CANADIAN NATIONAL CONSULTATION ON ACCESS TO SCIENTIFIC RESEARCH DATA

Michel Sabourin[1]* and Bernard Dumouchel[2]

**[1]Dept of Psychology, Université de Montréal, C.P. 6128, Succursale Centre-Ville*
Montréal QC Canada H3C 3J7
Email : Michel.Sabourin@umontreal.ca
[2] Former Director General, Canada Institute for Scientific and Technical Information, National Research Council, Canada
Email: bernardj.dumouchel@sympatico.ca

ABSTRACT

In June 2004, an expert Task Force, appointed by the National Research Council Canada and chaired by Dr. David Strong, came together in Ottawa to plan a National Forum as the focus of the National Consultation on Access to Scientific Research Data. The Forum, which was held in November 2004, brought together more than seventy Canadian leaders in scientific research, data management, research administration, intellectual property and other pertinent areas. This article presents a comprehensive review of the issues, and the opportunities and the challenges identified during the Forum. Complex and rich arrays of scientific databases are changing how research is conducted, speeding the discovery and creation of new concepts. Increased access will accelerate such changes even more, creating other new opportunities. With the combination of databases within and among disciplines and countries, fundamental leaps in knowledge will occur that will transform our understanding of life, the world and the universe. The Canadian research community is concerned by the need to take swift action to adapt to the substantial changes required by the scientific enterprise. Because no national data preservation organization exists, may experts believe that a national strategy on data access or policies needs to be developed, and that a "Data Task Force" be created to prepare a full national implementation strategy. Once such a national strategy is broadly supported, it is proposed that a dedicated national infrastructure, tentatively called "Data Canada", be established, to assume overall leadership in the development and execution of a strategic plan.

Keywords: Access, Data, National, Open, Preservation, Research, Scientific, Strategy

1 INTRODUCTION

In June 2004, an expert Task Force was appointed by the National Research Council of Canada (NRC) to plan a National Forum that would focus on how to ensure open access to, and the long-term preservation of, publicly funded research data. Based on the conclusions of the OECD March 2003 Report on "Promoting Access to Public Research Data for Scientific, Economic, and Social Development", as well as on the June 2004 International Declaration on Access to Research Data from Public Funding, this initiative received strong support from the Canada Foundation for Innovation (CFI), the Canadian Institutes of Health Research (CIHR) and the Science and Engineering Research Council of Canada (NSERC). An urgent need was felt for action to propel Canada into a new and transformational data-intensive paradigm to properly orient all future Canadian research efforts. Under the able leadership of Dr. David Strong, as Chair of the Task Force, and of Dr. Gordon Wood, as Project Management Chair, and their respective groups, the Forum, entitled "National Consultation on Access to Scientific Research Data (NCASRD)", took place in Ottawa on November 22 and 23, 2004.

The main objectives established for the Forum participants were threefold: (1) 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; (2) to recommend actions to preserve historically significant data as an historic record, and as a scientific and cultural asset for current and future research; and (3) to ensure the recommendations made in the NCASRD Report were aimed at generating workable solutions to the technological, institutional, cultural, legal, financial and behavioural barriers to open access.

It must be noted that the actions recommended were to apply only to digital data. Issues of open access to research findings and published research results were excluded from the consultation.

2 BACKGROUND

Ever since governments have been 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. There has been an intense debate concerning these issues.

On the one hand, some believe that the rate of scientific discovery is gated by access to the data and finding of others, and see 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 the valuable intellectual property rights derived from the research.

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 advances of IT. There is thus a need for systematic data archiving to make them readily accessible and available for reuse.

In this increasingly public debate, many feel that more open access would be highly beneficial to the efficiency of the research endeavour. Nevertheless there are always those who feel threatened by changes to their accepted and proven practices.

Two key documents helped inform the deliberations of the NCASRD group. The first was the *OECD Ministerial Declaration on Access to Public Research Data (30 January 2004).* Thirty-four countries, including all the G8 members, adopted this Declaration regarding their commitment to work towards the establishment of access regimes for digital research data stemming from public funding. The premise on which the Declaration was based is that publicly funded research data should be openly available to the maximum extent possible.

The other important document resulted from the Canadian Social Sciences & Humanities Research Council (SSHRC) *Consultation on National Data Archive* in June 2002, which examined issues about data archiving infrastructures within the context of the SSHRC research community. Its 2002 report, titled " Building infrastructure for Access to and Preservation of Research Data" was a very worthy effort, but it failed to address many of the key issue domains identified later in the OECD March 2003 Report that led to the Ministerial Declaration. There was a definite need to go further.

Because few nations and disciplines have made considerable and widespread progress in implementing data sharing, NCASRD provided an excellent opportunity to assert leadership in establishing appropriate physical, operational, systemic and policy solutions.

3 FORUM PARTICIPANTS

The NCASRD established a 14 member Task Force, chaired by Dr. David Strong (president of University Canada West), to plan the content and format of the Consultation. A Project Management Group of 14 members, a majority being employees of the NRC, but also including representatives of the three sponsoring organizations, was formed to concretely implement and logistically manage the Consultation. This group was chaired by Dr. Gordon Wood, the immediate Past Chair of the Canadian National Committee for CODATA (CNC CODATA) sponsored by the Canada Institute for Scientific and Technical Information.

Seventy-four expert participants were identified and invited to attend the two-day NCASRD meeting. These individuals were the members of the Task Force and the Project Management Group, as well as 51 other scientists coming from all parts of Canada and representing different expertise or disciplines, and different work areas

(university, government, industry). These invitees included senior academic researchers and academic administrators, members of Federal Networks of Centres of Excellence, research laboratory senior scientists and administrators, senior staff from Data Libraries and Granting Councils, and researchers from Statistics Canada. A good cross-section of the Canadian research community was thus obtained.

Some emphasis was placed on enhancing the participation from the university sector, because those working in this sector are usually the ones who mostly think of themselves as personally "owning" their data. At the same time, data generated by government scientists are covered by policies of their ministries – policies which themselves may result in sub-optimal dissemination and use (e.g. selling geographical data at high cost), and constitute a different, but related problem.

All meetings were held at Library and Archives Canada, which, in view of the issues to be discussed, seemed the most fitting setting.

4 FORUM STRATEGY

The Task Force decided to orient the discussion toward a Vision of the Research World and of the Canadian Data Scene in 2020. This was the challenge that the Forum participants were asked to consider. The participants were asked to think in terms of what would be the perspective in 2020, in other words, how things would be done then, and how the Canadian research community managed to get there. In the view of the Task Force, developing such a grandiose, far reaching and optimistic strategy was better than getting bogged down in details about how bad the situation is perceived to be now.

In this purposely "grandiose" vision, which was put forward by Dr. Carty in his Opening Address, "*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*."

Given that vision, what could emerge as an end result to this process? In other words, how can the dream come true? What is needed to make it come true? From this starting point, the discussions tried to anticipate what the situation would be in the year 2020. 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.

The discussions also generated a "Mind-Map" (see Figure 1 below), which was used eventually in the Final Report (NCASRD, January 2005) to link the ideas into the set of recommendations. At the meeting, the Mind-Map was a sheet of paper (about 6 meters long and approx. 1.3 meters high) on which the various ideas were grouped together in real time by the discussion facilitator. It turned out to be a very effective tool for identifying issues and ideas for the report.

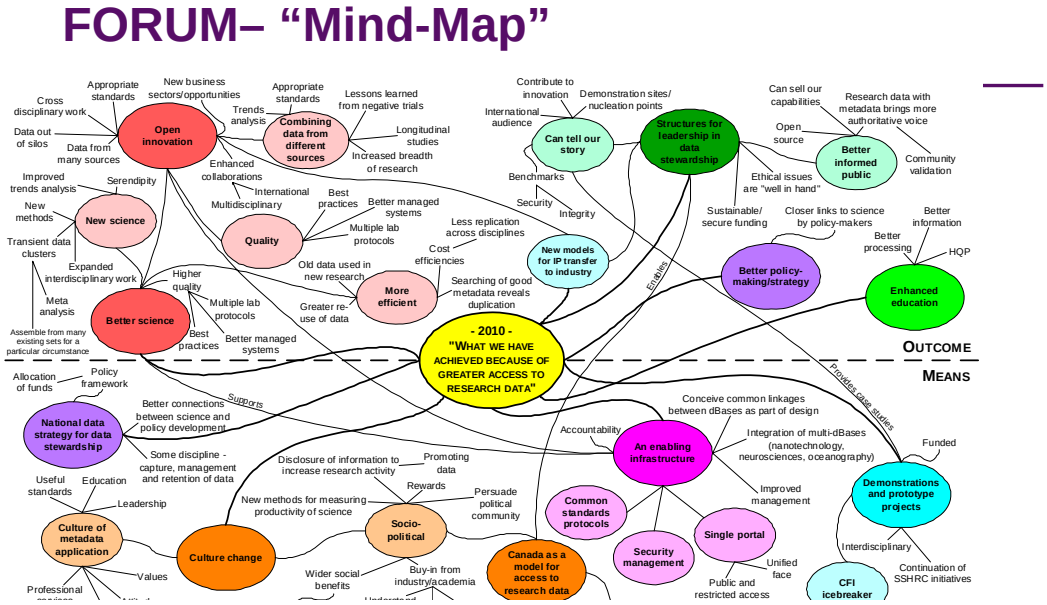

Figure 1. Mind-Map generated by the NCASRD discussion

The remainder of this article presents the main findings regarding (1) impact areas and (2) challenges to open access, as well as the recommendations that came from this exercise.

5 FINDINGS – IMPACT AREAS

5.1 New Science

The NCASRD participants anticipated many new areas of science that access to and novel intersections of current and prospective global data sources will enable. They derive principally from new intersections of disciplines and data. This would open up: (1) new methods of research based on the intermeshing of data that cannot be brought together at this point; (2) complexity reduction only achievable by interdisciplinary data integration and mining; (3) the study of transient data and continually captured data; and (4) unexpected and unknown relationships among data which are found serendipitously.

5.2 Better Science

Historically, the data on which scientific research was based frequently did not have the precision, accuracy, nor volume for many research theses to be validated with a sufficiently high degree of assurance. This was usually caused by the substantial time and cost of data generation and validation. This situation began to change dramatically in the digital era, however with the exponential increases in data volumes in most areas of research.

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 or data. Ever larger and more complex datasets will be created. Open access will allow superior longitudinal analyses, better experiment replication, and better informed peer-review

5.3 Leadership in Innovation

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. The interdisciplinary barriers that have inhibited the development of new business sectors and business opportunities will disappear as open access takes hold. For example, the analysis of genetic, social and environmental predisposition to particular diseases can enable the very rapid expansion of the "wellness" sector of the economy.

5.4 Superior Policy and Strategy

The ability to find, access, and combine data from an array of trusted databases, both current and archived, will allow 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 will become possible through the understanding of pollution models and pollution effects on health.

5.5 More Efficient Research

The anticipated changes will reduce the effort and time in acquiring data of the requisite precision, quality and scale. This, in turn, will reduce time and cost in duplicating experiments and will facilitate access to the computing tools and resources required to handle the demands of complex analysis.

5.6 Enhanced Education

The traditional approach to science education has been restricted to discipline-specific data. The opening of data access across disciplines will change the teaching habits. New teaching processes will allow students to gain a better understanding of the power of databases as discovery and problem solving tools.

6 FINDINGS – CHALLENGES TO OPEN ACCESS

NCASRD participants identified a substantial number of inhibiting factors to the broad acceptance and implementation of a Canadian open access and preservation system. These factors range from political issues and priorities to the Canadian research culture and reward systems. Countering these inhibiting forces successfully is an essential part of the NCASRD strategy.

6.1 Priority of Need

Universities, granting agencies, and other institutions in the research enterprise have many high priority issues, such as funding. For change to occur, open access and preservation of data must be given a high priority status. Funding mechanisms for the construction, maintenance and preservation of scientific databases are rare. .

6.2 Champions for Change

Studies, discussions and the signing of Declarations are useful, but insufficient. The time has come for the banner to be carried by influential champions within the Canadian research community and the governmental agencies.

6.3 Culture

The cost recovery policies for data access of certain governmental agencies constitute a barrier to use, which can prevent particular lines of research from being pursued. The present culture and reward systems are solely based on individual recognition. A culture of collaboration and of team contribution to the public good needs to be established. Recognition and reward systems can be changed to promote such behavioural and cultural goals.

6.4 Training

Few researchers have had specific training in database development and preservation. Researchers are frequently reticent to assume responsibility for database management beyond their immediate interests. There is also insufficient expertise to implement accessibility, security and preservation requirements. There thus is a definite need to develop an effective training capacity.

6.5 Archival Expertise

In Canada today, there is an insufficient number of trained archivists to serve the growing demand. There is not only a need to train researchers, but also a need for the supporting expertise to guide the development of sophisticated databases and to address problems of database access, reuse and preservation. Data archivists can become valued research partners and consultants.

6.6 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. There is a need to review existing databases to determine what the technical impediments are to making the data widely accessible. Are the applications of standards a solution? These problems can be mitigated by the adoption of a national data strategy to prevent future erratic development of databases.

6.7 Responsibilities, Systems and Tools

A better intellectual property regime needs to be defined. Large open access databases present new challenges. In the Canadian context, the issues associated with data ownership, custody and control are new to a majority of researchers. Data contributors and database curators should share the responsibilities related to preservation. A mechanism to transfer such responsibilities to capable successors should be defined. 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 supported for the entire period of the usefulness of the data.

6.8 Other Challenges and Opportunities

Unclear ownership, unreadable media data storage, inaccurate and incomplete metadata, losses of valuable historical data, owner liability and IP control, were all identified by NCASRD participants as other potential impediments to data access.

In view of the findings reported and the discussions held during the Forum, the NCASRD concluded that open access to scientific research data will, among many other benefits: (1) transform the 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; (2) accelerate the pace of knowledge development through the reuse of data and the interlinking of diverse datasets; (3) permit the study of far more complete systems and system interactions; (4) open the opportunities for a substantial increase in national and international research collaboration; and (5) result in myriad 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. To arrive at these positive outcomes, the NCASRD participants made 18 recommendations that should be implemented in the near future and assigned the responsibility for their implementation to different actors.

7 RECOMMENDATIONS

7.1 Responsibilities of Data Force

Recommendation 1 - Organizing

The first and most important recommendation deals with the organizational aspects.

It is recommended that the NCASRD Sponsors establish a task force (*Data Force*) to prepare a thorough national implementation strategy. Data Force should have representation from all the major members of the Canadian research community, in particular the Sponsors/Partners of this initiative. Its mandate would be to guide and oversee a small implementation secretariat to (1) commission a pilot data access project (*Data Project*); (2) plan and supervise the formation of a permanent Canadian data access organization (*Data Canada*); (3) to secure the long-term commitment to federal financing of Data Canada; and (4) to develop a data access strategic plan (*Data Plan*). There is thus a need for an ad hoc body of finite lifetime to " get the ball rolling"; this body should be sufficiently broadly representative to be credible and viable, yet not so large as to be unwieldy – this is a potentially challenging balancing act.

Recommendation 2 – Educating

Data Force should immediately begin fostering awareness among political, institutional and public opinion leaders of, *inter alia*, the educational benefits derived from the ability to place learning in a real-life context and enabled by open data access.

Recommendation 3 - Funding

Sufficient funding should be provided to Data Force to support the implementation of an open access database pilot project as soon as possible.

7.2 Responsibilities of Data Canada

Recommendation 4 – International Participation

Data Canada should establish a management capacity that can monitor and intervene in international open access to protect Canadian interests, and assist the international community in the promotion of open access.

Recommendation 5 – Ethics

Data Canada should establish consultations to identify legal barriers to access to scientific data.

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 to be of national importance, should be deposited and secured at Library and Archives Canada.

Recommendation 8 – Liability

Data Canada should 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.

Recommendation 9 – Anonymization

An expert panel should be appointed to examine the legal issues surrounding data anonymization and secure data practices that would prevent infringement of an individual's privacy, if the data are made accessible for other research.

Recommendation 10 – Databases at Risk

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

7.3 Responsibilities of Funding Agencies

Recommendation 12 – Training Researchers

All organizations that fund scientific research should 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 other similar organizations should 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.

Recommendation 14 – Resources

Federal and provincial government departments, agencies and ministries that fund scientific research should establish long-term stable, non-competitive core budget allocations to provide research institutions, organizations, and agencies with the resources to preserve all important databases.

Recommendation 15 – Peer Review

Databases and datasets in use or expected to be used in multiple research initiatives, including their metadata, should 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 should 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.

In general, the recommendations made concerning the funding agencies represent a fairly major shift in current emphases of these organizations – activities related to data have typically been seen as relatively unimportant and

have not attracted grants. This is one of the reasons these bodies were recruited as sponsors of the NCASRD. These agencies are key to raising the awareness of university researchers about the importance of making their data available – it is a well-known phenomenon that dollars usually triumph over appeals to altruism and the public good as a motivator!

7.4 Responsibilities of Universities and Researchers

Recommendation 17 – Rewards

University faculties and other academic research units should extend the recognition and award 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 should 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.

Excellence in data-related work has to be valued and given tenure "points"; it should not be regarded as second-class science as is too often the case. Just as the need for bio-informatics experts was recognized in the 1990s and new programs were initiated, so is there now a need for data preservation specialists.

8 CONCLUSION

Since the publication of the NCASRD Final Report at the end of January 2005, efforts have been made by Canada's National Science Advisor and the director general, Canada Institute for Scientific and Technical Information to identify a strong opinion leader of the Canadian scientific community to help with the implementation of the preceding recommendations, or at least, the most important ones (including the first recommendation, from which almost all others derive). These efforts have thus far failed. The Canadian National Committee for CODATA, which was initially at the core of the NCASRD project and whose mandate is to address all data management issues of importance to the Canadian scientific community, will keep trying to promote this much-needed implementation. The original partners of NCASRD, the National Research Council of Canada, the Canada Foundation for Innovation, Canadian Institutes of Health Research and Natural Sciences and Engineering Research Council (NSERC) need to be re-engaged to support the implementation of the recommendations.

In view of the very positive initiatives recently reported by the OECD and other countries (see the other articles in this special issue of the CODATA Data Science Journal), these examples might appropriately serve to reinforce the need to have a follow-up in Canada to the NCASRD Report.

9 REFERENCE

The Final Report of the NCASRD is available online at http://ncasrd-cnadrs.scitech.gc.ca/NCASRDReport_e.pdf. Print copies are also available through NRC's CISTI Help Desk at 1-800-668-1222 (Canada and US) or (613) 998-8544, or by e-mail at info.cisti@nrc-cnrc.gc.ca.