A Proposal to Strengthen Data Literacy in the Canada Research Data Ecosystem.

Submitted to NDRIO
December 14, 2020

Ernie Boyko, boykern@yahoo.com , Wendy Watkins, wendy.watkins@carleton.ca

Introduction

Paul Bernard was an internationally renowned Sociologist from the Université de Montréal.and a member of the Social Conditions Advisory Committee at Statistics Canada as well as the National Statistics Council. He was passionate about the democratization of data access. He offered the following observations:

> *"The genuine exercise of democracy increasingly requires that citizens get access to complex information and have the skills required to understand it … Concerning such issues, the public must have appropriate knowledge and not only hypothetical access to the data. Paradoxically, indeed, contemporary societies offer a wealth of information, but workers and citizens can be totally mystified, surrounded as they are by data whose flow and codes they do not master. ... All institutions producing specialized information must involve themselves in the transformation of these data into knowledge actually usable in economic activity and democratic debates. (1991)*[1]

It was this quote that inspired Wendy Watkins to propose the Data Liberation Initiative. [2] This initiative has given Canadian students and researchers full access to Statistics Canada's public data for the past 24 years. In 2000, Statistics Canada and Canadian universities launched the Canadian Research Data Centre Network[3] to offer secure access to confidential micro-data from censuses and surveys. The CRDCN has recently celebrated 20 years of success.

While these two initiatives have improved the data access issues that Dr Bernard cited, the skills portion of the challenges he identified has not followed. In previous years, a course in statistics was mandatory for graduation in the social sciences. Low student ratings led many institutions to drop the requirement. A member of the recently held meeting of the Canada National Committee on CODATA observed that many senior graduate students and new researchers were lacking the quantitative skills to work in a data intensive research environment. As well, a study by the Analytical Studies Branch of Statistics Canada identified a need to strengthen the data literacy skills of civil servants.[4] One can only wonder what the skill levels and needs in other sectors are.

Canada's Emerging Research Data Infrastructure.

The four key elements (as identified by NDRIO) of Canada's digital research infrastructure will be:

- digital network for research and education, allowing researchers to share data and collaborate across Canada and around the world
- data management (DM), allowing researchers to manage, find and access data
- research software (RS), enabling researchers to analyse the data they require

---

[1] Bernard, Paul, "Discussion Paper on the Issue of Pricing Statistics Canada Products", February, 1991.
[2] https://ihsn.org/sites/default/files/resources/IHSN-WP006.pdf
[3] https://crdcn.org/about-crdcn
[4] https://www150.statcan.gc.ca/n1/pub/11-633-x/11-633-x2019003-eng.htm

- advanced research computing (ARC), involving super computers that allow researchers to analyze massive amounts of data

Human skills are an essential component of this infrastructure.

Strengthening the Data Literacy Function in Canadian Research

In September 2020, Statistics Canada released a competencies profile for data literacy.[5] This is one of the key skill sets required to support and carry out research. The authors of this paper contend that there is an urgent need to improve the data literacy competencies in various sectors in the research data ecosystem to be successful in these endeavours.

What follows is a list of competencies outlined by Statistics Canada:

Data literacy competencies *are the knowledge and skills you need to effectively work with data.*

Data analysis

The knowledge and skills required to ask and answer a range of questions by analyzing data including developing an analytical plan; selecting and using appropriate statistical techniques and tools; and interpreting, evaluating and comparing results with other findings.

Data awareness

The knowledge required to know what data is and what are different types of data. This includes understanding the use of data concepts and definitions.

Data cleaning

The knowledge and skills to determine if data are 'clean' and use the best method and tools to take necessary actions to resolve any problems to ensure data are in a suitable form for analysis.

Data discovery

The knowledge and skills to search, identify, locate and access data from a range of sources related to the needs of an organization.

Data ethics

The knowledge that allows a person to acquire, use, interpret and share data in an ethical manner including recognizing legal and ethical issues (e.g., biases, privacy).

Data exploration

The knowledge and skills required to use a range of methods and tools to learn what is in the data. The methods include: summary statistics; frequency tables; outlier detection; and visualization to explore patterns and relationships in the data.

Data gathering

[5] https://www.statcan.gc.ca/eng/wtc/data-literacy/compentencies

The knowledge and skills to gather data in simple and more complex forms to support the gatherer's needs. This could involve the planning, development and execution of surveys or gathering data from other sources such as administrative data, satellite or social media data.

Data interpretation

The knowledge and skills required to read and understand tables, charts and graphs and identify points of interest. Interpretation of data also involves synthesizing information from related sources.

Data management and organization

The knowledge and skills required to navigate internal and external systems to locate, access, organize, protect and store data related to the organization's needs.

Data modeling

The knowledge and skills required to apply advanced statistical and analytic techniques and tools (e.g. regression, machine learning, data mining) to perform data exploration and build accurate, valid and efficient modelling solutions that can be used to find relationships between data and make predictions about data.

Data stewardship

Knowledge and skills required to effectively manage data assets. This includes the oversight of data to ensure fitness for use, the accessibility of the data, and compliance with polices, directives and regulations.

Data tools

The knowledge and skills required to use appropriate software, tools, and processes to gather, organize, analyze, visualize and manage data.

Data visualization

The knowledge and skills required to create meaningful tables, charts and graphics to visually present data. This also includes evaluating the effectiveness of the visual representation (i.e., using the right chart) while ensuring accuracy to avoid misrepresentation.

Evaluating data quality

The knowledge and skills required to critically assess data sources to ensure they meet the needs of an organization. This includes identifying errors or problems and taking action to correct them. This also includes awareness of organizational policies, procedures and standards to ensure good quality data.

Evaluating decisions based on data

The knowledge and skills required to evaluate a range of data sources and evidence in order to make decisions and take actions. This can include monitoring and evaluating the effectiveness of policies and programs.

Evidence based decision-making

The knowledge and skills required to use data to help in the decision-making and policy making process. This includes thinking critically when working with data; formulating appropriate business questions; identifying appropriate datasets; deciding on measurement priorities; prioritizing information garnered from data; converting data into actionable information; and weighing the merit and impact of possible solutions and decisions.

Metadata creation and use

The knowledge and skills required to extract and create meaningful documentation that will enable the correct usage and interpretation of the data. This includes the documentation of metadata which is the underlying definitions and descriptions about the data.

Storytelling

The knowledge and skills required to describe key points of interest in statistical information (i.e., data that has been analyzed). This includes identifying the desired outcome of the presentation; identifying the audience's needs and level of familiarity with the subject; establishing the context; and selecting effective visualizations.

Target-audiences for Data Literacy Training

We see two main groups for whom data literacy competencies are essential for ensuring that the goals of the Canadian research are met: the general population and the participants in the research process.

- **The general population:**
    - The COVID19 pandemic and the ensuing research findings and public discussion were not always embraced by the general population. While there may be many reasons for disbelieving the findings of scientific research, a lack of understanding of science and data were possible obstacles. At a minimum, the general population must be data aware, be able to discover data to help form opinions and actions, be able to interpret basic data and charts, have an appreciation for factors influencing data quality and draw conclusions and make decisions based on the research and information that has been provided by the scientific community.
    - The Conference Board of Canada released the results of a study on adult numeracy in 2014 that are concerning:[6]
        - *Overall, Canada earns a "C" grade on inadequate numeracy skills.*
        - *Fifty-five per cent of Canadian adults have inadequate numeracy skills—a significant increase from a decade ago.*
        - *No province earns above a "C" grade for inadequate numeracy skills.*

- **Research Participants:**
    - The research cycle involves numerous players ranging from people involved in research design, data/information gathering and preparation, data analysis, data management and publishing and communication.
    - The competencies required here involve a wide range including data discovery/gathering/exploration, data cleaning (a time consuming but essential task), data analysis/modelling, interpretation, management, stewardship, evaluation and communication.
    - Along with these skills is the requirement that they be able to use a wide range of computing and statistical tools. The players involved here may be researchers, research assistants, data stewards, data librarians and data scientists.

Addressing the Gaps

One could argue that the NDRIO network should not have to be concerned or devote much attention to the data literacies of the general population alone. The fact that Statistics Canada is devoting efforts in this direction and is

[6] https://www.conferenceboard.ca/(X(1)S(ajxc52kunyem0y3kgnsalot0))/hcp/provincial/education/adlt-lownum.aspx?AspxAutoDetectCookieSupport=1

finding knowledge gaps would suggest that the agency could be a partner in such an effort. We are suggesting two activities NDRIO could undertake to assess the needs cited above.

- Engage with the granting agencies and the NDRIO Researcher Council to determine which disciplines require enhancements to their data literacy competencies.
- Assess the need for a series of boot camps/summer schools/grants for graduate students, researchers, data stewards and research assistants who will be working with researchers.

Conclusions and Recommendations

It would be a mistake to assume that the state of numeracy and literacy in the general population does not apply to the academic sphere. A lack of data literacy skills can lead to costly errors in interpretation and decision-making. We need satisfy ourselves that all parts of the system (including the general public) have sufficient data literacy skills to meet Canada's research objectives.

The authors offer the following recommendation for NDRIO to consider:

- Lobby university Boards of Governors at Canadian universities to promote teaching data literacy competencies in all faculties.
- Sponsor a series of awards for student papers in the social sciences and humanities which demonstrate data literacy competencies.