CISTI'S ACTIVITIES IN SUPPORT OF SCIENTIFIC DATA MANAGEMENT IN CANADA 2008-2010

Mary Zborowski[1]

[1]*National Research Council Canada, Strategy and Development Branch (NRC-SDB), 1200 Montreal Road, Ottawa ON Canada K1A 0R6*
Email: mary.zborowski@nrc-cnrc.gc.ca

ABSTRACT

In the Canadian research environment, it is difficult for researchers to effectively discover, access, and use data sets, except for those that are the most well known. Several recent reports have discussed the issues around "lost" data sets: those which are intended to be shared but cannot be identified and utilized effectively because of insufficient associated metadata. Both problems are approaching critical levels in Canada and internationally, a situation that is unacceptable because these data sets are often generated as a result of public funding. Solutions may involve providing support and training for researchers on how they can best collect and manage their data sets or developing gateways to scientific data sets. NRC-CISTI is the largest comprehensive source of scientific, technical, and medical information in North America, with a mandate to serve as Canada's national science library. Through its publishing arm, NRC Research Press, it is also Canada's foremost scientific publisher. NRC-CISTI is an organization with demonstrated expertise in metadata management, which, until recently, focused primarily on library and publishing contexts. However in November 2007, it formally committed to expand its agenda to address the management of scientific research data and the related critical needs of the research community. This paper presents NRC-CISTI's activities in this area. NRC-CISTI has begun by hosting forums in which the critical players (including the granting agencies) mapped out targets and approaches. It has strengthened its own internal expertise regarding metadata and management of scientific data sets. Finally, NRC-CISTI is developing a gateway Web site which will provide access to Canadian scientific data sets and related metadata, tools, educational resources, and other informative and collaborative tools urgently needed by Canadian and international researchers. NRC-CISTI is the sponsoring body for the Canadian National Committee for CODATA and is committed to promoting and supporting CNC/CODATA's initiatives.

1   INTRODUCTION

The National Research Council's Canada Institute for Scientific and Technical Information (NRC-CISTI) is Canada's national science library and leading publisher of scientific information. It provides Canada's research and innovation community with tools and services for accelerated discovery, innovation, and commercialization by:

- providing access to scientific, technical, and medical (STM) research publications that inform the research program;
- publishing peer-reviewed articles resulting from that research; and
- providing customized information research, analysis, and competitive technical intelligence to support both the researcher and the entrepreneur.

NRC-CISTI is a dynamic and client-centred organization which continuously reviews its existing products and services and develops new ones according to national needs within an increasingly changing environment. As part of a larger research organization (the National Research Council of Canada), NRC-CISTI also engages in research activities to support its mandate and mission and to increase the quality of its products and services.

The issues around the management of data generated from scientific and technical research activities have risen to the top of NRC-CISTI's agenda. The needs of Canadian researchers, including NRC's own, are becoming critical – a concern which is widespread within national and international organizations such as CODATA. NRC-CISTI is now firmly embarked upon a course that will present it as a major Canadian player in this area, providing critical resources and support to researchers and innovators. This document highlights some of the activities and initiatives in which NRC-CISTI is participating.

2   THE ENVIRONMENT

The importance of research data has grown at NRC-CISTI and in Canada due to events both within the country and internationally. In 1996, two important events took place. First, Statistics Canada – the Canadian government agency responsible for the census and other related activities – started the Data Liberation Initiative (DLI). This initiative, which involved 43 post-secondary institutions, allowed for the release of previously inaccessible census and related data from Statistics Canada. While of primary importance to the social science and humanities research community, it set an early precedent on the importance of data access. Second – and more important – a report from the Royal Society of Canada (1996) made the significant assertion that the progress of research could be limited by policies that restrict access to data.

Since the Royal Society report, a number of subsequent reports have more closely examined the nature and importance of data in various research communities. These include *Report on the Advancement of Research using Social Statistics* (Joint Working Group, 1998), the *National Data Archive Consultation* (Social Sciences and Humanities Research Council, 2002), and the *National Consultation on Access to Scientific Research Data* (Strong & Leach, 2005). Although these studies have all had impact on the research community, a recent study (Perry, 2008) probing the understanding of data archiving roles and responsibilities by researchers funded by Canada's Social Science and Humanities Research Council (SSHRC), revealed that 72% were not aware that SSHRC had a mandatory data archiving policy. In addition, 90% were not aware that Canada had been a signatory to the OECD declaration on access to publicly funded data. Most researchers surveyed agreed that a National data archive would be a good idea (with an increase of agreement evident from 2001 to 2006) although 87% felt they would not alter their behaviour if SSHRC enforced its data archiving policy.

Recently, the funding councils in Canada have been moving towards policies offering stronger support to data archiving. In 2007, the Canadian Institutes of Health Research (CIHR) announced its Policy on Access to Research Outputs. SSHRC and Canada's Natural Science and Engineering Research Council (NSERC) are also in the process of similarly updating their policies.

3   NRC-CISTI'S ROLES

NRC-CISTI is, first and foremost, the library serving NRC and its researchers, providing quality information products and resources in support of critical national research programs. This activity was mandated by Canada's Parliament through the *National Research Council Act* in 1916 (Phillipson, 1991) and for many years both widely justified (to show the relationship of research and development to economic prosperity (Tyas, 1970)) and broadly announced, both nationally and internationally (Campbell, 1975). The Act also mandated NRC to serve Canada in the area of scientific research, development, and international collaboration; create a national science library for Canada; publish scientific and technical information; and develop a national scientific and technical information system. NRC-CISTI became committed to promote and support initiatives to provide seamless and permanent access to the world's STM information for Canadian research and innovation. Today, NRC-CISTI's knowledgeable staff serve NRC's approximately 4,300 employees located at research facilities across Canada.

NRC-CISTI is also Canada's largest and most significant publisher of peer-reviewed scientific and technical journals through its publishing arm the NRC Research Press. Since 1929, NRC has published not only its own (currently 16) journals, but also an extensive monograph series. It also offers publishing services to organizations and societies that lack sufficient resources.

Since its inception, NRC-CISTI has been an invaluable resource to all Canadian sectors (Brown, 1967; Ember, 1973), which has led to certain coordination and outreach roles for Canada, including holding the role of Canada's DOCLINE coordinator. By this, NRC-CISTI provides training, technical assistance, client support, and generally acts as the liaison between Canadian participants, the (United States) National Library of Medicine, and other DOCLINE libraries. NRC-CISTI, as the National Science Library at the time, has been recognized as the national library of health sciences for Canada since 1966 (Ember, 1969).

Additionally, in 2005, NRC-CISTI established the Partnership Development Office to explore new ways of working with organizations and of building sustainable collaborations especially with regards to Canada's information infrastructure. Most recently, NRC-CISTI has taken the lead in the development of a Federal Science eLibrary, which will enable the sharing of critical electronic resources among federal scientific and health-related government departments and agencies.

Finally, NRC-CISTI has sponsored and supported the work of the Canadian National Committee for CODATA (CNC/CODATA) since its inception, sharing common concerns and objectives. NRC-CISTI has become an acknowledged international collaborator, playing strategic and influential roles in organizations, including CODATA and the International Council for Scientific and Technical Information (ICSTI) – a scientific associate of the International Council for Science (ICSU).

4   NRC-CISTI, POSITIONED TO LEAD

Since its inception, NRC-CISTI has been serving the NRC (Brown, 1965; Morton, 1975), which is essentially a microcosm of the Canadian research and innovation community. NRC faces challenges similar to other Canadian research organizations in government, academia, and industry, functions within similar layers of complexity, and struggles with similar environmental and political issues. In serving NRC, NRC-CISTI has gained a unique understanding of similar organizations elsewhere in Canada and the world.

NRC has achieved excellence through well-established and fledgling data sets. For example, NRC's Herzberg Institute for Astrophysics' astronomy data sets are considered among the best in the world. NRC is also a participant in the Neptune Canada project, which when completed will be the world's largest cable-linked seafloor observatory, expanding the boundaries of ocean exploration to allow a new way of studying and understanding our planet. The data amassed by these two activities could be in the order of petabytes per year.

Supporting this work has allowed NRC-CISTI to link easily with other NRC collaborators, whether directly to small and medium enterprises (SMEs) through its Industrial Research Assistance Program (NRC-IRAP), through numerous guest work agreements, or through its Industrial Partnership Facility which provides facilities and technical infrastructure for newly established SMEs. In addition, NRC-CISTI works closely with NRC's lesser-developed data management offices, helping them improve their competencies in this area.

Since 1964, NRC-CISTI has also maintained NRC's Directory of Unpublished Data, which is a system that accommodates material that may be considered as supplementary to journal articles. Detailed calculations, numerical data on which graphs are based, detailed descriptions of methods, or extensive tabular material not essential to the text are examples of material suitable for deposit. A footnote or text reference indicating availability of such material from NRC-CISTI permits interested readers to obtain this material.

In addition to being part of Canada's leading federal research organization, NRC-CISTI's 80 years of expertise in the delivery of scientific information products to researchers and both national and international innovation communities has resulted in tremendous knowledge and expertise in the management of bibliographic data, which by extension supports the management of research data. As a publisher, NRC-CISTI is active in both aspects of the information world (production and consumption) and understands the increasing need to link research data to the publications which they support, not only for our own authors but for the broader STM community.

NRC-CISTI's demonstrated credibility, mandated role for Canada, and international stature, make it ideally positioned to take a leading role in the management of Canadian scientific research data.

5   NRC-CISTI'S SCIENTIFIC AND RESEARCH DATA MANAGEMENT ACTIVITIES

In its *2005-2010 Strategic Plan,* NRC-CISTI formally announced its intention to increase its focus on scientific data issues and presented new objectives regarding scientific and research data management. Of the four identified strategic goals in the *Plan*, two relate directly to data issues as well as to scientific, technical, and medical (STM) information in general: *To provide universal, seamless, and permanent access to information for Canadian research and innovation*; and *To lead STM information communities across Canada to become a national force for innovation*. NRC-CISTI has the expertise, the national presence, and the mandate to champion a coordinated approach to support access to data. The benefits of a coordinated national approach for STM will also apply to the social sciences.

To attain these goals, NRC-CISTI has become involved in a number of strategic activities, which are highlighted below.

Research Data Strategy Working Group

In 2007/2008, NRC-CISTI championed the formation of the (Canadian) Research Data Strategy (RDS) Working Group, a collaborative initiative with diverse partners, including libraries, universities, institutes, individual researchers, and Canada's three federal granting councils (NSERC, SSHRC, and CIHR). The RDS Working Group will address the challenges and issues surrounding the collection, access, and preservation of data arising from Canadian research. Members have grouped themselves to address (initially) three strategic concerns: Policies, Funding, and Research; Infrastructure and Services; and Capacity building (improving skills, providing a training program, and identifying a rewards system). These task groups will build on the foundation of past work in this area, including the recommendations of reports cited above and, most recently, the Canadian Digital Information Strategy recommendations on data and e-research.

The RDS Working Group held its inaugural meeting in January 2008 with representation from the United States' National Science Foundation in attendance. At the time of CODATA's 26th Annual Conference, the Working Group will have had its third full meeting. Its successes to date, as well as the report of the Task Group studying Policies, Funding, and Research issues, are posted on the Research Data Canada website.

Gateway to Scientific Data Sets
NRC-CISTI has launched a project known internally as the "Gateway to Scientific Data Sets" whose objective is to provide NRC and all Canadian researchers information about and access to a vast assemblage of scientific data sets and related resources (such as data management or visualization tools). In addition to the inherent need for such a project, it was launched to leverage both NRC-CISTI's metadata expertise and some of NRC-CISTI's significant inroads in this area to date. For example, NRC-CISTI supports two datasets through CNC/CODATA: LogKOW (Sangster, 1997) and Mycotox, described recently in CODATA Newsletter 96. Data records regularly reported through CNC/CODATA's *Report on Data Activities in Canada* will figure strongly in the early version, and the Gateway's long-term success will be linked to NRC-CISTI's synergy with CNC/CODATA and the participation of its members and observers.

NRC-CISTI's core expertise
CISTI already has extensive metadata expertise, including metadata creation, the development and application of controlled vocabularies and taxonomies, in-depth knowledge of national and international metadata standards, and the development and maintenance of metadata application profiles.

NRC-CISTI is currently participating in the development of several taxonomies for NRC, including the enterprise taxonomy for NRC web sites. As well, NRC-CISTI has a representative on ISO Technical Committee 229, Nanotechnologies. A delegate to Working Group 1, Terminology and nomenclature, this NRC-CISTI representative is leading a project to build a taxonomic framework of core nanotechnologies concepts.

NRC-CISTI is also expanding its expertise in the curation and management of data sets and the development of metadata practices for these sets. In the area of the Semantic Web research, NRC-CISTI is active in this area, important for both data management and publishing.

NRC Publications Archive (NPArC)
The NRC Publications Archive (developed by NRC-CISTI and set to launch in December 2008) will be a searchable, web-based archive, providing access to NRC's record of science, which will increase the access to NRC-authored publications, guarantee long-term access to NRC's research output, and serve as a valuable resource for NRC researchers, collaborators and the public. As part of this initiative, NRC has established a policy making it mandatory, starting in January 2009, for NRC institutes to deposit copies of all peer-reviewed, NRC-authored publications and technical reports in NPArC. Wherever possible, NPArC will provide access to the full text of these publications.

Trusted Digital Repository
"A trusted digital repository is one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future" (Research Library Group, 2002). Now under development, NRC-CISTI's trusted digital repository will store and preserve all NRC publications, including publications issued by NRC Research Press (NRC-CISTI's publishing arm) and documents created by NRC researchers. The repository may also contain publications issued by Canadian scientific societies and small Canadian scientific publishers. As well, the repository will store journal articles issued by commercial publishers (such as Elsevier and Springer) for which storage and access rights have been negotiated. With time, the repository may include data sets and related metadata for NRC research. Repository content will be made available through a search interface.

WorldWideScience.org
NRC-CISTI is a founding member of this initiative to provide access to the world's scientific literature. Organizations representing 38 countries participated in its launch in Seoul, Korea, formalizing their commitment to sustain and build upon the online gateway to the world's scientific information. Visitors to the site may, with a single search query, consult 32 national scientific databases and portals from 44 countries, searching over 400 million pages of scientific and technical information which is not typically accessible by other popular search engines.

NRC-CISTI's technology base
NRC-CISTI's technological infrastructure is continually being strengthened in order to support not only NRC-CISTI's own multi-million dollar enterprise but also other endeavours such as managing and developing (through a partnership with CIHR) PubMed Central Canada (PMC Canada). NRC-CISTI will provide more details once the final step in the agreement process (a signed agreement between NRC-CISTI/CIHR and the US National Library of Medicine to "officially" enter into the PubMed Central International (PMCI) network) has been completed.

ICSTI 2009 Conference – Ottawa
NRC-CISTI has agreed to organize the 2009 ICSTI conference, scheduled for June 9 and 10 in Ottawa, Canada. The conference theme is "Managing Data for Science" and will be followed by 2 days of meetings of the ICSTI General Assembly. There will be international speakers and attendees, and significant involvement from both CODATA and CNC/CODATA is anticipated.

Collaboration with NRC Institutes, Branches and Programs
NRC is increasingly recognizing the need to validate the great amounts of data that it produces. This is particularly critical given NRC's role in new and often ground-breaking areas of research, the success of which largely depends on NRC's ability to validate research results. To this end, NRC-CISTI has initiated a project that will involve interviewing NRC researchers to determine how NRC-CISTI can best support this activity. NRC-CISTI will be following established best practices, such as those described in the publication *Conducting a Data Interview* (Witt & Carlson, 2007).

NRC-CISTI Lab 2.0
This is NRC-CISTI's experimental website for demonstrating and evaluating prototype software and services for next-generation digital library applications developed by NRC-CISTI staff and research partners. Using a wiki, the site has been designed to provide the NRC-CISTI Lab developer community with a way to obtain feedback from early adopters of their applications as well as a way to encourage collaborations that will lead to the development of innovative software systems.

Other collaborations
In addition to its links with ICSTI, CODATA, and to the Open source and various research communities, NRC-CISTI represents NRC at the World Wide Web Consortium (W3C), ensuring that NRC and NRC-CISTI remain at the forefront of web research and development. NRC-CISTI is also a member of the Coalition for Network Information (CNI).

NRC-CISTI cannot address data management issues in isolation. They will be solved only with the support and collaboration of many organizations, such as those represented by the RDS Working Group. Accordingly, NRC-CISTI is working to build connections across jurisdictional and organizational boundaries to promote and support initiatives that improve seamless and permanent access to STM information, including data, for Canadian research and innovation. With CNC/CODATA, for example, NRC-CISTI is aligning its activities closely with those of this committee, which involve the development of a training workshop for junior researchers on the management of research data. Other collaborations may include those with non-Canadian organizations, such as one currently being explored with the Massachusetts Institute of Technology (MIT) in the United States. Other possibilities exist through NRC-CISTI's planned collaboration with ICSTI's German colleagues on the Digital Object Identifier (DOI) project.

6   CONCLUSION

Although NRC-CISTI has just begun to actively address these issues, it is aware that there is a relatively small window of time (from 3 to 5 years) to prepare while the Canadian infrastructure is being developed. NRC-CISTI is already engaged at a variety of levels: internally through its efforts with NRC; externally via vehicles such as the RDS Working Group; and internationally, through its support of CODATA, ICSTI, and others. Although NRC-CISTI's short- and long-term activities will not be defined until its late-2008 planning meeting, NRC-CISTI has already made significant progress to prepare it for the challenges anticipated during the next 5 years.

Given the scope of the data challenges faced by all scientific and research communities, it is clear that collaborations must be part of the solution. These issues must be addressed collectively from both national and international perspectives with the participation of all parts of the research community, including researchers themselves who create and use this data. Only with the ability to access and integrate research data from many diverse sources will researchers be optimally positioned to translate research into discoveries and innovations to address important scientific, social, and economic issues. NRC-CISTI invites its international colleagues to follow its progress in the upcoming years and welcomes suggestions for areas where we can collaborate.

7   ACKNOWLEDGEMENTS AND RESOURCES

CISTI web site: http://cisti-icist.nrc-cnrc.gc.ca
CISTI Strategic Plan: http://cisti-icist.nrc-cnrc.gc.ca/about/stratplan_e.html
CISTI's Media Room (links to News, Press Releases): http://cisti-icist.nrc-cnrc.gc.ca/media/newsroom_e.html
CISTI Contacts
Ms. Pam Bjornson, Director General, NRC-CISTI: http://cisti-icist.nrc-cnrc.gc.ca/media/bio_bjornson_e.html
Mr. François Dubé, Director, NRC-CISTI Technology & Research :
http://expert.nrc-cnrc.gc.ca/profiles/cisti-icist/dube_francois_1333_e.html

Canadian National Committee for CODATA: http://www.codata.org/canada
Documents and links inspiring the work of the RDS Working Group: http://data-donnees.gc.ca/eng/docs_links.html
RDS Working Group Task Groups and Leads: http://data-donnees.gc.ca/eng/groups/taskgroups.html

8   BIBLIOGRAPHY

Brown, J. (1965) The National Science Library: Information Centre for Industry. *Industrial Canada*, 65(1). NRC 8439.

Brown, J. (1969) The CAN/SDI Project: The SDI Program of Canada's National Science Library. *Special Libraries*, 60(8):73-77.

Campbell, H. C. (1975) National Planning for Canadian Science and Social Science Information Systems. *Library Trends*, 17(3):280-288.

Ember, G. (1969) The Health Science Resource Centre: A new information service of the National Science Library of Canada. In *Proceedings of the 3rd International Congress on Medical Librarianship*, pp. 383–389, Amsterdam.

Ember, G. (1973) Dissemination of Scientific and Technological Information in Canada. *Journal of Chemical Documentation*, 13(1):4-7.

Joint Working Group of the Social Sciences and Humanities Research Council and Statistics Canada (1998) *Report on the Advancement of Research using Social Statistics*. Retrieved from the WWW, December 1, 2008: http://www.statcan.ca/english/rdc/pdf/report.pdf

Morton, E. (1975) Cooperation in Canada. *Library Trends*, pages 399–415.

Perry, C. (2008) Archiving of publicly funded research data: A survey of Canadian researchers. *Government Information Quarterly*. Vol. 25, no. 1 2008:133-148. Retrieved from the WWW, December 1, 2008: http://www.verney.ca/apw2006/presentations/107.pdf

Phillipson, D. (1991) The National Research Council of Canada: its history, its chronology, its bibliography. *Scientia Canadensis*, 15(2):177-200.

Research Library Group (2002) Trusted Digital Repositories: Attributes and Responsibilities: An RLG/OCLC Report. Retrieved from the WWW, December 1, 2008: http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf

Royal Society of Canada, Canadian Global Change Program, Data and Information Systems Panel (1996) Data Policy and Barriers to Data Access in Canada: Issues for Global Change Research. Etrieved from the WWW, December 1, 2008:http://www.globalcentres.org/cgcp/english/html_documents/publications/data/toc.htm

Sangster, J. (1997) LOGKOW: A Databank of Evaluated Octanol-Water Partition Coefficients. *GDF Databanks Bulletin*, 1(1). Retrieved from the WWW, December 1, 2008: http://logkow.cisti.nrc.ca/logkow/

Social Sciences and Humanities Research Council (2002) Final Report, National Data Archive Consultation: Building Infrastructure for Access to and Preservation of Research Data. Etrieved from the WWW, December 1, 2008: http://www.sshrc.ca/web/about/publications/da_finalreport_e.pdf

Strong, D. & Leach, P. (2005) Final Report, National Consultation on Access to Scientific Research Data. Retrieved from the WWW, December 1, 2008: http://data-donnees.gc.ca/docs/NCASRDReport.pdf

Tyas, J. (1970)Knowledge, The Master Resource: The Future of Scientific and Technical Information in Canada. *Special Libraries*, 61(2):73-77.

Witt, M. & Carlson, J. (2007) *Conducting a Data Interview*. Purdue Universities, Libraries Research Publications. Retrieved from the WWW, December 1, 2008:
http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1092&context=lib_research

9   INTERNET REFERENCES

http://cisti-icist.nrc-cnrc.gc.ca/
http://www.statcan.ca/english/Dli/reports.htm
http://www.sshrc.ca
http://cihr-irsc.gc.ca
http://cihr-irsc.gc.ca/e/34848.html
http://www.nserc-crsng.ca
http://laws.justice.gc.ca/en/showtdm/cs/N-15
http://pubs.nrc-cnrc.gc.ca
http://safstl-asbstf.gc.ca
http://www.codata.org/canada/
http://www.icsti.org
http://www.icsu.org
http://www.neptunecanada.ca
http://irap-pari.nrc-cnrc.gc.ca
http://ipf-ipi.nrc-cnrc.gc.ca
http://cisti-icist.nrc-cnrc.gc.ca/about/stratplan_e.html
http://www.collectionscanada.gc.ca/cdis/index-e.html
http://data-donnees.gc.ca
http://logkow.cisti.nrc.ca/logkow/
http://www.codata.org/canada/mycotox
http://www.codata.org/resources/newsletters/newsltr96A4.pdf
http://dac.cisti.nrc.ca/data_activities.cfm
http://cisti-icist.nrc-cnrc.gc.ca/media/press/nparc_e.html
http://worldwidescience.org/news.html
http://cisti-icist.nrc-cnrc.gc.ca/our_research_e.html
http://w3c.org
http://cni.org