NCASRD

National Consultation on Access to Scientific Research Data

Final Report

January 31, 2005

David F. Strong
and Peter B. Leach

NCASRD
National Consultation on
Access to Scientific
Research Data

Government of Canada Gouvernement du Canada

Canada

For further information or to obtain a copy of this *Report*, contact:

CISTI Help Desk

Canada Institute for Scientific and Technical Information
National Research Council Canada
M-55, 1200 Montreal Road
Ottawa, Ontario, Canada K1A 0R6

Tel: 1-800-668-1222 (Canada and US) or (613) 998-8544
Fax: (613) 993-7619
E-mail: info.cisti@nrc-cnrc.gc.ca
URL: http://cisti-icist.nrc-cnrc.gc.ca

National Consultation on Access to Scientific Research Data

Final Report
January 31, 2005

David F. Strong
and Peter B. Leach

Library and Archives Canada Cataloguing in Publication

Canada. Task Force for the National Consultation on Access to Scientific Research Data

Report of the Consultation on Access to Scientific Research Data :
final report, January 31, 2005

Issued also in French under title: Consultation nationale sur l'accès
aux données de recherche scientifique.
Issued by theTask Force for National Consultation on Access to Scientific Research Data.
Available also on the Internet.
ISBN 0-662-39966-8
Cat. no. NR16-84/2005E

1. Research – Canada – Information services.
2. Science – Canada – Information services.
3. Science and state – Canada.
4. Technology and state – Canada.
5. Communication in science – Government policy – Canada.
6. Information storage and retrieval systems – Research.
I. Canada Institute for Scientific and Technical Information.
II. Title.

Q180.C3T37 2005 001.4'0971 C2005-980108-5

Paper – English: NR16-84/2005E, 0-662-39966-8
PDF – English: NR16-84/2005E-PDF, 0-662-39967-6

Table of Contents

Foreword

On behalf of the Task Force for and participants in the National Consultation on Access to Scientific Research Data (NCASRD), I am pleased to present this Final Report. I would also like to thank the National Research Council Canada (NRC), the Canada Foundation for Innovation (CFI), the Canadian Institutes of Health Research (CIHR), and Science and Engineering Research Canada (NSERC) for the insight that gave rise to the Consultation, the commitment to its successful completion and the financial support to permit its execution.

The NCASRD sees the very urgent need for action to propel Canada into a new and transformational data-intensive paradigm for Canadian research. Not only will the proactive recommendations keep Canada at the very leading edge of global research, but the resulting economic, industrial, social, environmental, ecological and technological advances will enhance our global competitiveness, improve the quality of life for all Canadians, and help Canada address the threats of environmental degradation and ecological damage.

I urge immediate and pressing consideration of our Report, and recommend the earliest possible implementation, as a national priority, of the step-by-step approach proposed that will lead to early and effective implementation of a national plan for open access to publicly funded scientific research data.

David Strong, *PhD, DSc, LLD, FRSC*
Chair
NCASRD Task Force

(Note: This Report was very substantially and ably drafted by Peter B. Leach FEIC, MA, Dip OR, President, Leach Technologies Ltd.)

Acknowledgements

As Chair of the Task Force for the National Consultation on Access to Scientific Research Data (NCASRD), I would like to express my gratitude to all those involved from conception to completion. The continued support, assistance and commitment of the NCASRD Task Force and Project Management Group members, our partner organizations and their staff, as well as the Forum participants, resulted in its successful conclusion. As Canada goes on to make progress in ensuring open access to, and the long-term preservation of, publicly funded research data, the following organizations and players deserve to be recognized and thanked for their contributions:

- ❑ All Forum participants and their respective organizations (*see section 11.3*).
- ❑ All NCASRD Task Force members (*see section 11.4*).
- ❑ All NCASRD Project Management Group members (*see section 11.5*).

Special thanks to the National Research Council's (NRC) Canada Institute for Scientific and Technical Information (CISTI), the host and organizer of the NCASRD. Among those at CISTI who were involved are Bernard Dumouchel, Director General, and his Administrative Assistant, Marie Charette; CISTI's Research Group, including former Group Leader Gordon Wood and his Administrative Assistant, Carolyn Ahern, as well as Glen Newton, Head, Research Unit; and the CISTI Communications team, including Catherine Betz, Linda Campeau, George Guillemette, Stéphane Levesque and Alexandra Talbot. Jac van Beek and Stephanie Delorme from NRC Corporate Services also significantly contributed to this initiative.

On behalf of all participants, I would like to personally thank the keynote speakers: Arthur Carty, National Science Advisor to the Prime Minister; Patricia Kosseim, former Director, Ethics Office, Canadian Institutes of Health Research; Claire Morris, President, Association of Universities and Colleges of Canada; Eliot Phillipson, President and CEO, Canada Foundation for Innovation; Steve Shugar, Director, Policy and International Relations, Science and Engineering Research Canada (NSERC); Marie Tobin, former Director General, Innovation Policy Branch, Industry Canada; and Ian Wilson, Librarian and Archivist of Canada. Your knowledge and vision initiated constructive discussions at the two-day Forum.

Finally, I thank all the partners in this initiative:

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

NSERC
CRSNG

Canada Foundation for Innovation
Fondation canadienne pour l'innovation

NRC·CNRC

The tabling of this report marks the completion of the NCASRD and the beginning of the process for making its recommendations a reality. To all those who have helped point the way forward, whether in the spotlight or behind the scenes, I am truly grateful.

Regards,

David Strong, PhD, DSc, LLD, FRSC
Chair
NCASRD Task Force

1 Executive Summary

Almost fifty years ago, the great writer and futurist H. G. Wells was very close to capturing both the development of the Internet and World Wide Web, and their potential impact on knowledge and research.

> *"Few people as yet, outside the world of expert librarians and museum curators and so forth, know how manageable well-ordered facts can be made, however multitudinous, and how swiftly and completely even the rarest visions and the most recondite matters can be recalled, once they have been put in place in a well-ordered scheme of reference and reproduction."*
>
> — H. G. Wells
> "*World Brain: The Idea of a Permanent World Encyclopaedia*"
> Contribution to the new *Encyclopédie Française*
> August 1937

That day has arrived, and Canada must seize it.

In mid-June, an expert Task Force, appointed by the National Research Council Canada (NRC), came together in Ottawa to plan a national Forum as the focus of the National Consultation on Access to Scientific Research Data (NCASRD). The Forum brought together more than seventy leaders Canada-wide in research, data management, administration, intellectual property and other important areas.

This Report is a comprehensive review of the issues, opportunities and challenges identified at the Forum, complemented by a selection of the supporting documents presented as Appendices.

1.1 The New World

Complex and rich arrays of scientific databases are changing how research is done, speeding discovery and creating new concepts. Increased access will accelerate these changes, creating a new world of research **and a whole new world**. When these databases are combined within and between disciplines and countries, fundamental leaps in knowledge can occur that transform our understanding of life, the world and the universe.

For example, in the analysis of human genetics, the technology to capture enormous amounts of data and to mine them for new information is already showing the genetic make-up of life and the understanding of numerous diseases and syndromes. We will soon be able to analyze such complexities as the pre-disposition to disease in animal and plant populations based on genetics, social and environmental conditions, and demographics, so that all these factors can become part of new disease prevention strategies. With the ability to access and integrate data compiled in different fields, totally new knowledge regimes are being opened in ways that have historically been impossible.

1.2 Canada

For Canada to be a leader in the knowledge economy, the country must be a leader in the new world of research. For Canada to lead in this research transformation, it is essential to take swift action on the recommendations of this Report. For Canada to benefit economically and socially, substantial changes are required in our scientific enterprise, including:

- research culture and behaviour;
- research institute management, policies and strategies;
- legal and policy frameworks;
- financing and budgeting of research; and
- data technologies and computing infrastructure.

Some of our OECD competitors are moving on these challenges much faster, posing an ultimate and very real threat to Canada's economic and social well-being.

While Canada is involved in many global database and research initiatives, these are individual cases not bound by any national strategy or standards. As a result, much of the data on which our knowledge is being built today is hard to access by other Canadian research communities, and is often not ideally structured to be as useful or as open as possible, even within the discipline for which it is being constructed. The vanguard of Canada's national activity and international presence in access to scientific data is through the shared leadership of the Canadian National Committee for CODATA (CNC/CODATA), the Canada Institute for Scientific and Technical Information (CISTI) and the Canadian Association of Research Libraries (CARL).

Member institutions of CARL are already active in preserving some of the country's scientific heritage in digital format, and provide much of the knowledge that is being used in the data capture, access and preservation processes of many national and international scientific database projects. CARL members are 27 of Canada's major academic research libraries, together with CISTI, Library and Archives Canada (LAC) and the Library of Parliament.

1.3 The Past

However, no national data preservation organization exists, nor does Canada have any national data access strategy or policies. Participants at the NCASRD expressed considerable concern about the loss of data, both as national assets and definitive longitudinal baselines for the measurement of changes over time. These losses occur as a result of storage media degradation, media and metadata loss, and software and hardware obsolescence, as well as privacy policies and decisions (e.g., by research ethics committees), and a lack of planning or attention to preservation beyond the individual researcher or organization.

1.4 Action

So, action is urgently needed to stop such degradation and loss of the country's research heritage; action that could concomitantly thrust Canada into a leading position in this new paradigm for research and development.

We recommend the creation of a task force, dubbed **Data Force**, to prepare a full national implementation strategy, and mount a pilot project to show the value and impact of multi-person and multidisciplinary access to research data. Once such a national strategy is broadly supported and has obtained appropriate funding commitments, we propose the establishment of a dedicated national infrastructure, tentatively called **Data Canada**, to assume overall leadership in the development and execution of a strategic plan. The plan would encompass and presumably extend the NCASRD's recommendations.

1.5 The Future

With Data Canada implementing the recommendations of this Report, we believe the country will be able to achieve the NCASRD's Vision of Canada's place in the global research enterprise of 2020. We envision that by then:

Canada is the centre of a global knowledge grid. It has become the desired nation with which to partner in research, because of its national system of open access to research data. Through this system and the collaborative culture it has generated, Canadian creativity and innovation are best in class worldwide. Open, but secure, access to powerful and globally assembled data has transformed scientific research. Researchers routinely analyze problems of previously unimaginable complexity in months, rather than decades, leading to revelations of knowledge and discovery that have enriched quality of life, transformed healthcare, improved social equality, provided greater security, broadened decision perspectives for social, environmental, and economic policy and advancement, and transformed the advancement of human knowledge.

Canada is not alone in having such lofty aspirations and delays in action will cause cumulative damage to its potential leadership role. With other countries already taking progressive action, delay is a destructive option.

2 Recommendations

These recommendations are grouped according to the organization(s) that we presume would take primary responsibility for their implementation. There is no priority or sequence implied in their order of appearance. Section 8 provides further explanations for the reasons behind each recommendation. The actions to implement the recommendations are linked in time across all the organizations responsible for their execution, while vigorous targets for timing and sequencing are shown in Section 9.

Responsibilities of Data Force

Recommendation 1 – Organizing

The Sponsors establish a task force (Data Force) to prepare a thorough national implementation strategy. Data Force should have representation from:

- Canadian scientific research community;
- Canada's research granting councils (CIHR, NSERC and SSHRC);
- Canada's research infrastructure foundations and trusts (CFI, Genome Canada and equivalent provincial trusts);
- universities, colleges and research institutes that manage Canada's research infrastructure;
- government departments and research laboratories (both federal and provincial) that set public research policy;
- CARL;
- Association of Universities and Colleges of Canada (AUCC);
- CNC/CODATA;
- CISTI;
- LAC;
- Statistics Canada;
- CΛNΛRIE Inc.;
- Privacy Commissioner of Canada;
- representatives of student and professional organizations that participate in planning research training curricula; and
- representatives of the general public who will ultimately benefit from the public good of publicly funded research.

The mandate of Data Force would be to guide and oversee a small implementation secretariat to:

- commission a pilot data access project (Data Project) to illustrate the concepts and values of this Report;
- plan and supervise the formation of a permanent Canadian data access organization (Data Canada);
- secure the long-term commitment to federal financing of Data Canada;
- develop a data access strategic plan (Data Plan).

Recommendation 2 – Educating

Data Force, together with the leaders of sponsoring organizations, immediately begin fostering awareness amongst political, institutional and public opinion leaders of the:

- paradigm shift enabled by access to massive data resources that is occurring in scientific research globally and within Canada;
- need for Canada to be among the global leaders in the transformation of the research enterprise, in order to retain and strengthen our economic competitiveness and scientific excellence over the long term;
- social, medical, ecological, environmental and economic benefits that will accrue in accelerating the pace of scientific discovery;
- educational benefits derived from the ability to place learning in a real-life context and enabled by open data access; and
- need for concerted action to drive the cultural, legal, managerial and political changes essential to establish Data Canada.

Recommendation 3 – Funding

Sufficient funding be provided to Data Force to support the implementation of an open access database pilot project in 2005. This project should be designed to show how databases, when linked, can lead to substantial knowledge breakthroughs. It would also include solutions to the challenges posed by such areas as policy, technology, infrastructure, system management, data and metadata quality, integrity and security. This does not mean a new scientific research project, but rather, the compilation and integration of a compelling example already in progress or completed.

Responsibilities of Data Canada

Recommendation 4 – International Participation

Data Canada establish a management capability that can monitor and intervene in international open access fora to protect Canadian interests, and assist the international community in promoting agreements, standards and policies that support best access, sharing and preservation practices compatible with Canadian needs.

Recommendation 5 – Ethics

Data Canada initiate consultations among the privacy commissions, the National Council on Ethics in Human Research, the Canadian Association of Research Ethics Boards, data librarians and archivists, and Statistics Canada to identify legal barriers to access to scientific data. Such consultations should result in proposed modifications to information privacy laws or their legal interpretation, to ensure high-value, publicly-funded data, properly protected for confidentiality, are preserved, secured and made accessible with appropriately managed access controls. Consequently, the data may be used for research purposes other than their original intent and for which informed consent may not have been sought.

Recommendation 6 – Privacy

Data Canada should initiate a review of the *Personal Information Protection and Electronic Documents Act*, as well as other related legislation to identify inconsistencies that would prevent international data sharing with countries whose collaborative research projects and database sharing practices are expected to be high – early analysis should focus on the US and EU. It should also work with Canadian privacy legislators to align such legislation to permit fully compliant data sharing between specific countries.

Recommendation 7 – Archiving

Databases and datasets, determined by Data Canada (e.g., through a peer review panel or special committee, or relevant professional society or association) to be of national importance, be deposited and secured at LAC.

Recommendation 8 – Liability

Data Canada establish an expert panel to examine the Canadian and international legal frameworks concerning responsibility and liability for databases and datasets, and task them to propose a new Canadian legal framework compatible with evolving international legal frameworks. The aim would be to balance the liability of data custodians and their institutions against the social benefits resulting from open access to such data, in order to protect the custodian against liability derived from unexpected future uses of the data.

Recommendation 9 – Anonymization

An expert panel be appointed to examine the legal issues surrounding data anonymization and secure data practices that would prevent infringement of an individual's privacy, if made accessible for other research. The panel should identify limits to the applicability of informed consent, when no possible identification, or deduction of the individual or small group remains feasible. Should opportunities be found to permit anonymized data re-use, the panel will propose appropriate changes to legal and regulatory practices.

Recommendation 10 – Databases at Risk

Data Canada establish a fund to preserve, and improve the accessibility of existing high-value, "at-risk" and/or critical databases identified by peer review panels as having significant current, future or historical value.

Recommendation 11 – Criteria and Quality

Data Canada work with its research partners to establish a function within Data Canada (and its international counterparts) to formalize assessment criteria for data quality, as well as define processes to measure data quality and integrity.

Responsibilities of Funding Agencies

Recommendation 12 – Training Researchers

All organizations that fund scientific research provide specific funding for the training of all principal investigators in best practices of database selection, management, rights management and data curatorship, metadata standards and other important issues, so access and preservation can be built in to the data acquisition and storage plans from the outset.

Recommendation 13 – Data Management Plans

Research councils, and all other public-sector research funding agencies and departments require that project and grant applications include a data management plan, as well as specifically identified funding that will ensure quality, integrity, accessibility and accountability. A funding condition should be the inclusion of a well-constructed plan for data acquisition, management, access and preservation. Adherence to such plans should also become a non-competitive performance metric for the project and gateway for subsequent grant applications. Councils should recognize these as added costs to the main thrusts of research projects.

Recommendation 14 – Resources

Federal and provincial government departments, agencies and ministries that fund scientific research establish long-term stable, non-competitive core budget allocations to provide research institutions, organizations, and agencies with the resources to preserve all important databases (historic, current or potential high value). The federal government also provides additional and sufficient funding to LAC, ensuring the long-term archival preservation of all important databases and datasets.

Recommendation 15 – Peer Review

Databases and datasets in use or expected to be used in multiple research initiatives, including their metadata, be subject to peer review, with the evaluation becoming part of the metadata.

Recommendation 16 – Time Limits

In collaboration with Data Canada, funding agencies and departments set limits for the length of time data custodians may deny open access to their databases. This time should be fair and reasonable in the prevailing circumstances. After the specified period, the database must be made publicly accessible subject only to constraints imposed by law, or international protocols and agreements.

Responsibilities of Universities and Researchers

Recommendation 17 – Rewards

University faculties, the professoriate, and other academic research units extend the recognition and reward systems for researchers to include excellence in contributions to scientific data, and the development of tools for improved data management and use, as an important performance indicator.

Recommendation 18 – Creating Specialists

Post-secondary institutions increase their intake of students in Information Science, and the teaching of database access and preservation to address the shortage of trained digital librarians, managers, curators and archivists.

3 Objectives

The objectives of the National Consultation on Access to Scientific Research Data (NCASRD) are to recommend to Canada's primary research funding agencies and organizations the actions necessary to maximize, through open access, the research and economic value, and public benefit of data gathered at public expense, as well as actions to preserve historically significant data as an historic record, and as a scientific and cultural asset for current and future research. The recommendations in this Report aim to generate workable solutions to the technological, institutional, cultural, legal, financial and behavioural barriers to such access.

The NCASRD was designed to complement the *National Data Archive Consultation for the Social Sciences*, completed by the Social Sciences and Humanities Research Council (SSHRC) and the National Archives of Canada (now known as Library and Archives Canada - LAC).

The NCASRD has been commissioned to recommend actions that only apply to digital data. We have excluded consultation on the issue of open access to research findings and published research results, even though the publication of results is often closely linked to open access to scientific research data. The issue of open and possible free access to research results and scientific papers is highly contentious but should become the focus of a dedicated national consultation in the near future. This may soon impact the development of research collaborations, especially with medical research initiatives in the US and UK.

4 Background

Introduction

Since governments in all developed countries became involved in the funding of scientific research, there has always been a question about the value that the public should obtain from such funding versus the commercial interest of industry, and the related financial interests of researchers, research laboratories and academic institutions. With the rapid expansion of knowledge globally and the dependence of new knowledge on both the prior observations and data of others, and the new databases that underlie these new scientific conclusions, the debate has intensified.

On the one hand, there is a growing community that believes the rate of scientific discovery is gated by access to the data and the findings of others, and sees open access as being a fundamental accelerator of scientific knowledge. On the other hand, there is a fear that such access will undermine the scientific publishing industry, and compromise the management and return on investment of valuable intellectual property derived from the research. In the realm of medical and social knowledge particularly, this debate is further complicated by patient rights set against the need for longitudinal observation of health and social change, frequently involving data previously collected by others for different purposes. The medical domain is also complicated by the Charter rights to personal security and privacy, in both publicly and privately provided healthcare, which is in substantial conflict with the commercial interests of the healthcare industry and the benefits to the broader community.

Also, many researchers have noted that early data on which much of our knowledge has been built have already been lost and continue to be at an accelerating rate, despite the adoption of information technology (IT). Indeed, IT is responsible for much of the loss, as storage technology has given a false sense of security against loss and obsolescence. Furthermore, while the data might still exist somewhere, there are very few cases where the data have been systematically archived, with relevant metadata about their applicability and creation, so as to make them readily accessible and available for reuse, regardless of whether this is provided on a free or commercial basis.

When considering the broader picture of open access to scientific papers outlining the results, conclusions and intellectual property of research analysis, the debate is far more complex. This raises intellectual property management and control issues, as well as concerns surrounding rights – those of the publishing industry versus the public's, the researchers and the institutions benefiting from their creativity, the institutions to use research results for education and subsequent research, etc.

In this increasingly public debate, there are proponents who feel that more open access both to research data and research findings would be highly beneficial to the efficiency of the research endeavour. In contrast, there are others who see open access as a threat to a wide range of accepted and proven practices. However, in the more restricted area of access to data, this discussion has reached sufficient intensity to warrant the signing of the *International Declaration on Access to Research Data from Public Funding*[1] by most developed nations, committing them to a more open data access regime.

The *Declaration*'s premise: publicly funded research data should be openly available to the maximum extent possible. While there is general consensus about increasing access for scientific data, the degree to which data should be openly available remains highly disputed. The NCASRD Forum was strongly in favour of access to research data being as open and affordable as possible, but recognized that the degree to which this would be achieved must evolve over time.

The *Declaration* does not make any commitment to the issue of open or free access to research papers, an already hotly debated matter among the political, academic and commercial arenas. In the US, the National Institutes for Health (NIH), backed by Congress, is currently planning to introduce a policy requirement whereby the results of all publicly funded medical research must be placed in a publicly accessible information system (i.e., PubMed Central), within six months of their initial publication. Advocates in the UK are also actively pursuing a similar proposal, although the government recently came out against it. Meanwhile, the UK's Wellcome Trust, an independent charity funding human and animal health research, is planning to make such disclosure a requirement of all Trust-funded research, as well as launch a European mirror of the US PubMed Central.

Issues and answers around data archiving have been addressed by the Social Sciences and Humanities Research Council (SSHRC), within the context of its research community. These recommendations can be read in the *Final Report – National Data Archive Consultation*[2] . While its discussion is broadly applicable to the larger Canadian context, it only covers issues relating to infrastructure. Therefore, it fails to address many of the following key issue domains identified in the OECD's follow-up *Final Report*[3] :

- technological, including interoperability and quality management;
- institutional and managerial, including the necessary diversity of institutional models, and discipline-tailored data and archiving management;
- financing of capital and operations, including long-term preservation;
- legal and policy, including international collaboration agreements on sharing practices, rights legislative frameworks; and
- cultural and behavioural, including reward structures, discipline communication barriers, ownership pride and data structuring compliance.

The objectives of the NCASRD are to examine all the issues that must be addressed to successfully implement a more open data access infrastructure and culture, and recommend actions that will accelerate system and culture change, enhancing the efficiency of Canada's publicly funded research endeavour.

With a variety of initiatives underway worldwide, there is already a growing community of researchers familiarizing itself with all aspects of open access and collaborative database design and operation. Examples include:

- medical communities – PubMed Central, its EU counterpart, Bio-molecular Interaction Network Database (BIND) and International Genome Database (Genbank);
- astronomical communities – Digital Palomar Observatory Sky Survey (DPOSS), Hubble Deep Field – South (HDF-S) and the Canadian Astronomical Data Centre (CADC);
- earth observation communities – Global Earth Observation System of Systems (GEOSS); and
- environmental communities – Global Biodiversity Information Facility (GBIF).

However, the infrastructures, systems and processes are significantly heterogeneous, as there has been little technical and managerial liaison among these global initiatives. They have proven of great value, though, collaboratively assembling data that could not be created otherwise, and allowing new scientific approaches to be used with much higher resolution and reliability. Although they are not easily transferable across or within disciplines, they still handle greater complexity, and achieve broader applicability and superior authority.

On another note, few nations and disciplines have made considerable, widespread progress in implementing data sharing. Therefore, there is a valid opportunity for Canada to assert global leadership in establishing appropriate physical, operational, systemic and policy solutions. We hope and expect this Report will mark the first step towards the development of a broadly appealing Canadian solution, which could then be offered worldwide.

In this report, the terms "data"[4], "database"[5] and "dataset"[6] are used with specific meanings. Their meanings can be found in the notes located at the end of this section.

Ministerial Declaration on Access to Public Research Data

On January 30, 2004, Canada and 33 other countries, including all G8 members, adopted a declaration of their commitment to work towards the establishment of access regimes for digital research data stemming from public funding. The *Declaration* seeks to achieve the following objectives and principles:

- openness that recognizes and balances the interests of open access to increase the quality and efficiency of research and innovation with the need to protect social, scientific and economic interests;
- transparency that makes the source, documentation (metadata) and conditions of use available and accessible internationally;
- legal conformity that addresses the national legal requirements concerning national security, privacy and trade secrets;
- formal responsibility that promotes rules covering authorship, producer credits, ownership, usage restrictions, financial arrangements, ethics, licensing terms and liability;
- professionalism that builds rules for management, based on professional standards and values;
- protection of intellectual property that provides open access under differing legal regimes applicable to databases;
- interoperability that will meet international standard requirements for use;
- quality and security that will ensure good practices are employed in generation, storage and accessibility management, guaranteeing authenticity, originality, integrity, security and liability;
- efficiency; and
- accountability.

The premises on which the *Declaration* is based include:

- international exchange of data contributes decisively to the advancement of scientific research and innovation;
- open access promotes scientific progress and training of researchers;
- access and reuse will maximize the return from public investment in data collection;
- IT enables a significant increase in the scope and scale of research endeavours from public investments;
- opportunity for substantial benefits to society could be compromised by undue restriction on access and use of such data; and
- access to research data will enhance the participation of developing countries, contributing to their social and economic development.

The *Declaration* concludes with an invitation to the OECD to develop guidelines based on commonly agreed principles to facilitate cost-effective access to digital research data from public funding. Responses to this invitation will most likely be discussed at a future OECD meeting, when open access data implementation is set as an agenda item.

SSHRC Consultation on a National Data Archive

The Executive Summary of the SSHRC consultation states:

> *"In October 2000, the Social Sciences and Humanities Research Council and the National Archivist of Canada established a Working Group of research and archival experts and asked them to assess the need for a national research data access, preservation and management system. After compiling extensive evidence for the need of such a service to support the knowledge creation work of Canada's social sciences and humanities research community, the Working Group now offers recommendations for the creation of a new national research data archival service. This service would have three core functions:*
>
> - *Preserving research data that are compiled by researchers, and preserving data compiled by government agencies, polling firms and other organizations that can be used by researchers to generate new knowledge;*
> - *Managing the data held, including ensuring quality; selecting data for retention; developing and applying standards for metadata, authenticity and security; and migrating data across technologies;*
> - *Providing access to research data, including Web-based delivery systems, cataloguing services, user and depositor agreements to protect confidentiality and intellectual property rights, and connections to other data depositories around the world.*
>
> In addition, the Working Group recommends that a new National Research Data Archive Network undertake a number of other functions, including providing advanced training in data handling techniques, representing Canadian interests in the development of international data standards, promoting data sharing as a best practice in research, undertaking research in information and archival sciences and acting as a central hub and coordinating body for a network of data services in Canadian research institutions.
>
> Digital information compiled for research purposes is playing an increasingly important role in today's knowledge economy. In many ways, data are the fuel driving innovation and our capacity to address complex social and economic problems. Although billions of dollars are spent each year collecting data, Canada lacks the necessary infrastructure to ensure these data are preserved and made publicly available. This limits the return that can be made on our public investment in research and undermines good public stewardship.
>
> Many of the building blocks necessary for the creation of a National Research Data Archive are already in place. University data services, high speed transmission networks, legal and ethical guidelines and frameworks, potential partner institutions, various data depository and access portal initiatives, and an active data-producing research community already exist. The missing element is a preservation, coordination and management service.

Almost all developed countries have recognized the need for a national research data service, and some have more than a generation of experience in their operation. Canada is in a position to learn from this experience while developing a research data service that fits our unique institutional and cultural context. We now have the technological capacity and expertise to create a "trusted system" that provides Canadians with an accessible and comprehensive service empowering researchers to locate, request, retrieve, and use data resources in a simple, seamless and cost effective way, while at the same time protecting the privacy, confidentiality and intellectual property right of those involved. The start-up infrastructure costs for this service could be funded through the Canada Foundation for Innovation. The annual operating costs for a comprehensive facility and network are benchmarked in the area of $3 million.

The Working Group offers three options for the creation of a National Research Data Archive Network:

1. Through federal legislation, create a National Research Data Archive Network as a modified version of a Separate Statutory Agency. This is the ideal approach to building a full-service, trusted agency, composed of a central data preservation and management facility and a series of access and service nodes located in research institutions. It takes full advantage of existing research infrastructure, has long-term stability, a direct connection to research data users and producers and the capacity to represent Canada's interests in the development of international data standards.
2. Create a National Research Data Archive Network under the auspices of the Social Sciences and Humanities Research Council. This approach captures the characteristics of the first model, but does not require legislation. It benefits from a direct, immediate connection with the researchers and established accountability and funding structure.
3. Create a Special Operating Agency with the National Archives of Canada. As a stand-alone division within the National Archives, this approach takes full advantage of existing archival infrastructure and expertise. This has not been the preferred approach in other countries, because the core mission of a national archive and a national research data service are fundamentally different. Nevertheless, as a Special Operating Agency, the service could potentially have both stability and the capacity to develop a trusted research data preservation, management and access system."

Within the context of the broader Canadian requirement, the recommendations of the SSHRC report, with their primary focus on infrastructure, is a very useful position from which to expand the dialog to the broader research community and to consider the other aspects that are instrumental to the commitments of the Ministerial Declaration and the framework outlined by the OECD."

[1] *Declaration on Access to Research Data from Public Funding,* dated January 30, 2004 in Paris, France. Governments signing: Australia, Austria, Belgium, Canada, China, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Russian Federation, the Slovak Republic, the Republic of South Africa, Spain, Sweden, Switzerland, Turkey, the United Kingdom, the United States of America. The material content of this declaration is provided in Appendix 11.5.

[2] *National Data Archive Consultation* – Final Report, dated June 2002. "Building Infrastructure for Access to and Preservation of Research Data" submitted by the NDAC Working Group to the Social Sciences and Humanities Research Council of Canada and the National Archivist of Canada. The material content of this report is included in the Appendices, section 10.

[3] OCED Report "Promoting Access to Public Research Data for Scientific, Economic and Social Development" by the OCED Follow-up Group on Issues of Access to Publicly Funded Research Data, March 2003

[4] Data are known facts, measurements, or discrete elements of information used as a basis of inference or reckoning – Adapted from *Oxford English Reference Dictionary*

[5] A Database is an ordered or structured set of data, usually held in a computer, designed to be accessible in a variety of ways and which can be manipulated in full or in part to assist in the process of inference or reckoning.

[6] A Dataset is either a subset of a database, or a set of data created by merging two or more databases. It is also used for inference or reckoning but is uniquely assembled for a specific inference or reckoning.

5 Vision

Arthur Carty, National Science Advisor to the Prime Minister, in his opening comments to the NCASRD Forum, entitled “The Power of Vision in Science: Opening Access to Canadian Scientific Research Data”, stated:

> *“Science is coming to a crossroads in the way it views and manages its rapidly growing store of scientific research data… scientific convergence and the arrival of the information commons are opening up powerful new forms of scientific collaboration in all corners of the globe. E-science is transforming the way the world’s scientific community works and shares its intellectual, analytical and investigative output… E-science is also transforming both the value we attach to and the way we work with data.”*

These observations give true meaning to the Vision of the NCASRD.

Our Vision of the research world in 2020:

Canada is the centre of a global knowledge grid. It has become the desired nation with which to partner in research, because of its national system of open access to research data. Through this system and the collaborative culture it has generated, Canadian creativity and innovation are best in class worldwide. Open, but secure, access to powerful and globally assembled data has transformed scientific research. Researchers routinely analyze problems of previously unimaginable complexity in months, rather than decades, leading to revelations of knowledge and discovery that have enriched quality of life, transformed healthcare, improved social equality, provided greater security, broadened decision perspectives for social, environmental, and economic policy and advancement, and transformed the advancement of human knowledge.

This vision is far reaching, but quoting Arthur C. Clark, “The world needs uninhibited thinkers, not afraid of far out speculation; it also needs hard-headed conservative engineers who can make their dreams come true”. The NCASRD brought together both uninhibited thinkers and conservative researchers, and this report lays out a roadmap to this forward-thinking vision, defining the preliminary, bold steps to start Canada’s ascendancy in the new data-enabled research milieu.

6 Opportunity and Impact

Forum participants identified a wide range of opportunities and impacts that would result from the implementation of open access to scientific research data, and solutions to locating the best, most relevant and broadest diversity of data sources for each particular problem. Consequently, a vista of new research directions and knowledge emerged. These opportunities and impacts are grouped into subject areas, so cumulative impacts can be better understood. However, the overall Impact Statement is presented first.

Ten-Year Impact Statement

In 2015, Canada's participation as one of the global leaders in the sharing of scientific research data has resulted in transformative systemic, cultural, environmental and economic achievements in our scientific research enterprise, measures of which are encapsulated in the following:

- all institutions and publicly funded research laboratories have fully operational scientific data capture, storage, access, reuse, and archiving processes, procedures and infrastructure that fully meet the Canadian National Scientific Data Access and Preservation Policy, as well as long-term stable budgets to maintain them as state-of-the-art capabilities;
- the Prime Minister often cites Canada's National Scientific Data Access and Preservation Policy, and the investment in its implementation, as the most important transformative change that has allowed Canada to reach its target as the fifth most research-intensive country in the world;
- Canadian data access, mining and preservation companies are amongst the global leaders in data management solutions;
- Canada is an acknowledged global leader in the practices, processes and standards of scientific data management, demonstrated through the adoption of Canadian scientific data management systems by most countries that have acquired such systems;
- Canada's granting councils report that all research grant holders are actively involved in the creation, expansion, assembly, and maintenance of databases and datasets that form the underpinning of Canada's scientific knowledge base, resulting in Canada's per capita scientific research productivity being amongst the top three in the world;
- Canada's public sector research laboratories and universities report the highest-ever number of international research collaborations, with a greater proportion of international collaborations than any other country;
- almost all Canadian academic and industrial research papers cite scientific data reuse from Canadian and international open access data sources;
- all peer-reviewed scientific data of national importance have been preserved in the National Data Archive, and sustainable funding is embedded in the federal government's long-term core budget, with all-party political consensus to its continuity; and
- the performance of academic and publicly funded laboratory researchers includes the assessment, within their institutions and by granting councils, of their contribution to national and international open access databases as a consideration equal to other research contributions.

New Science

The NCASRD Forum participants anticipated many new areas of science that access to and novel intersections of current and new global data sources will enable. These derive principally from new intersections of disciplines and data, opening up:

- new methods of research based on the intermeshing of data that cannot be brought together in 2005;
- complexity reduction only achievable by interdisciplinary data integration and mining;
- the study of transient data and continually captured data; and
- unexpected and unknown relationships among data which are found serendipitously as other research challenges are undertaken.

Forum participants are confident that major scientific breakthroughs will be achieved, due to the existence of accessible databases and datasets of a size and scale previously impossible to build, owing to the constraints of team size, discipline boundaries, database management capabilities and metadata standards. The impact of open access has transformed the pursuit of knowledge purely through the scale and quality of data accessible to researchers.

Examples of 2000-2005 advancements and challenges

Data-enabled research acceleration

The **genetics and genomics** communities have already established a proven model of data access and data sharing. Repositories for DNA and protein data have already revolutionized the way biology is carried out. This model is the foundation for the acceleration of research in bio-informatics. The continuing development of dynamic processes to automatically build the raw data directly derived from DNA sequencing machines will further accelerate bio-informatics research. As a result of the accumulation of very large datasets, the clustering and analysis of the multiple experiments being undertaken internationally become feasible for the first time. Subtleties that would otherwise have been undetectable or gone unnoticed will now become evident and lead to new insights. These disciplinary advances are just the start of a new wave of opportunity. The interactions and discoveries which become possible when huge databases – medical, environmental, social, economic and demographic – can be co-analyzed, will enable researchers to begin truly understanding the many contributing factors that impact universal understanding, including life and health.

Challenges of legal constraints

In the **health community**, research data access and reuse have been heavily restricted, due to privacy issues and the *Personal Information Protection and Electronic Documents Act*. New methods for ensuring anonymization, like secure access management under strictly enforced protocols that certify the correct application of informed consent and effective data encryption, will eliminate the need for the data destruction practices prevalent in health research today. These methods allow the assembly of substantially larger datasets for research, which, in turn, permit research into the causes, treatments, and treatment alternatives (and combinations), prognoses and outcomes of both common and less-common health problems, and provide the long-term studies that are often impossible or statistically unreliable due to data loss or small sample sizes available to today’s practices. When such changes to the data infrastructure are in place, researchers will have the ability to connect their new data to many other databases so that other factors can be investigated and identified.

New knowledge from multi-disciplinary linked databases

In the **environmental research community**, there are already many national and international databases, and many of these are linked. Our understanding of the complex interactions of the oceans, earth and air, as well as the actions of humans, is increasingly driven by global databases of sensor-generated data from these domains and from broad-spectrum satellite imaging data. Such linkages will uncover totally new understanding of our environment and the changes that are taking place. The efforts to mitigate such changes will, for the first time, become measurable and observable. Also, this new knowledge will create commercial opportunities for new products and services for Canadian companies.

Better Science

Open access to research data will ensure better science. Historically, the data on which scientific research was based frequently had neither the precision, accuracy, nor volume for many research theses to be validated with high degrees of assurance. As a result, the knowledge and hypotheses generated have always remained open to re-evaluation and reinterpretation. Much of our knowledge has been based on assumptions of the minimum necessary precision and lowest observation frequency of data, resulting in inconsistent findings and unreliable science. This has usually been caused by the time and cost of data generation and validation. Furthermore, in such circumstances, unless the original data have been kept in some form that can be re-used, all re-evaluation and re-interpretation must be done only on newly created or captured data and the historic baseline cannot be verified and re-analyzed.

Better science will result from the ability to observe, re-evaluate and re-analyze the original data and consider them in the context of new knowledge and/or new data. It will also come from the much larger global datasets that will be created as teams work together. Both the increase in observation frequency and the parallel improvement in data validation will reduce errors and encourage more comprehensive explanation and understanding of data deviations. Open access, together with systematic archiving and improved and standardized metadata, will also allow superior longitudinal analyses, better experiment replication, and better-informed peer review.

Examples:

There exist some notable examples of better science that have come out of the global need to provide open access to scientific data on specific issues such as:

- the 1991 Bromley Principles that instituted full and open exchange of global environmental change research data and allowed the creation of GEOSS; and
- the 1996 Bermuda Principles on the Release of Human Genome Sequence Data that resulted in GenBank.

In these and numerous other cases, the open access protocols have led to leaps in knowledge, which have surprised even the researchers. Perhaps the ultimate example at present is the mapping of the human genome.

Leadership in Innovation

Open access to scientific research data will stimulate a serial change in scientific research culture, with each cultural shift encouraging further expansion of accessibility. As the ability to work across disciplinary boundaries increases, and exposes the enormous value of the new knowledge that will result, the rate of innovation will surge. While Canada will not be alone in experiencing such a surge of innovation, the integrated strategy that we are recommending, and the early adoption of open access as a national priority, will guarantee Canada's leadership position among research-intensive countries.

The NCASRD Forum noted, in particular, that the interdisciplinary barriers that have inhibited the development of new business sectors and business opportunities will disappear as open access takes hold. One of the greatest areas for opportunity is in the prevention of disease. Analysis of genetic, social and environmental predisposition to particular diseases will allow the very rapid expansion of the "wellness sector" of the economy. This breakthrough will enable pre-emptive treatment in the cases of pre-disposition to diseases, and either eliminate their onset or minimize their impact. Other sectors will similarly be stimulated by open access, including, for example:

- the computer software industry that will develop improved search, data-mining and data-management tools and systems;
- the sensor technology industries that will capture data both for scientific research and for commercial, industrial, environmental, social, healthcare and wellness use.

Superior Policy and Strategy

The ability to find, access and combine data from an array of trusted databases, both current and archived, allows policy advisors and strategic planners to examine the impact of policy and strategy alternatives in a much broader context than is possible today.
For instance, the impact of transportation policies on long-term health-care costs and workplace productivity becomes possible through the understanding of pollution models and pollution effects on health, as well as the resultant health costs and related impacts on productivity. Being able to derive understanding of the broader impacts of policy and strategy alternatives, through open-access-enabled modeling, has the potential for huge pay-offs to society. Such dividends cannot be generated, nor even computed today, because assembly of the essential data is well beyond the capacity of policy analysts, strategic planners or, indeed, the researchers themselves.

Many other examples can be projected, such as global efforts to create, and integrate international and national environmental policies and strategies for the future. In the past, such initiatives have only been marginally successful, because the necessary data have not been readily accessible or consistent:

- The dynamics of environmental change, although based on current environmental data, are still not adequately understood and the impact of human activity is subject to interpretation. As a result, the public has a justifiable scepticism about both its importance and validity – as do many scientists. This has led to a diversity of views in governments about the scale of the challenge and its timeframe. Bringing together international and national data and resources, and relating these to other data, such as demographic, land use and economic databases, will enable researchers and policy makers to increase their understanding, reduce interpretive uncertainty and strengthen the case for better-defined action.
- The economic, health, population and social impacts of environmental change continue to be speculative and questionable, because the data used for such projections are rarely well matched to the general public need and difficult to access, and foreign national databases

are seldom consistent with their Canadian counterparts. Overcoming these challenges will permit Canadians and others to reach conclusions that have less unexplained error, increased confidence, and lead to better political and public support for whatever remedial policies and strategies may be required.

More Efficient Research

Scientific research is generally constrained by the:

- challenges of acquiring or constructing data with adequate precision, accuracy and scale;
- participation of cross-disciplinary teams that bring essential expertise to the research; and
- assembly of the computing resources that have the power and speed to model hard problems.

Opening access to scientific data will progressively address: the acquisition challenges; the access to essential expertise, through the standardization of metadata and enabled partnerships with researchers in other disciplines; and the increasing ability to use networked resources will address computing power.

These changes will:

- reduce the effort and time in acquiring data of the requisite precision, quality and scale;
- enable data access and help align researchers with valuable domain knowledge and expertise; and
- permit access to the computing tools and resources to handle the demands of complex analysis.

Such new capabilities will accelerate research, while reducing the investment necessary to develop new knowledge and solve problems of far greater complexity. The effort applied to data acquisition in some discrete projects with negative outcomes will also become accessible to other research teams, generating value from those unsuccessful research efforts and creating opportunities for the development of new knowledge.

The effort of building large multi-purpose databases and the reward processes for those involved will become more clearly understood and allow superior decisions to be taken in optimizing investment between:

- data acquisition, accessibility and archiving; and
- theory postulation, verification, knowledge creation and benefit derivation.

Enhanced Education

The process of scientific teaching has traditionally been discipline- and theory-based, and only at the post-graduate stages have scientific observation and data been introduced as a principle component. Even here, the traditional approach has been restricted to discipline-specific data. The opening of data across disciplines will encourage teachers and professors to re-examine their teaching. Both teaching and research will be changed due to open access that will permit the use of evolving and archived data, in a way which will complement and reinforce the knowledge that teachers are trying to impart. In addition, open access will permit research projects to be implemented with greater ease at the undergraduate and even at the secondary school levels. These younger researchers will be able to use real data, thus building on the practical value of the knowledge being imparted. Students will be able to expand their horizons beyond those in the required curriculum and even undertake self-directed research.

With learning based on accessible scientific data, new teaching processes will allow students to gain a better understanding of the power of databases as discovery and problem-solving tools, as well as the ability of effective data management to produce valuable and useful results. Therefore, not only will the value to society of the graduating students be greatly enhanced, but with their new knowledge and respect for scientific data, they will help impart such respect to preceding and succeeding generations.

Scientific discovery from undergraduate and secondary school students is relatively rare, but not unknown. With open access, it will become more likely for exceptional students to have the ability to make scientific discovery, adding to the general growth of knowledge and possibly providing them with exceptional experiences that may inspire careers in scientific research.

7 Challenges to Open Access

NCASRD participants identified a substantial number of inhibitors to the broad acceptance and implementation of a Canadian open access and preservation system. These range from political issues and priorities to the Canadian research culture and reward systems. Without actions to address such inhibitors, the ability to institute open access, as the enabling vehicle for the transformation of the Canadian research enterprise, will be compromised and Canada will fall behind other advanced economies as a result. Moreover, any later systematic approach will be complicated by the existence of open access silos stemming from single initiatives, independent granting council actions, distinct discipline or institutional approaches, or large international projects which have adopted different standards, processes and infrastructures. Re-engineering a large number of differing approaches will be highly disruptive, expensive, time consuming and far more resource intensive than doing it right earlier, under a national strategy.

Priority of need

Universities, granting agencies and research laboratories have other high priority issues (e.g., funding, curricula, community expansion, ageing infrastructure). For change to occur, the importance of open access and preservation of data must become a high priority within the entire research and academic communities.

At the political level, there is little public comprehension of this issue, except the growing national concern over personal data privacy and its possible misuse, which itself is perceived to be the antithesis to open access. Also, it remains unclear whether open access to research data in universities is the jurisdiction of the federal or provincial governments.

Funding mechanisms for the construction, population, maintenance and preservation of scientific databases only rarely exist outside the activities of Statistics Canada and CISTI. CFI has and will continue to provide funds for the capital cost of databases for strategic research initiatives, but cannot fund the data acquisition, maintenance or preservation of them. Exceptional cases exist where the research communities themselves have had no alternative but to commit resources to build and sustain the large databases essential for the complexity of the research questions set out to resolve (e.g., astronomical, genomic, proteomic, environmental, etc.). However, preservation has only been addressed by the research libraries under the leadership of CARL and little funding has been made available for such preservation needs.

Existing NSERC and CIHR grant application rules and selection criteria generally exclude project funding to sustain research capability, such as databases or to establish such databases, unless the underlying technology is a research initiative in its own right. While recent SSHRC rules allow for such funding, few grants have actually had funds approved for this purpose.

Champions for Change

With the two granting councils, CFI, Industry Canada and NRC being Sponsors of the NCASRD, we see a growing concern for the need to promote open access to scientific research data. Another encouraging indicator is the Government of Canada's ratification of the OECD Declaration on Open Access. Nevertheless, we believe studies, discussion and signatures are insufficient. The time has come for the banner to be carried by influential champions within the Canadian research community, the government bureaucracy, the House of Commons and the Senate, so swift action is forthcoming.

Culture

Federal departments and agencies have generally instituted a practice of charging for access to data in order to recover the cost of building and sustaining the data under their control. Such practices may place the data beyond the reach of many researchers who could otherwise add value and make new discoveries from such access. In some instances, this barrier may prevent a particular line of research from being pursued, with the following consequences:

- the detriment of Canadian research and its global competitiveness;
- the lack of broader use by industry and commerce; and
- the loss of perceived value, which eventually leads to the decision to discontinue sustaining investment and subsequent data loss.

In large part, Canadian research is driven by the curiosity of individual researchers and is often manifested by self-interest and an ethos that lauds competitiveness over collaboration. Consequently, a reward system almost exclusively based upon individual recognition has emerged. In pursuing personal goals, institutional reward and peer recognition, the culture of collaboration and team contribution to the public good regularly takes second place. Therefore, the motivation to contribute to openly accessible data is a low priority for most researchers. In such a culture, data primarily become the means to individual recognition. Contrarily, the open access framework requires collegial contributions to higher value data in order to garner motivation driven by recognition equivalent to the peer recognition and reward currently attributed to the development of new knowledge. Open access and preservation require modified recognition and reward systems so as to meet the career aspirations, particularly of some younger researchers, through excellence in quality data generation or collection.

Training

Few researchers have had specific training in database development and preservation, and there is a reticence amongst many to assume responsibility for database management beyond their immediate interests. There is also little expertise, within scientific research institutions and agencies devoted to database management, which meets accessibility, security and preservation requirements. Clearly, there is a need to develop this training capacity, as training needs to be provided to scientific researchers and resident expertise needs to be available to complement such training.

Standards and Processes

Widespread interdisciplinary open access to scientific data requires adherence to standards and, therefore, such standards will have to be implemented independently of any one field. In contrast to this, most existing databases use discipline-specific approaches and architectures, permitting their effective use by researchers pursuing new knowledge in that immediate field, but representing a potential hurdle to any other potential users. While such databases could become openly accessible, the value to other research communities would likely be significantly diminished. However, re-architecting, restructuring and rebuilding existing databases, and re-engineering their interfaces to analysis, mining and application tools – not to mention their metadata – will require exceptional efforts of conversion for those established communities which have already developed and maintain large databases. This problem can be mitigated by the earliest possible adoption of a national data strategy, and consequent compliance with it through the migration of existing databases and datasets.

In many medical and social research areas, personal privacy legislation, including the ethical rules developed to prevent its breach, has inculcated a practice of destroying data, after being used for their intended purpose. The impact of this legislation is further compounded by concerns surrounding such issues as ethnicity, gender, lifestyle choice, wealth, etc., and, in all scientific domains, research that covers any aspect of national security and intellectual property. Database security, management, access controls and authorizations in most situations require a level of knowledge and expertise typically found only in security agencies. Processes will have to be developed, and possibly legislated, so ethics boards and security agencies can grant permission to the preservation and eventual archiving of such databases, if they meet specific standards of anonymity.

Archival Expertise

In Canada today, there are insufficient trained archivists (for the growing demand) with knowledge of data cataloguing, metadata standards and processes, preservation management and data value assessment. There is not only a requirement for the training of researchers in the architecture, management, standards and best database practices, but also for the supporting expertise to guide the development of sophisticated databases, and to address problems of database access, reuse and preservation.

Data archivists are currently accepted as valued research partners and consultants in an increasing number of projects that rely on large databases and datasets. However, they are still not widely regarded as essential to the research enterprise in research institutes and remain vulnerable to budget pressures, even more so when such “library overheads” require budget increases. The number of data archivists supported within institutes has been growing, but not in correlation to the rate of demand. The change to data-intensive research processes will accelerate as Canadian researchers promptly embrace this new research paradigm, while striving to remain competitive with their peers. This will exacerbate the shortfall of data archivists and managers, and challenge both the supply of these skilled resources and institutional budgets required to employ them.

Responsibilities, Systems and Tools

Ownership responsibilities for data acquired for a single project are conceptually simple, although not well covered by law. Of the various management regimes for intellectual property (IP) protection and management, databases have their own rights law in most countries. These laws generally resemble copyright law, but fall within their own domain. While the need to define a better IP regime has been identified, progress is rather sluggish at the World Intellectual Property Office (WIPO) and its allied national offices. Currently, there are no deadlines set to address key data IP issues, whether at the Committee on Copyrights and Related Rights' table or at the international meetings of WIPO. In these discussions, issues pertaining to large open access databases will represent an entirely new challenge and it may, therefore, take time for the international community to reach a consensus. Meanwhile, in the Canadian context, data ownership, custody and control are new to a majority of researchers. The NCASRD concluded that, in the immediate future, the only logical solution for database curatorship and custody is for the data contributor and database curator to assume shared responsibility of these functions. This can be transitioned to the international protocol framework, when relevant national and international data IP regimes are established.

At some time, either at the conclusion of active use of any scientific database, or whenever the data curator is no longer willing or able to exercise custody and control, there must be a mechanism to transfer such responsibilities to capable successors, whether other competent researchers, institutional libraries and archives, or provincial or national agencies or archives. Also, in cases where databases are deemed national assets, secure archives of such databases must be undertaken by a credible, trusted entity, such as Library and Archives Canada, to eliminate the risk of contamination, destruction or other loss.

To make databases truly accessible, metadata descriptions, technological platforms, access and mining tools, and management and maintenance systems must be stable and in common use. They should also be supportable for the entire period of the usefulness of the data or moved to an alternative environment, once stability, regular use or support can no longer be guaranteed. Similar protection for archival access must be applied, and should include the obsolescence and/or breakdown of storage media and storage drives, as well as access and maintenance programs.

Other Challenges and Opportunities

NCASRD participants also highlighted challenges in several other areas:

There are large reservoirs of existing data not in current use, but which either have substantial potential value for historical analysis and longitudinal studies (both to establish historic patterns and baselines for future research) or are considered to be of national historical significance. As the scale of these data collections is large, there is concern about the magnitude of the challenge in recovering and archiving such important artefacts. Complications arise because of unclear ownership, inaccessibility to the knowledge necessary to create the required metadata, unreadable media data storage or inaccurate information regarding completeness, as well as lost information as to their location. As valuable historic data have already vanished, and continue to do so at an ever-increasing rate, there is urgency in commencing recovery and archiving. This is also further complicated by the lack of archivist resources noted above.

From the legal perspective, owner liability, and IP control and infringement are also identified as impediments to open access. There appears to be no legal precedent to indicate the liability that any data owner or custodian assumes when the data are used by other researchers for purposes that are not within the control of the custodian. In general, it is believed that databases have seldom been registered as copyright, but it is self-evident that owners or custodians of valuable databases and datasets will increase the use of IP protection processes. Control through copyright and open access are compatible, in that open access does not preclude the granting and withholding of access permission in selected circumstances, nor does it preclude the pursuit of infringement, etc. However, they are not interchangeable, as they embrace different philosophies: exclusive access only to those with permission versus open to all. Liability, control and IP infringement issues must be resolved to eradicate this inhibitor to open access to scientific research data.

8 Consultation Recommendations

The NCASRD's recommendations directly result from the Forum. It is highly relevant to note that they have close congruence with the recommendations of the International Council for Science's (ICSU) *Scientific Data and Information*[7] report, published in December 2004, by the Assessment Panel of the Committee on Science Planning and Review.

Data Force Responsibilities

8.1 Getting started

The NCASRD has concluded that open access to scientific research data will, amongst many other benefits:

- transform the very processes of scientific discovery, through the ability to quickly access much larger, and more rigorous, complete and diverse datasets that have already been, or will be, assembled through public funding;
- accelerate the pace of knowledge development, through the reuse of data and the interlinking of diverse datasets;
- permit the study of far more complex systems and system interactions;
- open the opportunities for a substantial increase in national and international research collaboration; and
- result in a myriad of economic, environmental, ecological and social benefits in all domains of science, many of which will be unexpected owing to new interrelationships yet to be recognized.

The NCASRD initiative was designed to produce recommendations to the Sponsors on the justification for and steps to establish a Canadian scientific data access and preservation system. The first recommendation of the NCASRD is to create an organizational structure to begin the implementation of this initiative:

Recommendation 1 – Organizing

The Sponsors establish a task force (Data Force) to prepare a thorough national implementation strategy. Data Force should have representation from:

- **Canadian scientific research community;**
- **Canada's research granting councils (CIHR, NSERC and SSHRC);**
- **Canada's research infrastructure foundations and trusts (CFI, Genome Canada and equivalent provincial trusts);**
- **universities, colleges and research institutes that manage Canada's research infrastructure;**

- **government departments and research laboratories (both federal and provincial) that set public research policy;**
- **CARL;**
- **Association of Universities and Colleges of Canada (AUCC);**
- **CNC/CODATA ;**
- **CISTI;**
- **LAC;**
- **Statistics Canada;**
- **CΛNΛRIE Inc.;**
- **Privacy Commissioner of Canada;**
- **representatives of student and professional organizations that participate in planning research training curricula; and**
- **representatives of the general public who will ultimately benefit from the public good of publicly funded research.**

The mandate of Data Force would be to guide and oversee a small implementation secretariat to:

- **commission a pilot data access project (Data Project) to illustrate the concepts and values of this Report;**
- **plan and supervise the formation of a permanent Canadian data access organization (Data Canada);**
- **secure the long-term commitment to federal financing of Data Canada;**
- **develop a data access strategic plan (Data Plan).**

8.2 Building national support

There is growing international momentum for making publicly funded scientific research data as accessible as possible to enable the realization of substantial economic, ecological, environmental, social and societal benefits, and considerably accelerate the growth of scientific knowledge. Nevertheless, the NCASRD identified insufficient public understanding and political will as key inhibitors to the establishment of scientific data access as a national priority. National support is necessary to implement the many systemic, legal, managerial, governance and financial changes required for Canada to join the most aggressive OECD nations as a leading player in the new data-intensive research paradigm. Clearly, Canada will want to be among the leaders, as well as to benefit from the advantages setting them apart in their rate of innovation-driven, national economic growth. The current lack of will is evidenced by the delay in the implementation of the joint SSHRC-LAC *National Data Archive Consultation* recommendations. Success in establishing accessible scientific research databases has only been achieved in a relatively few major national and international initiatives, such as Genome Canada, as part of the global genome initiative, the global astronomy research community (with DPOSS, GEOSS etc.), the Centre for Global Research and Education in Environment and Health, Bio-molecular Interaction Network Database (BIND) and the Canadian node (CBIF) of the international bio-diversity project (GBIF). It is important to build on such pathfinder initiatives, harnessing existing knowledge across these communities, as well as in the Canadian research libraries, CISTI and Statistics Canada.

Recommendation 2 – Educating

Data Force, together with the leaders of sponsoring organizations, immediately begin fostering awareness amongst political, institutional and public opinion leaders of the:

- **paradigm shift enabled by access to massive data resources that is occurring in scientific research globally and within Canada;**
- **need for Canada to be among the global leaders in the transformation of the research enterprise, in order to retain and strengthen our economic competitiveness and scientific excellence over the long term;**
- **social, medical, ecological, environmental and economic benefits that will accrue in accelerating the pace of scientific discovery;**
- **educational benefits derived from the ability to place learning in a real-life context and enabled by open data access; and**
- **need for concerted action to drive the cultural, legal, managerial and political changes essential to establish Data Canada.**

8.3 Pilot Project

The implementation of a national system that can support the capture, storage, classification, access, and archiving of all important research databases and datasets, within an integrated management process and governance structure spanning many diverse institutions and organizations, is clearly too big a feat to be tackled at once. Without a demonstration of Canada's ability to implement a project with all the required elements, showing the ensuing benefits and leverage, full implementation will not be possible. This is especially true when considering the possible costs, resource requirements, organizational linkages, benefits, national and global standards, legal and international collaboration frameworks, as well as technological solutions and their evolution, all of which have yet to be explored and clarified. This complexity of the "big picture" requires a more evolutionary, learn-as-you-go approach, which was recommended by the NCASRD.

Further, the rollout of Data Canada requires sound coordination. In addition to the strategic responsibilities and overall coordination, initially carried out by Data Force and later transferred to Data Canada, there are areas where specific expertise is required to agree, approve and coordinate activities. These areas include:

- national scientific data policy development;
- technology solutions for an evolving heterogeneous infrastructure of computing and data storage environments, and access and management systems, practices and protocols;
- management protocols for full-life data management;
- evolving metadata standards and practices; and
- data quality and integrity standards, and assessment processes and procedures.

Recommendation 3 – Funding

Sufficient funding be provided to Data Force to support the implementation of an open access database pilot project in 2005. This project should be designed to show how databases,

when linked, can lead to substantial knowledge breakthroughs. It would also include solutions to the challenges posed by such areas as policy, technology, infrastructure, system management, data and metadata quality, integrity and security. This does not mean a new scientific research project, but rather, the compilation and integration of a compelling example already in progress or completed.

This recommendation is designed to illustrate both the Canadian capacity to establish the required infrastructure and integrated management processes, and the power of trans-disciplinary database integration as a much more efficient methodology for scientific research. It would be highly beneficial if the project chosen complemented existing data-intensive research activities. Many have already created relatively open, accessible databases, and the pilot project would help some of these prior investments. The NCASRD Project Management Group (PMG) suggests a project targeting environmental data, bringing focus to the data management challenges in environmental sustainability, could be chosen. Such a project could involve all granting councils, CFI and NRC. Such a project would complement GBIF and build on its international experience. It may also encourage the modification of this existing project and demonstrate Canadian leadership.

The PMG also believes a budget in the order of $200,000, provided by the three granting councils, CFI, NRC and LAC, would permit the set-up of such a demonstration project with funding for a one-year term. It also recommends funding for the first year of the order of $100,000 for a small secretariat to run this project, to support Data Force in meeting its three above-mentioned obligations and to establish the basis for the operational management of Data Canada.

Data Canada Responsibilities

8.4 Management Capability

NCASRD participants identified the need for Canada to be proactive in the growing international scientific data open access and preservation community. With Canada producing a mere 4.1 per cent[8] of the world's scientific knowledge – annual per cent of research papers produced – and outputting 5.6 per cent[9] of the top 1 per cent of papers ranked in terms of citations, our capacity to benefit from scientific discovery partnerships, based on global data, is greater than most of our larger international competitors. Growing Canadian knowledge contributions can only be achieved by accessing all the available relevant global data and using them to support the complex problems we now need to investigate. This change in practice will increase our reliance on the establishment of global standards. Further, history has shown that Canada can help influence standards adoption, as we are predominantly seen as a competent and technically knowledgeable arbitrator in competing approaches, standards and protocols. The ICSU Priority Area Assessment on Data and Information report, *Scientific Data and Information*, strongly argues in favour of international action on a long-term strategic framework for scientific data and information. It recommends the formation of a Scientific Data and Information Forum (SciDIF) to oversee the development of this framework. Canada, through Data Canada, in cooperation with CNC/CODATA, should be represented on the SciDIF.

Within the global open access community, it is vital for Canada's research community to be engaged in the establishment of international agreements on topical issues, including: sharing protocols; metadata standards; and private versus public versus non-governmental organization control of databases and datasets produced at public expense. We also need to ensure that international agreements, standards and practices address Canada's uniqueness, such as our research culture, legal frameworks, data access practices, personal information and national security.

Recommendation 4 – International Participation

Data Canada establish a management capability that can monitor and intervene in international open access fora to protect Canadian interests, and assist the international community in promoting agreements, standards and policies that support best access, sharing and preservation practices compatible with Canadian needs.

8.5 Privacy Protection

Some legislation, especially the safeguarding of personal information, like *PIPEDA*, –carries serious implications for access to crucial data collected at public expense. Such data must be used only with explicit consent, for a specified purpose, and are usually required to be destroyed once that purpose is fulfilled. However, these may be key data and, when fully anonymized, will have significant value, particularly in health, demographical and social research. Due to *PIPEDA* and similar legislation, as well as their interpretation by enforcement bodies, numerous data resources are being lost; data that may have helped provide lifesaving research or other valuable knowledge. Thus, NCASRD participants believe that fully anonymized data should be preserved, whenever possible, and made accessible, under appropriate controls, for research qualifying as an enhancement to the quality of life and health of Canadians, as well as the global population at large.

Recommendation 5 – Ethics

Data Canada initiate consultations among the privacy commissions, the National Council on Ethics in Human Research, the Canadian Association of Research Ethics Boards, data librarians and archivists, and Statistics Canada to identify legal barriers to access to scientific data. Such consultations should result in proposed modifications to information privacy laws or their legal interpretation, to ensure high-value, publicly-funded data, properly protected for confidentiality, are preserved, secured and made accessible with appropriately managed access controls. Consequently, the data may be used for research purposes other than their original intent and for which informed consent may not have been sought.

This recommendation is premised on the ascendancy of the interests of the broader public over the residual rights of any individual in the anonymized data that have been gathered within the constraints of specific informed consent. This recommendation applies only to digital information and not to any other information or physical instantiation.

8.6 International Data Sharing

Legal and ethical concerns have effectively prevented international collaboration and data sharing of databases held by the US functional Magnetic Resonance Imaging Data Center (fMRIDC). The OECD Follow Up Group Report on *Issues of Access to Publicly Funded Research Data* states "fMRIDC has been hesitant to accept data from non-US settings because of concerns regarding Institutional Review Board compliance." It also claims "researchers submitting or requesting data across national boundaries may find it especially difficult to act in accordance with the various ethical guidelines that exist in different countries." It is clear that pressure will mount to harmonize privacy rights legislation between countries with similar privacy concepts, at which point broader sharing of critical health and other data gathered under rules of informed consent can be assembled and accessed across international boundaries, without compromising regulations and laws designed to protect personal information. Undoubtedly, Canada will have to resolve its own questions vis-à-vis the sharing of health and demographic data deemed to be private or a threat to national security, if it wants to avoid impeding its rate of scientific discovery.

Recommendation 6 – Privacy

Data Canada should initiate a review of the Personal Information Protection and Electronic Documents Act, as well as other related legislation to identify inconsistencies that would prevent international data sharing with countries whose collaborative research projects and database sharing practices are expected to be high – early analysis should focus on the US and EU. It should also work with Canadian privacy legislators to align such legislation to permit fully compliant data sharing between specific countries.

8.7 Nationally Important Data

There are known instances where nationally important databases and datasets have been lost or contaminated, due to the failure of the data custodian or archivist (if any) to take adequate data preservation measures. Hackers, computer viruses and worms, and other malicious electronic attacks may have also caused this loss. To heighten the security of such national assets, it is highly desirable to place a back-up copy of the database in an archival system with increased protection.

Recommendation 7 – Archiving

Databases and datasets, determined by Data Canada (e.g., through a peer review panel or special committee, or relevant professional society or association) to be of national importance, be deposited and secured at LAC.

8.8 Liabilities and Benefits

Key databases and datasets compiled by researchers or research teams must be subject to review, prior to being certified as reliable data sources. While such review underpins the willingness of the research community to use such data, their reuse may give rise to protectable IP and to new knowledge, potentially resulting in the development of products, goods or services. Such use, under current Canadian law, could pose a liability on the data custodian and, perhaps, the granting agency, the government providing the resources for any such agency and the employer of the researcher(s). International collaboration in database and dataset construction may also create a liability for foreign contributors under Canadian law and extend Canadian accountability to the legal systems of other countries.

Furthermore, the owners of databases and datasets, compiled at any particular time, cannot anticipate the reuse of such data, an uncertainty which will raise concerns about possible future, unanticipated liability from unexpected sources. It is, therefore, crucial that the responsibility and liability of data custodians are well defined in law. To the greatest extent possible, this legislation should also encourage Canadian leadership through access to scientific data.

Recommendation 8 – Liability

Data Canada establish an expert panel to examine the Canadian and international legal frameworks concerning responsibility and liability for databases and datasets, and task them to propose a new Canadian legal framework compatible with evolving international legal frameworks. The aim would be to balance the liability of data custodians and their institutions against the social benefits resulting from open access to such data, in order to protect the custodian against liability derived from unexpected future uses of the data.

8.9 Accessibility of Anonymized Data

As noted in 8.5 – Privacy Protection, data generated from personal information or through specific access to other confidential or secure databases must usually be destroyed, once they have fulfilled their intended purpose, granted by way of informed consent or other explicit permission. This destruction of potentially valuable data represents a substantial cost to the research enterprise, prevents future longitudinal impact studies and denies the research community potentially valuable data. While personal information must be carefully protected, data collected at public expense should be available for the greatest possible public good. In the future, data that can no longer be traced to the person, or from which any individual or small group cannot be deduced, should become accessible for other research purposes

Recommendation 9 – Anonymization

An expert panel be appointed to examine the legal issues surrounding data anonymity and secure data practices that would prevent infringement of an individual's privacy, if made accessible for other research. The panel should identify limits to the applicability of informed consent, when no possible identification, or deduction of the individual or small group remains feasible. Should opportunities be found to permit anonymous data re-use, the panel will propose appropriate changes to legal and regulatory practices.

8.10 Databases at Risk

Institutions and organizations manage huge volumes of data, which are growing exponentially. Some databases already contain upwards of hundreds of terabytes and some institutional repositories are now managing petabytes. Data and their associated metadata exhibit the full spectrum of quality, integrity, ownership control, accessibility, version control and preservation management. While Data Canada should consciously focus on establishing a competent data access and preservation service as it goes forward, there is also a clear need to recover or upgrade existing databases and datasets to meet open access and preservation requirements. Urgent action within the existing research environment is required to address the vulnerability and poor accessibility of both active and non-active databases.

Recommendation 10 – Databases at Risk

Data Canada establish a fund to preserve, and improve the accessibility of existing high-value, "at-risk" and/or critical databases identified by peer review panels as having significant current, future or historical value.

In order to understand the magnitude of this recommendation, it will be necessary to identify existing databases with the highest value, either as national assets or sources of baselines, for other research initiatives, and also assess the degree to which they are "at risk." With the rate of loss increasing rapidly, it is important to commence such identification and prioritization as soon as possible, before more national data assets are lost.

8.11 Standards

Standards compliance and quality reviews will normally be initiated by the data custodian and funded as part of the research grant application's data management segment. However, to satisfy the broader community, reviews may also be initiated by any user group upon application to Data Canada. The direct costs associated with externally requested reviews should be borne by Data Canada. Further, in the case of dynamically changing databases and datasets, standards compliance and quality reviews may be triggered whenever any of the following conditions apply:

- the method or process of data capture is changed;
- new contributors are added; and
- more than a specified proportion of the data has been altered or added.

Any database that fails to meet peer review standards of quality and integrity shall be so designated in its metadata, and all known prior users shall be notified of the qualification.

Recommendation 11 – Criteria and Quality

Data Canada work with its research partners to establish a function within Data Canada (and its international counterparts) to formalize assessment criteria for data quality, as well as define processes to measure data quality and integrity.

Responsibilities of Funding Agencies

8.12 Provision for the Training of Researchers

Open access databases place a number of additional requirements on researchers. New responsibilities will exist, due to the architectural structuring of the data, custody and control responsibilities/liabilities, access management, metadata creation and maintenance, peer review and preservation. Few researchers – mainly those involved in the largest current data-intensive mega-projects – have the necessary knowledge to undertake these custodial responsibilities. In particular, only in the largest projects has it been possible to involve data archivists from an early stage, so few researchers have been able to learn from those with such extensive knowledge.

This inadequacy needs to be addressed, by providing researchers, especially principal investigators, with education about data management, ownership responsibilities, metadata standards and preservation solutions.

Recommendation 12 – Training Researchers

All organizations that fund scientific research provide specific funding for the training* of all principal investigators‡ in best practices of database selection, management, rights management and data curatorship, metadata standards and other important issues, so access and preservation can be built in to the data acquisition and storage plans from the outset.

* where agencies do not provide funding for training, such agencies should work with their government departments to secure funds and implement training programs for all principal investigators who are recipients of or apply for their grants.

‡ including other researchers that are involved in the creation of databases and datasets.

8.13 Data Management Plan

While a limited number of major health, biology, geology, material science and astronomy projects have included the creation of discipline-targeted accessible databases, only SSHRC grant applicants are required to present a data management plan. And, even where plans are presented to SSHRC, the human and financial resources to implement such plans have been systematically denied. All research council and funding agencies must rectify this shortfall by making it an application requirement to provide a data management plan, and for the review and decision panel to include the quality, integrity, accessibility and cost effectiveness of the data management plan as decision criteria for acceptance. Data management plans designed to create, or complement national databases or contributions to international ones should be awarded higher ratings than those only serving the interests of independent research projects. Performance against the data management plan must also become a reporting requirement and subsequent grant decisions should consider prior satisfaction as a decision criterion.

Recommendation 13 – Data Management Plans

Research councils, and all other public-sector research funding agencies and departments require that project and grant applications include a data management plan, as well as specifically identified funding that will ensure quality, integrity, accessibility and accountability. A funding condition should be the inclusion of a well-constructed plan for data acquisition, management, access and preservation. Adherence to such plans should also become a non-competitive performance metric for the project and gateway for subsequent grant applications. Councils should recognize these as added costs to the main thrusts of research projects.

8.14 Core Funding

Projects building databases and datasets, with value beyond their immediate and initial requirements, will demand the resources to make them accessible and preservable. Moreover, the custodians of multi-use databases will require ongoing resources to sustain and update these databases. Therefore, it is clear that long-term stable and non-competitive funding is required to ensure such databases are supported and remain accessible over long periods, from institutional repositories or LAC.

Recommendation 14 – Resources

Federal and provincial government departments, agencies and ministries that fund scientific research establish long-term stable, non-competitive core budget allocations to provide research institutions, organizations, and agencies with the resources to preserve all important databases (historic, current or potential high value). The federal government also provides additional and sufficient funding to LAC, ensuring the long-term archival preservation of all important databases and datasets.

8.15 Peer Review

A perpetual challenge with databases is the assurance of their quality and integrity. When scientific research uses data compiled solely for a single project, the data schema is generally designed for that project only. The quality of the findings is fully dependent on the quality of the data, and both are part of the peer review process of scrutiny and validation. However, in the circumstance where the data are part of a public database, there will be an inferred assumption that the data are of the best quality, and free of systematic or random inaccuracies. While it is evidently the researcher’s responsibility to seek assurance that the data are suited to and of sufficient quality for the purpose of the research, the research community as a whole, in addition to the data custodian, has the underlying function of ensuring data quality and integrity.

Recommendation 15 – Peer Review

Databases and datasets in use or expected to be used in multiple research initiatives, including their metadata, be subject to peer review, with the evaluation becoming part of the metadata.

8.16 Time Limits

NCASRD participants recognized that databases will be developed which provide new critical insights to researchers. They should be given the opportunity to have exclusive access to such data for a reasonable, but limited, time, so new knowledge can be prepared for publication before other researchers can start their own discovery processes. Exclusive access must be limited in time with the data becoming publicly accessible shortly thereafter. This eventual public accessibility should, however, be controlled and restricted when sharing would infringe Canadian law or international protocols and agreements.

Recommendation 16 – Time Limits

In collaboration with Data Canada, funding agencies and departments set limits for the length of time data custodians may deny open access to their databases. This time should be fair and reasonable in the prevailing circumstances. After the specified period, the database must be made publicly accessible subject only to constraints imposed by law, or international protocols and agreements.

Universities and Researcher Responsibilities

8.17 Reward Systems and Recognition

Academic performance and reward systems are based on excellence in teaching, research and public service. Research performance is evaluated on the quality of research, the publication and citation record, patents awarded, success in winning research grants and securing industrial partnerships. In this current system, there is no recognition for leadership in the compilation of, or major contribution to, high value, open access databases and datasets, nor in the development of tools that enhance the value of data (like, for example, database combination as well as access and mining capabilities.) To change the research culture, the academic reward system must include recognition for substantive contributions to scientific research data and their utilization.

Recommendation 17 – Rewards

University faculties, the professoriate, and other academic research units extend the recognition and reward systems for researchers to include excellence in contributions to scientific data, and the development of tools for improved data management and use, as an important performance indicator.

To highlight this new performance indicator, universities, the granting councils, CFI and other scientific research funding agencies should consider collaborating with industry and professional associations to establish awards of excellence in contribution to scientific research databases. Awards of this type demonstrate to the research community the value of such activities and demonstrate to the general public the value that the recipients have created through database development and the resulting acceleration of new knowledge.

8.18 Training of Specialists

With growing consensus on the need for specialists in scientific database design, management, access and preservation, the training currently offered to experts in the field is inadequate. It will not meet the rapidly increasing demand arising from the implementation of open access databases. Immediate action is essential to provide these expert resources at the earliest possible time.

Recommendation 18 – Creating Specialists

Post-secondary institutions increase their intake of students in Information Science, and the teaching of database access and preservation to address the shortage of trained digital librarians, managers, curators and archivists.

[7] "Scientific Data and Information". A report of the Committee on Science Planning and Review of the ICSU dated December 2004 – ISBN 0-930357-60-4 – which can be downloaded from: http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/551_DD_FILE_PAA_Data_and_Information.pdf

[8] UK office of Science and Technology (Department of Trade and Industry) Survey of National Research Productivity Metrics – PSA target metrics for the UK research base – Oct 2004: see (http://www.ost.gov.uk/research/psa_target_metrics_oct2004.pdf); Section 2.03

[9] *Ibid*, Section 3.07

9 Roadmap to Access

Phase	Action	Target Date	Responsibility
0	Concurrence of CFI, CIHR, LAC, NRC, NSERC, SSHRC to create “Data Force”	Apr 2005	NCASRD
1	Data Force in place with budget	Jun 2005	NCASRD
1	Pilot Open Access Database Project approved and funded by Data Force	Oct 2005	DF
2	Major “at-risk” historic databases identified and prioritized	Dec 2005	DF
2	Universities announce increased intake targets for Information Science	Feb 2006	U
2	Pilot Data access and management training courses/ seminars for researchers completed (planning completed Nov 2005)	Mar 2006	GC
2	Grant application Data Management Plan requirement in place	Mar 2006	GC
3	Data Canada established and initial funding budgeted by Industry Canada	Mar 2006	DF
3	Data Canada active participation in International Open Access Data Fora	Mar 2006	DC
3	Data Canada Strategic Plan approved by Federal Government	Mar 2006	DC
3	Highest priority “at-risk” database recovery projects approved	Mar 2006	DC
3	Exclusive access policies for data owners established	Jul 2006	GC/DC
3	University/faculty reward systems to include recognition of database construction and management as a valued contribution to performance	Jul 2006	U
3	University increased intake for Information Science achieved	Sep 2006	U
3	National Researcher Data Access, Management and Preservation Training Program started	Sep 2006	GC/DC
3	Privacy Protection consultations initiated	Sep 2006	DC
3	International Privacy Protection harmonization alignment initiated	Oct 2006	DC
3	Data Owner Liability evaluation initiated	Oct 2006	DC
3	Pilot Open Access Database Projects completed and accessible	Oct 2006	DC
3	Catalogue of “National Asset” databases completed	Dec 2006	DC
4	Data Canada funding committed in Federal Budget	Feb 2007	IC
4	Data Owner Liability recommendations to Federal Government	Apr 2007	DC
4	National Researcher Data Access, Management and Preservations Training Program transition to sustaining state	Sep 2007	GC/DC
4	Privacy Protection recommendations to Federal Government and Privacy Counsellors	Sep 2007	DC
4	International Privacy Protection harmonization alignment recommendations to Federal Government	Dec 2007	DC
4	LAC replication of “National Asset” databases completed	Dec 2007	DC/LAC

Note: Abbreviations in Table

NCASRD – Chair of NCASRD in collaboration with the Presidents of the Granting Councils, CFI, NRC AUCC, and with the assistance of the National Scientific Advisor to the Prime Minister.

DF – Data Force
U -Universities

DC – Data Canada
IC – Industry Canada

LAC - Library and Archives Canada
GC – Granting Council

10 Appendices

10.1 Framework of a Vision for 2010 – Enhanced Access to Scientific Research Data (November 2004)

Enhanced Access to Scientific Research Data
Framework of a Vision for 2010

Background Paper for the
National Consultation on Access to Scientific Research Data (NCASRD)
November 2004

Introduction

This document is a collection of short articles or descriptions of the nine thematic achievements or opportunities that emerged from the 11 June 2004 meeting of the Task Force overseeing the *National Consultation on Access to Scientific Research Data* (NCASRD). These descriptions are best understood in conjunction with the report on that meeting which may be found at the NCASRD website (http://ncasrd-cnadrs.scitech.gc.ca/home_e.shtml).

The purpose of the Task Force meeting was to preview the consultation process for the NCASRD Forum planned for 22-23 November 2004. In particular, one of the meeting goals was to gain a broad understanding of the opportunities presented to Canada by ensuring access and stewardship of scientific research data. To do this, Task Force members were asked to envision the year 2010 and the achievements that would have had to have been accomplished by then to reach the desired state of access to data. These achievements or opportunities, depending on how they are viewed, were summarized by the various writers or teams identified below.

1	*International Leadership in Data Management*	Peter Pennefather Paul Uhlir
2	*Removal of Discipline Boundaries*	John Spray
3	*New Methodologies*	Ellsworth LeDrew
4	*Enhanced Education*	John ApSimon
5	*Data Stewardship*	Chuck Humphrey
6	*Capacity Building for Data Archiving*	Chuck Humphrey
7	*Network Effects*	Guy Baillargeon
8	*New Discoveries*	Robyn Tamblyn (Health) Steven Jones (Genetics/Genomics) Andrew Pollard (Eng/Instrumentation) Ellsworth LeDrew (Environment)

1. International Leadership in Data Management

Beyond Best Practices for Stewardship of Access to Data Generated by Publicly Funded Scientific Research in Canada: Leading the way for global access to public research.

Introduction.

Overview.

Knowledge produced by national research systems, if disseminated widely, is a global public good (Pang et al Bull, WHO 81: 815-820, 2003). Canada's publicly supported scientific research funding agencies have a duty beyond simply distributing research funds in a systematic way. They need to be able to articulate a vision for their programs, ensure that identified priorities are advanced, and demonstrate explicitly that the results of those programs serve the public. These stewardship roles are essential for continued public support.

A key stewardship issue is ensuring access to data and information produced by publicly funded research. Technological barriers to accessing and sharing data globally continue to fall with continued national and multinational investment in cyberinfrastructure dedicated to the production, management and use of data and international frameworks are being developed to deal with concerns and problems that emerge from promoting open access to data (Arzberger et al. Science 303: 1777-78, 2003).

Here we will discuss how the Canadian research councils can foster international leadership by Canadian researchers in stewardship of the data that their funding programs help to generate. We contend that this stewardship must extend beyond the stewardship of the upstream, unprocessed data to their integration into downstream information (data received and understood) and other knowledge products.

We propose that new opportunities are emerging for managing primary research data and that finding new ways of managing those data on a research system-wide scale is an area where Canadians can excel. This could help Canada build a capacity to set standards of excellence for managing access to data resulting from publicly funded scientific research and to demonstrate explicitly how such management can enhance return on public investment in research. Supporting a national strategy for new scientific data exchanges and the collaborative development of information based on those data, could seed emergence of innovative goods and services focused on data interpretation, integration and reuse that are valued nationally and internationally.

Background.

Canada is a small nation with relatively few sources of public research funding. These sources nevertheless support a broad spectrum of research projects that generates an ever-increasing stream of data. These data are being continuously transformed into globally useful information and applications. This skill at transforming data into information could be applied more broadly if there were international standards for making global data generally available and usable. It thus is within the ability and interest of Canada's scientific research systems to provide leadership in demonstrating how research data could be made more widely available.

With the *National Consultation on Access to Scientific Research Data* (http://ncasrd-cnadrs.scitech.gc.ca/ncasrdfall_e.shtm), there is now an opportunity to define a Canadian perspective on how research data, and the information they generate, are currently being managed here and around the world. There is also an opportunity to explore new digital approaches (especially those involving the internet) for developing and establishing standards for shared infrastructure for enhancing access and use of that global heritage.

Such infrastructure is needed not only for enhancing return on Canada's investment in publicly funded research, but also for quantifying the information output of research enterprises in Canada and indeed globally. The latter is necessary for ensuring that research system priorities are in fact being advanced and for identifying gaps between the needs perceived by government and the public and what is actually being produced.

There are also new opportunities emerging that demand new ways of doing things. Up to now, much of the publicly funded research in Canada has been driven by individual researchers (PIs) who could be called upon to explain what they had done and where their data could be found. However, new models of collaborative, targeted, and programmatic research are being made possible through rapid advances in Internet communication technology.

Examples include the open-source software movement, open public-domain data archives and federated data networks, community-based open peer review, collaborative research Web sites, collaboratories for virtual experiments, virtual observatories, and open access journals, among others. Taken together, these emerging capabilities represent aspects of a broader trend toward both formal and informal peer production of information in a highly distributed open networked environment. Such activities are based on principles that may be more accurately characterized as a "public scientific information commons," rather than as proprietary information authored by individual PIs, and that reflect the cooperative ethic that imbues much of public science. These developments have given rise to unprecedented opportunities for accelerating research and creating wealth based on adding value to data and non-proprietary information produced through public funding (Uhlir, 2003).

Although some of these efforts are entirely driven by unpaid volunteers, the major roles played by employees of HP and IBM in the open source software development are a part of those companies business development strategy. Responsibility for both the production and storage of primary data resulting from e-science projects is also often distributed over many organizations. This presents new challenges and opportunities for describing and disseminating research results (Hey and Trefethen 2004, http://www.ecs.soton.ac.uk/~ajgh/DataDeluge(final).pdf).

It is becoming clear that even for PI-driven research, access to research data and the derived information does not simply mean making the PIs' interpretation available in the form of a publication or storing raw data in a file cabinet. The primary data itself is being transformed from analog artifacts digitised after production to data artifacts that are digital from the start. It is a challenge to ensure that the media in which the raw digital data are stored remain accessible.

Also for data to be usefully accessed and for information to be transmitted to those who could benefit from it, there needs to be appropriate documentation of what data are available, why and how they were produced, and what benefits might accrue from their use. In many cases, stewardship of such documentation (metadata) is even less developed than stewardship of the data themselves, but equally important. The increasingly digital nature of primary data raises new opportunities for the collection of metadata (http://www.ifla.org/II/metadata.htm)

International Initiatives on Public-Sector Research Information.

A review of the data access regimes in each country is beyond the scope of this background paper. At the international level, however, the access regimes for public research data typically have not been particularly open. Nonetheless, there have been some notable examples of general policies on open access, including the 1991 "Bromley Principles" on the full and open exchange of global change research data in the U.S., the 1996 Bermuda Principles on the Release of Human Genome Sequence Data, the 1997 ICSU-CODATA *Principles for Dissemination of Scientific Data*, and the 2004 OECD Ministerial initiative on access to data from publicly-funded research.

Notable examples where access is open and that have been particularly successful include the GeneBank, the International Virtual Observatory (IVO), and the Global Earth Observation System of Systems (GEOSS). But these are driven by highly motivated, well integrated research communities who all see a clear benefit to individual research programs from pooling data and who have established an extensive infrastructure dedicated to supporting their shared goals.

The open access movement in the publication of scientific papers also is stimulating a re-examination of how scientific information generated by publicly funded research is distributed. Although this is a much larger issue than can be considered adequately here, it is useful to note some of the most recent legislative and policy initiatives that provide a window on where policy development is going with regard to the sharing of research results.

For example, the NIH has recommended to Congress a new plan that if passed into law would mandate that all papers generated using any NIH funds must be made freely available through PubMed within 6 month of publication. The UK is looking at another solution where institutional libraries will be mandated to support access to self-archived versions of papers published by scientists at those institutions. These proposed legislations are putting pressure on publishers to adjust their business models and allow authors more discretion in circulating their papers. (see http://www.earlham.edu/~peters/fos/newsletter/08-02-04.htm).

There are now over 1200 scholarly journals provided under open access conditions on the Internet, including some notable initiatives such as the Public Library of Science and BioMed Central. Policy principles on open access to publicly funded journals were issued in both the United States and Europe in 2003 through the "Bethesda Principles" and the "Berlin Declaration." In 2004, many professional society journal publishers produced the "DC Principles," (Farnham and Brinkeley Scientist, 18 (13):8, 2004) which also recognized the imperative of broad access to the scholarly literature produced from publicly funded research. Some related initiatives also have been established already in the US for pre-prints and e-prints of journal articles (e.g., the Cornell arXiv, originally established for high-energy physics and now expanded to include other areas of physics, mathematics, computer science, and computational biology), for individual research articles and other information resources (e.g., the Social Science Research Network, the MIT D-Space initiative), and for university educational material (e.g., MIT's OpenCourseWare).

There is thus a movement away from merely entering conclusions and interpretations into the public record through publication. These initiatives are focusing on how to make papers broadly accessible though electronic repositories. This means that there is an opportunity to link those now freely accessible papers to repositories containing the primary data used in generating the published interpretations of those data.

Open access initiatives in developed countries frequently are being designed with the needs of developing countries expressly considered. In addition, new open access journals are being established within developing countries themselves. Here, Canada is already leading the way. For example, the work of John Willinsky at UBC in Vancouver, which created the Open Journal system for running a peer-reviewed journal, has great promise for supporting such efforts. The Ptolemy and Bioline projects at the University of Toronto are trying to increase the exchange of research papers and data between resource rich and resource challenged nations.

There are also initiatives that deserve mention that are aimed at linking data and papers to interpretation so as to reduce the research-theory, research-policy, and research-practice gaps. Examples include the Cochrane and Campbell initiatives that focus on medical and educational research papers, respectively. Canada boasts a number of leading centres of research on the topic including: the Global Centre for e-Health Innovation in Toronto, the Public Knowledge Project in Vancouver, and the Center for Collaborative Research in Effective Diagnostics in Sherbrooke.

A Leadership Role in Global Stewardship of Access to Research Data for Canada

What Does Canada Have to Offer?

The UK Office of Science and Technology has recently published a useful survey (http://www.ost.gov.uk/research/psa_target_metrics.htm) of national research productivity metrics. It shows that Canadian scientists exceed performance standards of scientists in other G-8 nations. When normalized to GDP, Canadian research papers are second only to those from the UK in term of citations. Rates of citations and publications per researcher are again second only to those of UK researchers and much better than those of US scholars. Canada ranks first in G-8 nations in terms of the impact of highly cited papers.

When normalized to GDP, Canada's return on investment (in terms of publications and citations) is admirable. Its output, however, is only a tiny percentage of the absolute global research product. Canada generated only 4.1% of the world's ¾ million annual research papers and 5.6% of the output of the top 1% of papers ranked in terms of citations. Canadians make up only 0.5% of the world's population and generate only 2.4% of the $32 trillion global GDP. With a national GDP of around $0.8 trillion, Canada invests $15 billion (1.9%) in gross expenditures on research and development (GERD). This investment is only 7% of that of the US GERD.

Canada spends the same fraction of its GDP on research within public institutions (0.8%) as the US, but an additional 2% of the US GDP goes to private research, twice the Canadian percentage and more than 30 times in actual dollars ($7.5 vs. $220 billion). This discrepancy is challenging. If Canadian publicly funded research systems want to ensure that their investments in research capacity development lead to international leadership and national advantage, they need to encourage development in targeted areas where innovation can compensate for the relatively small size of this public research investment. We suggest that stewardship of research data and associated information is one area where Canada has the potential for international leadership and for obtaining considerable return on any such public investment.

What Are Canadian Interests?

Canadians can be justifiably proud of the excellence of individual Canadian scientists. But at the same time its needs to be recognized that Canada is a small country whose distinct national interest can be buried by those of its much larger neighbour to the south. Because of the ready access to the US scientific arena, Canadian scientists are motivated to define success in US terms. They compete with US colleagues for “air-time” in the same US-based high-impact journals and for cracking conundrums promoted at large US-based scientific meetings. Measuring output in terms of citations will bias the enterprise towards US interests since a large percentage of papers come from US-based scientists and they will quite reasonably cite other papers that are relevant to their interests. Canadian research funding programs have a responsibility to ensure that the steering effect of Canada’s proximity to the US does not compromise the need to create a system that serves Canadian interests.

Despite Canada’s outstanding research productivity, disturbing trends are evident in the UK-OST statistics. Publications per researcher and global publication share are falling. Many indicators of research productivity have a long lag between inputs and outputs, however, so these trends may not have bottomed out. Many scientists are being rendered inactive by an inability to fund even excellent grant proposals. Very good proposals are now almost routinely rejected due to lack of funds. Although Canadian research programs distribute studentships, fellowships and salary awards based on an ability to publish papers and write very good research proposals, and although Canadian research institutions recruit professors based on similar criteria, Canada is losing a capacity to support these valuable researchers adequately.

This breakdown of our research capacity-building “farm system” perhaps reflects our attempt to emulate research methodologies prevalent in the US, where the funds available are orders of magnitude higher. Up to now, established scientists have managed to do more with less and keep up. But there is a disturbing trend in Canada to support big projects by individual PIs so as to allow some Canadian scientists to “compete” at a level comparable to their US colleagues. In the US, big dollars are delivered to those individuals who can make a big impact (as defined by the US funding agencies and the US scientific community).

To be applied in Canada with its limited funding resources, this “overwhelming force” approach needs to be highly selective. This selective approach weakens the ability to make full use of the installed base of Canadians researchers holding publicly (provincially) funded positions created with the expectation of being able to receive research support from federal granting agencies. By developing an infrastructure—broadly defined—that permits secure, but distributed, collaborative production and development of scientific information, the Canadian research funding agencies could help fulfill their research capacity stewardship responsibility. This would permit development of more collaborative research programs that take advantage of Canadian distributed research infrastructure resources.

New Internet-based collaborative research environments may be one highly significant area whose development could provide considerable advantage to Canadians and a higher return on public research investments. Such infrastructure would support both individual and collaborative research within Canada while at the same time providing an environment for recruiting researchers from around the world to share in driving forward research areas deemed important to Canadians. If Canada can set standards of excellence, scientists may secure access to data being generated around the world and apply themselves to the hard but less expensive work of interpretation.

Such a strategy will require development of a national strategy for enhancing collaborative and rationalized (accredited) access to research data and information. This need not involve loss of control by the individual scientist over access to their own data. Rather, it would simply implement standards and infrastructure for ensuring that all results of research funded by the Canadian public are suitably cared for in a way that consumes no more public funds than necessary, and yet enables collaborative or distributed use if appropriate.

How Canada Might Assist in the Development of the Global Research Enterprise?

In addition to promoting development of Canadian research priorities, a less US-centric approach could also provide leadership in addressing the tremendous dichotomy between nations in terms of their contributions to the global research enterprise. While 10% of the world's population controls 90% of the GDP, the same 10% control 98% of the world's research output (as defined by peer-reviewed publications). The significance of viable national research systems for international development, especially in terms of overlap with national health, environmental protection, and economic development systems is now widely recognized and promoted by international agencies like WHO, COHRED, and OECD, among others.

Canada could play a major role in global scientific development by promoting innovation and dissemination of research product stewardship strategies that work anywhere and in a distributed manner, and therefore can link Canadian expertise to both local and international decision makers and experts. Canadian leadership in the development of internationally accepted protocols and innovative mechanisms for more effectively managing access to data and information generated by publicly funded research, coupled with the related expertise of Canadian researchers, could benefit from a concerted effort by Canadian research councils since there will be common information stewardship issues that will cross disciplinary boundaries.

How to Promote a Canadian Brand of Stewardship for Information Generated by Publicly Funded Research Data?

The highly successful IKEA model may provide insights. This starts with searching the world for cost effective solutions to everyday household needs. The product scan is then followed by efficiently translating those ideas into useful IKEA-style solutions. Standardised IKEA portals (stores) are accessible around the world and where all these products are strategically laid out in standardized floor plan and in a simple way that makes their utility clear. This encourages repeat visits and facilitates matching particular solutions for particular needs. This "portal" design was prototyped on sensible Swedish shoppers and then exported to the world.

Canada has the potential for developing a position of international leadership in public research data/knowledge stewardship and in standardizing ways of sharing that data/knowledge. Canada's small and highly distributed research base, the excellent quality of its researchers, its highly educated population, and its expertise and infrastructure for telecommunication, knowledge management and resource extraction could be leveraged to develop a uniquely Canadian solution to the problem of rationally regulating access to the products of publicly funded research. Solutions developed in Canada could then be exported to the world in a way that provides the Canadian private sector with an advantage in linking up with and servicing important new emerging markets.

A First Step: A Survey of Research Data and Information Management Strategies Used by Canadian Scientists.

Before a "National Strategy for Stewardship of Access to Scientific Research Data" can be developed, there needs to be a comprehensive survey of current best practices in Canada and throughout the world for managing the data and information generated by publicly funded research. Because scientific data are a fundamental infrastructure component of modern research, a global perspective is appropriate. This survey and the subsequent actions that may be taken will require considerable "buy-in" and participation by individual scientists if a thorough and accurate understanding is to be developed. This in turn will require appropriate incentives to be developed.

Scientists expend significant resources in managing the stream of information that they generate. With new guideline being promoted by a variety of sources on the responsibilities of scientists for being able to account for and archive their data, the cost of this activity to the individual scientist will increase substantially in the near future. The "Principles and Recommendations for Sharing Publication-Related Data and Materials" (Science Editor, 26:92-93, 2003) generated by the "Committee on Responsibilities of Authorship in the Biological Sciences" of the US National Academy of Science in the US is one example of guidelines that may soon become requirements. These principles and guidelines are reasonable and carefully thought out. However, implementation may require considerable investment on the part of researchers in terms of training, time, and resources expended.

In helping Canadian researchers meet responsibilities of this type, the national research funding agencies could provide a powerful stimulus and incentive for participation in a general and centrally managed solution to this problem. Thus, the survey should not only document current specific solutions but also explore what features of a general research data stewardship strategy would encourage enthusiastic participation. The survey should also assess the general awareness of scientists as to their responsibilities and options for making research data accessible.

The survey would have to recognize the many types and forms of data that are characterized by both content and context descriptors. Starting from a discussion paper by Paul Wooters we have developed a matrix (see Appendix A) that could help guide construction of the survey instrument. The topics inserted into the matrix cells are meant to highlight key aspects of that cell.

The goal of this survey should be to identify a set of achievable standards that could evolve into a national data stewardship program. Participation in such a survey may eventually need to be made a condition of receiving public funds, but it would be advisable initially to first recruit a set of volunteers to work out how to best implement this survey in a minimally intrusive fashion. The ultimate solution will require optimizing the management and policy regimes not only of the data themselves, but also for the metadata describing the products of the research.

One perhaps controversial suggestion as to how to acquire such information cheaply and accurately would be to use a benign form of the spyware that currently is plaguing the computers of many scientists. This software would have to be open source and entirely transparent as to its function. At the same time, measures to protect confidentiality would need to be built in, but we are assured that such an instrument can be easily developed. Commercial versions of such metadata management programs such as NuGenesis (http://www.nugenesis.com) are now widely implemented. However, those systems are built on expensive proprietary software platforms.

Since almost all Canadian scientists rely on digital data collections, representation, and storage strategies running on Internet accessible computers, it would be a straight-forward matter to develop a transparent open-source middleware product that delivers transactional metadata generation and feeds that information to a secure but searchable database. This database would collect information on what data are being generated by whom, for what reason, using what methodology, and where those data are stored. Because data acquisition and storage is a dynamic process it will be necessary to update information on a quarterly basis for at least 18 months.

STEWARDSHIP of DATA/INFORMATION ACCESS: ACTIVITY GRID (Catagories modified from Paul Wouters; http://www.kbn.gov.pl/GRV/abstracts/4-3.pdf)

		CONTEXT				
CONTENT		**Technological**	**Institutional & Managerial**	**Financial & Cost Centre**	**Legal & Policy**	**Cultural & Behavioural**
Actors (participants in exchange)	Peer-to-Peer	Transactional DB	Program	Minimal Local	Barter	Collegial Discipline Specific
	Archives/ Repositories	Transactional DB Memory Farm	Program Natl Institute	Substantial Local/National	Regulated Curated	Heritage Efficiency
	Coordinated Exchange	Transactional DB Memory farm Transaction map	Program Natl Institute Natl Network	Substantial National	Contractual Outcome-Driven	Common Goals Incentives
Agents	(Re)Generators of Information (protocol, reagents, programs, computers)	Virtual Computer Reagent Banking	Standards SOPs Block Purchases	Substantial Local/National	Mutual Interest Mandated	Responsible Research Professional Ethics GLP
Medium	Face-to-display-to-Face (Lecturing &	Interpretive Visualization	Educational Scholarly	Advertisement Return on	Contractual Summative	Worldviews Languages Knowledge Base
	Knowledge Mapping) Mediated (by ICTs)	Data Mining Software Hardware	Evaluative ISP Security	Investment Broad Band Costs	Privacy Property	Global Village
Type of Information	Hypothesis Testing	Large number of specific data sets	Investigator driven			
	Observational and the "Omics"	Limited number of large data sets	Consortium Driven			
	Informal/Amateur					
	Metadata	Links to data sets	Standards Driven			
Purpose of Information	Research/ Discovery	Novel approaches			IP	
	Regulatory/ Development	Established approaches			Documentation	

2. Removal of Discipline Boundaries

The creation of a National Research Data Archive will facilitate the following advances by 2010:

Cross-disciplinary research through shared data

For example, statisticians and the medical profession working together can better assess population disease frequency and distribution as functions of patient age, location (environment), and climate through time (epidemiology). Infectious diseases can be tracked in terms of spread mechanisms, rate and target specificity. As another example, consider the effect on urban planning. Study of the change of population density in different parts of Toronto over the last decade, coupled with forward planning by the City, has resulted in a new rail line being constructed to a particular suburb before the road transport system became grid locked. This was achieved by the census office sharing data with computer experts, city planners, engineers and the rail company.

International database resources as part of larger computing environment

Closer collaboration between countries can aid our understanding of population demographics. This assists in understanding migration trends (e.g., from Third World to Western World). The introduction of extraneous sicknesses (e.g., chicken flu) from incoming populations/individuals can be tracked through shared international database resources (i.e., collaboration between immigration/customs agencies in different countries). Shared international databases can also aid national and international security in our ability to track individuals and groups.

Increased collaboration

The availability of data will forge new joint research projects between disciplines (e.g., physics and chemistry, math and engineering). Many data sets are of use to more than just the researcher/source user that generated the database. For example, an analysis of the distribution of lead and mercury in soils in a particular region by a geologist can be used by an epidemiologist to track certain chronic symptoms in humans. Without a common database, the medical profession would not necessarily be able to access and use this information.

Integrated science and data environments

Science data need to be formatted such that a common mode of access and classification is developed. By 2010, this has been done using various standardized formats. Software has been designed to reformat input into template suites, a range of which has been created to accommodate the majority of data formats as used by different disciplines. This has been achieved by setting up a designated task force to design globally acceptable formats and for IT experts to build novel software to convert the different formats into a more limited number of templates. International collaboration is underway to make data presentation and access uniform across the globe.

New studies in Metadata

The integration of the data generator and the data product has been streamlined through new metadata protocols, largely arising from integration activities described above. This has spawned a new arena of computer studies involving the classification and description of seemingly unrelated objects and commodities. This has implications for the computer control of complex systems (i.e., the discovery of commonality amongst seemingly disparate objects).

Researcher as a citizen

By 2010, the private citizen has access to most databases via her/his home computer. This will also facilitate the participation of private citizens in contributing to living databases. For example, health agencies may invite people to participate in various studies related to exercise, diet, environment and job through time (years, decades). Participants will be able to observe the results as they evolve through time. This may influence lifestyle choices.

Currency of data

One issue that is raised by many of the above comments is whether the data available are up-to-date. For many predictive operations, it is necessary to have data trends manifest through both a reasonable length of time (months, years, decades), depending on the subject, and through to the present day. For certain systems (e.g., population demographics) it is best to have the most current data available. Automation of information updating should therefore be encouraged (e.g., people who change houses are automatically logged as a change of address/location, etc., into a national computer database).

3. New Methodologies

Technological

The evolving focus in data methodologies will be data mining using geographically and thematically linked databases distributed throughout several research centres within a country and between different countries, joined with high speed data networking capabilities. Current leading-edge research projects are including 'just in time' modeling where the output from one iteration on a cell is completed in time for input to another cell iteration for a different process, at a different physical location. This idea focuses on a 'proxel' or raster based data space that integrates observed data assimilated in real time, processes appropriate to that space (including physical, social, psychological) and simulated data based upon the evolution of those processes (Ludwig et al., 'Web-based modeling of energy, water and matter fluxes to support decision making in mesoscale catchment", Physics and Chemistry of the earth, 28 (2003) 621-634). The challenges include processing and data access in a language used in all relevant disciplines, time sensitive data assimilation and archive, massive connectivity between disparate nodes, and inclusion of stakeholders in complex physical and conceptual modeling.

Other Technological opportunities include:

- suites of on-line tools for data extraction and preliminary analysis that will be transparent to the non-technical user, student and media.
- temporal synchronization of instrument and computer processes at many nodes
- replication of software at many nodes to ensure conformity in analysis
- quality control of observed data assimilated in real time, simulated data, and non-metric data

Institutional and managerial

New methodologies must include transparency in processing language (such as the Unified Modeling Language) between disparate disciplines, 'push' data management technologies, coordination of data exchange protocol, encoding for large image data archive transfer, and, most importantly, fiscal commitment to an evolving technical and scientific environment.

Financial and Budgetary

Future methodologies will be trans-national, such as relationships between the Canadian Space Agency and ESA and NASA. The issues of data ownership and technological IP become major access barriers associated with the inherent conflict between private sector support by the government versus civil services by the government. For example, scientists working with Canadian RADARSAT data looking at ice motion over the entire Arctic Basin have budget-breaking acquisition expenses, while European colleagues have ready access to similar data.

Legal and Policy

The government will continue to be a major supplier of environmental information. Some protocols, such as agreements with the World Meteorological Organization, are effective for data sharing, but new data types, such hyperspectral imaging for geological exploration, will be, at the least, cost recovery, at the worst, proprietary. This may be the major hurdle for effective environmental data analysis.

Cultural and Behavioural

Data of various types and structures will need to be integrated for decision support. The range will include high level processed satellite imagery with geocoded precision as well as subjective assessment such as elder knowledge of changes in local fishing practices in response to climate change. New paradigms of collaboration, and effective communication, will have to be developed to include these various data types in a systematic analysis process. Data archives may have to include indices of 'competence' of holdings for various types of applications.

Canadian Content and Context

Canadians have made substantial and innovative contributes to data management methodologies. The Canadian Information System for the Environment (CISE http://www.cise-scie.ca/english/notices.cfm) is a notable recent example that has cut across several traditional boundaries. The SHARCNET program (http://www.sharcnet.ca/) opens new opportunities in distributed data archives and analysis. These efforts will be compromised by the Canadian policy of full cost recovery and privatization of environmental satellite data. We have lost considerable goodwill in the international research community.

4. Enhanced Education

The establishment of a common Canadian approach to the access to publicly funded scientific research data will open up enormous educational possibilities particularly in the Post-Secondary Educational sector.

Traditional discipline-based programs involving the use of scientific and engineering data will be augmented by educational experiences relating to the cross cutting (integrative) approach to research data availability. Students will have ready access to data hitherto only poorly accessible and the potential for new and innovative program development will be widespread. Research supervisors will be able to recommend a wider range of research problems with the new tool.

Universities (and colleges) have entrenched systems for admission, program development and courses of study for all degree and diploma programs. The introduction of a new set of standards for access to data will necessitate re-examination of some of these practices. Universities are usually reluctant to change processes unless there is a clear demonstration of obvious advantages to their students, researchers and, of course, all this without the allocation of new resources. Unless significant new resources are available it is likely that we will see no new programmatic changes unless they are revenue neutral.

It is therefore up to the Task Force to demonstrate the benefits to all parties. Some of these benefits are described below, together with some thoughts on an engagement strategy.

The Student

Clearly, any process that provides for a broader and more efficient educational experience is to be desired. Ready access to research data will have a direct effect on project selection and will expose students at an early stage in their research careers to the power of accessible data as a research tool. Exposure to other researcher’s data will then strengthen the network of researchers at all levels and, hopefully, lead to greater student mobility, whether real or virtual.

Recognising the much broader availability of useful research data, Universities should capitalize on the potential in their recruitment strategies for Graduate students (in appropriate departments), in the appointment of thesis/advisory boards and the vigorous pursuit of joint program opportunities between sister institutions and government laboratories. For example, the NRC Canadian Bioinformatics Resource an excellent data rich research tool that could form the basis for joint University-NRC research projects.

Students would have ready support of their institutions to identify research mentors from a variety of other institutions. Clearly, new opportunities need preparations. Currently, students entering graduate school face an already daunting array of requirements. Often, especially in science and engineering, a course-heavy Masters leads to the Ph.D. program. Somewhere in this route many universities have orientation sessions for new and ongoing students. I see a need for inclusion of the ‘power of accessible data’ aspects of their programs to be provided at this stage. Procedures vary from institution to institution. Orientation sessions can be run at the Laboratory, Department, Faculty or University level (or not at all, which is what I suspect is often the case). A data-access and archiving component for graduate students should become part of the orientation process for all graduate students with a nation wide syllabus prepared for such a course (½-1 day). Either an on-line or a face-to-face version could be led by CISTI and Task Force institutional partners. Granting Councils, as part of their widely touted HQP strategies could require institutions to present such a briefing to all incoming students and, where necessary, to in-course graduate students. Post -doctoral fellows would also profit from such an exposure, as well as senior undergraduates involved in final year research projects. Such a workshop would, of necessity, be rather broad but at least the opportunities would be clear, particularly if couched in terms of easing some research burdens and broadening scope.

Oft-times students learn new skills by osmosis (at the feet of their mentors) but this can be a frustrating exercise. The existence of a National policy, fully supported by all research institutions and agencies, could provide the leadership for easing the burden on capitalizing on research data access and understanding the need for archiving. For instance, besides institutional orientation workshops mentioned above there is the opportunity of other valuable activities to merge such as summer Schools (cf. NATO Advance Study Workshops); short internships for students with Government and industrial partners both nationally and internationally. The idea of a 'Data Access for Dummies' short course given by a consortium of Universities or the granting councils or an arms length institution (e.g. Vitesse, see http://www.vitesse.ca). Governments possess massive amounts of data, much of which is supposedly available publicly. One could envision internships and/or joint research projects accessing such data for research needs.

The Research Community

The production of a cohort of graduates with an ease and understanding of accessing archived research data will have an immediate effect on the Canadian research community producing a sensitised and competent group of next-generation faculty and researchers in all walks of life ensuring that the Canadian tax-payers receive even greater value for money from publicly funded research. Politically this will also be a powerful tool for the search for continued and extended research funding in our public institutions.

The Task Force 'Achievement 2010 ' blueprint identifies many sectors all of which will depend on the next generation of researchers in data-rich disciplines being able to fully understand the ramifications and potential of a data access policy. Comments in the rest of this paper on these areas reveals the intricate intertwining needed for experts and therefore the educational opportunities.

Some Thoughts on an Engagement Strategy Related to Educational Opportunities

This section is presented in point fashion and could overlap with the communication strategy document. I make the difference here between communication and engagement. 2010 is not that far away if one needs to influence many agencies and organizations and Governments.

1. Brief Granting Councils and request 'buy-in'. Relate to emerging and existing HQP strategies. Seek sponsorship for 'Data-use workshops.
   (NSERC, CIHR, SSHRC, Genome Canada, CFCAS, CFI, NCE program)

2. Brief Federal and Provincial Govt. departments dependent on data management. Seek support for workshops, internships, targeted summer placements etc.
   (EC (EPS, MSC); HC; AAFC; NRCan)
3. University strategy.
   a. Brief VP's research (Regional organizations)
   b. Brief Graduate Deans (Regional bodies for program approvals and National, CAGS)
   c. Interact with US equivalent, Council of Graduate Studies.
4. Brief National Science Advisor with respect to impact on HQP production.
5. Prepare a standard presentation package for "champions"

Conclusion

An exciting array of possibilities emerges to produce Canadian researcher sensitive and expert in the area of research data management. It is essential that possibilities be explored as soon as possible if the expectations for 2010 are to be achieved.

5. Data Stewardship

Data stewardship consists of the norms and behaviours of researchers that underlie the practices of sharing and preserving research data. By 2010, several developments have occurred in Canada that corresponds to a cultural shift within the research community that accentuates data stewardship and promotes the researcher as “good citizen”.

This cultural shift coincided with a fundamental change in educating the next generation of researchers. As part of the ethics training in graduate programs across Canada, students have been taught the importance and value of sharing data in science. The challenge to open access to scientific research posed by ownership claims to data has been transformed into an understanding of data stewardship and its importance both to data access and to the preservation of data. Students have learned that the sharing of data is a fundamental principle of open science and that this enables replication, exploration, discovery, and the creation of new data from old.

Sharing data is widely recognized in the research community as being equivalent to an accountant’s “open book.” The evidence is available for all to examine. Furthermore, sharing publicly funded research data is recognized as the appropriate ethical choice because such data are part of our country’s digital heritage and as such belong to society.

The responsibilities for the long-term care and preservation of data are part of a national initiative of the National Data Archive to introduce digital curatorship in data management practices. Through a life-cycle model of research data management, researchers with the assistance of the National Data Archive are establishing a chain of authority for the care and preservation of the data they are managing over the life of their research and the life of the data subsequent to this research.

Universities incorporate the importance of data creation, management, and preservation in their tenure and promotion procedures. Included in the academic reward system is the acknowledgement of the intellectual contributions associated with data creation and management. Guidelines of best practices in data creation, management, and preservation are available to assist universities in assessing the performance of researchers and their contributions in this area. Among the factors used in rewarding researchers is the secondary use of the data from their research. That is, data sharing has become part of the reward structure of excellent scholarship.

Researchers are now using existing data in more creative ways and consequently, returning greater value for the original investment in the data. Disciplines are awarding new achievements accomplished through the innovative uses of existing data as well as creative uses of new and old data.

Research Ethics Boards routinely receive background training in the best practices of data collection and management. Members of these Boards understand the importance of preservation both for the protection of public investment in the data but also for the implications of privacy protection when human subjects are involved in the data. These Boards understand that proper preservation of data provides greater protection for human subjects than the outright destruction of data. Proper archiving ensures that human subjects will not be identified or disclosed while also ensuring that the data are available for further scientific evaluation to protect subjects from poor or bad research.

Standards in anonymizing data with human subjects have emerged and best practices have been established. The National Data Archive provides researchers with workshops on approaches to and methods in anonymizing data to promote best practices and to ensure privacy protection.

6. Capacity Building for Data Archiving

Students are encouraged across disciplines to consider data archiving as a career choice. The National Data Archive requires specialist knowledge from various fields of study and staff are hired to fill these skilled positions. Special degrees are granted across disciplines that incorporate a substantive background in a field with data archiving through new programs that have been established in Library Information and Archival Studies. Students graduating with Canadian degrees in this hybrid field are sought for their training and expertise by countries around the world.

In addition to new degree programs to help build capacity in data archiving, the granting councils in Canada have offered students incentives to train in this area. Tri-council support has been mobilized through a jointly funded research initiative into the preservation of research data. This initiative supports graduate students working on their degree as well as funds research to generate new knowledge about data archiving. Exchanges with other national data archives are supported to increase knowledge of and experiences in working with major international data sources.

7. Network Effects

The network effect arises when a good or a service is more valuable to a user the more users adopt the same good or service. Classical examples are the telephone and the fax machine. The purchase of a fax by one individual indirectly benefits others who already own a fax by making all of them more useful. Beyond a certain number of early adopters, known as critical mass, a network effect will become significant in attracting more users because the value to the user will exceed the cost they have to pay. The value of a network depends on the number of network users (Metcalfe's law). The utility of large networks, particularly social networks, can scale exponentially with the size of the network (Reed's law). Both combined explain partially the explosive popularity of the Web in the nineties.

A few examples of network effects in the context of the year 2010 target:

- Sharing of large volumes of environmental data from studies of trans-disciplinary environmental impact and change;
- Ability to contact data-builder scientist / experts to access their data and knowledge;
- Expanded awareness of research issues and methods in other disciplines and the potential serendipitous impact on a researcher's ‘worth’;
- More cross-disciplinary research teams;
- New business start-ups based on access to and commercial exploitation of public data;
- Greater collaboration in the science discovery enterprise due to enhanced data access.

Knowing that network effects could have a synergistic and accelerating impact on the adoption of new methodologies and practices, how could we put them to use to overcome the challenges of implementing a system of collection, storage, processing, distribution and preservation of scientific research data with due diligence to intellectual property rights and adequate access controls?

In the context of the year 2010 target and of the achievements that would have had to have been accomplished by then to reach a desired state of **Enhanced Access to Scientific Research Data** how could network effects play a role? Could they be induced or triggered by targeted actions?

Network effect considerations include:

Technological

- Canada benefits already from a robust and flexible technical infrastructure that needs to be maintained and further expanded.
- Much work still needs to be done in the matter of interoperability of software and protocols and adoption of standards for metadata and for data exchange and data quality. This needs to be done by active and concerted participation of Canadians in international fora.

Institutional and managerial

- Canada lacks a national agency to preserve, catalogue and provide systematic, efficient and convenient access to scientific research data and assist other institutions in developing discipline specific policies and methodologies for transparent and open access to their data.
- Institutions should put in place policies that limit the tendency to withhold data from public circulation

Financial and budgetary:

- On the simple principle of operational efficiency, granting agencies and government should strive to maximize their return on investment by promoting reuse of data and providing proper documentation, specialists and effective data management facilities. The funding of long term data management should received as much attention as the funding of the research itself.

Legal and policy:

- Restriction on reuse of public data by the research community must be eliminated or minimized.
- Current Canadian intellectual property rights policies frequently prohibit the reuse of data.
- Cost recovery practices of federal departments and agencies should be revisited in light of what is a happening in other countries that have less restrictive data access policies

Cultural and behavioural:

- Improve reward structure and mechanisms to promote open access and sharing by individual researchers. By enhancing the attractiveness of joining and contributing to the network individuals will see their own benefits from sharing. Open access to primary data in persistent data stores will also lead to more long-term and sustained error-correction procedures.

8. New Discoveries

Genetics and genomics.

In many ways the genomics community has proven to be a model for data access and data sharing. Data repositories for DNA and protein information such as Genbank (USA) and the EMBL (Europe) have revolutionized the way that biology is now done and without these resources it is difficult to see how the burgeoning field of bioinformatics would have been able to have formed. Continuing this paradigm further, more recently Genbank has formed a trace repository allowing the deposition of the actual raw data derived from DNA sequencing machines. The trace repository continues to attract new uses and scientific analyses for its data which certainly were not envisaged during its conception.

It is also inevitable that the fields of Genetics and Genomics will significantly benefit further through a strategy of continued data deposition and archiving. Another recent step is the creation of an archive to store experimental results of gene expression experiments and, equally importantly, an accurate and standardized description of the actual experiment. The underlying nature of these data being such that having access to large datasets allows more reliable clustering and analysis of everyone's subsequent experiments. Subtleties may become apparent which would not otherwise have been detected, e.g. slight perturbations in gene levels in response to the addition of pharmaceutical agents previously unnoticed may lead to insights into the mechanism of a particular adverse drug reaction. Likewise, information on human genetic variation, often a by-product of other studies, can be easily deposited for use by others for genetic mapping, functional assays or for bioinformatics studies. Such analysis could easily be envisioned to help rapidly identify genetic variations or mutations which confer susceptibility or resistance to genetic diseases or infections.

As is common in the study of complex systems, such as those encountered in biology, many experiments fail and the results remain unpublished. The ability to distribute the negative results can be seen as a significant change in the way that scientific results are disseminated. The economic impact of being condemned to independently repeat the same experiment because there was no obvious way to communicate the result to the research community must surely be significant and also not limited to the field of genetic and genomics.

The bioinformatics community has also been a strong advocate of the open source software. However, just because the software is open does not relieve the problems of how to access a specific version that an analysis was performed on and also the continued availability of the software from the reliable and trustworthy source.

Since many data and file formats require software for reading and analysis it is also worthy to note that an archive for software will be as equally valuable as the data it supports.

Health

In the health research community, there are at least three major groups who could make considerable and exciting scientific advances with the availability of a national public research data repository.

First, the population health researchers. They are dedicated to understanding population health dynamics—the risk factors and outcomes of disease. Their research has identified the importance of smoking in lung cancer and cardiovascular disease, DDT exposure in cancers, fluoride treatment in the prevention of dental caries, and risk factors for infant prematurity and its adverse consequences.

Second, the clinical epidemiology researchers. They are typically clinicians involved in the treatment of patients with particular health problems. Their research usually involves studies that determine the prognosis of illness, the best methods of diagnosing and treating disease, and the effectiveness of new and old medical treatments. Their results directly enhance the benefits for their patient populations.

Finally, the health policy and health service researchers. They study the effects of policies and health care system delivery on access to care, quality of care, and patient outcome; often issues that are of major concerns for Canadians and health care managers. The impact of private-public partnerships on the quality and accessibility of health care, co-payments by consumers on the use of drugs and medical services, the impact of waiting lists on quality of care and outcome, and public screening and vaccination programs for breast cancer and hepatitis are some of the many issues they address.

Despite the diversity of research, these groups of scientists have one common requirement. They need to assemble and follow groups of people for an extended period of time to determine the effects of a particular agent, test, policy, health care practice on health outcomes. While some of this work is experimental, the majority of scientific investigation is observational. It relies on the collection of personal and health information over time, typically from a variety of sources, and the analysis of natural variations in the population to address relevant research questions. To provide timely and efficient answers to the many questions that need to be addressed, researchers began to assemble cohorts—populations of people that had a common exposure (e.g working in occupations with high levels of exposure to potentially carcinogenic agents), common health condition or disease (e.g. pregnancy, diabetes, multiple sclerosis, stroke, cancer), or common interventions (e.g. transplant). With the exception of cancer, where many provinces legislate comprehensive reporting of information, there were no guidelines or standards for this form of assembling populations for ongoing research in other population groups (usually referred to as a disease or population registry).

As a result, there has been a proliferation of local disease and population registries, each with its own costly infrastructure. Few are large enough to study key questions in a timely manner. Research on less common health problems is often not done because the effort required to assemble the population takes much more time than is possible for a single research group within restricted research funding envelopes (e.g. amyotrophic lateral sclerosis). All patient and population registries have difficulties in maintaining high quality research data because of the on-going costs and the varying standards required for consent and data management.

A national data archive would go a long way to resolving many of these problems. It will allow national patient registries to be developed to investigate the potential causes, treatment, prognosis and outcomes of both common and uncommon health problems, as well as new procedures (e.g. islet cell transplant). It will facilitate higher quality research by standardizing data capture and coding requirements. It would allow far more robust solutions to be applied to data access and security, particularly in health, where privacy and confidentiality are critical issues. At the present time, research groups rarely have sufficient resources to implement state-of-the-art data security infrastructure and processes.

With the sustained standardized archival storage of national population registries, researchers would have access to populations that are large enough to develop new methods of identifying patient clusters who, for example, may be more likely to respond to a particular treatment, or experience complications. Scientists will be able to develop prediction rules that will allow a much better match to be made in the choice of treatment or preventive strategy for a given individual, dramatically increasing the likelihood of success. The timing is right for action. The rapid advance of genomics provides the opportunity to investigate genetic and complex gene-environment determinants of disease expression and treatment response, as well as a model for managing highly sensitive data (e.g. UK biobank). The emergence of the electronic patient record in health care provides the computer-based mechanism by which complex prediction rules for treatment choice can be implemented rapidly, with minimal effort required by the clinician, as treatment recommendations can be generated by automated calculations with patient-specific data. The Canadian interoperable electronic health record initiative, led by Canada Health Infoway Inc., will create distributed clinical repositories of standardized health data to enhance the safety and quality of care for Canadians from coast to coast. Synergistic opportunities exist to capitalize on this investment by creating the parallel infrastructure that would allow scientists to use these data to create new knowledge, and integrate new knowledge back into the care delivery process. A national archive of de-identified longitudinal health records for Canadians will also allow new knowledge and methods to be developed to enhance quality assurance and patient safety, public health surveillance, outbreak investigation, and health planning. Indeed, the absence of accessible, high quality, health care data has hindered the development of a vibrant scientific community that could have generated knowledge to address the many ailments of our out-of-date 20th century approach to health care delivery. It is time to act.

Engineering and Instrumentation

The drivers for many data streams can be found in engineering and instrumentation (which includes physics, chemistry, all branches of engineering including materials science). The evolution in interdisciplinary research, and what can be expected in 2010, is being revolutionized by, for example:

1. mixing biology with micro-fluids (engineering and physics) into the development of “labs on chips”;
2. Materials science is now more than “metallurgy”: access to sophisticated high energy beam lines, such as those found at the Canada Light Source, will enable new materials and material properties to be understood;
3. “There is plenty of room at the bottom” according to Richard Feynman and the nanotechnology revolution owes its emerging impact to multi-physico-chemical science, computing and manufacturing;
4. Earthquakes, particularly land-based on Canada’s west coast are an ongoing concern. Data collection today exceeds 1 tera-byte per year (http://www.seismo.nrcan.gc.ca/cndc/index_e.php) and increased resolution of instruments or the distribution density will further exacerbate data storage issues;
5. The North-East Pacific Time-series Undersea Networked Experiments project will be the

world's largest cable-linked ocean observatory. NEPTUNE will give scientists, educators, policy-makers and the general public a new way of studying and understanding issues critical to our survival, such as earthquakes, fish conservation, climate change, and energy sources;

6. The Herzberg Institute on Vancouver Island maintains the data bases for many large scale experiments (http://cadcwww.hia.nrc.ca/), including those data to be acquired from NEPTUNE, ATLAS etc. and are expected to exceed 100 terabytes per year;
7. The world renowned Sudbury Neutrino Observatory (SNO) collects data from its 9600 photomultiplier tubes continuously and produces about a terabyte of raw data per year. These data are one-off and therefore must be preserved for current and future generations of scientists to reprocess as detailed understanding of the detector improves or new theories are developed.

New discoveries require new technology, new instruments (sensors, actuators, algorithms) that either build on existing scientific principles or require in themselves new discoveries. For example SNO-LAB is being established to broaden the experiments to answer questions of, for example, "dark matter" and which require the development of new detectors or sensors (http://www.sno.phy.queensu.ca/public/queens/subatomic.html). Additional examples abound, and are neatly summarised at http://www.ecs.soton.ac.uk/~ajgh/DataDeluge(final).pdf.

While acquiring data, either through sensors or other instruments, reprocessing of those data is a critical part of science and technological advancement. Thus, electronically stored data become the evidence for claims made in scientific papers, health records etc. The access to and re-assessment of claims made in light of new ideas, particularly against data that are "one-off" is very important. For example, aircraft now are monitored in flight, transmit vital engine data to the engine manufacturers via, say, SITA (http://www.sita.aero) and thus either the health of the engine can be continuously monitored or in the event of a failure, trace the events that lead to that failure. (See, for example, http://www.aviationnow.com/content/publication/om/200104/om77.htm or http://www.scientificmonitoring.com/resources/papers/State%20of%20Play_ATEM%20EMS%20Oct03.pdf.) In the future, this could be extended to on-line monitoring and prognoses that will enable interaction between ground and airborne-based computers to automatically alter the flight to achieve a safe outcome.

Environment

Technological

Environmental modeling has always been at the mercy of computing cycles and storage access. GCM (Global Climate Models) are evolving towards including a fully coupled oceanic model, and more effective boundary parameterization for land surfaces. The future will include full feedbacks with the biosphere. On the horizon is the inclusion of non-numeric data such as economic models, precision farming practices and psychological indices that determine land use changes and are affected by atmospheric trends such as enhanced green house gas concentrations. Models will also include current observations through real-time data assimilation. Regional models that are approaching these 'ideals' are running with integration times that are close to the actual time of processes, which is incredibly restrictive from a long-term analysis perspective. It is clear that new analysis technologies, and data access technologies have to be catastrophic, not incremental, in the scale of improvement.

Many of these complex models are using distributed nodes of analysis, both in space and in expertise. A current experiment uses calculations at time x in one location by an atmospheric physicist, for example, and requires calculations at time x + y at another physical location by a plant ecologist in a different institute, and so one, before the atmospheric physicist proceeds with his/her next iterations. The hardware, software and data quality control issues are significant.

Institutional and managerial
Simulation using input from institutions separated by political boundaries is currently a reality. Issues of data ownership, intellectual property, and adherence to standards have yet to be addressed but can break the types of experiments that need to be made.

Cultural and Behavioural
Whilst the atmospheric community has a long tradition of collaboration, we still have the NCAR model, the Hadley Centre model, etc that are based upon social traditions and government control rather that rational effective direction of funds and expertise.

Canadian Content and Context
Notwithstanding the above, Canada has an enviable international record in environmental modeling and monitoring. The climate modeling group at U Victoria has made international-scale contributions to coupled atmosphere, ocean, ice and land modeling. We have a very strong remote sensing technology and analysis reputation, being a world leader in ground station technology, radar satellite technology and emerging hyperspectral imaging technology, for both satellite and airborne platforms. We have one of the larger supercomputers for modeling. We have not made any innovative advances in data management and data sharing. The focus on industrialization and cost recovery by the government is a major anchor.

10.2 OECD Final Report March 2003 (without appendices)

Promoting Access to Public Research Data for Scientific, Economic, and Social Development

OECD Follow Up Group on Issues of Access to
Publicly Funded Research Data

Final Report

March 2003

Acknowledgements

Participants from the United States are being funded through NSF Grant ACI-9619020 to help coordinate this group’s activities and develop a series of case studies to explore data access issues. Additional support for the group’s activities comes from Netherlands' Ministry of Education, Culture and Science, and governments and agencies of the Follow-up Group’s members.

The CODATA Secretariat hosted a meeting of the Follow-up Group in March 2002 in Paris. CODATA is the interdisciplinary Committee on Data for Science and Technology of the International Council of Scientific Unions (ICSU).
The Polish State Committee for Scientific Research hosted the 4th Global Research Village Conference (October 2002), which provided a forum to discuss the Follow-up Groups’ Interim Report.

The European Science Foundation (ESF) is a partner in this activity. ESF hosted a meeting of representatives of member funding agencies in October 2002 in Strasbourg, France. The meeting provided a forum for discussion of the Interim Report of the Follow-up Group, and resulted in continued discussions at the funding agencies themselves and, in some cases, new policies (see Box 1). In addition, ESF helped solicit input on agency policies for this report.

Finally, The Ministry of Education, Culture and Science from The Netherlands funded a web survey of current data policies as well as a number of case studies. These efforts were directed by Paul Wouters of NIWI-KNAW, The Netherlands. Their publication in the series The Public Domain of Digital Research Data also contributed to the activities of this Follow-up Group on Issues of Access to Publicly Funded Research Data.

Executive Summary

It is now commonplace to say that information and communications technologies are rapidly transforming the world of research. We are only beginning to recognize, however, that management of the scientific enterprise must adapt if we, as a society, are to take full advantage of the knowledge and understanding generated by researchers. One of the most important areas of information and communication technology (ICT)-driven change is the emergence of e-science, briefly described as universal desktop access, via the Internet, to distributed resources, global collaboration, and the intellectual, analytical, and investigative output of the world's scientific community.

The vision of e-science is being realised in relation to the outputs of science, particularly journal articles and other forms of scholarly publication. This realisation extends less to research data, the raw material at the heart of the scientific process and the object of significant annual public investments.

Ensuring research data are easily accessible, so that they can be used as often and as widely as possible, is a matter of sound stewardship of public resources. Moreover, as research becomes increasingly global, there is a growing need to systematically address data access and sharing issues beyond national jurisdictions. The goals of this report and its recommendations are to ensure that both researchers and the public receive optimum returns on the public investments in research, and to build on the value chain of investments in research and research data.

To some extent, research data are shared today, often quite extensively within established networks, using both the latest technology and innovative management techniques. The Follow Up Group drew on the experiences of several of these networks to examine the roles and responsibilities of governments as they relate to data produced from publicly funded research. The objective was to seek good practices that can be used by national governments, international bodies, and scientists in other areas of research. In doing so, the Group developed an analytical framework for determining where further improvements can be made in the national and international organization, management, and regulation of research data.

The findings and recommendations presented here are based on the central principle that ***publicly funded research data should be openly available to the maximum extent possible***. Availability should be subject only to national security restrictions; protection of confidentiality and privacy; intellectual property rights; and time-limited exclusive use by principal investigators. Publicly funded research data are a public good, produced in the public interest. As such they should remain in the public realm. This does not preclude the subsequent commercialization of research results in patents and copyrights, or of the data themselves in databases, but it does mean that a copy of the data must be maintained and made openly accessible. Implicitly or explicitly, this principle is recognized by many of the world's leading scientific institutions, organizations, and agencies. Expanding the adoption of this principle to national and international stages will enable researchers, empower citizens and convey tremendous scientific, economic, and social benefits.

Evidence from the case studies and from other investigation undertaken for this report suggest that successful research data access and sharing arrangements, or regimes, share a number of key attributes and operating principles. These bring effective organization and management to the distribution and exchange of data. The key attributes include: openness; transparency of access and active dissemination; the assignment and assumption of formal responsibilities; interoperability; quality control; operational efficiency and flexibility; respect for private intellectual property and other ethical and legal matters; accountability; and professionalism. Whether they are discipline-specific or issue oriented, national or international, the regimes that adhere to these operating principles reap the greatest returns from the use of research data.

There are five broad groups of issues that stand out in any examination of research data access and sharing regimes. The Follow Up Group used these as an analytical framework for examining the case studies that informed this report, and in doing so, came to several broad conclusions:

- Technological issues: Broad access to research data, and their optimum exploitation, requires appropriately designed technological infrastructure, broad international agreement on interoperability, and effective data quality controls;
- Institutional and managerial issues: While the core open access principle applies to all science communities, the diversity of the scientific enterprise suggests that a variety of institutional models and tailored data management approaches are most effective in meeting the needs of researchers;
- Financial and budgetary issues: Scientific data infrastructure requires continued, and dedicated, budgetary planning and appropriate financial support. The use of research data cannot be maximized if access, management, and preservation costs are an add-on or afterthought in research projects;
- Legal and policy issues: National laws and international agreements directly affect data access and sharing practices, despite the fact that they are often adopted without due consideration of the impact on the sharing of publicly funded research data;
- Cultural and behavioural issues: Appropriate reward structures are a necessary component for promoting data access and sharing practices. These apply to both those who produce and those who manage research data.

The case studies and other research conducted for this report suggest that concrete, beneficial actions can be taken by the different actors involved in making possible access to, and sharing of, publicly funded research data. This includes the OECD as an international organization with credibility and stature in the science policy area. The Follow Up Group recommends that the OECD consider the following:

- Put the issues of data access and sharing on the agenda of the next Ministerial meeting;
- In conjunction with relevant member country research organizations,
    - Conduct or coordinate a study to survey national laws and policies that affect data access and sharing practices;
    - Conduct or coordinate a study to compile model licensing agreements and templates for access to and sharing of publicly funded data;
- With the rapid advances in scientific communications made possible by recent developments in ICTs, there are many aspects of research data access and sharing that have not been addressed sufficiently by this report, would benefit from further study, and will need further clarification. Accordingly, further possible actions areas include:
    - Governments from OECD expand their policy frameworks of research data access and sharing to include data produced from a mixture of public and private funds;
    - OECD consider examinations of research data access and sharing to include issues of interacting with developing countries; and
    - OECD promote further research, including a comprehensive economic analysis of existing data access regimes, at both the national and research project or program levels.

National governments have a crucial role to play in promoting and supporting data accessibility since they provide the necessary resources, establish overall polices for data management, regulate matters such as the protection of confidentiality and privacy, and determine restrictions based on national security. Most importantly, national governments are responsible for major research support and funding organizations, and it is here that many of the managerial aspects of data sharing need to be addressed. Drawing on good practices worldwide, the Follow Up Group suggests that national governments should consider the following:

- Adopt and effectively implement the principle that data produced from publicly funded research should be openly available to the maximum extent possible;
- Encourage their research funding agencies and major data producing departments to work together to find ways to enhance access to statistical data, such as census materials and surveys;
- Adopt free access or marginal cost pricing policies for the dissemination of research-useful data produced by government departments and agencies;
- Analyze, assess, and monitor policies, programs, and management practices related to data access and sharing polices within their national research and research funding organizations.

The widespread national, international and cross-disciplinary sharing of research data is no longer a technological impossibility. Technology itself, however, will not fulfill the promise of e-science. Information and communication technologies provide the physical infrastructure. It is up to national governments, international agencies, research institutions, and scientists themselves to ensure that the institutional, financial and economic, legal, and cultural and behavioural aspects of data sharing are taken into account.

1. Preface

At its March 2001 meeting, the OECD Committee on Scientific and Technology Policy (CSTP) accepted a proposal from The Netherlands to establish a working group on issues of access to research information. The plans of the working group were presented at the October 2001 CSTP meeting. Subsequently, the Committee narrowed the scope of activities to access to and sharing of research data produced from public funding.[1] Participation in the group was broadened to include Australia, Canada, Denmark, Finland, Germany, Japan, Poland, the Netherlands, the United Kingdom, and the United States. The CSTP asked the working group to:

- Report on current practices concerning access to and sharing of research data and their underlying principles on the basis of case studies;
- Report on the effects of selected current data sharing practices on the quality of research and the progress of science;
- Suggest principles for making policy on data sharing within the relevant national and international policies and regulatory frameworks.

The report's core principle is that ***publicly funded research data should be openly available to the maximum extent possible***. Adoption of this principle will promote good stewardship of public knowledge, strong value chains of innovation, and maximize benefits from international cooperation (see Box 1). The report's findings and recommendations are addressed to: CSTP members as representatives from the governments of OECD member countries that carry responsibilities for national and international science policy and the functioning of research funding agencies; research institutes; and professional and scholarly associations. The objective is to contribute to a better understanding of the importance of research data access and sharing, and to offer suggestions on how the new digital challenges should be met.

Box 1: This core principle guides many public scientific institutions and scientists. However, it remains unevenly implemented. Most recently, it was adopted by the United Kingdom's Medical Research Council. After a workshop hosted by the European Science Foundation, the MRC drafted the following statement: MRC promotes the creation of a diverse range of datasets, many of which are rich in informational content, unique and cannot be readily replicated. Sharing allows scientists to extend the value of these datasets through new, high quality, ethical research and exploitation. It also reduces unnecessary duplication of data collection. Building preservation systematically into routine data management is part of good research practice: it strengthens quality, enables replication and audit, and provides a sound basis for data sharing. [2]

Building on a number of case studies and a great deal of other research, the report focuses on issues related to the access and sharing of publicly funded research data, in digital form, across all disciplines in the natural, health, and social sciences. Attention is paid to the international aspects of access and sharing relevant to scientific cooperation among OECD member states. Three significant topical areas fell outside the charge of this working group, however, and will require separate follow-up: issues particular to developing countries; issues related to data produced by a mixture of public and private funding; and the issue of national security restrictions in light of recent global events since 11 September 2001.[3]

2. Introduction

2.1. The changing information technology context for scientific research and innovation

Information and communication technologies (ICTs) are rapidly transforming research and the broader society: witness the growth in the number of Internet hosts per person, in the percentage of computers per household,[4] and in the continued rate of growth of chip, storage, and network technology capacity.[5] Concurrently, there has been an explosion in the amount of data produced across all types of scientific endeavour.[6] Continuing ICT advances, such as the development of grid computing, large-capacity optical transmission networks, wireless networks of sensors and devices, and complex imaging systems, promise to push these transformations farther and faster. ICT-dependent research, such as geographic information systems, data visualisation systems, and realistic modelling, are adding tremendously to our ability to study and understand the world in which we live. These developments provide researchers in OECD countries, and increasingly in developing countries, with the opportunity not only to be more efficient, more effective and better connected, but also to dramatically expand the scope and nature of their investigations.[7] Together they create the possibility of an "e-science infrastructure."[8] The growing activities in data collection, storage, processing, distribution, and preservation are, however, only loosely connected. They require systematic planning to realize the full potential of the emerging e-science infrastructure.

2.2. The benefits of data access and sharing in public research

Within this new technological context, more widespread and efficient access to and sharing of research data will have substantial benefits for public scientific research (see Box 2). Open access to, and sharing of, data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new data sets when data from multiple sources are combined.

Sharing and open access to publicly funded research data not only helps to maximize the research potential of new digital technologies and networks, but provides greater returns from the public investment in research.[10]

BOX 2: ACCESS to international data has helped produce a better understanding of public health issues and worldwide disease prevention and control. For instance, research on cholera outbreaks and their relationship to numerous environmental factors relied upon data drawn from epidemiology, NASA remote sensing, marine biology, microbiology, genomic data, and social science data. This research—an example of 'biocomplexity' studies supported by the U.S. National Science Foundation—would have been impossible without access to numerous databases. The effect of this interdisciplinary and international research project is an increased scientific and sociological understanding of cholera outbreaks and their prevention. [9]

Improving and expanding the open availability of public research data will help generate wealth through the downstream commercialisation of outputs, provide decision-makers with the necessary facts to address complex, often trans-national problems, and offer individuals the opportunity to better understand the social and physical world in which we all live (see Box 3).

As a key link in the value chain of investments in research, open access to factual data plays an increasingly important role in all these areas.

BOX 3: A recent analysis demonstrated the economic benefits of providing open access to government meteorological data without any restrictions on re-use. 11 The "value adding" meteorological information industry in the United States has revenues in excess of $500M annually. The public meteorological data also support a rapidly growing weather risk management industry that underwrites financial risk management instruments valued at approximately $8B. In contrast, the private-sector value adding industry for meteorological information in the European Union is very small, largely attributable to the highly restrictive data policies of most national governmental meteorological services. What are harder to measure, but certainly occur, are the countless lost opportunity costs for researchers, students, and various other potential public users who find the high costs of the public data to be too great to use.

2.3. Roles and responsibilities of governments

If researchers throughout the world are to take full advantage of ICTs to improve and expand access to, and sharing of, research data, existing technological, institutional and managerial, financial and budgetary, legal and policy, and cultural and behavioural aspects must be addressed comprehensively and in an integrated way. To date, these aspects have often been treated on an ad hoc, project-specific basis. Given that OECD countries spend tens of billions of dollars each year collecting data that can be used for research, and for other social and economic benefits, ensuring that these data are easily accessible so that they can be used as often and as widely as possible, is a matter of sound stewardship of public resources (see Box 4).

BOX 4: Poor stewardship and lost opportunity for data access is exemplified by the case of Statistics Canada, which attempted to recover costs for its data management by charging data users. The effect of this form of management of these public data was a dramatic decrease in their use. In a study of the case, it was found that "Cost recovery was supposed to introduce a market type discipline on the demand for and supply of goods and services provided by the government. Since in economic terms Statistics Canada's outputs are public goods, the type of discipline envisioned by this policy is impossible to attain. Instead we have users who complain, refuse to pay and generally attempt to find alternative sources for their information needs. This policy fails the improved management of resources test."[12]

Scientists, research institutions, and research funding agencies around the world are increasingly engaging in large-scale, data-intensive projects. Such projects require data-management infrastructure, data-exchange protocols and policy frameworks, and a broad professional understanding that more extensive availability and use of the data is both necessary and desirable. Over the past decade, numerous studies, disciplines, research programs, and agencies have begun to address the complexities and benefits of open data access and sharing arrangements.[13] As scientists become better connected with each other, particularly through the Internet, and as research focuses on issues of global importance, such as climate change, human health and biodiversity, there is growing need to systematically address data access and sharing issues beyond national jurisdictions and thereby create greater value from international co-operation. The goal should be to ensure that both researchers and the broader public receive the optimum return on public investments, and to build on the value chain of investments in research and research data.[14]

3. Core Principle and Premises

The findings and recommendations that follow are based on the central principle that:

Publicly funded research data should be openly available to the maximum extent possible

As a general principle, publicly funded research data should be as open as possible and available at the lowest possible access cost, subject only to legitimate restriction and considerations. Restrictions may be necessary for reasons of national security, for the protection of privacy of citizens, or the confidentiality of trade secrets. Access to research may be limited by the respect for private intellectual property rights. Finally, there may be reasons for granting temporary exclusive access to those who collected the data. But the guiding principle should be openness.

In order to derive the maximum benefit from public investments in research data, access, use, management and preservation must be an integral part of the research process. Conversely, data should not be considered an expendable by-product of research. In many cases, data have value beyond the project and anticipated use for which they were originally collected. The reuse of publicly funded data for research and other types of applications should be promoted and not restricted.

The accessing and sharing of data is not merely a technical matter, but also a complex social process in which researchers have to balance different pressures and interests. Purely regulatory approaches to data sharing are not likely to be successful without consideration of these factors. Various approaches to data access and sharing are therefore necessary, including the establishment of regulations and incentives, and the dissemination of best practices.[15]

The following three premises complement and support the core principle of this report:

3.1. Data from publicly funded research are a public good produced in the public interest
Both the data from publicly funded research and research itself have strong public good characteristics that support their open availability to the public, and especially to other researchers.[16]

3.2. Factual data are central to the scientific research process
The production, open dissemination, and unfettered use of factual data are essential attributes of, and inputs to, modern systems of scientific research and technological innovation. Recognizing the role of digital data as fundamental to the value chain of science, technology and innovation will enable an optimum return on public investments.

3.3. Data access and sharing issues are international in scope
To more fully exploit the possibilities of global digital networks, and to capture their benefits for the global community, policy issues concerning access to and sharing of publicly funded scientific research data must be addressed, not only at the institutional and national levels, but also at the international level.

4. Data Access Operating Principles and Attributes

Data access and sharing requires effective organization and management. The necessary components that make for this organization and management may be characterized as “data access regimes.” In their ideal form, these regimes enable all participants in the scientific research process to freely and efficiently access and share data. Adequate data access regimes require dispersed, as well as centralised, responsibilities across different management domains that include the technological, institutional and managerial, financial and budgetary, legal and policy, and cultural and behavioural.

No single approach to developing an effective data access regime is possible; however, a list of operating principles for and attributes of effective data access regimes and resources can be offered. This list of attributes and operating principles is based on a broad set of experiences, and supported by the case studies conducted for this report. Key attributes are listed below, and illustrated with an example from the case studies. [17]

4.1. More explicit access regimes

There is a universal need for the formalisation of institutional rules and data management policies. This formalisation follows from the growing complexity and scale of scientific research and the increasing expenditure on research data. At the moment, it is not clear who is authorised to distribute data across the globe. To reach the necessary transparency in the tasks and responsibilities of those involved, terms of access and use of data that rest on tacit agreements will have to be made explicit and formalised. A systematic and institutionalized approach is needed to help address operating characteristics of data access, and to take advantage of the opportunities arising from publicly funded research.

4.2. Operating Principles

4.2.1 Openness.

Open availability of publicly funded research data to the maximum extent possible is the core principle of this report.

4.2.2. Transparency of access and active dissemination.

Open data access requires actively disseminating where the data can be found, what the context and structure of the data collection is (metadata), how long the resource will be accessible, and what protocols and standards are employed. In short, this principle refers to the systematic visibility and traceability of data resources.

4.2.3. Assignment and assumption of formal responsibility

Formal responsibility for tasks associated with data access must be assumed by the appropriate participants in the global science system. The various individuals and institutions involved in the chain of data-related activities all have specific manifest and latent duties and obligations. These are founded in formal legal and professional normative standards and in the regulations of various agencies. Responsibility must also be assumed for various rights in the data supply, such as authorship, producer credits, ownership, financial arrangements, licensing terms, and, where appropriate, restrictions on use.

4.2.4. Professionalism

Codes of conduct, and related normative standards, of professional scientists and their communities can help to promote good practice and simplify the regulatory aspect of access regimes.

4.2.5. Interoperability

Technical and software standards and protocols are required to ensure the access and usability of data. These should be clear to the user and adopted by as many data management organizations as possible.

4.2.6. Quality

Quality refers to the proper description of uncertainties surrounding the production of the data (e.g., the techniques employed in their collection and archiving, and the measuring instruments and their calibration), the ability to ensure that the cited source and value are authentic, that the data retain integrity (complete and absent from introduced errors), and that they are secure against loss, destruction, modification, and unauthorized access.

4.2.7 Operational Efficiency

Open access to data increases the efficiency of research by avoiding unnecessary duplication of data collection and permitting the creation of new data sets by combining data from multiple sources. Coupled with open access, comprehensive documentation of data sets and how to access them provides a more efficient use of resources.

4.2.8. Flexibility

In general, scientific communities will approach data management requirements more consistently within their discipline internationally, than they will across other disciplines on a national level. Data access regimes need to be sufficiently flexible to take account of this variation.

4.2.9. Property

Institutional intellectual property rights as well as the individual rights of researchers are considerations of property interests. Unlike the private sector, public research operates on a principle of collective property interests, which are promoted by the open access and sharing of data resources.

4.2.10. Legality

Legal restrictions may limit access to and use of data.[18] Restrictions will apply primarily to 'secondary' data sets compiled for purposes other than scientific research. In some cases, the sensitive parts of data sets can be left out without rendering them useless. Specific types of legal restrictions include: national security, privacy and the protection of trade secrets.

4.2.11. Accountability

Accountability involves measuring the cost, benefit, and performance of data access and sharing regimes and taking appropriate actions in response to the results.

4.3 Building a Data Access Regime: the Global Biodiversity Information Facility (GBIF)

The Global Biodiversity Information Facility (GBIF), which began under the auspices of the OECD Megascience Forum, has sought to implement these principles as a means to achieve the larger goal of providing worldwide access to biodiversity data. GBIF's goal is to make "the world's scientific biodiversity data freely available to all [**openness**]."[19] The fundamental motivation for GBIF is to enable access to a vast amount of biodiversity data housed in databases distributed in numerous countries and institutions. By bringing all these data into one interoperable network, and producing a registry of biodiversity information resources, GBIF will produce systematic visibility and traceability of data resources [**transparency**].

Formal responsibilities of different participants involved in the task of building GBIF's organisation and legal relationships have been put forth in GBIF's Memorandum of Understanding. GBIF's Secretariat is responsible for carrying out work programmes that are approved by the Governing Board, which consists of representatives of GBIF's Participants. This structure enables GBIF to have a legal identity as an international body, manage financial contributions and work programmes, while drawing upon efforts and resources from Participants. In his reflection on the establishment of GBIF submitted to the OECD, Eric James attests: "The way in which these legal requirements are met may be the most important factor determining the structure of the organisation that is created."[20] The establishment of GBIF's activities occurred in and through contact with existing scientific and political bodies to maintain and establish professional codes, gain consensus about scientific outcomes, and negotiate with government representatives about GBIF's larger social and economic roles [**professionalism**]. The review will evaluate GBIF's progress toward data availability and interoperability, its responsiveness to user needs, and the professionalism of the Secretariat.

Participants will provide stable gateways, or "nodes," to databases that contain primary or meta-level biodiversity data. These nodes must provide documentation and metadata about the data in the databases, vouch for data **quality**, ensure data authenticity and security. GBIF will help develop standards for database **interoperability** through one of its 4 work programmes, Data Access and Database Interoperability (DADI). GBIF aims to develop an interoperable network of distributed databases by coordinating and leveraging existing national and international programs and projects, which allows for **operational efficiency** and more cost-effective basis for making biodiversity data freely and easily available to a heterogeneous user community.

The databases and the data accessed through GBIF are in most cases owned and developed by other organisations and thus will not entail any assertion of IPRs by GBIF itself [**property**]. GBIF aims to provide best practices on how to deal with IPRs, particularly since it will be drawing from databases hosted by different institutions and countries with different legal frameworks, with a view to promoting open access and sharing to the maximum extent possible.[21] GBIF also asserts in its MOU that biodiversity data will be properly used and acknowledged by its participants [**legality**]. Further, its efforts do not conflict with the Clearing House Mechanism, and they abide by the Global Taxonomic Initiative of the Convention on Biological Diversity concerning the proper and equitable use of biodiversity data and the resources to which they refer.

During the establishment of GBIF, the OECD provided the forum to assess the level of support for this new scientific collaboration, to bring together related proposals and to develop detailed plans that could then be taken up by interested countries. GBIF will have a third-year review of the effectiveness of its MOU, its scientific efforts and the "transparency of its dealings with politically sensitive issues"[22] [**accountability**].

5. Data Access Management: Five Domains

Efficient data access can only take place with the proper administration and organization of different management domains within data access regimes. These domains include technological, institutional and managerial, financial and budgetary, legal and policy, and cultural and behavioural considerations (see Figure 1). These domains provide a framework for locating and analyzing where improvements to data access and sharing can be made.

The five domains differ in character across the traditions and practices of specific scientific disciplines, e.g., astrophysics, biology. Thus, data access regimes may vary in significant ways. There is no single model for how data access should take place. The implementation of the core principle of open availability, however, requires a systematic approach that recognizes the necessity of implementing improvements across the interdependent management domains. This approach also requires the involvement of actors from various levels: governments, funding agencies, and research institutions and professional and scholarly societies, as well as individual scientists themselves.

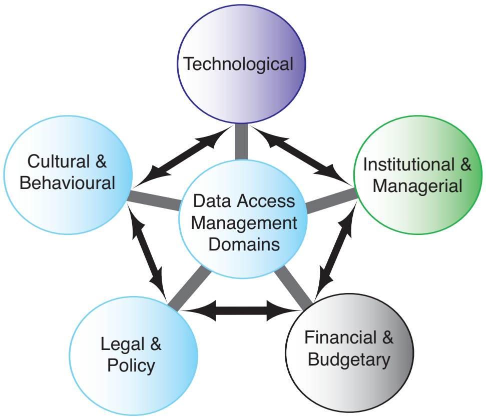

Figure 1. Components of a Data Access Regime

5.1 Technological domain: *Broad access to research data, and their optimum exploitation, requires appropriately designed technological infrastructure, broad international agreement on interoperability, and effective data quality controls.*

A technical infrastructure that supports user needs is necessary to derive maximum benefits from data access and sharing. This infrastructure must be robust enough for long term use and, when appropriate, for diverse uses. It also must be flexible enough to respond to the continuous and rapid changes in scientific research and technology. While there are many technical issues to be resolved to take full advantage of past, current and future investments in ICT infrastructure, the main barriers to effective data access and sharing are no longer technical, but are institutional and managerial, financial and budgetary, legal and policy, and cultural and behavioural.

Data Preparation and Metadata: ICPSR

In 1995, the Inter-University Consortium for Political and Social Research (ICPSR) initiated the development of the Data Documentation Initiative (DDI), an international criterion and methodology for the content, presentation, transport, and preservation of metadata about datasets in the social and behavioural sciences. DDI, which is in XML format, helps enhance users' ability to acquire and use data while it assists producers' in packaging and disseminating them. After a period of beta-testing with participating international organisations, DDI is now in use by a number of organisations, including Networked Social Science Tools and Resources (NESSTAR), Health Canada, and ICPSR. ICPSR continues to assist data producers in preparing their data through its "Guide to Social Science Data Preparation and Archiving," a guide with broad appeal for individuals and organisations searching for easy and effective ways to technically manage and prepare data so that they can be easily and effectively placed into network environments.[23]

Technical operating principles for data access regimes include **interoperability** (of protocol and software to ensure the access and usability and multiple use of the data); and **quality** (including technical components of **authenticity**, **integrity**, and **security**) of data.

5.2 Institutional and managerial domain: *While the core open access principle applies to all science communities, the diversity of the scientific enterprise suggests that a variety of institutional models and tailored data management approaches are most effective in meeting the needs of researchers.*

Because scientific data have many different characteristics and uses, there is no monolithic institutional and management approach that can be applied universally.[24] Key characteristics of data production and use include whether the data are (1) government-generated or generated at a research institution using public funds; (2) useful only within the discipline or across many disciplines; (3) useful over the very long term or only within short-term horizons; (4) have public-policy implications; or, (5) have significant broader economic and social value, among other factors.

Institutional and managerial operating principles for data access regimes include **transparency** (systematic visibility of the data source); **responsibility** (explicit formal institutional rules on data management); and **accountability** (rendering public account for the performance of data access regimes).

Negotiated collaborations: CERN

The European Organisation for Nuclear Research, CERN, is one of the world's largest scientific laboratories, presently financed by twenty European countries. CERN overtly subscribes to the core principle and premises outlined in this report, but leaves it to individual 'collaborations' of scientists to devise experiment-specific regulations to ensure compliance. Negotiations between different collaborations are necessary to enable data sharing, including agreement on definitions and standards. The type of data produced and the method of processing used will play a large part in deciding upon the most effective management model to adopt. This flexibility of management approach is a key factor in the data sharing environment at CERN.

5.3 Financial and budgetary domain: *Scientific data infrastructure requires continued, and dedicated, budgetary planning and appropriate financial support. The use of research data cannot be maximized if access, management and preservation costs are an add-on or after-thought in research projects.*

In many areas of public research, there are indications of discrepancies between the funding of the specific research itself and the related data-management requirements (which do not necessarily benefit the individual scientist, but which are necessary for data reuse). Generally, research organizations fund the former well, but pay scant attention to the latter. In the digital environment, scientific data sets must be viewed as a key element of the broader research infrastructure and as an investment in the future capacity to innovate and solve pressing problems. Adequate support is essential for data-management functions, such as the development of sufficient explanatory documentation for each data set (i.e., metadata), conversion of old formats onto new media, adaptation to new standards, and long-term preservation, archiving, and maintenance.

Budgetary operating principles for data access regimes include **operational efficiency** (maximizing the return on investment by promoting re-use of data, and providing proper documentation, specialists, and effective data management facilities).

Funding schemes "on a rolling basis:" the European Bioinformatics Institute (EBI)

The official mission of the EBI is to ensure that the growing body of information from molecular biology and genomic research is placed in the public domain and is freely accessible to the scientific community in ways that promote scientific progress. Like other scientific bodies, the EBI has a major problem in the funding for its building, maintaining and making available databases and information services even though they represent only a small fraction of the total research costs. The key issue is that funding for data sharing infrastructures needs to be constructed "on a rolling" or on-going basis to maintain effective data management. These funding requirements are very different from the funding schedules of research, which are usually project oriented. These differences in budgeting constitute the main threat to the EBI's commitment to maintaining the public availability of its data.

5.4 Legal and policy domain: *National laws and international agreements directly affect data access and sharing practices, despite the fact that they are often adopted without due consideration of the impact on the sharing of publicly funded research data.*

Intellectual property laws, information policies, institutional guidelines, and contracts at the national and international levels often impose terms and conditions on data access and sharing practices. Laws and policies governing data access and sharing practices may vary among different countries, resulting in barriers to scientific cooperation and progress. Based on a recent Web survey, most of the national research organization managers who responded expected that data sharing will become a major policy issue in the next five years. This situation requires greater attention by the science policy community at all levels. In particular, restrictions on re-use of public data by the research community must be eliminated or minimised as much as possible. Research grant provisions and licensing templates for promoting open access and unrestricted re-use of public research data already exist, but have not yet been broadly adopted.

Policy interconnections: functional MRI and the Institutional Review Boards

The functional Magnetic Resonance Imaging Data Center's (fMRIDC) principal endeavour is to promote data sharing in brain mapping. The Western tradition of informed consent in bio-medicine operates according to the principle that the 'most specific consent is the best consent.' When data are to be gathered for submission to databases, the specificity of consent may run counter to the goals of meta-analysis or re-analysis by third parties, to investigate issues different from those for which the data was originally gathered. The creation of infrastructures for data sharing, therefore, has to conform to the rules of regulatory bodies, such as institutional review boards (IRBs), whose approval must be obtained to share data. As such, these bodies function as gatekeepers to the circulation of data. International coordination may also be necessary. Researchers submitting or requesting data across national boundaries may find it especially difficult to act in accordance with the various ethical guidelines that exist in different countries. The fMRIDC has been hesitant to accept data from non-US settings because of concerns regarding IRB compliance.

Legal and policy operating principles for data access regimes include **property** (balance intellectual property rights of investigator and institution versus public good); and **legality** (lawful data management, respecting national security, privacy and trade secrets).

5.5 Cultural and behavioural domain: *Appropriate reward structures are a necessary component for promoting data access and sharing practices. These apply to both those who produce and those who manage research data.*

Although formal policy frameworks and regulations are necessary to make research data publicly available, they need to be supplemented by appropriate community-based norms and incentives for researchers to share and provide access to their data and for appropriate recognition of their data-related work. In many cases, there is a general lack of reward structures and mechanisms to promote open access to, and sharing of, data from public research.

Cultural and behavioural operating principles for data access regimes include quality (trust that data are what they purport to be); professionalism (build on codes of conduct and ethics of the scientific community); flexibility (there is no single model on how data access must be provided.)

Incentives: the Protein Data Bank

To publish in scientific journals, U.S. scientists involved in the field of crystallography must deposit their data in the Protein Data Bank (PDB) and acquire an accession number. “By requiring everyone to submit data, the community is assured of having the most up to date information possible. Now, increasingly, under our regime, a lot of [data] depositors have come to realize that the practice that we use has some advantages for them in that we check things and we find errors and inconsistencies. That actually improves the quality of the product they produce.”[25]

6. Possible Action Areas

Our findings from the case studies and from other research indicate a number of action areas by the different actors involved in making possible open access to, and sharing of, publicly funded research data. In this section we recommend possible action areas for the OECD and national governments.

OECD

As an international organization with credibility and stature in the science policy arena, the OECD, through the CSTP, can play a crucial role in promoting access to, and sharing of, data from publicly funded research. Central to this role is the gathering and sharing of information on data related activities and policies. At the international level, only a small handful of organizations have undertaken to do this, usually in the context of a specific discipline or research program. The recent, and vast, expansion of research data assets and the trend towards issue-based, interdisciplinary research, however, suggests that all countries and all fields of science stand to benefit from greater attention and an organized and coordinated approach to effective policy actions

1. **The OECD should put the issues of data access and sharing on the agenda of the next Ministerial meeting.** ICT advances have created the ability to transform science. New tools allow researchers to find data in seconds that would have taken months just a few years ago. Effective data access and sharing requires a comprehensive policy approach for implementation by public research institutions. Monitoring progress and devoting attention to the public research data issues and activities would assist decision-makers and research support agencies in developing appropriate policies and allocating resources

Areas in Conjunction with Relevant Member Country Research Organizations

2. **The OECD should consider conducting or coordinating a study to survey national laws and policies that affect data access and sharing practices.** This relatively simple undertaking could determine what policies exist, how accessible they are, and result in listing of the web sites where these policies are posted. This study would be of considerable benefit to science policy-makers, research administrators, and information resource managers in all countries, both within OECD and beyond. The study could look at the feasibility of developing a central and easily accessible repository of national laws and policies that affect data access and sharing practices. Such a compilation does not currently exist, and could be useful to facilitate international research collaborations.
3. T**he OECD should consider conducting or coordinating a study to compile model licensing agreements and templates for access to and sharing of publicly funded data.** Depending on the context, numerous factors need to be considered in data access and sharing arrangements. Nevertheless, many contractual models already exist that have been developed by research funding organisations, research program managers, university administrators, librarians, and others. The OECD, as a global organization, is ideally suited to span national domains where examples do exist, and thereby bring an international perspective. The study could compile and review existing agreements and models to find exemplary approaches. Having readily available models on hand would be of considerable benefit to researchers, universities, and research institutions, as well as data centers and archives, and could facilitate international research collaboration

Areas for Further Examination

4. **Governments within the OECD should expand their policy framework of research data access and sharing to include data produced from a mixture of public and private funds.** Collaborative public/private research projects, and the resulting data, have their own unique set of characteristics and issues. As more national governments promote public-private partnerships in research, these issues will be of increasing importance to both public researchers and the companies that are involved. A further examination of the state of data sharing and access in these types of research arrangements needs to be made to develop sound science policy guidance.
5. **The OECD should consider examinations of research data access and sharing to include issues of interacting with developing countries.** The increase of participation in the research enterprise benefits the global science system and innovation. Providing developing countries with access to data from publicly funded research increases their participation in science. Further, as United Nations Education, Scientific and Cultural Organization (UNESCO), the International Council of Scientific Unions (ICSU), private foundations, and other organizations have emphasized, access to scientific knowledge by developing countries is vital to the progress of the entire world. This access is particularly important in the context of global issues such as population health, environmental change, and food production. Of course, open access to data from publicly funded research in developed countries can provide a valuable resource for economic development, education, and scientific capacity building. Many efforts are already underway to improve access for researchers in developing countries (e.g. providing free or below costs access to data and scientific information) as well as establishing optimal data regimes for developing countries to share their data (e.g. addressing issues of data repatriation). A systematic examination of barriers and best practices would provide both a picture of the current situation and a set of guidelines for further action.
6. **The OECD should promote further research, including a comprehensive economic analysis of existing data access regimes, at both the national and research projectd or program levels**. To date, no one has yet undertaken a comprehensive, economic analysis of different data access regimes. Several key issues have not been closely examined, including the relative costs of providing data openly, the impact of cost recovery on the use of those data, and the positive externalities and network effects from providing open access to publicly funded research data. The OECD should consider conducting this type of analysis or encouraging member country research organizations to fund such studies.

National Governments

Although the OECD, UNESCO, ICSU, and other international bodies can play a role in improving the current situation regarding research data access and sharing, it is at the national level that many important decisions and actions must be taken. National governments provide the resources for making data accessible, establish the overall policies for data management, regulate matters such as confidentiality and privacy, and determine restrictions based on national security. Most importantly, it is national governments that are responsible for the major research support and funding organizations, and it is here that many of the managerial aspects of data sharing need to be addressed.

The national governments of OECD countries should consider:

1. **Adopting, and effectively implementing, the principle that data produced from publicly funded research should be openly available to the maximum extent possible.** The public investments made in research data collection can only be maximized if the data are preserved, managed, and made accessible. This requires coordinated attention by governments at all levels, and adequate policy and financial support. The starting point for these actions, however, is the affirmation that data collected using public funds should be openly accessible to all.
2. **Encouraging their research funding agencies and major data producing departments to work together to find ways to enhance access to statistical data, such as census materials and surveys.** Many countries have taken steps to facilitate access to census and survey materials by developing catalogues, user-friendly repositories, off-site research facilities, training programs, and regulatory frameworks for providing appropriately guarded access to confidential information. Such steps have proven enormously effective in maximizing the use of national surveys and producing insights into the functions of economies and societies.
3. **Adopting free access, or marginal cost pricing, policies for the dissemination of research-useful data produced by government departments and agencies.** The use of information collected through public funding should be freely accessible for research purposes. This maximizes the use of such information for public policy and public knowledge development.
4. **Analyzing, assessing and monitoring policies, programs, and management practices related to data access and sharing policies within their national research and research funding organizations.** This information would be useful to national governments so that they may assess the implementation of the previous three considerations. The resources, support programs, policies, and regulations related to research data sharing are, in large part, developed and implemented by research funding organizations. The operations of these organizations play a crucial role in determining the degree to which data are made accessible and shared between researchers. Many organizations, such as NSF and NIH in the United States, Social Sciences and Humanities Research Council in Canada, and the European Science Foundation are now developing, or have developed, policies, regulations and support programs that promote data sharing. Issues such as establishing protocols for the collection and release of confidential information, developing technical infrastructure, agreeing on metadata standards, requiring data preservation strategies within individual research projects, and including data management costs as eligible expenditures in grant applications have been dealt with by one or more of these agencies. It would benefit the global scientific community if decision-makers within national governments had a clear understanding of where their respective agencies stood in relation to those in other countries.

7. Conclusion

Improving access to and sharing of publicly funded research data is an issue that touches on all aspects of the research enterprise and the development of knowledge, and involves all participants in the conduct of research. For the individual researcher, the sharing of data, particularly prior to publication[26], can be burdensome, time consuming, and unrewarding if the necessary measures are not taken to provide funding, facilities, and a social context that emphasises its value to the research community and to society.

Advances in ICTs, the internationalisation of science, and the trend toward issue-based research hold great potential for the advancement of knowledge and for the benefit of all people. This potential will not be fully realized unless all of the major elements of data access regimes identified in this report are properly developed. To do so will take considerable discussion, understanding, and commitment on the part of all those involved in research, particularly at the policy level.

Agreement among OECD governments on a set of general principles to shape specific data access regimes, as well as adoption of the recommendations set forth above, would be enabling for scientists, empowering for citizens, and provide an important contribution to fulfill the promises of e-science.

[1] In this report, we define "access to data" as the act of making the data available for use by others; by "sharing" we mean a researcher allowing one or more other individuals to use data, typically with the implicit, if not explicit assumption that it is on a reciprocal basis. The sharing of data involves providing specific access, whereas the act of providing access by itself does not necessarily involve any sharing arrangement. Data sharing focuses on data exchanges between individual researchers rather than institutions, while access may be provided at any level. Sharing also reflects the cooperative norms of public science as practiced within many disciplines by many researchers in OECD countries. We define data as in the U.S. National Institutes of Health definition of final research data: “the recorded factual material commonly accepted in the scientific community as necessary to validate research findings”.

[2] See http://www.mrc.ac.uk/index/strategy/strategy-science_strategy/strategy-strategic_implementation/strategy-data_sharing/strategy-data_sharing_policy-link

[3] CODATA, the interdisciplinary Committee on Data for Science and Technology of ICSU, is currently examining barriers to data access and sharing that are particular to developing countries. CODATA, however, does not normally examine issues related to social science and humanities research. Related to issues of national security, see “NAS Censors Report on Agricultural Threats,” *Science* 20, p. 1973-1975, on the several scenarios that were left out of a public report of the U.S. National Academy of Sciences.

[4] See NSF 2002 Science and Engineering Indicators, http://www.nsf.gov/sbe/srs/seind02/start.htm

[5] Gary Stix (2001), “Triumph of Light,” *Scientific American*, January, available at http://www.sciam.com/2001/0101issue/0101stix.html

[6] Examples range from genetic sequence and protein structure data in bioinformatics, to various types of brain imagery in neuroscience, to sky surveys and virtual observatories in astronomy, and geospatial data such as Global Spatial Data Infrastructure.

[7] Examples include combining data from multiple data sources to gain a greater statistical power to resolve hypotheses (see the Biomedical Informatics Research Network, http://www.nbirn.net); and obtaining real-time global measurement on environmental observations.

[8] John Taylor, Director General of (UK) Research Councils (UK), www.research-councils.ac.uk/escience/. “E-Science will refer to the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualisation back to the individual user scientist. . . . Besides information stored in Web pages,

scientists will need easy access to remote facilities, to computer – either as dedicated Teraflop computers or cheap collections of PCs – and to information stored in dedicated databases. The Grid is architecture to bring all these issues together." See also Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue Ribbon Report on Cyberinfrastructure, http://www.cise.nsf.gov/evnt/reports/atkins_annc_020303.htm .

[9] Rita Colwell (2002), "A Global Thirst for Safe Water: The Case of Cholera," Abel Wolman Lecture at the National Academy of Sciences, January 25, 2002, available at http://www7.nationalacademies.org/wstb/2002_Wolman_Lecture.pdf. Other examples of the impact of access to and sharing of international data in the control and elimination of worldwide diseases include the World Health Organisation's network of collaboration centres. In the worldwide programme of epidemiological surveillance of influenza, these receive epidemiological information on outbreaks of influenza from national institutions throughout the world. They also receive new strains of the virus for characterization and give advice as to their possible use in vaccine preparation. The centres then distribute the necessary reagents, antigens and anti-sera to national laboratories, and high-yielding recombinant viruses for to vaccine producers. See http://whqlily.who.int/general_infos.asp.

[10] For more benefits of data sharing, see National Academy Press (1985), *Sharing Research Data*, available at http://books.nap.edu/catalog/2033.html

[11] Peter Weiss (forthcoming 2003) presentation in Proceedings of the Symposium on the Role of Scientific and Technical Data in the Public Domain, National Academies Press. See also, European Union Green Paper (1998), "Public Sector Information: A Key Resource for Europe," COM 585, and PIRA International, "Commercial Exploitation of Europe's Public Sector Information, Final Report for the European Commission (2000),"Directorate General for the Information Society," which provide similar comparisons of such policies in other information sectors.

[12] Ronald C. McMahon (1996), "Cost Recovery and Statistics Canada," in Government Information in Canada, Volume 2, number 4 (spring 1996), retrieved from http://www.usask.ca/library/gic/v2n4/mcmahon/mcmahon.html, February 2003

[13] Studies include: National Research Council (1997), Bits of Power: *Issues in Global Access to Scientific Data*, National Academy Press, Washington, D.C.; National Research Council (1999); and *A Question of Balance: Private Rights and The Public Interest in Scientific and Technical Databases*, National Academy Press, Washington, D.C.; Stephen Hilgartner (1996), "Access to Data and Intellectual Property: Scientific Exchange in Genome Research" in *Intellectual Property Rights and the Dissemination of Research Tools in Molecular Biology: Summary of a Workshop held at the National Academy of Science, February 15-16, 1996*; and National Research Council (1995),*On the Full and Open Exchange of Scientific Data*, National Academy Press, Washington, D.C.; and National Research Council (2002), C*ommunity Standards for Sharing Publication-Related Data and Materials*, National Academy Press, Washington, D.C. The European Bioinformatics Institute, the Global Change Program, the Global Biodiversity Information Facility, the European Social Survey, the International Union of Crystallography, the international Ocean Drilling Program; The European Organization for Nuclear Research, otherwise known as CERN, provide good examples of research programmes with effective data policies. Funding agency statements include: NSF at http://www.nsf.gov/sbe/ses/common/archive.htm) and "NIH Draft Statement on Sharing Research Data" at http://grants2.nih.gov/grants/policy/data_sharing/. Sites that discuss how to develop a data policy include Smithsonian Environmental Research Center at http://www.serc.si.edu/datamgmnt/policy1.htm and the Ecological Sciences Network at www.esnet.edu. The policy of the Long Term Ecological Research (LTER) network is at http://www.lternet.edu/data/netpolicy.html.

[14] For more on the Role of Governments in the Digital Age, see Stiglitz, Orzag and Orzag at http://www.ccianet.org/govt_comp.php3. In particular note the following three:

Principle 1: Providing public data and information is a proper governmental role

Principle 2: Improving the efficiency with which governmental services are provided is a proper governmental role

Principle 3: The support of basic research is a proper governmental role

[15] As one researcher put it, "Incentives for data sharing need to be offered that offset the investigators' loss of control over their databases. Usually, this is some form of added scientific value. By sharing data, an

investigator may gain access to more data or other tools. Ultimately, there has to be a procedural framework that makes sharing sensible, efficient, and value-added. If all those pieces are in place, fewer external or coercive forces are needed to convince researchers to share." From minutes from an NIMH meeting, see Paul Wouters, Data Sharing Policies, 10 June 2002. Networked Research and Digital Information, NIWI-KNAW on http://dataaccess.ucsd.edu

[16] In economics, a good is considered a "public good" if it is "non-rivalrous" and "non-excludable." The former means that the marginal costs of providing the good to an additional person are zero. The latter means that once the good is produced, the producer cannot exclude others from benefiting from it. See, Inge Kaul, Isabelle Grunberg, and Marc Stern (1999), "Defining Global Public Goods," in *Global Public Goods: International Cooperation in the 21st Century*, eds. Both publicly funded basic research and the data produced from it and disseminated on digital networks are non-rivalrous. They are not purely excludable, however, although their excludability, especially for other researchers, is neither economically efficient nor desirable as a matter of public policy, absent countervailing and superseding reasons to the contrary.

[17] These operating principles evolved from the document produced by Hans Franken, Access to Publicly Financed Research, Conference Conclusion. Global Research Village III Amsterdam 2000. For other principles on data access and sharing see http://www.codata.org/data_access/principles.html . Examples of successful guidelines based on a systematic set of principles are the OECD Guidelines on the Protection of Privacy and Trans-border Flows of Personal Data (1980) and the Principles and Guidelines for the Sharing of Biomedical Research Resources (1999) from the US National Institutes of Health (NIH) and the OECD Guidelines for Security of Information Systems and Networks (2002).

[18] Examples include *National security*: Data sets from some oceanographic or geological surveys may be (partly) classified and not accessible; Privacy: Data from human subjects are vulnerable to breaches of confidentiality and privacy and therefore should only be obtained by fair and lawful means, with knowledge or consent of the data subjects; and *Trade secrets*: Data potentially relevant to prospective patenting or commercial opportunities may contain (partly) confidential information.

[19] See www.gbif.net

[20] Eric James, "Establishing International Scientific Collaborations: Lessons Learned from the Global Biodiversity Information Facility," submitted to Sixth Meeting of the OECD Global Science Forum, available at http://www.oecd.org/pdf/M00027000/M00027203.pdf

[21] Research on issues of IPR, particularly for natural history museums, is being conducted by European Natural History Specimen Information Network, see. http://www.nhm.ac.uk/science/rco/enhsin/details.html and "Beset with pitfalls-specimens and databases, intellectual property and copyright," Simon J. Owens and Alyson Prior, from the 2000 meeting of the Taxonomic Databases Working Group, November, 2000; Senckenberg Museum, Frankfurt, available at http://www.tdwg.org/tdwg2000/ipr.htm.

[22] Eric James, "Establishing International Scientific Collaborations: Lessons Learned from the Global Biodiversity Information Facility," submitted to Sixth Meeting of the OECD Global Science Forum, Section 10.

[23] For more information on ICPSR and DDI, http://www.icpsr.umich.edu/DDI/ORG/index.html. For ICPSR's Guide, see www.icpsr.umich.edu/ACCESS/dpm.html. For information on the importance and development of DDI, see "Providing Global Access to Distributed Data through Metadata Standardisation -- The Parallel Stories of NESSATAR and DDI", submitted by the Norwegian Social Science Data Services to the Conference of European Statisticians, UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, 22-24 September 1999, at http://www.nesstar.org/papers/GlobalAccess.html.

[24] For example, mathematics presents a special feature in that published material never becomes obsolete, so that the data needed by the working mathematician is ideally the full collection of published papers, past and present. With the development of internet access, this is not an impossible objective. New papers are almost always produced in electronic form, and therefore could be stored and accessed. The amount of past literature to be scanned and digitized is estimated to be around 50 million pages. Under the umbrella of the International Mathematical Union, an attempt is made to coordinate national efforts to insure permanent accessibility at a reasonable cost for the users to both new and digitized papers. Without this, research will be limited to rich parts of the world, where libraries can be heavily funded. See http://www.mathematik.uni-bielefeld.de/~rehmann/DML/ and http://www.library.cornell.edu/dmlib/.

[25] Berman, Helen. Director, Protein Data Bank. Personal communication.

[26] Rapid Data Release Policy: "Ever since the 1996 Bermuda Principles provided guidelines on the rapid release of data from large-scale sequencing projects, access to the pre-publication sequence data that has been made freely available in public nucleotide sequence databases has accelerated biomedical research. However, in 2002, it became clear that new strategies and other advances in large-scale DNA sequencing necessitated a re-examination
and updating of the data release policies originally developed to implement the Bermuda Principles for pre-publication sequence data. At its February 10-11, 2003 meeting, the National Advisory Council for Human Genome Research (NACHGR), the main advisory group to the National Human Genome Research Institute (NHGRI) on genetics and genomic research, discussed the subject of pre-publication release of large-scale sequencing data. NACHGR approved a draft policy that would reaffirm and extend the rapid data release policies developed to implement the 1996 Bermuda Principles, and recommended that NHGRI publicize the draft policy statement for the purpose of obtaining comment from the scientific community." (http://www.genome.gov/page.cfm?pageID=10506376). For a reaffirmation and extensions of the NHGRI rapid data release policy, see http://www.genome.gov/page.cfm?pageID=10506537. For community discussion see Sacrifice for the greater good? Nature 421, 875 (2003), and Draft guidelines ease restrictions on use of genome sequence data, Nature 421, 877-878 (2003).

10.3 National Data Archive Consultation Final Report June 2002 (without appendices)

National Data Archive Consultation

Final Report

Building Infrastructure for Access to and Preservation of Research Data In Canada

Submitted by the NDAC Working Group to:
the Social Sciences and Humanities Research Council of Canada and
the National Archivist of Canada

Executive Summary

In October 2000, the Social Sciences and Humanities Research Council and the National Archivist of Canada established a Working Group of research and archival experts and asked them to assess the need for a national research data access, preservation and management system. After compiling extensive evidence for the need of such a service to support the knowledge creation work of Canada's social sciences and humanities research community, the Working Group now offers recommendations for the creation of a new national research data archival service. This service would have three core functions:

- Preserving research data that is compiled by researchers, and preserving data compiled by government agencies, polling firms and other organizations that can be used by researchers to generate new knowledge;
- Managing the data held, including ensuring quality; selecting data for retention; developing and applying standards for metadata, authentication and security; and migrating data across technologies;
- Providing access to research data, including Web-based delivery systems, cataloguing services, user and depositor agreements to protect confidentiality and intellectual property rights, and connections to other data depositories around the world.

In addition, the Working Group recommends that a new National Research Data Archive Network undertake a number of other functions, including providing advanced training in data handling techniques, represent Canadian interests in the development of international data standards, promote data sharing as a best practice in research, undertake research in information and archival sciences, and act as a central hub and co-ordinating body for a network of data services in Canadian research institutions.

Digital information compiled for research purposes is playing an increasingly important role in today's knowledge economy. In many ways, data is the fuel driving innovation and our capacity to address complex social and economic problems. Although billions of dollars are spent each year collecting data, Canada lacks the necessary infrastructure to ensure these data are preserved and made publicly available. This limits the returns that can be made on our public investments in research and undermines good public stewardship.

Many of the building blocks necessary for the creation of a National Research Data Archive are already in place. University data services, high-speed transmission networks, legal and ethical guidelines and frameworks, potential partner institutions, various data depository and access portal initiatives, and an active data-producing research community already exist. The missing element is a preservation, co-ordination and management service.

Almost all developed countries have recognized the need for a national research data service, and some have more than a generation of experience in their operation. Canada is in a position to learn from this experience while developing a research data service that fits our unique institutional and cultural context. We now have the technological capacity and expertise to create a "trusted system" that provides Canadians with an accessible and comprehensive service empowering researchers to locate, request, retrieve and use data resources in a simple, seamless and cost effective way, while at the same time protecting the privacy, confidentiality and intellectual property rights of those involved. The start-up infrastructure costs for this service could be funded through the Canada Foundation for Innovation. The annual operating costs for a comprehensive facility and network are benchmarked in the area of $3 million.

The Working Group offers three options for the creation of a National Research Data Archive Network:

1) Through federal legislation, create a National Research Data Archive Network as a modified version of a Separate Statutory Agency. This is the ideal approach to building a full-service, trusted agency, composed of a central data preservation and management facility and a series of access and service nodes located in research institutions. It takes full advantage of existing research infrastructure, has long-term stability, a direct connection to research data users and producers, and the capacity to represent Canada's interests in the development of international data standards.
2) Create a National Research Data Archive Network under the auspices of the Social Sciences and Humanities Research Council. This approach captures the characteristics of the first model, but does not require legislation. It benefits from a direct, immediate connection with researchers and established accountability and funding structures.
3) Create a Special Operating Agency within the National Archives of Canada. As a stand-alone division within the National Archives, this approach takes advantage of existing archival infrastructure and expertise. This has not been the preferred approach in other countries, because the core mission of a national archive and a national research data service are fundamentally different. Nevertheless, as a Special Operating Agency, the service could potentially have both stability and the capacity to develop a trusted research data preservation, management and access system.

As a next step, the Working Group recommends that SSHRC and the National Archivist create a Steering Committee to select the appropriate approach to setting up a National Research Data Archive Network, or research data archiving service, further define the characteristics and funding requirements for such a service, and promote its establishment.

Report Contents

Sections

Building Infrastructure for Access to and Preservation of Research Data in Canada

1. Introduction

The network of institutions and agencies that make up the infrastructure supporting Canada's knowledge economy currently has a serious gap. Canada lacks a national agency to preserve, catalogue and provide systematic, efficient and convenient access to research data. This digital information enables researchers to substantiate existing knowledge, replicate and verify research findings and explore and create new knowledge. Effective access to, and use of, research data can play a central role in Canada's innovative capacity. The necessary infrastructure, however, must be in place.

In October 2000, the Social Sciences and Humanities Research Council and the National Archivist of Canada mandated a Working Group of research and archival experts to consult with the research and archival communities and assess the need for a national data archiving service or function. After completing this assessment, and compiling extensive evidence for the need of such a service, the Working Group investigated research data archives in other countries and explored possible approaches to building such a core research facility in Canada.

The Working Group now recommends the establishment of a Canadian agency to close this gap in the infrastructure of the Canadian knowledge economy – the creation and long-term, stable support of a National Research Data Archive.

Today, almost all research takes place in a digital environment. Complex multi-layered statistical databases, digital maps and images, and encoded texts are now commonplace tools for researchers. Although these resources have dramatically expanded the scope of research, and increased its efficiency, the institutional structures required to preserve, manage and make accessible that digital information have not kept pace. This situation undermines the innovative capacity of Canadian researchers and places tens of millions of dollars worth of highly valuable research data at risk.

To build a knowledge society, to foster innovation, and to deal with pressing, complex social, political and economic problems depends in large part on the discovery of knowledge through research. In order to be responsive and efficient, while incorporating multiple perspectives, researchers require access to, and sharing of, a wide variety of research data. For this to happen, infrastructure is necessary. Today, many elements are in place – university research libraries and data services, research support councils, high-speed data transmission networks – but one vital element, a facility for storing, distributing and preserving research data is missing.

Good public stewardship demands that public investment in research data realise maximum returns. In order to maximise returns, research data should be used as many times, and in as many different situations, as possible. This can only happen if we put in place effective research data infrastructure. The cost of inaction not only puts our investments in science at risk, it undermines one of the core responsibilities of government.

2. What is Research Data Archiving?

Unlike many forms of traditional archiving, research data archiving is not about keeping records for legal, historical or cultural purposes; it is about meeting the needs of researchers operating in today's digital environment. The core mission of a research data archive is not to preserve the recorded memory of a group, organization or nation, but to provide a vital service to the research community.

Although there are many emerging institutional needs related to digital materials, the Working Group examined only the data access, management and preservation needs of the research community, mainly the social sciences and humanities. From this perspective, the Working Group defined the process of research data archiving as preserving, managing and making publicly accessible digital information structured through research methods with the aim of producing new knowledge. This process provides stewardship for those outputs of research that exist between initial information and published results. Acquisitions would include digital information produced by researchers and of interest to researchers, subject to the limitations of financial resources and retention protocols developed by research data archivists and the research community itself.

National research data agencies and archives in other countries provide a broad range of access, preservation, and management services to their respective research communities, including on-site and off-site storage, access to catalogues and data sets through the Internet, retention protocols, metadata creation, migration of data across software and hardware systems, training and developing international standards. In offering these services, they play an active and crucial role as information and knowledge brokers.

I see a National Data Archive as an institution that is trusted and recognized as having the Canadian mandate to preserve research data, to work with other governmental and non-governmental agencies in ensuring that their data management practices incorporate preservation standards, to work closely with other Canadian institutions charged with preserving Canada's heritage to guard against gaps in responsibilities, to co-ordinate and represent Canada in international research data exchanges and in the development of related standards, to provide access to these data, to educate Canadians about the use of research data, to contribute to new research by helping create new data from archived data, to help safeguard privacy in Canadian society in light of massive amounts of stored digital information on individuals, and to conduct research and development into all aspects of data preservation.

—Charles Humphrey, Data Librarian,
University of Alberta, NDAC
Working Group Member

3. The Need for a National Research Data Archive

As one of the Working Group members 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. [1]

Research information in digital form is extremely fragile yet capable of being collected in huge quantities. Today, we are only beginning to understand how to preserve and manage this information effectively. Although there are no easy short-cuts for dealing with such issues as media obsolescence, digital "rust", 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.

In the initial phase of the National Data Archive Consultation, the Working Group sought input from a broad range of stakeholders who use, manage and produce research data related to the social sciences and humanities. The objective was to assess the need for a national research data archival service or function (see Appendix E). This assessment brought to light a number of structural gaps:

- Currently, there is no national institution preserving, managing and making research data publicly accessible on the scale required to support the Canadian research community. The National Archives of Canada does not have the resources to do so;
- University research data services have neither the resources nor the responsibility to act as nationally-oriented research data archives. Although they are struggling to fill the gap left by the absence of a national data archival service, university data services are, in general, only mandated to provide local patrons with access to readily-available data;
- The SSHRC Data Archiving Policy, which directs the researchers it supports to deposit their data with university data services, has not achieved its objectives. In fact, over an eleven-year period only 10 data sets have been deposited with the university data depositories listed in the SSHRC Guide. Although some researchers are reluctant to share their data, it would be unethical for SSHRC to enforce this policy in the absence of a facility that would allow researchers to abide by the regulations;
- Canada has no co-ordinated voice in setting international research data standards, in metadata schemes such as Data Documentation Initiative, in tools for data access such as the Networked Social Science Tools and Resources (NESSTAR) project, and in collaborative international infrastructure projects such as the European Union Frameworks. As well, Canada lacks national representation on the International Federation of Data Organizations or participation in the initiatives of the Council of European Social Science Data Archives;
- One of the paramount problems researchers face today is difficulty in locating data relevant to their research. There is no 'union list' or catalogue of data sets held by data producers, distributors or other researchers. As a result, researchers may needlessly replicate costly studies, rely on anecdotal rather than empirical evidence, or use substitute data from other countries. Potentially, a national data service could place information about data sources, as well as the data itself, directly on the researchers' desktops, thereby saving time, money.

[1] Numerous organizations are currently wrestling with research data archiving policies and structures, including the Library of Congress, the Economic and Social Research Council of the United Kingdom, the U.S. National Institutes of Health, the U.S. National Archives and Records Administration and the National Research Council, the International Council for Scientific and Technical Information, and the International Council of Scientific Unions.

As well, a National Research Data Archive could serve fundamental needs by:

- Ensuring the authenticity of research data, a growing concern among both research data producers and data users. Authentication procedures embedded in the process of creation, transmission, receipt, use, maintenance and preservation of data files are the most effective way to ensure the authenticity of data over time. Currently, we have neither national standards of this kind nor any agency to oversee their application;
- Reformulating and articulating, at the national level, security standards that protect data adequacy and consistency. These standards should address: (1) methods for identifying data assets and risk-management procedures for assessing vulnerabilities; (2) identification of legal, statutory, regulatory and contractual requirements, 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.

If Canada were to build a National Research Data Archive, would it be used? Ample experience in other countries shows that data usage is growing in number of users and frequency (see Appendix C).

Among the top ten most popular data sets requested by users of the UK Data Archive in the 2000/01 fiscal year, four of these titles were from government departments, two were co-sponsored by government departments, and four were sponsored by a major research granting agency.

UK Data Archive Annual Report 2000/01

Two of the most common measures of activity levels of data archives are the size of their collections and the number of patrons whom they serve. For example, last year, the ICPSR at the University of Michigan added 1,835 data files to its collection, an eight-percent increase from the previous year. At the same time, it disseminated five thousand gigabytes of data to its patrons. During the same period, the UK Data Archive processed over 500 acquisitions and served 1,000 patrons who had placed 2,000 orders for a total of almost 9,000 data files. Over a three-year period, this was an increase of 2,000 data files delivered to users.

Several data archives record use statistics based on Web traffic. The Oxford Text Archive, for example, reported over 18,000 downloads of electronic texts during1999/2000. This electronic usage outnumbers Oxford Text Archive offline orders by a factor of 39. In addition to file downloads, the number of user contacts is also captured from Web statistics. For example, the ICPSR reported a substantial growth in patron contacts as a result of more users relying on the Internet for research and teaching. Over the past three years, during which more ICPSR resources were made available online, the agency reports an increase of more than one thousand gigabytes of data being accessed.

Data archives also maintain use statistics for other services. For example, the ICPSR training program consistently supports a yearly enrolment of between 500 and 540 participants. In another example, the Norwegian Social Science Data Service (NSD) maintains statistics about researchers' use of their service to investigate projects for legal compliance. NSD reports that this service has grown as much as 65% in a given year. Reference services usually maintain their own statistics. During 2000/2001, UK Data Archive staff fielded 332 post-order inquiries for assistance with data files, which represents just one aspect of reference services. During 2000 the Archaeology Data

Service reported 174 total inquiries, with questions touching upon catalogue use, technical assistance, and general archaeology information. The History Data Service received approximately 480 general reference inquiries during this same period. As well as reference support, the Oxford Text Archive provided technical assessments for 125 grant applications.

Another statistic used by some data archives is the volume of licenses for software that their service develops and distributes. For example, NSDstat, which is developed and distributed by NSD, is licensed to approximately 2,000 institutions in Norway and 200 organizations internationally. While an exact number of individual users per license is unknown, experience indicates that several individuals have access to NSDstat through a single copy of the license.

Larger data archives also record statistics about their international activities. For example, the German Central Archive for Empirical Social Research (ZA) reports that they consistently have 50 international scholars each year doing on-site research with data at the ZA EUROLAB. The ZA also integrates the data and documentation for a number of international projects, including the International Social Survey Program for 38 countries and the Eurobarometers for the European Commission.

Overall, data archives that offer comprehensive services (including training, software development, and online access to data files) demonstrate significant use by researchers of a national and international scope. In every case, this use is growing.

There are many reasons to share data from NIH-supported studies. Sharing data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new data sets when data from multiple sources are combined. By avoiding the duplication of expensive data collection activities, the NIH is able to support more investigators than it could if similar data had to be collected de novo by each applicant.

National Institutes of Health (US),
Policy Statement on Sharing Research Data

4. The Building Blocks of a National Research Data Archive

Over the past several years, the Government of Canada has taken major steps towards building a comprehensive and coherent research infrastructure and research support system in Canada. Measures such as the creation of the Canada Foundation for Innovation and the building of CA*Net3 have gone a long way towards filling existing gaps. One of the few gaps remaining, however, is a facility or institution with the responsibility for ensuring preservation of, and access to, research data. Nevertheless, Canada already has many building blocks for this agency in place.

University Data Services – Perhaps the most important of these building blocks are the existing university data services. Although limited resources prevent them from acting as full-service agencies, the university data services have the potential to be nodes of a National Research Data Archive. This has been strengthened enormously through the experience of the Data Liberation Initiative, where librarians and data archivists from 66 universities have come together to form a consortium to improve access to Statistics Canada data. These dedicated professionals remain in close contact with each other, sharing best practices, information about data sources, ways to improve services for their clients, and the latest advances in technical capacities and standards. With sufficient resources, university data services could form a comprehensive, nation-wide network of contact points for researchers who wish to access research data collected by others, deposit data they collected themselves, seek training in advanced statistical and data handling skills, and obtain advice on how to conform to data standards and best practices. Perhaps more importantly in the long run, the network of university data services personnel could act as a feedback system from users, helping to shape and improve the services provided by a National Research Data Archive, and ultimately the knowledge created by Canada's researchers.

Canadian Archival Institutions – As with university data services, those Canadian archival institutions with a specific research mandate offer other potential nodes in a National Research Data Archive network. They exist in local, regional and institutional environments, either as independent entities, as part of a parent institution, or within municipal, provincial and federal levels of government. Furthermore, they exist in many communities that do not host universities. While Canadian archives have, until recently, dealt primarily with non-digital records, their community infrastructure, descriptive standards, best practices, extensive experience with privacy protection and copyright, etc. all provide a firm basis from which to develop the knowledge and skills to participate in a national research data network.

International Representation – Although lacking national authority, some university data services staff currently provide one of Canada's principal connections with numerous international bodies and agencies charged with the management of research data and the establishment of international standards for metadata creation, data sharing, and preservation. The creation of these standards, agreements and common practices are vital in a scientific world that increasingly works beyond national borders. Employing their experience and expertise in a co-ordinated effort will mean that Canada's interests are represented when key decisions, with long-term implications, are being made.

Data Transmission Infrastructure – Connecting university data services is CA*Net3, and soon, CA*Net4, the ultra-high speed national optical data transmission network, built by CANARIE Inc. Now linking all of Canada's major research institutions, CA*Net3 provides the extensive pipeline necessary for the nation-wide distribution of research data. The huge capacity of this network allows for the rapid, efficient and reliable transmission of very large, complex data sets. This is crucial for the future. Research data sets are increasing in both size and complexity at an amazing rate.

Management Frameworks – The management frameworks for the use of research data are just as important as the digital pipelines and access nodes. Because of the sensitive information contained about individuals, social science data in particular must be managed within a comprehensive ethical framework, as well as Access to Information and Privacy legislation. The Tri-Council Guidelines on Research Involving Humans provides one of these frameworks. These guidelines spell out in general terms the principles by which a National Research Data Archive should treat privacy and confidentiality. Along with the university-based Research Ethics Boards, we have both the rules and the institutional capacity to ensure that information on individual citizens is protected. These Boards determine the conditions under which sensitive data can be deposited and released and so constitute a built-in, first stage screening process for a National Research Data Archive.

Research and Development – In our rapidly developing digital world, many aspects of handling research data are done without sufficient knowledge. Ensuring the quality, authenticity and security of research data are examples. A National Research Data Archive will be positioned to capitalise on the knowledge emerging from cutting-edge research in this field, including, for example, the SSHRC-funded InterPares project.

Partner Institutions – Various institutions can play an important role in the operations and services of a National Research Data Archive. Both the National Archives of Canada and the National Library of Canada have, over the years, developed significant expertise with their respective records and in the transition of those records to electronic form. Storage environments, descriptive standards, physical and logical format migration, and protection of copyright are just some of the areas where knowledge could be shared and joint projects undertaken.

Research Data – The central building block of a research data service is the research data itself. Not all research data sets should be preserved, of course. Some will be of limited use beyond the project for which they were collected; some will contain personal identifiers that cannot been effectively removed; some simply re-produce data collected elsewhere. Determining what should, and what should not, be preserved, however, lies at the core of archival science, and is critical to an effective partnership between researchers and data archivists.

The existence of plentiful research data is not in question. In the first phase of the consultation, the Working Group determined that SSHRC-funded researchers produce, on average, some 400 data sets each year. Since SSHRC is able to support only a fraction of the Canadian social sciences and humanities research community, the total number of data sets produced each year could be three or four times this number. This does not include those data sets produced by natural scientists, health scientists or research engineers, but it is not unreasonable to estimate that some 4,000 to 5,000 are produced annually, all of which are supported by public funds. Although impossible to know in precise detail, this represents a public investment of tens of millions of dollars annually.

Government Research Data – The Working Group's investigations of data archives in other countries revealed that, in the social sciences, government-produced research data are often more widely used than data produced by researchers themselves. One valuable role for a new agency would be to provide a preservation facility, catalogue and access conduit for government collected research data. The Working Group heard testimony on numerous occasions that accessing such information is, at best, difficult and time-consuming, and at worst, impossible. Yet, it has been estimated that departments such as Statistics Canada, HRDC, Health, Natural Resources, Environment, Justice and many others spend upwards of $1 billion annually on collecting data. Finding effective and efficient means for researchers to utilise this data is a matter of good public stewardship. [2]

Preservation Services for Other Research Agencies – Today's information technologies greatly facilitate our ability to access, manipulate and apply digital information to research questions of fundamental importance to Canadians. However, the long-term preservation of digital research materials is one area, from both a technological and institutional perspective, that has not kept pace. In the research world, the current emphasis is on compiling and providing access to information, predominantly through the Internet. Inter-agency cataloguing and preservation services are often considered of secondary importance or ignored altogether. The Canadian Institute for Health Information, the Canadian Centre for Justice Statistics, the Canadian Information System for the Environment, GeoConnections, and the recently announced Community Social Data Strategy of the Canadian Council on Social Development all provide excellent data access systems, but lack a well considered, adequately supported, long-term data preservation strategy. One of the most important roles that a National Research Data Archive can play is providing the preservation services and expertise for these, and many other, research data access initiatives.

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.

Questionnaire Respondent

[2] Canadian Global Change Program, Data and Information Systems Panel, "Data Policy and Barriers to Data Access in Canada: Issues for Global Change Research", (Royal Society of Canada, 1996), p.7.

5. Towards an Agency Model: Lessons Learned in the International Arena

The Working Group examined all existing national research data archives focusing on the social science or humanities (see Appendix C). This investigation included face-to-face interviews with data agency directors, comparative analysis of policies and regulations, examination of services, mandates, budgets and governing structures. Chief among the lessons learned are the following:

- Many countries have long recognized the need for a research data archive to assist and support the work of the research community. Several of the data archives examined have been in existence for 30 years or more;
- Although many services of a research data archive, particularly those related to access and training, are best distributed among a number of locations, for reasons of economy, practicality and effectiveness, preservation, network management and standards development functions are best performed within one facility;
- No two research data archives are the same. Each was established within a specific national or disciplinary context that reflected the particular needs of the research community it serves. They range in size from small, disciplinary specific, limited service organizations to large, multidisciplinary, full service, internationally networked, R&D focused, national institutions;
- Successful research data archives are directly attached to a country's research infrastructure, rather than to its archival community. They are characterised by a service orientation that emphasises access to, and preservation of, the most useful data for research, rather than capturing records of the past;
- Research data archiving is a complex and highly technical business. Successful data archives employ dedicated, professional data experts and place considerable emphasis on training the next generation of research data managers. Developing highly qualified personnel serves the needs of both the research community and many other areas of the public and private sectors that have to deal with large volumes of data;
- There is a direct correlation between the funding stability of a research data archive and its success in supporting the research community. By its very nature, archiving is a long-term enterprise. The most useful data archives are those that are assured of their continuing existence;
- Although research data archiving requires long-term funding commitments, the institutional costs are always only a very small fraction of the costs of data collection;
- Building trust with both users and producers of research data is vital. If users cannot rely on the timely and efficient delivery of high quality data, and if depositors are not convinced that their intellectual rights and the protection of their participants will be upheld, no one will trust or use the services provided;
- The most successful data archives have both institutional independence and flexibility. They work in close co-operation with numerous government departments and universities but are not dependent upon any particular one for financial stability or decision-making. Independence is necessary to ensure that the data access needs of the research community remain the first priority, rather than the record keeping needs of government departments or traditional cultural archives. Flexibility is important for the adoption of new technologies and the ability to respond to the changing needs of researchers.

The Working Group's detailed survey of 36 institutions produced three generalised approaches to preserving and providing access to research data. Each represents the organizational characteristics of today's national data archiving services:

- **A small scale, specialised topical data archive**, usually hosted by a university department, with limited data handling capability, employing off-the-shelf technology. Clientele are often restricted to one, or a small group, of research disciplines, and annual operating budgets range from between $200K to $400K.
- **A medium sized, agency-based data archive**, whose parent organization is usually a national research institute or government department. Often located on a university campus to better serve its core research clientele, these archives base their mandate, and subsequent collection activities, on that of their parent agency. Services are moderately extensive, and staff sometimes take leadership roles in relevant national and international organizations. Annual budgets range from $500K to $1.5M.
- **A comprehensive research data archive**, servicing a wide variety of communities, including academic researchers, NGO and government policy analysts, public archival agencies, and individual citizens. Often established through legislation, such data archives are recognized as a national institutions responsible for the general principles and specific duties outlined in their founding Acts. Through one or more physical locations, and extensive use of the Internet, a comprehensive range of services are provided, often including specialised training, educational outreach, technical support and R&D. Data management capabilities are extensive and often developed in-house. Such agencies have established working relationships with other national institutions and government departments, and staff are often leaders of international associations and actively engage in international data exchanges. Annual budgets range from $3M to $6M.

Benefits of Depositing and Archiving Data:

* *Reinforces open scientific inquiry;*
* *Encourages diversity of analysis and opinions;*
* *Promotes new research and allows for the testing of new or alternative methods;*
* *Improves methods of data collection and measurements through the scrutiny of others;*
* *Reduces costs by avoiding duplicate data collection efforts;*
* *Provides an important resource for training in research;*
* *Ensures the safekeeping of data;*
* *Allows owners to avoid the administrative tasks associated with external users and their queries;*
* *Fulfils grant obligations regarding making funded research available to the research community;*
* *Enables researchers to demonstrate continued use of the data after the original research is completed.*

Inter-University Consortium for Political and
Social Research Web Site, University of Michigan

---

6. Core Principles and Assumptions

The Working Group concluded that a National Research Data Archive should operate according to a set of core principles. The overall objective should be to create a "trusted system" that provides the research community with an accessible and comprehensive service empowering end users to locate, request, retrieve and use data resources in a simple, seamless and cost effective way. Such a system should follow these core principles:

1) A National Research Data Archive should support the creation of knowledge by being an integral part of the research process and should aid discovery and decision-making in Canada, including the formation of public policy, by preserving and making accessible sources of evidence;
2) Whenever possible, access to research data should be as open as possible and free of charge;
3) Ensuring confidentiality, privacy and the protection of human research participants should be paramount in all operations;
4) Data collected with the use of public funds should remain publicly available, subject only to conditions of fair prior use by the depositor and the ethical and legal provisions under which the data were collected.

The Working Group heard on numerous occasions, and from many authoritative and experienced sources, that establishing trust is the key factor in building a successful research data access and preservation system. This can only be accomplished if the institution's users and depositors know that the archive is an integral part of their research processes, that it will provide useful services, and that it will add value to their work. Moreover, the data service must support and actively uphold established regulations and guidelines regarding protection of confidentiality, privacy and intellectual property. Most importantly, in order to be a trusted system, a new agency must have long-term stability, both in its institutional structure and financing. This is one of the hard lessons learned by many data archives around the world. The source of mandate, governance, accountability and a stable, long-term commitment to providing the necessary financial resources determine success or failure.

7. Options for Canada

Drawing on these lessons and consultations with the research community, the Working Group concluded that Canada would be best served by an agency with the following general characteristics:

- A comprehensive mandate derived from, and responsive to, the needs of a wide variety of stakeholders;
- Dedication to society and the individual as the core subjects and scope of the target data;
- A service orientation that emphasises both preservation and access;
- Protection of privacy and confidentiality as a core element of its operating principles;
- The ability to process data according to international standards, engage in international data exchanges, and represent Canadian interests in international negotiations;
- The capacity to conduct advanced research and development in archival and information sciences;
- Application of the latest information and communications technologies to maximise access to research data while reducing the time and cost burdens on researchers;
- The capacity to educate and train both the producers and users of research data and the next generation of data management professionals;
- Established, on-going working relationships with other national agencies and organizations, such as the National Archives and National Library, as well as extra-governmental agencies such as CANARIE Inc.;
- Institutional memberships and other formal data exchange agreements with major data archives outside Canada, such as the ICPSR in the United States and the European CESSDA network;
- Public funding, on a long-term sustained basis, as its principal source of support. This could be supplemented by the sale of value-added data products and consultation services to for-profit organizations, but should not constitute core funding.

The Canadian context, however, shows that a National Research Data Archive must also have the following specific traits:

- A fully bilingual service;
- Access to research data produced by all levels of government, while respecting federal and provincial jurisdictional boundaries in areas such as education and health;
- Respect for, and assistance in developing, Canadian intellectual property, copyright, privacy and confidentiality legislation, regulations and guidelines;
- Close working relationships with major Canadian data producers such as Statistics Canada and provincial statistical agencies;
- Use and support of existing research infrastructure, research support services and funding support programs, including existing university data services and research libraries, the research support councils, the Canada Foundation for Innovation, the Canadian Centre for Justice Statistics, the Canadian Institute for Health Information;
- Interest in research data from both the social sciences and humanities, and, where appropriate, the natural sciences, health sciences and engineering.

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

A Canadian National Research Data Archive should meet data preservation and access needs, as well as push the boundaries of information and archival science. It should build on existing research infrastructure while learning the lessons provided by a generation of data archiving experience in other countries. Most importantly, it must be successfully adapted to fit the Canadian social and institutional context, while meeting the public need for accountability and effective governance.

In exploring how a National Research Data Archive could be created, the Working group examined existing federally-funded university research centres, sought advice from the Privy Council Office and used the guidelines provided by the Treasury Board's Framework for Alternative Program Delivery. The Working Group considered six possible options and discussed each in detail; reviewing institutional and governance structures, requirements for start-up and long-term stability, and both strengths and weaknesses from the perspectives of data users and producers (See Appendix D).

The Working Group first explored the option of creating a new division or Special Operating Agency within an existing national institution such as the National Archives or National Library of Canada. While the mandate of the National Archives is broad enough to extend to unpublished research data, its current level of funding could not support a move into such a new area of service, while it simultaneously responds to the government-wide challenges of information management in the era of e-government, the transition of its records into electronic form and the extensive digitization of its existing holdings. Furthermore, the failure of an earlier attempt to create a data archives division within the National Archives (1973-1986) suggests a disjunction between the broad cultural preservation role of the National Archives and the specific service role that a National Research Data Archive would be called on to play within Canada's research infrastructure. These differences extend from acquisition strategies to available staff expertise, current descriptive practices and the needs of clientele.

The National Library of Canada does collect a limited number of research data sets that meet the definition of "publications". These are, however, a small sub-set of the research data sets requiring preservation in Canada. As with the National Archives, preserving, maintaining and providing access to the two institution's current holdings do not require the extensive knowledge of quantitative research methodology, statistics and advanced computing skills necessary to meet the needs of those who would use a National Research Data Network. The Working Group believes that the unique requirements of the research community, and the research data they use, could marginalize the activities of a research data archive within these existing institutions, thus undermining the long-term stability needed for success.

Another option that the Working Group examined was the creation of what the Treasury Board refers to as a Public Partnership. This involves establishing an agency as a partnership between federal and provincial levels of government. Although this route has certain interesting aspects, it does not lend itself to the building of direct connections with the university and non-governmental research communities. As national data archives in other countries have learned, this is a crucial element in building a trusted system.

A Separate Statutory Agency or Departmental Corporation has many of the characteristics necessary for a robust, full-service and effective National Research Data Archive. It would be a permanent institution secured by legislation. It would be an element within the policy framework of the Innovation Agenda, focused on research, capacity building, stewardship and international competitiveness. It would have clear lines of authority and accountability, and a ministerial champion. Funding would be secure, stable and from a single source. Like Statistics Canada, it would have the potential, and the means, to develop a reputation as a "trusted system", and could have official national representation status in the international arena. The one important element missing is a direct connection with the research data user community.

The final option discussed, a University-Based Centre, has this direct, immediate, on-site connection. With such a facility, a sense of ownership, operations and policies would be in the hands of the associated university members. It builds on existing data services, expertise and technology infrastructure within universities; it could use a hybrid centralized/de-centralized system, where the Centre takes care of preservation and data set processing and the associated members act as local facilities for access to data, deposit of data, on-site advice, and training activities; scope of the agency is scalable and could include NSERC and CIHR areas of science. Finally, digital archival research activities would take advantage of proximity to university-based information science researchers. The principal weakness of this option is that it lacks long-term stability. A second weakness is that it would not necessarily have the authority to act as a national voice in the international arena.

After examining and discussing all these options in detail, the Working Group concluded that the nature of the Canadian federal system of government, new communication and information technologies, the particular characteristics of the research community, and the emerging needs of Canada's knowledge economy, present a unique opportunity for institutional innovation – the creation of a hybrid agency that combines the stability of a separate statutory agency and the user community connections of a university-based research centre.

8. Recommendations for the Implementation of a National Research Data Archive Network

In order to build an effective national research data archiving service, one that best meets the needs of Canada's knowledge economy, fosters innovation, builds on the strengths of existing infrastructure, ensures effective public stewardship and gives Canada a voice in the international arena, the Working Group recommends that the Government of Canada undertake the following:

- Legislate the creation of a National Research Data Archive Network as a modified version of a Separate Statutory Agency;
- Require that this agency report to Parliament through either the Minister of Industry or the Minister of Canadian Heritage, or – preferably – a combination of the two, in an accountability structure similar to that of agencies such as the Climate Change Secretariat;
- Enable this agency to operate at arm's length, in the same manner as the federal research support councils;
- Allocate operating funds directly to both the central facility and the nodes by annual vote in Parliament, within the regular federal budget process, or alternately, flow funding through participating federal research support councils, as occurs with the Networks of Centres of Excellence.

Regarding the structures and operations of a National Research Data Archive Network, the Working Group further recommends:

- That the new agency develop a comprehensive service network, with a central facility responsible for data management, standards development and preservation and a series of nodes, located within university research data services and other institutions responsible for providing access, depository, training and consultation services for researchers. It is suggested that institutions wishing to become nodes form a consortium to seek initial infrastructure funding from the Canada Foundation for Innovation and various provincial matching-fund research agencies;
- That a Management Board be created to govern the National Research Data Archive Network, composed of representatives from the various regions of Canada and the various stakeholder groups that manage, use and produce research data;
- That the agency develop, over time and in response to the identified needs of the research community, a suite of research data access, management and preservation services;
- That the agency develop the capacity to further our knowledge and understanding of information management sciences, ethical and legal frameworks, knowledge management practices, and promote a culture of research data sharing within the research community;
- That the agency enter into formal co-operative working relationships with other national institutions such as the National Archives and the National Library, and data access and preservation agreements with major data producers such as Statistics Canada and provincial statistical agencies;
- That the agency be given the authority to act on behalf of the Government of Canada in international negotiations related to research data management standards and common practices.

Although the Working Group is convinced that the model outlined above would place Canada at the forefront of data archiving and information science, and would substantially increase the competitive advantage of the Canadian research community, the members are also aware that the best or ideal solution is not always the most practical or feasible. With this in mind, we suggest two alternative routes to establishing a National Research Data Archive.

1) **A SSHRC National Research Data Archive Network** – following the approach taken by the Economic and Social Research Council in the UK, this would involve establishing a university-based facility and network under the auspices of SSHRC. The agency would be accountable to the SSHRC Board of Directors. It would have the same management and network structure, and range of services, as the option outlined above. Such an agency would not require enabling legislation, since it would fall within the research support function of the SSHRC mandate, but it would also lack the long-term stability that legislation provides. It would benefit from a direct connection with the research community, as well as from SSHRC's working relationships with major data producers such as Statistics Canada. Conceivably, it would be scalable to include all areas of scientific and humanities research, and could take advantage of SSHRC-funded research in information and archival sciences.

2) **A Special Operating Agency within the National Archives of Canada** – although on the surface this may seem to be the most logical route for establishing a National Research Data Archive, it should be noted that the Working Group heard very few voices recommending this course of action. Moreover, the investigation of research data archives in other countries revealed that only one – the Danish Data Archive – is directly attached to a national archive, and anecdotal evidence suggests that this arrangement is having a detrimental effect. Nevertheless, the creation of a Special Operating Agency within the National Archives could provide a simple solution. A Special Operating Agency would be able to draw on the archival experience of the National Archives staff, use existing facilities, as well as technical and administrative infrastructure, and have the stature and authority to act as Canada's voice in the development of international standards and practices. As a Special Operating Agency it would have a degree of autonomy within the management structure of the National Archives, while still being accountable to the National Archivist. This would provide greater stability than that of the now defunct Machine Readable Archives Division. The most significant disadvantage of this approach is that the agency would not have a direct, immediate connection with the research community, either through its management structure or through the university data services. Although this could be built, the agency would still have to exist within a federal government body whose core mission is to preserve the national memory and the records of government, not service the data needs of researchers.

Researchers are in agreement that the infrastructure to allow for sharing of research data is long overdue in Canada and that we need to have a coherent infrastructure to collect, document, share, and preserve digital research data. In particular, it is critical to reduce the high costs of data collection and make files available for secondary analyses.

Submission from the University of Calgary, Office of the Vice-President (Research)

9. The Cost of a National Research Data Archive Network

The amount of funding required to establish and maintain a national research data service depends on the size and scope of its operations and on the range of services it provides. The key consideration is to define the minimum level of funding that would be required to provide an adequate level of services. Too small a funding base would not only restrict the range of services that could be provided but might threaten the continuation of funding. There is a danger that if the agency had to exist on too low a level of funding it would become too narrowly focused on a limited number of disciplinary areas or types of services. This in turn might arouse resentment from the users not being serviced and thus jeopardise the continuation of funding.

Funding for a National Research Data Archive Network -- both the central facility and its nodes -- should come through the Federal Government of Canada. Although supplementary funding can be secured through other routes, such as R&D grants, the sale of value-added data products and charges for speciality consultation services, general agency operations should be funded this way. This is the only effective means to ensure that the research data archive serves all Canadians, across all regions, has long-term stability, meets the needs of a broad range of researchers and research data producers in academia, government, NGOs and the private sector.

A detailed costing of a National Research Data Archive Network, along the lines recommended here, is beyond the resources currently available to the Working Group. The international study of existing archives, however, provides a solid benchmark for the levels of funding necessary to provide certain levels of services.

In current Canadian dollars, and once fully operational, the low end of an annual operating budget for a full-service research data agency is approximately $3 million. As pointed out by the Irish Data Archive feasibility study, approximately 40% would be devoted to acquiring, processing, cataloguing and preserving data, while the remaining 60% would be spent on processes involved in servicing user needs. [3]

Initial infrastructure costs would depend on a range of factors, including the location and size of the central preservation and processing facility, the number of nodes that join the network, the distribution of specific functions between the nodes and the central facility, and the overall capacity and complexity of the computing hardware. If the agency were to be attached to Canada's research institutions, rather than the National Archives or other federal government department, infrastructure funding could be sought through the Canada Foundation for Innovation and the various provincial matching-fund agencies.

The services provided by the network, and therefore its operational costs, could be scaled up over time as both deposits and usage grows. This has been the usual route taken in other countries. The volume of data held does not significantly affect operational costs, since the price of digital storage is declining rapidly. Rather, the experience in other countries is that data handling, management, and value-added services grow as the research community uses the services and becomes aware of its real and potential benefits.

[3] The Data Archive, University of Essex, "The Irish Data Archive Feasibility Project", 1997, p.49.

10. Next Steps

In order to initiate the process of creating a National Research Data Archive Network, the Working Group recommends the following:

- The SSHRC Board and the National Archivist create a Steering Committee to initiate the implementation process, including seeking the support of AUCC, CARL, HSSFC and other relevant organizations, enlisting ministerial sponsorship if enabling legislation is required, and securing participation of stakeholders to further develop mandates and organizational structures;
- This committee should establish appropriate contacts with Justice and Finance Department officials, and officials from other relevant departments and agencies, to begin the process of implementation and further develop the operational details of the proposed new agency;
- This steering committee should be given the responsibility for developing selection criteria for the central facility and nodes of the National Research Data Archive Network;
- The committee should also advise on criteria for the composition of the Management Board, if one is called for;
- The Working Group strongly recommends that these activities begin as soon as is feasible in order to make the case that a National Research Data Archive Network can play a central role in furthering the Government of Canada's Innovation Agenda.

10.4 Declaration On Access to Research Data from Public Funding

Declaration On Access to Research Data from Public Funding

adopted on 30 January 2004 in Paris
The governments (1) of Australia, Austria, Belgium, Canada, China, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Russian Federation, the Slovak Republic, the Republic of South Africa, Spain, Sweden, Switzerland, Turkey, the United Kingdom, and the United States

Recognising that an optimum international exchange of data, information and knowledge contributes decisively to the advancement of scientific research and innovation;

Recognising that open access to, and unrestricted use of, data promotes scientific progress and facilitates the training of researchers;

Recognising that open access will maximise the value derived from public investments in data collection efforts;

Recognising that the substantial increase in computing capacity enables vast quantities of digital research data from public funding to be put to use for multiple research purposes by many research institutes of the global science system, thereby substantially increasing the scope and scale of research;

Recognising the substantial benefits that science, the economy and society at large could gain from the opportunities that expanded use of digital data resources have to offer, and recognising the risk that undue restrictions on access to and use of research data from public funding could diminish the quality and efficiency of scientific research and innovation;

Recognising that optimum availability of research data from public funding for developing countries will enhance their participation in the global science system, thereby contributing to their social and economic development;

Recognising that the disclosure of research data from public funding may be constrained by domestic law on national security, the protection of privacy of citizens and the protection of intellectual property rights and trade secrets that may require additional safeguards;

Recognising that on some of the aspects of the accessibility of research data from public funding, additional measures have been taken or will be introduced in OECD countries and that disparities in national regulations could hamper the optimum use of publicly funded data on the national and international scales;

Considering the beneficial impact of the establishment of OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (1980, 1985 and 1998) and the OECD Guidelines for the Security of Information Systems and Networks (1992, 1997 and 2002) on international policies for access to digital data;

DECLARE THEIR COMMITMENT TO:

Work towards the establishment of access regimes for digital research data from public funding in accordance with the following objectives and principles:

Openness: balancing the interests of open access to data to increase the quality and efficiency of research and innovation with the need for restriction of access in some instances to protect social, scientific and economic interests.

Transparency: making information on data-producing organisations, documentation on the data they produce and specifications of conditions attached to the use of these data, available and accessible internationally.

Legal conformity: paying due attention, in the design of access regimes for digital research data, to national legal requirements concerning national security, privacy and trade secrets.

Formal responsibility: promoting explicit, formal institutional rules on the responsibilities of the various parties involved in data-related activities pertaining to authorship, producer credits, ownership, usage restrictions, financial arrangements, ethical rules, licensing terms, and liability.

Professionalism: building institutional rules for the management of digital research data based on the relevant professional standards and values embodied in the codes of conduct of the scientific communities involved.

Protection of intellectual property: describing ways to obtain open access under the different legal regimes of copyright or other intellectual property law applicable to databases as well as trade secrets.

Interoperability: paying due attention to the relevant international standard requirements for use in multiple ways, in co-operation with other international organisations.

Quality and security: describing good practices for methods, techniques and instruments employed in the collection, dissemination and accessible archiving of data to enable quality control by peer review and other means of safeguarding authenticity, originality, integrity, security and establishing liability.

Efficiency: promoting further cost effectiveness within the global science system by describing good practices in data management and specialised support services.

Accountability: evaluating the performance of data access regimes to maximise the support for open access among the scientific community and society at large.

Seek transparency in regulations and policies related to information, computer and communications services affecting international flows of data for research, and reducing unnecessary barriers to the international exchange of these data;

Take the necessary steps to strengthen existing instruments and - where appropriate - create within the framework of international and national law, new mechanisms and practices supporting international collaboration in access to digital research data;

Support OECD initiatives to promote the development and harmonisation of approaches by governments adhering to this Declaration aimed at maximising the accessibility of digital research data;

Consider the possible implications for other countries, including developing countries and economies in transition, when dealing with issues of access to digital research data.

INVITE THE OECD:

To develop a set of OECD guidelines based on commonly agreed principles to facilitate optimal cost-effective access to digital research data from public funding, to be endorsed by the OECD Council at a later stage.

(1) Including the European Community

11 Addenda

11.1 Acronyms

Acronym	**Full Text**
AUCC	Association of Universities and Colleges Canada
BIND	Biomolecular Interaction Network Database
CADR	Canadian Astronomical Data Centre
CARL	Canadian Association of Research Libraries
CIHR	Canadian Institutes for Health Research
CISE	Canadian Information Systems for the Environment
CISTI	Canada Institute for Scientific and Technical Information
CFI	Canada Foundation for Innovation
CNC CODATA	Canadian National Committee for CODATA
CODATA	ICSU Committee for Sciences and Technology
DNA	DeoxyriboNucleic Acid
DPOSS	Digital Palomar Observatory Sky Survey
EU	European Union
IC	Industry Canada
ICSU	International Council of Scientific Unions
IP / IPR	Intellectual Property / Intellectual Property Rights
fMRIDC	functional Magnetic Resonance Imaging Data Centre (USA).
GBIF	Global Biodiversity Information Facility
GDP	Gross Domestic Product
GERD	Gross Expenditure on Research and Development
GEOSS	Global Earth Observation System of Systems
HDF-S	Hubble (Telescope) Deep Field – South
HQP	High Quality People
IC	Industry Canada
ICSU	International Council on Science
IP / IPR	Intellectual Property / Intellectual Property Rights
ISO	International Standards Organisation
IT	Information Technology
LAC	Library and Archives Canada
NCASRD	National Consultation on Access to Scientific Research Data
NGO	Non-Governmental Organization
NIH	National Institutes for Health (USA)
NRC	National Research Council of Canada
NSERC	Natural Sciences and Engineering Research Council of Canada
OECD	Organisation for Economic Cooperation and Development
OST	Office of Science and Technology (UK)
PI	Principal Investigator
PIPEDA	Personal Information Privacy and Electronic Document Act
SciDIF	Scientific Data and Information Forum (ICSU planning forum)
SSHRC	Social Sciences and Humanities Research Council of Canada
UK	United Kingdom
USA	United States of America
WIPO	World Intellectual Property Office

11.2 Consultation Agenda
(Summary form as provided to Attendees)

Monday November 22 – MORNING – Session 1

Time	Activity	Description
7:30 AM	**Continental Breakfast**	Informal networking
8:30 AM	*Session 1:* Setting the Stage *Moderator* *David Strong*	Welcome and Introductions – **David Strong**
8:50 AM		“A Vision for Scientific Research Access “ – **Arthur Carty** *Introduction by* ***Ian Wilson, LAC***
9:10 AM		“The Power of Accessibility” – Chuck Hasel, *Genome Canada*
9:40 AM		“Implications and Possibilities of the Ministerial Declaration on Access to Publicly Funded Research Data” **Marie Tobin,** *Industry Canada*
10:00 AM		Coffee Break and informal discussions
10:30 AM		An Introduction to recent experiences in developing access to scientific research data – **Janet Halliwell**
10:40 AM		Lessons Learned: **Paul Uhlir** *(American perspective)* **David Moorman** *(Canadian social sciences perspective)* **Charlyn Black** *(CIHR perspective)*
12:00 PM	Lunch	Greetings from sponsor organizations **Patricia Kosseim,** *CIHR* **Eliot Phillipson,** *CFI*

Monday November 22 – EARLY AFTERNOON – Session 2

Time	Activity	Description
1:00 PM	*Session 2:* **Opportunities** *Moderator* *David Strong*	An Exploration of Opportunities *Facilitated session*
2:10 PM		Coffee Break
2:30 PM		Plenary Discussion of Opportunities *Facilitated Discussion*

Monday November 22 – LATE AFTERNOON – Session 3

Time	Activity	Purpose
3:00 PM	*Session 3:* **Challenges** *Moderator* *David Strong*	Identification of Challenges *Facilitated Session*
3:45 PM		Plenary Discussion of Challenges *Facilitated Discussion*
4:15 PM		Presentation of a Draft Impact Statement – **David Strong**
4:30 PM	**Assessing Our Progress**	Gauging community support “Homework”
6:30 PM	**Cocktails**	Informal Networking *Participants to re-convene at Canadian Museum of Civilization*
7:15 PM	**Dinner**	Key Note Speech – **Claire Morris**, *President, AUCC*

Tuesday November 23 – EARLY MORNING – Session 4

Time	Activity	Description
7:00 AM	**Continental Breakfast**	Informal networking
8:00 AM	*Session 4:* **Building a Vision for Access to Research Data** *Moderator David Strong*	Review, Preview, Big view
8:15 AM		An Exploration of the Future *Facilitated Session*
9:15 AM		Plenary Discussion of the Future *Facilitated Discussion*
9:45 AM		Coffee Break
10:00 AM		Inhibitors to Access – **Chuck Humphrey**

Tuesday November 23 – LATE MORNING – Session 5

Time	Activity	Description
10:30 AM	*Session 5:* **Identify Key Inhibitors** *Moderator David Strong*	Identification of Inhibitors *Facilitated Session*
11:30 AM		Plenary Discussion of Inhibitors *Facilitated Discussion*
12:00 PM	Lunch	Greetings from sponsor organizations **Steve Shugar,** *NSERC* **Michael Raymont**, *NRC*

Tuesday November 23 – AFTERNOON – Session 6

Time	Activity	Description
1:00 PM	*Session 6:* **Developing the Actions** *Moderator David Strong*	Actions to Address the Inhibitors *Facilitated Session*
1:45 PM		Plenary Review of Proposed Actions *Facilitated Discussion*
2:15 PM		Summarize Workshop and Outline Next Steps – **David Strong**
2:45 PM	**Close**	Closing Remarks – **David Strong**

11.3 Consultation Attendees

Name		Organization of Institution	Expertise
Cindy	Bell	Genome Canada	Genetics
Vijay	Bhargava	University of British Columbia	Electrical Eng
Charlyn	Black	University of British Columbia	Health Science
Sharon	Buehler	Memorial University	Health Science
Sheila	Chapman	Canadian Institutes of Health Research	Health Science
David	Crane	Independent Journalist	Journalism
Josef	Cihlar	Canada Centre for Remote Sensing Ottawa	Environment
Mark	de Jong	Canadian Light Source	Photonics
Bernard	Dumouchel	Canada Institute for Scientific and Technical Information	Information
Kenneth	Edgecombe	Queen's University	U. Corporate
Carole	Estabrooks	University of Alberta	Medicine
Alan	Evans	Montreal Neurological Institute	Neurosciences
Louis	Fortier	University of Laval	Biology
Dan	Gale	Canadian Microelectronics Corp.	Electronics
Thomas	Goldthorpe	University Heath Network (U of T)	Medical Informatics
Elizabeth	Griffin	NRC – Herzberg Institute of Astrophysics	Astrophysics
Chuck	Hasel	Genome Canada	Genetics
Vicky	Hipkin	University of Toronto	Astrophysics
Christopher	Hogue	Mount Sinai Hospital	Genetics
Gregory	Kealey	University of New Brunswick	U. Corporate
Richard	Keeler	University of Victoria	U. Corporate
Martin	Kryzwinski	Genome Sciences Centre	Genomics
Ian	Lancashire	University of Toronto	Social Science
Andreas	Laupacis	Institute for Clinical Evaluative Studies	Medicine
Marc	Lepage	Genome Canada	Genetics
Bryan	Lynch	St. Francis Xavier University	Chemistry
Susan	McDaniel	University of Windsor	U. Corporate
Paul	Melancon	University of Alberta	Biology
Ikechi	Mgbeoji	York University	Law - IP
Javad	Mostaghimi	University of Toronto	U. Corporate
Francis	Ouellette	UBC Bioinformatics Centre	Genetics
David	Phipps	York University	U. Corporate
Linda	Pilarski	University of Alberta	Life Sciences
Robert	Prince	York University	Physics
Jorg	Sack	Carleton University	Computer Sc
Dennis	Salahub	University of Calgary	U. Corporate
David	Schade	NRC – Herzberg Institute of Astrophysics	Astrophysics
Christoph	Sensen	University of Calgary	Genetics
Frances	Sharom	University of Guelph	Life Sciences
Pamela	Slaughter	Institute for Clinical Evaluative Sciences	Health Science
Randall	Sobie	University of Victoria	Physics
Elizabeth	Spangler	University of Prince Edward Island	Veterinary Med.

Dale	Swan	Environment Canada	Environment
Fraser	Taylor	Carleton University	U. Corporate
Lucy	Thompson	University of New Brunswick	Space Science
Kip	Tyler	Manitoba Land Initiative	Geography
Robert	Vet	Environment Canada	Meteorology
Donald	Weaver	Dalhousie University	Environment
Alan	Wildeman	University of Guelph	U. Corporate
Michael	Wolfson	Statistics Canada	Health Science
Tsoi	Yip	Environment Canada	Meteorolgy

11.4 Task Force Members

David Strong
NCASRD Task Force Chair
President, University Canada West

John ApSimon
Special Advisor to the President
Carleton University

Guy Baillargeon
Agriculture and Agri-Food Canada

Joyce Garnett
University of Western Ontario, and
President of the Canadian Association
of Research Libraries

Charles (Chuck) Humphrey
University of Alberta, and
U of A representative on the Canadian
Association of Research Libraries

Steven Jones
BC Cancer Research Centre

Ellsworth LeDrew
University of Waterloo

Peter Pennefather
University of Toronto

Andrew Pollard
Queen's University

Richard Rachubinski
University of Alberta

John Rodgers
Toth Information Systems

John Spray
University of New Brunswick

Robyn Tamblyn
McGill University

Paul Uhlir
US National Academies

11.5 Project Management Group Members

Gordon Wood
Project Management Group Chair
National Research Council Canada

Jac van Beek
National Research Council Canada

Catherine Betz
National Research Council Canada

Dennis Blinn
Science and Engineering Research
Canada (NSERC)

Tony Damiani
Industry Canada

Stephanie Delorme
National Research Council Canada

Peter Leach
Leach Technologies Ltd.

David Strong
NCASRD Task Force Chair
President, University Canada West

Alan Le Couteur
Science and Engineering Research
Canada (NSERC)

David Moorman
Social Sciences and Humanities
Research Council of Canada

Katy Nau
Canada Foundation for Innovation

Glen Newton
National Research Council Canada

Stephanie Robertson
Canadian Institutes for Health
Research

Alexandra Talbot
National Research Council Canada

ISBN 0-662-39966-8

9 780662 399667

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

NSERC
CRSNG

Canada Foundation for Innovation
Fondation canadienne pour l'innovation

NRC·CNRC