The Canadian Research Data Centre Network (CRDCN) in Canada's DRI Ecosystem[1]

In the context of Canada's digital research infrastructure, a defining characteristic of the CRDCN is its 20-year partnership with Statistics Canada. CRDCN serves as a vital conduit, providing social science and health researchers access to the rich reservoir of the agency's confidential and secure microdata in the 33 research data centres ("RDC's") located in host universities across the country. Since its inception, the CRDCN has gained unique expertise in the governance and management of secure data, while serving as a primary liaison between Statistics Canada, the university research community, and policymakers to ensure that the microdata assets are mined to advance knowledge and to inform public policy on economic, social and health issues of importance for Canada and Canadians. The CRDCN has been at the forefront of pioneering collaborative research programs with government and non-profit organizations to promote knowledge mobilization and the application of research results. To these ends, it has been supported by operating funding from SSHRC and CIHR, the host universities, some provincial governments and, since 2017, by CFI as a Major Science Initiative. The CRDCN brings to Canada's DRI ecosystem, and to NDRIO, a highly productive and successful 20-year history of governing and managing a major and unique national research platform, involving the effective coordination of multiple and multi-sector stakeholders.

The records in the RDCs are currently being analysed by more than 2000 researchers, about one third of whom are graduate students. The analyses relate mostly to the income, employment, education, housing, health, and well-being of individuals and families and, increasingly to businesses and workplaces. The research using these data files has resulted in more than 5300 publications that have led to major advances in our understanding of the socio-economic, cultural and health dynamics of Canadian society. Importantly, many studies have been the basis for assessing the impacts of federal and provincial policy measures; some have resulted from academic researchers collaborating with government departments and others from government agencies themselves using the RDCs to analyse data that are not otherwise available to them. As part of an ambitious change agenda the CRDCN has been working with Statistics Canada to develop the methodological and technological solutions that will make it possible for more researchers soon to have secure remote access. Further information about the CRDCN is available at www.crdcn.org.

Current Issues

- The main DRI tools, services and/or resources **currently used**

The CRDCN provides researchers with secure access to a wide range of datasets that relate to virtually all aspects of Canadian society, to the hardware and the statistical and other software needed to analyse them, and to support services provided by Statistics Canada analysts on site. In addition to hundreds of household surveys, researchers can access census records and a growing number of files that government ministries maintain for administrative purposes, e.g.,

[1] *This document was prepared by Dr. Byron G Spencer, Emeritus Professor of Economics, McMaster University, and has benefited from extensive consultation with members of the CRDCN Academic Council (composed of Academic Directors from 33 partner universities), its multi-sector national Board of Directors, and several CRDCN research collaborators. In addition, we have consulted with Health Data Research Network Canada (HRDN) in the preparation of its submission, and HRDN with us.* ***Contact for follow-up is Dr. Martin Taylor, CRDCN Executive Director.***

those relating to personal income (from the federal Canada Revenue Agency), to enrolment and performance in educational institutions and to the use of health care services (both from provincial ministries). The analysis of such records has immense social value.

- Do you have access to all the DRI tools, services and/or **resources you need** for your research? What are they? **What is missing?**

Data collected through surveys and censuses continue to be critical. In addition, researchers in the social sciences and health are giving increased attention to administrative records. The fact that such records have near-universal coverage, provide accurate information, and continue to be collected even during a pandemic makes it possible for analysts to drill down to small populations of interest (e.g., those with rare health conditions, or living in remote communities) in a timely fashion. However, those benefits are offset by the limited information that such records contain. For example, while education, family situation, and income are known to be important determinants of health status, the records that provide detailed information about patient diagnosis and treatment only rarely include socioeconomic information beyond age, sex, and place of residence. However, that disadvantage can be overcome through the linkage of health records with records from surveys, censuses, and other administrative sources.

Such linkages have been a major element in the continued rapid increase in data files available to CRDCN researchers. By way of example, individual-level responses from the cross-sectional Canadian Community Health Survey, which is conducted annually by Statistics Canada, have been linked to the longitudinal administrative hospital-stay records for respondents from all provinces except Quebec. Access to records of that sort makes it possible to address and answer questions of policy importance that simply could not be considered otherwise. But much remains to be done; great opportunities await.

The fact that the provinces and territories are responsible for the design and delivery of health care and education services has resulted in considerable diversity in the services provided and the ways in which they have been delivered. While challenges remain in addressing confidentiality concerns to gain access to the data and in working with measures that differ across jurisdictions, the diversity creates an enormous (potential) opportunity to assess the impact of different interventions – and to identify what works best. In relation to health services, a 2015 report for the Canadian Institutes of Health Research emphasized the "world-class research potential": if pooled across jurisdictions the available data would have "far-reaching effects for health care and the overall health of Canadians". In addition, the 2015 report of The Council of Canadian Academies Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation concluded that concerns could be addressed to provide "appropriate data access for bona fide public interest research".

In broad terms, while progress has been made, much more needs to be done in data access, sharing, and linkages across federal and provincial jurisdictions. What is missing and unprecedented, is privacy-protected linkage of provincial data (e.g., in health and education) with federal data (e.g. Census, personal income, immigration). This will enable transformative collaboration between the research community and policymakers to inform evidence-based decision-making. Applying such comprehensive data analytics would be world leading. As an especially diverse society, and one particularly rich in having the required longitudinal population-based data systems, Canada has the opportunity to establish a leadership role.

- What are your **biggest challenges accessing and using** the DRI tools, services and/or **resources that do exist** and are available to you?

The CRDCN was founded to address the critical lack of quantitative social statistics expertise in Canada and to enable exploration of Canadian issues using Canadian data. While our Network

contributes very substantially to training the next generation of quantitative social scientists, with more than 750 graduate students currently using Network resources and 22 completed their doctoral research in 2019, a continuing challenge is to increase and support the number of researchers experienced in the analysis of large and complex data files.

From the perspective of individual researchers, an on-going challenge is the need to work in an RDC during open hours. Having remote access 24/7 would be not only efficient but would also meet equity and diversity concerns of people with family obligations or physical disabilities. The CRDCN is currently working with Statistics Canada to find ways to make remote access feasible while addressing security concerns.

Access to business data remains a challenge. Until recently, research access to the full range of business data was only on-site at Statistics Canada in Ottawa. As a result of the SSHRC-funded Productivity Partnership, various workarounds have been created, including the provision of research assistance at Statistics Canada. While business data are now accessible in the RDCs, much more could be done if all the relevant files were accessible remotely. That will be technically feasible once the platforms currently under development by Statistics Canada and the CRDCN are in place in the next two to three years.

Future DRI State

- What is your **vision** for a cohesive Canadian DRI ecosystem [for CRDCN researchers]?

More than three-fifths of total government spending went to social protection, health care, and education in 2019. A major concern is that remarkably little is known about the impact of that spending on the well-being of Canadians or about how the benefits could be delivered more efficiently. Many within government recognize that having their administrative records analysed by academic researchers can help them to carry out their departmental missions; a cohesive DRI ecosystem would make such analyses possible.

The **first** element in our vision is to have a substantial increase in the number and range of (linkable) administrative records that federal and provincial government departments make accessible to academic researchers. Only through the analysis of such records can we have a full understanding of the impact of existing policies. Similar comments apply to certain data holdings outside of government, many assembled with considerable public funding. For example, research programs funded by Genome Canada have collected vast amounts of individual-level data of a highly specialized nature over a period of many years. The same applies to the Allergen NCE. However, the value of this massive investment is not being optimized because the data remain siloed. Only by linking such records with information about the socioeconomic characteristics of individuals in their studies can the full value of the data be realized. As has been demonstrated, Statistics Canada is able to secure such data files and make them research accessible. However, overcoming legislative and other barriers to facilitate these linkages remains a major hurdle.

The **second** element is to ensure comparability of data across jurisdictions through the development of standardized reporting procedures and by other means. Only then will we know, for example, how many Canadians are in nursing homes, retirement homes, or other places that provide a level of care.

The **third** element relates to the impact of research. Much academic work has relevance for policy but we need to continue to find more effective ways of engaging policymakers in co-creation of research questions, collaborative programs of research, the joint training of HQP, and the cross-sectoral mobilization of knowledge regarding research and policy results. The current models are challenging to implement at scale given the one-off nature of funding models for these initiatives and often misaligned incentives for academic researchers and government policy-

makers. The CRDCN, led by its Research Program Director, is piloting such collaborations working with Women and Gender Equality (WAGE) on programs of research related to their departmental priorities and with Crown-Indigenous Relations and Northern Affairs Canada (CIRNAC) on the provision of mini-grants to analyze the most recent Aboriginal People's Survey. The appointment of senior federal and provincial representatives to the CRDCN Board and committees ensures that the Network benefits on an ongoing basis from the wise counsel of senior government officials in framing research and policy collaborations and in defining the mechanisms to ensure their sustainability and impact.

The **fourth** element of our vision also relates to strengthening relationships among practitioners who engage in data-intensive research and policy analysis in the social sciences and health. CRDCN is encouraging communities of practice, among researchers and policy-makers across disciplines and sectors, but Canada remains "a will against its geography"; policy leadership and improved funding incentives are needed.

- What are the **types** of DRI tools, services and/or resources you would like to use, or envision using, in the future?

Following from the above, the two major DRI services are continued growth in the number and range of (potentially) linked data files, including business files, and remote data access. The tools that would be most helpful in this connection include metadata standards, more automation, collaborative space, and a federal/provincial data repository together with enhanced analytical tools. A practical concern, however, is that because of security concerns many analytical tools for machine learning cannot be used in the RDCs.

- What **challenges** do you foresee while using integrated DRI tools, services and/or resources?

Given the ever-increasing number of linkable data files to which researchers have access, a major challenge is to enlarge the researcher capacity to analyse them. Government agencies, the private sector, postsecondary institutions, and others are already having difficulty hiring researchers with the skills needed to analyse large and complex data files effectively. Having remote access could be part of the solution, in that more convenient access alone would result in an increase in researcher participation. To that end, Statistics Canada is now testing approaches to share the risks with researchers and their institutions of providing access to confidential microdata. This is an area in which experience gained elsewhere, especially in some European countries, can provide guidance.

A further challenge is to have government departments make administrative records available for analysis. In many cases, jurisdictional issues need to be resolved in order to provide opportunities to work collaboratively, perhaps through identifying consent approaches that permit future data linkage. The goal is to have a national digital research infrastructure that will allow researchers to harness federal and provincial data simultaneously. A related matter is the need to engage government departments with the research community to identify policy concerns and to ensure that the administrative records are interpreted appropriately. More generally, we need to align with policymakers' needs in order to mobilize their resources for collating and sharing administrative data. Data-centric program evaluation needs to be built into the policy development cycle so that analyses are provided in a timely and predictable way.

The big data revolution also affects commercial enterprises as firms now rely on massive datasets to understand and target customers. Important for the CRDCN, such datasets are finding their place in academic research. For example, Harvard economist Edward Glaeser has demonstrated that Yelp data provides information about economic trends that is more fine-grained and timely than similar business data collected by governments, enabling social scientists to

"nowcast" the economy in ways previously impossible. Partnering with firms to make their data available in a trusted and secure research environment would allow their data to be protected and used for the public good.

A potentially valuable advance for the CRDCN and other national data networks is to develop communities of practice whereby the lessons learned from the use of particular data sets and analytical techniques are shared with colleagues across the country. The CRDCN is in the early stage of building on the expertise resident within its extensive researcher community to create communities of practice in its thematic areas. The CRDCN's plans here extend beyond sharing technical expertise to include lessons learned in research collaboration and knowledge mobilization. Moreover, NDRIO could certainly play a role in promoting communities of practice that transcend the scope of a single data network to harness and increase our collective strengths. AI/machine learning would be at the top of the list.

How to Bridge the Gap

- What are the tools, services and/or resources NDRIO should leverage to achieve your desired future state? What other suggestions do you have?

Drawing on funding from CFI and working in close collaboration with Statistics Canada, the CRDCN will soon have the necessary computing power to make it possible for researchers from across the country to analyse effectively and efficiently a very large increase in the volume of data collected from households in surveys and censuses, from businesses in surveys, and from administrative records relating to both persons and businesses. As emphasised above, and as noted in the first Annual Report of the Canadian Statistics Advisory Council, in order to address the full range of matters of policy importance and research priority, the linkable government administrative records will need to be obtained and checked for consistency across jurisdictions and over time; they will also need to align with policy development cycles. That will take policy leadership, especially in dealing with the provinces, and NDRIO could help, perhaps drawing on the excellent CIHI example in working with the provinces and territories to provide standardized pan-Canada health data.

In addition, the granting councils have funded research initiatives that have resulted in the assembly of valuable data bases; mention was made above of research programs funded by Genome Canada. As one element of broader data management policies, we note that there would be significant social benefits in having such records linked with other records and analysed. NDRIO could work with the granting councils to consider ways to take advantage of this important resource.

Creating greater research capacity to analyse the growth in large and complex data files will also take resources. As noted above, the CRDCN is committed to training the next generation of quantitative researchers in the social sciences and health. Our intention is to create and expand ways in which university researchers can work with government agencies to understand the nature and impact of policy measures. The development of collaborative programs of policy relevant research is one avenue. Another is the creation of 'policy incubators' whereby undergraduate and graduate students and faculty could work with government colleagues to understand the intended purpose of policy interventions and assess their actual consequences. NDRIO could well provide policy leadership and funding (e.g. through pilot programs) to catalyze more rapid innovation in our data-intensive research ecosystem.

In conclusion, the CRDCN welcomes the opportunity to collaborate with NDRIO drawing upon its extensive 20-year DRI expertise and experience to share lessons learned and pilot new initiatives, especially as they apply to heightened and more strategic collaboration in the acquisition, analysis and application of confidential data to advance the frontiers of knowledge and inform vital areas of public policy in Canada and internationally.