out on Quebec families using the BALSAC database. BALSAC has been designed, built and validated over the last 30 years by interdisciplinary teams of researchers in the humanities and social sciences working together

within a framework that has now become the Institut interuniversitaire de recherche sur les populations (IREP).

Many researchers in the social sciences—historians, geographers, demographers, sociologists, anthropologists and jurists—and in the health sciences—geneticists and epidemiologists—have been drawn to the wealth of data in the BALSAC database. In most cases, they have been working in interdisciplinary teams, sharing their problems, methods and conclusions. Some researchers have linked the BALSAC data to their own nominal data, and as a result they have created substantial multi-source databases and have been able to conduct innovative research. In the social sciences and humanities, the research has covered such themes as the history and development of fertility, population migrations, social mobility, social cleavages, occupation of space, transformation of the rural world, literacy, labour history and urban history.

IREP has done an excellent job of making the BALSAC database available to the Quebec research community, but there is no established facility to ensure its long-term preservation, nor to advertise its existence across Canada. As important, there is no central depository for the numerous spin-off databases that have built upon the original work.

### **4.3 Contributing to Education and Research Training**

University professors on a daily basis are faced with students who are interested in conducting all kinds of research projects using data. For a graduate or undergraduate research essay on environmental policy in Canada, the ability to quickly locate relevant data sets and identify questions of interest permits the student to devote most of his or her time learning how to analyse data. Instead, in the present state of affairs, too often students are left with too little time to analyse data because of the time-consuming and frustrating process of figuring out whether there are actually any data out there, and, if there are, how one can access them.

Several respondents observed that students in the social sciences spend less time analyzing data than they should, and are at a distinct disadvantage when they pursue graduate education or employment when compared to American students who have ready access to the ICPSR. Many social science researchers contend that the ability of our graduate students to analyse data ranks far below that of their American counterparts. This stems, in part, from an uncoordinated effort to make data easily accessible. It could be rectified with a national data archiving function.

### **4.4 Building Capacity for Data Archiving**

In this country we lack expertise in the management of data and the preparation and preservation of metadata, the coding system used to describe and locate data sets. A national data archiving function could provide the training required to address our dearth of skills in the management of data files. The potential benefits would extend far beyond the research community, since all sectors of society, private and public, are facing serious data management and preservation challenges.

As well, at present there is little need or motivation for data producers to provide well-documented data files as there is no national repository for their work. A national data archiving function would provide training for scholars and other data producers in the creation of metadata and the management of data sets. These skills would serve the country well in the era of the new economy.

#### **4.5 Preservation and Access for Electronic Text Resources**

Written works like books, journals and manuscripts remain the primary means by which we transmit, study and store for future use scholarly work in the humanities and certain social sciences. Philosophers, historians, literary critics, art historians, political scientists and others in the social sciences and humanities use primary sources that themselves are texts and produce new works of research that are also texts. An increasing number of these texts are generated on a computer and are therefore originally in electronic form. Further, a significant number of Canadian scholars have created “tagged” or “marked” text research resources that can only be studied in electronic form with the appropriate tools. There is now a critical mass of research resources available as electronic texts, and in some cases, only as electronic texts. It is safe to say that a significant number of researchers now need access to well maintained electronic text services in order to conduct research and that, in the near future, the majority will require such access as computing methods and text services become routine in most disciplines.

The National Library of Canada plays a central role in the preservation of published texts, including those in electronic form, but it is not an archive and does not have the mandate or capability to deal with many of the complex technical issues involved in the handling of today’s textual databases. One of the greatest challenges here is the management and preservation of the often cutting-edge software tools developed for textual analysis. These tools come both separate and embedded within electronic manuscripts. Specific expertise is necessary to ensure compatibility with evolving hardware systems, the maintenance of links with other text databases, and in providing advice to researchers on how the tools can best be employed.

#### **4.6 Non-Reproducible Data**

Statistical survey data offer snapshots of a point in time that researchers can use in the future as benchmarks or reference points. Such data sets, by their very nature, cannot be replicated. If not properly preserved, they are lost for future research. A national data archiving function would ensure that data gathered to track our progress over a number of social indicators are available for future assessment and review. Such data would permit us to objectively gauge the ramifications of policies and would provide insight into those areas where we could bring about improvements. In this way, we could base social programs on evidence, which would allow us to make better use of our most important national treasure—our human resources.

Large-scale statistical survey data sets, such as the Canadian National Census, can be used for multiple levels of analysis, far beyond what Statistics Canada considers important or is capable of undertaking. Proper preservation of, and ready access to, such data sets would allow researchers to bring fresh questions to the analysis of the data as social and economic conditions change over time. At present it is difficult to examine the impact of social change. Most surveys give only a ‘snapshot’ of current conditions. Longitudinal surveys are seldom conducted, in part because they are prohibitively expensive. By allowing the researcher to make use of various data sets collected on similar topics, a national data archiving function would make diachronic research possible without incurring the costs and time involved in many years of data collection. In addition, Canada would have a vehicle with which to track the history of its current issues. This would fill a much-needed gap in both research and higher-education arenas.

#### **4.7 Building Models for the Public and Private Sectors**

Appropriate investment in the preservation of data created by or for researchers will not *only* keep valuable research resources accessible. It will *also* provide a model for other domains—from the health sciences to the insurance sector—that have significant investments in research.

Archiving data is in itself an area that requires extensive research. There are a great many aspects of data archiving that we simply do not know how to do, the most important of which is effective long-term preservation, both in terms of the technical aspects and the conceptual aspects, such as acquisition and retention protocols. Organizations around the world, both public and private, are struggling to preserve and manage the enormous volumes of data that are created on a daily basis. In addressing these needs, a national research data archive would be in the unique position of being able to draw on the research talents of computer scientists, engineers, information, library and archival scientists, management specialists, legal scholars, ethicists and the entire well-developed infrastructure of Canadian universities.

#### **4.8 Accessibility for Researchers to Expensive Resources**

Accessibility is a key issue for researchers. Conditions at the moment are far from ideal. Many spatial data sets, for example, required for research are available only on a cost-recovery basis. The high costs of such data sets and remote sensed images place them beyond the reach of most researchers and even most libraries because libraries, too, must often pay for access. There are instances where Canadian institutions and researchers have to obtain data on Canada from federal agencies in the United States because they cannot afford to buy the data from the original Canadian source. A study is currently underway suggest that the financial cost of marketing the data may be very close to the financial returns received. Researchers and industry alike have expressed serious doubts about the utility of such policies. Such data should be publicly accessible through a national data archive service. If it is not, Canada may well end up with a National Spatial Data Infrastructure which is all skeleton and no flesh.

*The principle of equal access should be followed in whatever scheme is recommended for data archives in Canada. The academic community has fought long and hard to break down the barriers of cost which produced such startling inequities in the research world and, though an arrangement has been reached with Statistics Canada for a graded level of equalized access, many of our researchers are still barred from access to files due to excessive fees.*

Questionnaire Respondent

### **5.0 Conclusion**

As one of the Working Group members aptly put it, an unprecedented firestorm is now incinerating Canada's digital research wealth. Although this may seem an overstatement, it is a deep-seated concern shared by many archivists, librarians and researchers around the world. Information in digital form is both extremely fragile and capable of being collected in huge quantities. Today, we are only beginning to understand how to effectively preserve, manage and make accessible this information. Although there are no easy short-cuts for dealing with such issues as media obsolescence, copyright, confidentiality, the creation of national and international standards, and the limitations of the current research culture, avoiding or ignoring them will prove costly in the long run.

Despite our massive investments in information technology, there is a critical and cumulative weakness in our information and knowledge infrastructure. The fact is, we are losing valuable Canadian knowledge resources because there is no national agency mandated to provide the necessary stewardship. As a result, we are needlessly wasting public money.

The value of Canadian research data lies in its present and future scientific worth, in the public investment expended to create it, and in its important contribution to the overall record of our society. Without the intentional collection, systematic preservation, intellectual organization, and purposeful management of our research data, we will lose this part of our heritage and undermine our future research capacity.

## **6.0 Recommendation**

Having completed the needs assessment, the members of both the Working Group and the Resource Group unanimously recommend the continuation of the National Research Data Archive Consultation into Phase Two, where the members will investigate the best possible arrangement for a national research data archiving system and propose a model for the effective stewardship of research data in Canada.

# Appendices

## Appendix 1

### A Survey of SSHRC-funded Researchers

A twenty-percent stratified sample<sup>3</sup> of all researchers receiving a grant from the SSHRC between 1998 and 2000 was drawn. Each person in this sample was sent a questionnaire asking about her or his experiences in creating and archiving research data or using research data created by others. Researchers were also asked about their opinion of fundamental issues that underlie archiving data and were asked to assess the importance of establishing national services for the preservation of research data.

A total of 116 responses were received,<sup>4</sup> which amounts to a 26 per cent response rate. Efforts were made to ensure a balanced response rate between anglophone and francophone researchers. Twenty-two per cent of all project titles funded between 1998 and 2000 are in French. The initial return-rate of the French version of the questionnaire was just ten per cent. After conducting a follow-up mailing of researchers with French project titles, this percentage was increased to 22 per cent.

To gain some introductory insight into those who returned a questionnaire, 64 per cent of the respondents completed their highest postsecondary degree between 1980 and 2000. This distribution may be more representative of the highest degree attainment of recent SSHRC grant recipients than of the wider population of researchers. Nevertheless, the distribution of highest-degree attainment does cover a forty-year span. The breakdown by decade was 34 per cent in the 1990's, 30 per cent in the 1980's, 28 per cent in the 1970's and nine per cent earlier than 1970 (see Table 4).

#### How Large of an Issue Is Data Creation and Data Archiving?

The findings from this survey help answer the question about how large of an issue data creation and archiving is for SSHRC-funded research. First, the number of researchers producing data from SSHRC-funded projects is substantial. Secondly, the data from very few of these projects are being archived. Furthermore, researchers who say that they have archived data often give sources that are not authentic archival services.

The questionnaire began with a set of questions to help determine the number of projects in which data have been created. First, a definition of research data and examples were presented.<sup>5</sup> Respondents were then asked if they had created data files or databases in their most recent SSHRC-funded research project that corresponded to the definition and examples that were given. Fifty-five per cent of researchers (see Table 1) reported that they had created data files or databases as part of their most recent project. Of this group, only seven percent said that they had deposited any of these data with an archive. Of the 93 per cent who have not deposited data, 18 per cent indicated that they plan to deposit data with an archive in the future. Researchers were next asked if they had created data in past research projects. Fifty-nine per cent stated to have created data in the past and 18 per cent of these said that they had archived

---

<sup>3</sup> See Appendix 1 for a description of the sampling method employed.

<sup>4</sup> This total of returns was as of April 6, 2001.

<sup>5</sup> "By research data, we are referring to digital information that has been structured by methodology for the purpose of producing new knowledge. Two examples of research data from among a wide variety of digital research data are files containing numerically coded information from questionnaires and files with text that have been coded with tags representing some form of structure."

data from this research. Twenty-two per cent of the respondents reported that they have plans to archive data from past projects.

These results provide ample information to estimate how large the issue of data creation and archiving in regards to SSHRC-funded research is today. By calculating a 95 per cent confidence interval for each of the key data questions mentioned above (see Table 2), an estimate of the range within which the number of researchers producing and archiving data is derived.

Conservative estimates of the confidence intervals were calculated using the number of actual returns rather than the original sample size. For example, the confidence interval for the percentage of researchers who created data in their current SSHRC project (Q1A) is estimated to be plus or minus 4.5 per cent using the 115 researchers who responded as the sample size. The range of this interval goes from a low of 50.5 per cent to a high of 59.5 per cent. Expressed in terms of the rate per thousand SSHRC-funded researchers, one would expect between 505 and 595 out of every 1,000 researchers to produce data files and/or databases in their projects.

<b>Table 1</b>				
<b>Distribution of Researchers Creating, Archiving or Intending to Archive Data</b>				
Survey Question	Percent 'Yes'	N's		
		Response	Valid N	Missing
Q1A. Create data/databases in current SSHRC project?	55%	63	115	1
Q1B. Ever deposit these data? (if no, go to Q1F)	7%	4	61	2
Q1F. Ever plan to deposit these deposit data?	18%	10	55	2
Q2A. Create data/databases in past research projects?	59%	68	116	0
Q2B. Ever deposit these data?	18%	12	67	1
Q2D. Ever plan to deposit these deposit data?	22%	15	67	0

If the original sample size is used to calculate the confidence interval, that is, instead of the valid response rate, the confidence interval shrinks to plus or minus 2.3 per cent or approximately half of the conservative estimates in Table 2. In this instance, one would expect between 530 and 570 researchers per 1,000 researchers producing data. Rather than become distracted by the precision of the interval width, the overall significance of these findings is that data are being created in a large proportion of SSHRC-funded research. The most conservative estimate identifies between 50 and 60 per cent of researchers engaged in data creation. Applying this estimate to the total number of SSHRC-funded researchers between 1998 and 2000, the number of researchers producing data is between 1,140 and 1,360.

<b>Table 2</b>						
<b>95 Percent Confidence Interval Estimates for Main Data Questions</b>						
	N	p	q	CI	CI low	CI high
Q1A Yes – created data	115	0.55	0.45	0.045	50.5	59.5
Q1B Yes – deposited data	61	0.07	0.93	0.016	5.4	8.6
Q1F Yes – plan to deposit	55	0.18	0.82	0.039	14.1	21.9
Q2A Yes – created data	116	0.59	0.41	0.044	54.6	63.4
Q2B Yes – deposited data	67	0.18	0.82	0.035	14.5	21.5
Q2D Yes – plan to deposit	67	0.22	0.78	0.041	17.9	26.1

The fact that many researchers are producing data in their research is offset by the alarming discovery that hardly any of the data from these projects have been archived or will be archived. Of the 55 per cent who say they have created data, only seven percent report having archived data from their project (see Table 1). Furthermore, a follow-up question (Q1C), which was asked of the seven percent, revealed that not all of them actually deposited data with an archive. One researcher deposited her or his data with a university data library service and one sent a copy to the National Library. One placed data on the Web, which in itself does not constitute an archival deposit, and the final researcher said that electronic access would be provided if a CFI initiative is funded. Using the confidence interval estimates in Table 2, an estimate of the number of researchers who archive data is between 50 and 80 per 1,000 researchers. A troublesome finding is that as many as half of these may not have deposited data with an actual archival service.

Compared to those who say they have archived data (Q1C), a little over twice as many researchers have good intentions of archiving the data from their projects (Q1F). As shown in Table 1, 18 per cent indicated that they expect to deposit data with an archive. These researchers were subsequently asked with which archive they would deposit data (Q1G). Two of ten identified university data services, two did not know where to deposit data, one mistakenly identified a Statistics Canada Research Data Centre, one provided the name of a non-existent archive, one said that she or he would start an archive if a national service is not established, one said the ICPSR in the United States, one named a university archive, and one said a Web site. Of the ten responses, less than half identified an actual archival service.

Researchers were then asked if they had ever created data in at least one previous project (Q2A). Fifty-nine per cent said that they had. The 95 per cent confidence interval for this estimate is 55 to 63 per cent of researchers. The percentage of researchers claiming to have deposited data from a previous project (18 per cent – Q2B) is higher than the percent reporting deposit on their current project (seven percent – Q1B). Intentions to deposit are slightly higher on data from previous projects (22 per cent in Q2D compared to 18 per cent in Q1F), although the confidence intervals for these two questions overlap suggesting that the differences may not be significant.

Of the 12 respondents who reported having deposited data from previous research<sup>6</sup>, two had deposited data with the U.K. Data Archive at Essex University. Four respondents said that they had given copies to a university data library. Thus, half of these respondents identified actual data services. Two respondents listed the name of a university as the place of deposit. One respondent mentioned a university archive but went on to note that this became a problem when the university's computer system changed. As far as she or he knew, no provision had been taken to read data in file formats that were no longer used. One respondent listed her or his centre's Web site, while another said her or his research laboratory. Finally, one respondent said that she or he deposited a paper copy of the data with an archive. These latter three cases do not constitute sound practices in archiving data.

The confidence interval estimate for the number of researchers who have archived data from past research is between 145 and 215 per 1,000 researchers. Based on the places of deposit just reported, this estimate of researchers likely includes a number who did deposit data with an archive service. There also is the likelihood that some did not deposit data with an actual archival service. The encouraging news that some researchers have deposited data, however, is offset by the large gap that remains between the number who created data in past research and the number who deposited data with an archival service. In terms of percentages, between 80 and 85 per cent of researchers who created data in past research have not made an archival deposit.

---

<sup>6</sup> These responses are taken from Q2C from Researcher's questionnaire.

Twenty-two per cent of respondents reported that they intend to deposit data arising out of past research with an archive (see Table 1 - Q2D). Forty per cent of this group are among the researchers who said they had deposited data from past research, that is, those who answered 'yes' to Q2B. Thus, a substantial percentage of those with good intentions also have experience with depositing data. These researchers were subsequently asked with which archive they intend to deposit data from their past research (Q2E). Since some of these respondents had experience in depositing data, some answers were similar to the ones given in Q2C. Three of fifteen respondents said they would deposit data with a university data library. Two said that they would rely on university archives, although one of these qualified her or his answer by saying, "*unless another alternative becomes available.*" Two researchers mentioned non-Canadian sites. One listed the Murray Research Center for the Study of Lives, Radcliffe College at Harvard University, while the other mentioned the ICPSR. Three respondents said that they were undecided or simply did not know where to deposit data. Two replied that they would rely on Web sites. Two wanted to find an appropriate archive and one said that until there is a Canadian archive, "*we will maintain our own archive, as we are not willing to archive it in a non-Canadian archive.*" Finally, one respondent listed the National Library or Canadian Museum of Civilization. These last eight responses indicate that a little over half of those answering this question were either searching for a data archive or did not understand what archiving data entails.

### What Kind of Support Exists Among Researchers for National Data Archiving Services?

The majority (60 per cent) of researchers see the importance of establishing national services in Canada to preserve research data. However, a number of researchers seem unsure of or undecided on issues that underlie the preservation of research data. These issues simply are not being discussed in Canada's academy either in the training of graduate students or in the dialogue about professional ethics. Furthermore, evidence suggests that many researchers are unaware and/or uninformed about what archiving research data entails. This low rate of awareness and knowledge likely stems from the absence of a national service dedicated to the issues of research data preservation. Results show that different research experiences seem to shape the attitudes of researchers about sharing and archiving data. For example, researchers who have analyzed data collected by other researchers are more supportive of attitudes endorsing data preservation.

A profile of researchers was determined on the basis of their work with data. Categorizing researchers in this fashion helped identify differences among the attitudes that researchers hold about data archiving. There are researchers who are actively producing and using data and others who clearly are non-data users. As shown in Figure 1, 76 per cent of respondents use or have used data in their research. Twenty-five per cent of all respondents answered 'yes' to the three questions—Q1A, Q2A and Q3A—about current and past creation of data and the use of data from others. Another 29 per cent were current data producers and either had created data in the past or used data from others. Twenty-two per cent did not create data in their current project but had either created data in the past and/or used data created by others. Finally, 24 per cent answered 'no' to all three questions.

---

<sup>7</sup> Using these four categories, a breakdown by type of data user by discipline of study, unfortunately, is not possible. The field of study for each respondent was not asked in this survey. Consequently, the percentage of non-data users who are in the humanities versus the social sciences cannot be determined. A few respondents did write on their questionnaires that they felt that the questions were less relevant to them because they are in the humanities instead of the social sciences. However, one cannot conclude that all of the non-data users are from disciplines in the humanities.

The year of highest degree attainment differs across these four types of data users (see the column percentages in Table 4). For the non-data group, roughly a third of the respondents attained their highest degree before 1980. Another third completed their degree during the 1980's and the final third received their degree in the 1990's. The decade in which non-data users were trained does not differ. For the middle two categories of data users, their breakdown is roughly 40 per cent prior to 1980, 20 per cent in the 1980's and 40 percent in the 1990's. The middle two groups were trained predominantly either prior to or after the 1980's. Finally, the breakdown for the most active data group is 31 per cent prior to 1980 (no one obtained their highest degree prior to 1970), 48 per cent in the 1980's and 21 per cent in the 1990's. Almost half of this group was trained in the 1980's.

<b>Table 3</b>				
<b>Use of Data Collect by Others (Q3A) and Sharing Data with Others (Q4)</b>				
Survey Question	Per cent 'Yes'	N's		
		Response	Valid N	Missing
Q3A. Used or tried to use data created by other researchers?	39%	45	116	0
Q3B. Denied access to data produced by others?	7%	3	42	0
Q4. Have you denied others access to data you created?	8%	9	114	2

Interesting cohort differences are shown by the decade in which respondents received their highest degree. Forty per cent of the 1980's degree-cohort answered 'yes' to all three data-use questions (see the row percentages in Table 4). The 1970's cohort has the largest percentage of researchers who are or have been data users with 84 per cent. The low percentage of the 1990's cohort who answered 'yes' to all three data-use questions (15 per cent) might be because these researchers are still becoming established and consequently, have not had the opportunity to produce data in the past. Finally, those who attained their highest degree prior to 1970 have the highest percentage of non-data users, although the size of this group is only 10.

To summarize, the pattern of data use by researchers who attained their highest degree over the past three decades shows a significant proportion producing and using data. Eighty-four per cent of 1970's cohort fall into this group. Seventy-four per cent of the 1980's and 1990's cohort similarly are active data users. Unfortunately, this survey does not permit comparisons of data usage between the humanities and the social sciences since a question about field of study was not asked.

The practice of secondary analysis has increased in recent years as more data have been made available through international data archives, such as the ICPSR, or through government data subscription services, such as Statistics Canada's DLI. Experiences in secondary analysis introduce researchers to the value of data collected by others. To assess the extent to which secondary analysis has been experienced, researchers were asked if they had ever analyzed data collected by other researchers. Thirty-nine percent of respondents said that they have used or tried to use data that were produced by other researchers (Q3A). Of this group, seven per cent reported that they had been denied access to data produced by others (Q3B). Eight per cent of all respondents said that they had denied other researchers access to data that they had produced (Q4). Use of data prepared by others seems to be an important experience when examining the attitudes researchers hold about sharing and archiving data. This will be described further below.

Eleven items (Q5A through Q5K) in the survey were used to gauge the attitude of researchers toward principles underlying data archiving. These items touch upon the legitimacy of secondary analysis as a research method, on the value of data as a by-product of research, on the issues of data ownership and data sharing, on research council funding to prepare data for sharing, and on the impact that ethics review boards have on data sharing (see List 1). Because the wording of five items (5a, 5d, 5e, 5h, 5j) does not support the principles of data sharing or archiving, the response categories for these items were reflected to correspond with the direction that supports these principles.<sup>8</sup>

Figure 2 shows the combined percentage of respondents 'agreeing' or 'strongly agreeing' on each item. The items in this Figure have been arranged in decreasing order of support. Eighty-one per cent agree that data should be a valued by-product of research, while only 21 per cent agree that data do NOT belong to the researcher as her or his intellectual property. This decreasing order of agreement represents an increasing difficulty in support of sharing and preserving research data. Six steps of item-difficulty can be seen in Figure 2. The first two items, data as a valued by-product and secondary analysis as a valid research method, are accepted by 81 and 78 per cent of the respondents, respectively. The second step consists of the items about research councils covering the costs to prepare data for sharing and about researchers serving as trustees of data that cannot be easily reproduced from respondents. Seventy-one and 68 per cent endorsed these items, respectively. The third step is made up of the item stating that data should be shared if it has been properly anonymized (64 per cent) and the item that ethics review boards need to be educated about the need to preserve data (62 per cent). A slightly bigger step occurs with the next two items. Fifty percent agree that spending resources to prepare the data from their research would not be a waste and 48 per cent agree with the statement that archiving data should be an integral part of conducting research. An even larger drop occurs with the fifth step. Twenty-eight per cent disagree that ethics review boards make it impossible to share confidential data on human subjects, while 27 per cent disagree that data should only be shared if the principal investigator decides to share it. As mentioned above, the smallest percentage of agreement (21 per cent) is that data do NOT belong to the principal investigator as her or his intellectual property.

A scale was constructed based on the total number of items in which each respondent agreed with the principles of data archiving (see Figure 4). Thus, a score of zero means that the respondent does not support any of the items endorsing data archiving, whereas a score of 11 represents someone who supported all of the items. As shown in Figure 4, 17 per cent are low supporters of data archiving (those with scores from zero to three), while 24 per cent are high supporters (those with scores from eight to 11). Fifty-nine per cent are in the middle. The correlation between this scale and the question asking how important it is for Canada to establish national services for the preservation of research data (Q6) is 0.498. This fairly high correlation corroborates the interpretation that this scale measures support for data archiving.

Figure 6 shows box-plots of this scale for each of the four data user groups described above. Clearly, the researchers in the group who answered 'yes' to all three data-use questions are the strongest supporters of data archiving. The median for this group is 7.5 while the next largest median (6.0) is for the group that did not create data in their current project but had either created data in the past and/or used data created by others. As mentioned earlier, those researchers who have not used data created by others but who have created data in their current SSHRC-funded project as well as past projects tend to be the least supportive of data archiving. For example, this group has a median of 5.0 on the data archiving scale. Furthermore, 89 per cent of this group agrees or strongly agrees that data belong to the principal investigator as her or his intellectual property.

---

<sup>8</sup> "Strongly agree" and "agree" are the responses supportive of these principles.

A number of researchers seem unsure about the items in the data archiving scale. Figure 3 shows the percent of respondents who are not sure or undecided on each item. Over 25 per cent of the respondents were unsure on six items (5c, 5f, 5g, 5h, 5j, 5k) and two items were just below 25 per cent (5d and 5i). The two items about ethics boards had the highest percentage of undecided responses with 37 and 35 per cent. This indicates that many researchers simply have not thought about or are uninformed about the issues represented by these items.

Finally, Figure 5 shows the distribution of respondents for the question asking how important it is for Canada to establish national services for the preservation of research data (Q6). Sixty percent agree that such services are 'important' or 'very important' for Canada, while only 12 per cent say they are 'unimportant' or 'not important at all'. Twenty-eight per cent are unsure, which is consistent with the percent that are unsure or undecided in several of the attitude items discussed above.

Support for national services differs greatly among the type of data users. Ninety percent of those who answered "yes" to all three data-use questions agree to the importance of national services to preserve research data. This is followed by 72 per cent of those who did not create data in their current project but had either created data in the past and/or used data created by others. The percentage of support by the remaining two data-user groups drops substantially with 44 per cent of non-data users agreeing and only 38 per cent of those who are current data producers and either had created data in the past or used data from others. These latter two groups have the largest percentage of unsure or undecided with 44 per cent each.

The data-user group consisting of those who are current data producers and either had created data in the past or used data from others also most strongly supports the statement that data belong to the principal investigator as her or his intellectual property (81 per cent). Sixty-five per cent of this group also endorses the statement that data should only be shared if the principal investigator decides to share it. These are researchers who seem to have a strong proprietary outlook about the data that they collected through their SSHR-funded project.

The decade in which respondents received their highest degree also shows differences between support for Q6 and degree-cohorts. The least supportive degree-cohort is the group prior to 1970, although this group is small in size consisting of only 10 respondents. Thirty percent of this cohort supports national services to preserve research data, while 50 per cent are unsure. The next least supportive degree-cohort is the 1990's group, in which 56 per cent agree to the importance of national services. Twenty-eight percent of this cohort are unsure. Similarly, 29 per cent of the 1980's degree-cohort are unsure but 60 per cent are supportive. Finally, the 1970's cohort offers the most support with 74 per cent supportive of national services. Only 19 per cent of this cohort are unsure. One-third of the 1990's cohort is composed of data-users who are the least supportive of Q6, which likely explains this cohort's lower level of support. It should be noted, however, that 56 per cent of this cohort nevertheless are still supportive.

Year of Highest Degree Attainment by Type of Data User

			TYpe of Data User				Total
			Non-Data Users	Produced Data in Past Research and/or Used Data from Others	Current Data Producer and Either Produced Data in Past or Used Data from Others	Current Data Producer, Past Data Producer and Used Data from Others	
Year	Prior to 1970	Count	4	1	5		10
		Row %	40.0%	10.0%	50.0%		100.0%
		Col %	14.3%	4.0%	14.7%		8.6%
1970 - 1979	Count	Count	5	9	9	9	32
		Row %	15.6%	28.1%	28.1%	28.1%	100.0%
		Col %	17.9%	36.0%	26.5%	31.0%	27.6%
1980 - 1989	Count	Count	9	5	7	14	35
		Row %	25.7%	14.3%	20.0%	40.0%	100.0%
		Col %	32.1%	20.0%	20.6%	48.3%	30.2%
1990 - 1999	Count	Count	10	10	13	6	39
		Row %	25.6%	25.6%	33.3%	15.4%	100.0%
		Col %	35.7%	40.0%	38.2%	20.7%	33.6%
Total	Count	Count	28	25	34	29	116
		Row %	24.1%	21.6%	29.3%	25.0%	100.0%
		Col %	100.0%	100.0%	100.0%	100.0%	100.0%

Table 4

Researchers Experiences in Creating Data Files or Databases or Using Data Files or Databases Created by Others

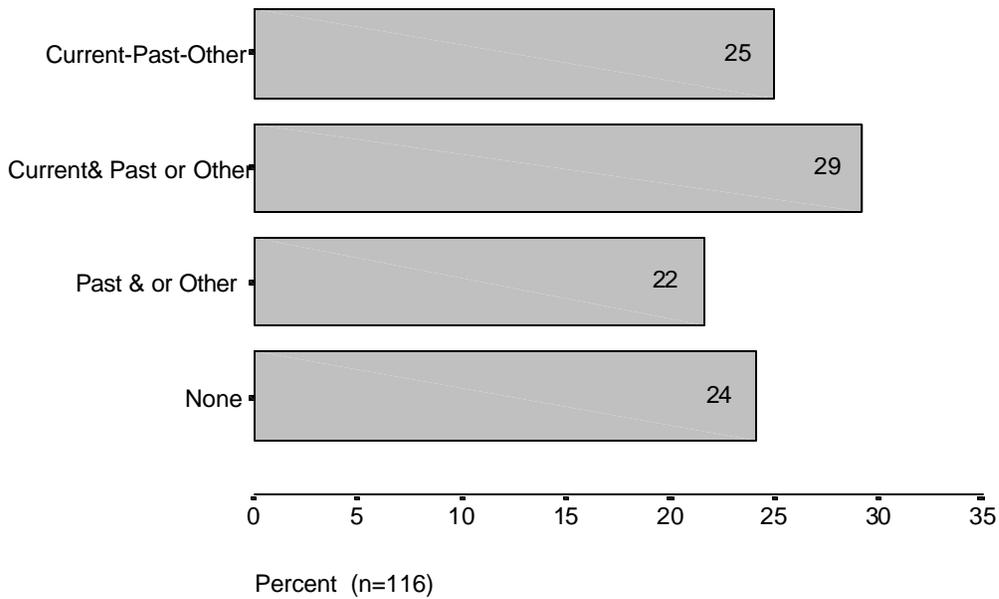


Figure 1

## List 1

### Attitudinal Items Underlying Support for Data Archiving

- 5a. Secondary data analysis is not a valid research method.
- 5b. Data should be considered a valued by-product of research.
- 5c. Data should be shared with other researchers, assuming it has been appropriately anonymized.
- 5d. Data belong to the principal investigator as her or his intellectual property.
- 5e. Data should only be shared if the principal investigator decides to share it.
- 5f. Archiving data should be an integral part of conducting research.
- 5g. Researchers who obtain information that cannot be easily reproduced from respondents are, to a degree, trustees of the data.
- 5h. Spending resources to prepare the data from my research so that other researchers can use it would be a waste.
- 5i. Research councils should include funds to cover the costs of preparing data for sharing.
- 5j. Ethics review boards make it impossible to share confidential data on human subjects.
- 5k. Ethics review boards need to be educated about the need to preserve data.

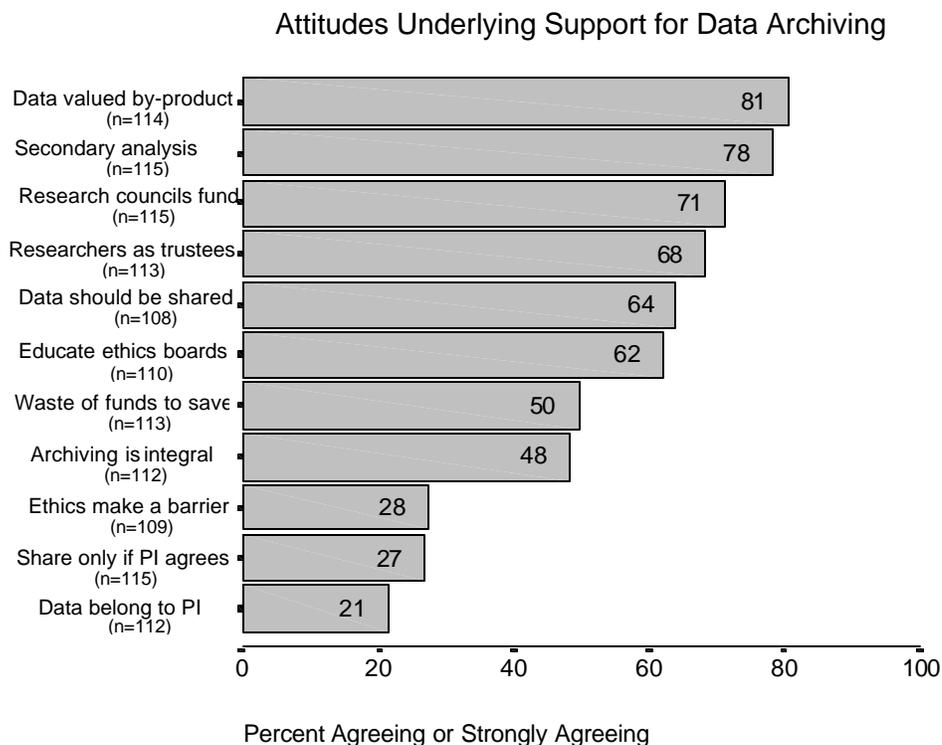


Figure 2

### Attitudes Underlying Support for Data Archiving

#### Percent Who Were Unsure

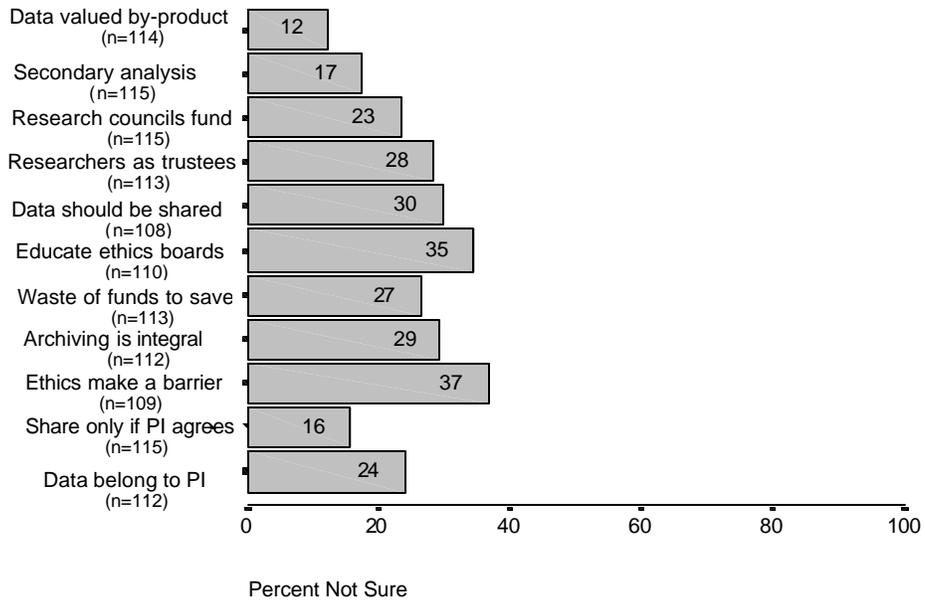


Figure 3

#### Number of Items Supported that Underlie the Principles of Data Archiving

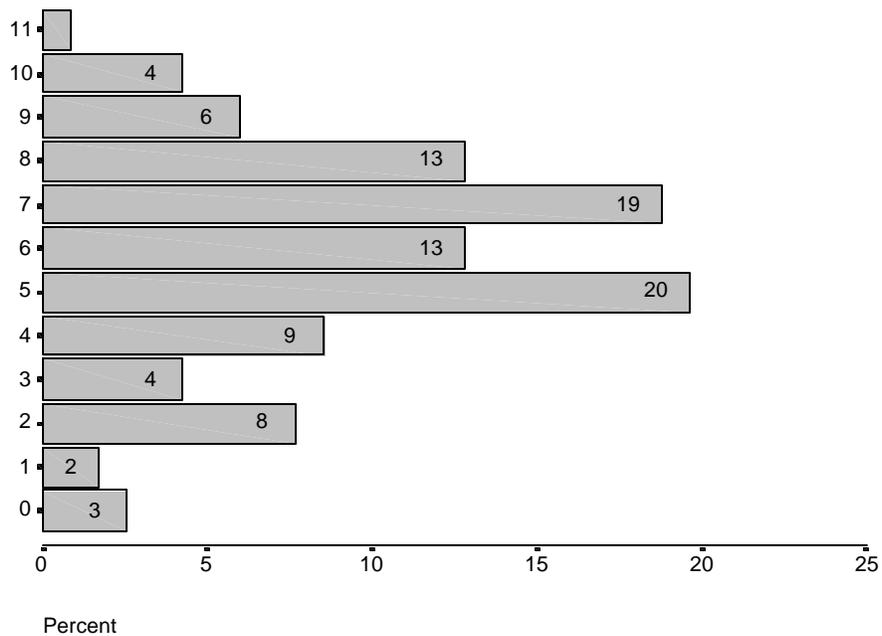


Figure 4

The Importance for Canada to Establish National Services  
to Preserve Research Data

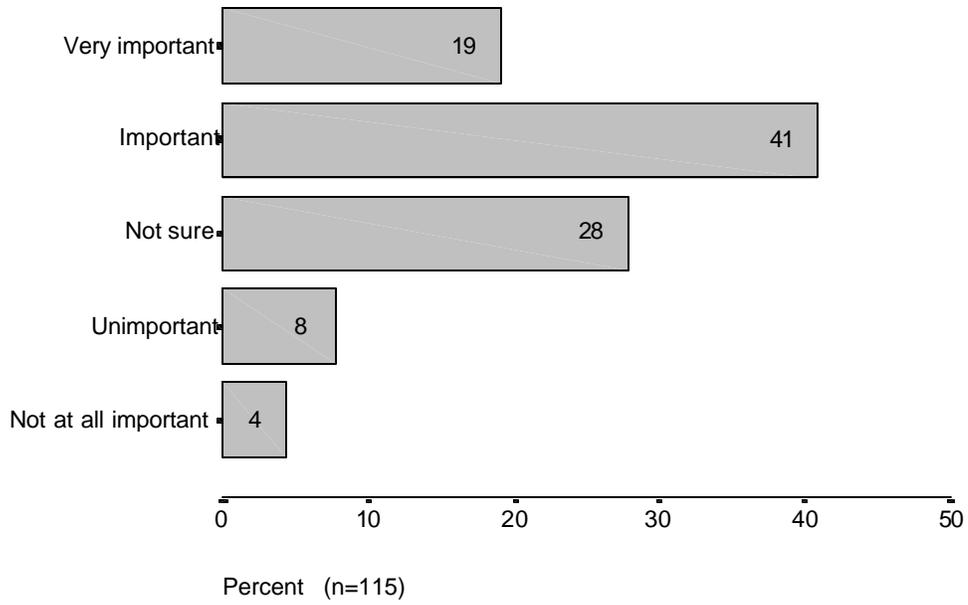


Figure 5

Data Archive Attitude Scale by Type of Data

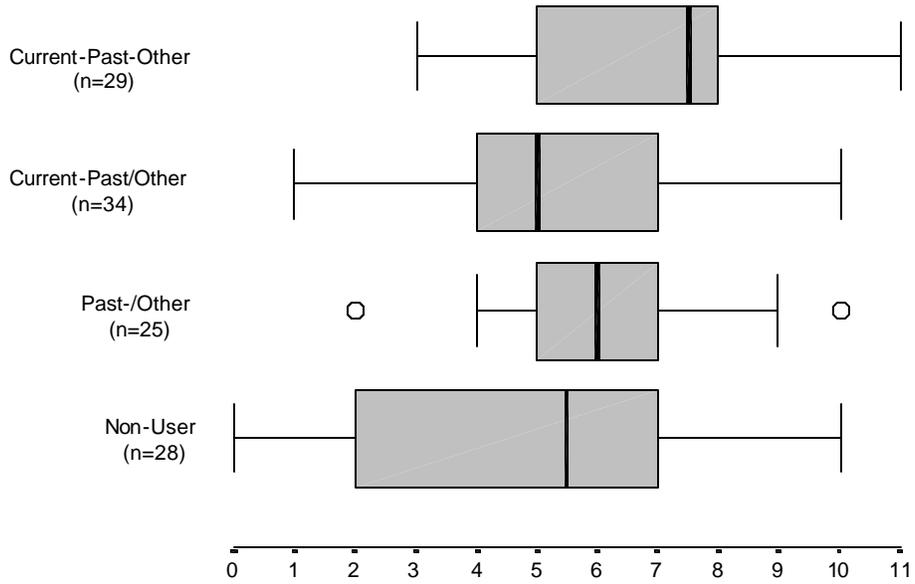


Figure 6

<b>Importance of National Services to Preserve Research Data (Q6) by Type of Data User and Year of Highest Degree</b>									
The Importance for Canada to Establish National Services to Preserve Research Data (Q6)*		Type of Data User				Year of Highest Degree			
		Non-data Users	Produced Data in Past Research and/or Used Data from Others	Current Data Producer and Either Produced Data in the Past or Used Data from Others	Current Data Producer, Past Data Producer and Used Data from Others	Prior to 1970	1970-79	1980-89	1990-99
Disagree	N	3	4	6	1	2	2	4	6
	Col %	11.1%	16.0%	17.6%	3.4%	20.0%	6.5%	11.4%	15.4%
Unsure	N	12	3	15	2	5	6	10	11
	Col %	44.4%	12.0%	44.1%	6.9%	50.0%	19.4%	28.6%	28.2%
Agree	N	12	18	13	26	3	23	21	22
	Col %	44.4%	72.0%	38.2%	89.7%	30.0%	74.2%	60.0%	56.4%

\* The responses for "Strongly disagree" were collapsed into "Disagree" and "Strongly agree" into "Agree".

Table 5

Below are quotations from researchers to an open-ended question asking for comments about the issues in this survey.

- A. Tricky issue—many complicated aspects and ethical concerns, all of which need to be fully deliberated before a decision to “share data” can be reached.
- B. These are on the whole questions I haven't needed to consider. The text databases I've worked on have been designed for my own use, and I don't think they'd be—in their raw form—useful to others, although I hope the conclusions I draw from them will be useful. If I worked in the social sciences, my perspective would no doubt be quite different.
- C. Archiving data in digital form should be encouraged rather than mandatory (e.g., thru supplemental funds earmarked for this purpose & only available for it). Researchers might reasonably be required to make data available to other legitimate researchers upon request (subject to reasonable exemptions re: confidentiality etc.) but this isn't the same as depositing it in a Central Archive.
- I would be very concerned about SSHRC imposing a uniform digital technology (e.g., some particular program &/or format) as an Archival requirement. SSHRC's track record here, e.g., the web-site for grant applications in 1999, makes some of us worry! Technological pluralism should be the order of the day—I'm glad this is the approach you have taken with this questionnaire—our e-mail systems do seem to be incompatible, so I'm glad to have option of fax/snail-mail!
  - I do think it's an ethical obligation for researchers to share data gathered partly at public expense, once they have published it. I remember being upset with one prominent scholar who refused to provide even his coding protocol.
- D. The Council could begin with a voluntary program and some funding for data archival projects on an experimental basis, with review after x years.
- E. I do not work in an area where this is a large issue at this time.
- F. As the abundance of “not sure” responses suggest, these are questions/issues of which I don't have direct experience, a situation I would assume common to scholars in the humanities.
- G. Does not apply to my field
- H. I do not do this type of research
- I. I have often thought that data (not just databases but research notes on computer disk) should be conserved, and that inactive scholars should be alerted to the need to give diskettes to an archive—whether university or government. I should think that data produced by public funding should particularly be so conserved.
- J. As you may be able to tell, I have had some experience with a non-Canadian data archive that has not been positive. The procedures of that archive have changed in the interim (at least, in part, as a result of my negative experiences, along with those of some other researchers, largely Canadians). In that they might revert to their previous practice (their current practice is linked to their funding source, not their own policy statements), I am unwilling to take the risk. Their practice was to charge Canadian researchers to use their own and other Canadian data and assign a lesser priority of their access to the data than that of their own nationals. Therefore, I am convinced that Canada must either have its own data archive or, at a minimum, be able to enter into National agreements with non-Canadian data archives. Many countries do have such agreements with the archive in question. In brief, my

experience was that while researchers based in the country of the archive or in those countries that had national data archive agreements with the archive in question did not have to pay for access and had priority access to our Canadian data.

- K. - Publicly-funded, properly-anonymized data should be easily accessible to all researchers
- Preparing data for use by a broad community of researchers both enhances our potential for knowledge creation and makes data creation less expensive (the more users per data set, the less it costs...)
  - This preparation takes time & money and must therefore be funded
  - We need a permanent Data Archive to safely store data, spearhead efforts to enhance the secondary use of data, and, most importantly, to foster a research culture of data sharing.
- L. Publicly-funded research should require that the data generated, research instruments employed, designs used and sampling frameworks etc. be archived and made available for other researchers. This would be very important to activities such as fostering collaborations, longitudinal studies, replication studies, comparative studies, creation of 'normative' question designs in certain areas of enquiry, and secondary analyses. Transparency, accountability and responsibility would be encouraged by requiring the archiving and access to data. Further, consideration of such data should become a more central attribute of planning 'new' primary research—less re-inventing the wheel and more imaginative and creative work might result. Thoughts—for what they are worth.
- M. As I stated related to some of the questions above, I think it is difficult to respond to many of these questions with a yes/no or agree/disagree as some are more complex and my response would differ based on different factors. However, overall I do feel that there is an important need to preserve research data—I just think the complexities of this issue and process need to be addressed.
- N. Because many of the issues and questions concerning the structure, ethics, and accessibility of data are discipline- or period-specific, attempting to write a single over-arching policy for all seems to me counter-productive. Issues of data preservation might be something for applicants to address and for assessors and committees to factor into their evaluations, without imposing a uniform and ultimately arbitrary policy. In my experience, structure and interpretation of data are in no sense independent of one another; more often than not the distinction made or implied here between data and interpretation is questionable.
- O. The two points I will raise are tangential to your concern with digitized material.
1. Much of my work is based on interviews with policy makers. In the past I have taken notes, but not made recordings. If I receive the grant which I have submitted to SSHRCC, I will record and transcribe my interviews for the project which I am undertaking. In this case it makes more sense to make these interviews available for colleagues. But it is of course much more expensive to work in this way.
  2. The way I keep "data" is in file boxes. Storage becomes a physical problem of renting space or having some library take over the material. This is a technologically less glamorous issue than what this study is dealing with but probably affects many more scholars and is of relevance to historical research in the future.
- P. **NOTE:** It was hard to answer some questions, those about rights of a researchers to data he or she generates for example. I answered in a way that reflects my views of the average situation but I can think of a host of exceptions, to do with confidentiality, the amount of creativity and effort it took to generate the data in question, etc.

- Q. The research in which I am presently involved would require community/research subject consent to archive data. Data will be accessible in the form of reports, etc. that will be part of a website. Research instruments will be shared. Research subjects will receive copies of their interview transcripts. In short, I think there are different ways of sharing data depending on the type of research being done.
- R. I wonder if this issue is of more importance to some areas of research than others. Archival data doesn't seem very important for research in my area but I could see how it could be important in other areas. Nonetheless, I think that the researcher who actually goes to the trouble of collecting the data must have some ownership of it.
- S. À mon avis les réponses à la question 5a. (L'analyse de données secondaires n'est pas une méthode de recherche fiable) et les questions suivantes demandent d'être nuancées. Je suis d'accord pour qu'un groupe de chercheurs puisse faire des analyses secondaires sur leur banque de données car ils/elles en connaissent les limites. Un groupe de chercheurs pourrait utiliser une banque de données d'un autre groupe de chercheurs seulement si la méthodologie de collecte des données était très simple, peu complexe.

Je ne crois pas que le conseil devrait dépenser beaucoup d'argent à créer une structure pour forcer la création de banque de données disponibles à tous et à chacun. L'analyse secondaire de données sera bonne seulement si les auteurs de la collecte des données participent à une seconde analyse. Le conseil pourrait aider financièrement la collaboration entre chercheurs pour une analyse secondaire de la banque de données.

- T. Il me semble qu'il est extrêmement difficile de traiter comme un bloc homogène l'ensemble des données pouvant être dérivées de la recherche subventionnée. J'admets que certaines données sensibles devraient être maintenues confidentielles. Mais cela n'est certainement pas le cas de l'ensemble des données. En règle générale, des résultats scientifiques obtenus avec des fonds publics devraient appartenir au domaine public.

## Appendix 2

### National Data Archive Consultation Working Group and Resource Group Members

#### Working Group

**Dr. John ApSimon**  
Special Assistant to the President  
Carleton University

**Prof. José Igartua**  
Département d'histoire  
*Université du Québec à Montréal*

**Prof. Gérard Boismenu**  
Département de science politique  
Université de Montréal

**Mr. Charles Humphrey**  
Data Librarian  
University of Alberta

**Prof. Ian Lancashire**  
Department of English  
University of Toronto

**Ms. Sue Bryant**  
Treasury Board Secretariat  
Senior Project Co-ordinator  
Public Key Infrastructure Secretariat

**Dr. Luciana Duranti**  
School of Library  
Archival and Information Studies  
University of British Columbia

**Dr. Michael Murphy**  
Director of the Rogers Communications Centre  
Ryerson University

**Prof. Matthew Mendelsohn**  
Department of Political Studies  
Queen's University

#### Resource Group

**Prof. Paul Bernard**  
Department of Sociology  
Université de Montréal

**Prof. Joseph Desloges**  
Department of Geography  
University of Toronto

**Prof. Fraser Taylor**  
Department of Geography  
Carleton University

**Prof. Geoffrey Rockwell**  
Department of Modern Languages  
McMaster University

**Ms. Wendy Watkins**  
Data Librarian  
Carleton University

**Mr. Michael Ridley**  
Chief Librarian  
University of Guelph

**Mr. Ernie Boyko**  
Director  
Library and Information Centre  
Statistics Canada

**Dr. Frits Pannekoek**  
Director, Information Resources  
University of Calgary

**Dr. Martin Brooks**  
Group Leader, Interactive Information  
Institute for Information Technology  
National Research Council

**Ms. Yvette Hackett**

Electronic Records Officer  
Government Archives and Records Disposition  
Division  
National Archives of Canada

**Mr. Douglas Hodges**

Information Technology Services  
National Library of Canada / National Archives  
of Canada

**Prof. Joseph R. Desloges**

Department of Geography and Program in  
Planning  
*University of Toronto*

**Dr. Timothy Jackson**

Associate Professor of New Media  
Ryerson University

**Ms. Wanda M. Noel**

Barrister & Solicitor

## Appendix 3

### National Data Archive Consultation List of Submissions

#### Letters from Deputy Ministers

**Peter Harrison**  
Deputy Minister  
Natural Resources Canada

**Shirley Serafini**  
Deputy Minister  
Indian and Northern Affairs Canada

**David Dodge**  
Deputy Minister  
Health Canada

**Alan Nymark**  
Deputy Minister  
Environment Canada

#### Stakeholder Submissions

1. **Canadian Association of Public Data Users (CAPDU)**
2. **Dr. A.W. Taylor**, Chair, Canadian Centre for Activity and Aging  
Faculties of Health Sciences, Medicine, and Dentistry  
University of Western Ontario
3. **Prof. John Wilson**, Department of Political Science  
University of Waterloo
4. **Dr. Tom Nesmith**, Associate Professor, Department of History  
Faculty of Arts, St. Paul's College  
University of Manitoba
5. **Mr. Doug Hodges**, Information Technology Services  
*The National Library of Canada*
6. **Dr. Michael Ornstein**, Director, Institute for Social Research  
*York University*
7. **Dr. Janice M. Morse**, Director, International Institute for Qualitative Methodology  
Faculty of Nursing, University of Alberta  
*Senior Scientist, Medical Research Council of Canada*
8. **Dr. Bryan Corbett**, President, Association of Canadian Archivists  
*AND*  
**Mr. Fred Farrell**, Chair, Canadian Council of Archives
9. **M. Robert Garon**, Conservateur  
*Archives nationales du Québec*

10. **M. Pierre Bordeleau**, Vice-recteur adjoint aux TIC et directeur général  
AND  
**M. Michel Lespérance**, Secrétaire général  
**Direction générale des technologies de l'information et de la communication,**  
**Université de Montréal**
11. **Ms. Miriam McTiernan**, Archivist of Ontario  
Management Board Secretariat, Archives of Ontario
12. **Ms. Elizabeth Krug**, Project Officer, Canada's Digital Collections  
Industry Canada
13. **Prof. Mary Jane Miller**, Department of Fine Arts  
Brock University
14. **Marc Lacasse**, Président  
*L'Association des archivistes du Québec*
15. **Douglas McLeod**, Director of Projects, NETERA Alliance  
Calgary, Alberta
16. **Gary Strike**, *Data Librarian, Data Library Services*  
University of Manitoba Libraries
17. **Dr. Rosemary Ommer**, Director, Calgary Institute for the Humanities  
The University of Calgary  
AND  
**Prof. Eric Sager**, Chair, Department of History  
University of Victoria

## Appendix 4

### Research Data Archiving Reports and Other Related Documents

- 1) *Preserving the Whole: A Two-Track Approach to Rescuing Social Science Data and Metadata*, Ann Green, JoAnn Dionne and Martin Dennis, The Digital Library Federation, Council on Library and Information Resources, Washington, DC, June 1999.
- 2) *Digital Electronic Archiving: The State of the Art and the State of the Practice*, Gail Hodge and Bonnie C. Carroll, Report sponsored by the International Council for Scientific and Technical Information, Information Policy Committee, and CENDI, Oak Ridge, TN, April 1999.
- 3) *Responsibility for Digital Archiving and Long Term Access to Digital Data*, David Haynes, David Streatfield, Tanya Jowett and Monica Blake, Joint Information Systems Committee of the Higher Education Funding Council, Digital Archiving Working Group, Great Britain, 1997.
- 4) *Data Policy and Barriers to Data Access in Canada: Issues for Global Change Research*, A Discussion Paper by the Data and Information Systems Panel of the Canadian Global Change Program, Royal Society of Canada, 1996.
- 5) *Preserving Scientific Data On Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources*, Steering Committee for the Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government, Commission on Physical Sciences, Mathematics and Applications, National Research Council, Washington, DC, 1995.
- 6) *Digital Archiving – Developing Policy and Best Practice Guidelines at the National Library of Australia*, Pam Gatenby, National Library of Australia, January 2000.
- 7) *Digital Archiving: Bringing Issues and Stakeholders Together*, International Council for Scientific and Technical Information, Conference Proceedings, January 2000, UNESCO House, Paris.
- 8) *An Overview of the Acquisition Policy of the National Archives of Canada*, prepared by Yvette Hackett, Government Records Branch, National Archives of Canada, November 2000.
- 9) *The Role of the National Archives of Canada and the National Library of Canada*, Report submitted to the Honourable Sheila Copps, Heritage Minister, Government of Canada, by Dr. John English, 1998.
- 10) *The National Library of Canada's Role in the Digital Environment*, Report prepared by Doug Hodges, Information Technology Services, National Library of Canada, November 2000.
- 11) *The Social Sciences Dream Machine: Resource Recovery, Analysis and Delivery on the Web*, Jostein Ryssevik (Norwegian Social Science Data Service) and Simon Musgrave (UK Data Archive), paper presented to the IASSIST Conference, Toronto, May 1999.
- 12) *The Paradox of Digital Preservation*, Su-Shing Chen, University of Missouri-Columbia, *Perspectives*, Institute of Electronic and Electrical Engineers, 2001.
- 13) *Digital Archiving: Approaches for Statistical Files, Moving Images, and Audio Recordings*, Oya Y. Rieger (Cornell University), *RGL DigiNews*, December 1998, vol.2, no.6.
- 14) M.J. Peterson, "Community and Individual Stakes in the Collection, Analysis and Availability of Data", *PS: Political Science and Politics*, September 1995, pp.462-4.
- 15) Gary King, "Replication, Replication", *PS: Political Science and Politics*, September 1995, pp.444-52.
- 16) *ESRC Green Paper on Data Policy and Data Archiving: Consultation Paper*, Economic and Social Research Council (Great Britain), October 2000.
- 17) *Preserving Digital Objects: Recurrent Needs and Challenges*, Mihael Lesk, Bellcore Corp., 2000.
- 18) *Ensuring the Longevity of Digital Information*, Jeff Rothenberg, Rand Corp., Santa Monica, CA, February 1999.
- 19) *The Data that Archiving Fails to Capture*, Peter Buneman, University of Pennsylvania, 2000, [www.cis.upenn.edu/peter/archive.htm](http://www.cis.upenn.edu/peter/archive.htm)

SOCIAL SCIENCES AND HUMANITIES RESEARCH COUNCIL OF CANADA

## We build understanding

The Social Sciences and Humanities Research Council of Canada (SSHRC) is an arm's length federal agency that promotes and supports university-based research and training in the social sciences and humanities. Created by an act of Parliament in 1977, SSHRC is governed by a 22-member Council that reports to Parliament through the Minister of Industry.

SSHRC-funded research fuels innovative thinking about real life issues, including the economy, education, health care, the environment, immigration, globalization, language, ethics, peace, security, human rights, law, poverty, mass communication, politics, literature, addiction, pop culture, sexuality, religion, Aboriginal rights, the past, our future.



Social Sciences and Humanities  
Research Council of Canada

Conseil de recherches en  
sciences humaines du Canada

Canada

350 Albert Street  
P.O. Box 1610  
Ottawa, ON K1P 6G4  
Canada

Phone: (613) 992-0691  
Fax: (613) 992-1787  
Internet: [www.sshrc.ca](http://www.sshrc.ca)