# Canadian Data Repository Inventory

White Paper for Canada's New Digital Research Infrastructure Organization (NDRIO)

December 14 2020

Contact: Peter Webster (peter.webster@smu.ca), Associate University Librarian, Saint Mary's University

Other contributors: Alicia Urquidi Díaz (aurquidi@oceannetworks.ca), Research Associate, International Technology Office of the World Data System. Jean-Baptiste Poline (jean-baptiste.poline@mcgill.ca), Associate Professor, McGill University

## Why is an inventory of Canadian data repositories needed?

**Summary.** A complete and up-to-date inventory providing basic information about Canadian data sources is needed. It is an essential tool for understanding the Canadian research data landscape. The benefits are outlined below.

An inventory was recently carried out as a voluntary effort by several Canadian organizations with the cooperation of the international data repository registry Re3data.org. This inventory has proved very valuable.

However, there are no ongoing arrangements to maintain or build on existing inventory work. An ongoing, simple and sustainable data source inventory process needs to be a basic feature of Canada's developed research data infrastructure.

- Data users—including students, community researchers, academic and public library users—all need a comprehensive reference tool which will identify available sources of Canadian data on particular subjects.
- Academic, government, non profit, and private sector researchers who are creating data need a source for identifying existing data in their subject discipline, and for locating repositories where their data may be deposited to be shared with others.
- An inventory provides essential data for analysing the subject distribution, software and metadata use, and other features of the Canadian data repository landscape.
- An inventory that is interoperable with other (national and supranational) repository registries and RDI catalogues will facilitate integration of Canadian research data into the Global Open Research Commons.
- For national funders and coordinating agencies, an inventory is a tool for targeting data infrastructure in different disciplines, promoting better standards, software and metadata integration and interoperability.

- An inventory identifies data repositories that can receive support from the funding and data coordination agencies, including assistance with certification, preservation, and other functions.

## The current state of Canadian data repository inventory

At present, a listing of Canadian data repositories has been compiled in the international registry of data repositories *Re3data.org*. Several Canadian organizations have worked to bring this listing up-to-date as of 2018. Re3data.org currently lists more than 250 Canadian repositories and is being actively used for several purposes:

- Re3data is actively used by libraries, researchers and data creators as a reference source for locating Canadian data repositories.
- This Canadian inventory in Re3data.org has been a key source used by the Federated Research Data Repository (FRDR) to identify repositories for metadata harvesting;
- Re3data standardized metadata has made it possible to identify groups of interoperable data repositories, which use the common software standards, such as CKAN, Dataverse, ESRI ArcGIS, and Socrata. (See Appendix 1)
- Standardized metadata in Re3data.org allows analysis of the types of institutions offering data, and the subject disciplines being covered.(See Appendix 1)

### FAIRsharing index

In addition to Re3data.org, the FAIRsharing.org repository indexing organization returns 85 databases or data resources when queried for the keyword "Canada". Both Re3data and FAIRsharing are indexers required by some "data journals" for instance, "Scientific Data" from Nature-Springer. FAIRsharing uses metadata such as Domain, Taxonomy, Related Database, Related Standard, Related Policy, etc.

## Challenges and gaps for Canadian data repository inventory

- **Responsibility for submitting inventory entries:** Existing international registries (e.g. re3data, FAIRsharing) places the onus on repository managers to self-identify and enter comprehensive, accurate metadata. Many repositories will not comply or will not provide full information. A centrally curated approach is needed to compile as complete as possible a list of data sources.
- **Staffing for ongoing maintenance:** Entries need to be kept to date as new repositories emerge and as existing repositories change names or cease to operate.
- **Scope data source coverage:** There is no easy definition of what a data repository is. How are repositories with many data sets different from large single database projects? How should independent data collection projects housed in the same institutional repository be identified.

- **Metadata standards and limits**: The extent of metadata about each repository needs to be determined. Consistent metadata is essential. In depth metadata is valuable, but overly complex metadata requirements can impede inclusion of some repositories, rendering the inventory incomplete.

- **Choice of online registry platform:** Re3data.org is being used as a Canadian registry at present. But a number of other options could be considered.

- **Integration of Canadian research data into the Global Open Research Commons.** An inventory of Canadian data repositories should have the goal of interoperable with other international repository inventories (re3data, Fairsharing etc) and should ultimate work seamlessly with a larger inventory of all DRI resources in Canada (e.g. software, ARC and other reusable research assets). This would be a step towards making this  inventory interoperable with other national, pan-national and domain specific catalogues (such as the EOSC marketplace, s. Barbot et. al. (2019)).

- **Rapidly expanding data landscape:** Interest in open data has led many government organizations to develop useful data catalogues. Many national, provincial, and  municipal data sources are already included in Re3data.org. It is likely that hundreds or even thousands of new data sources will emerge, as additional towns, cities, and non-government organizations see the benefits of hosting open data. Although it can be hoped that many existing data repositories will be consolidated,the work of maintaining an accurate inventory is likely to grow in scale.

## Data inventory background

Since the early 2000s, several organizations have recognized the need for a data source listing and have worked to collect Canadian data sources and data distributing repositories.

The Canadian Committee of CODATA compiled an annual list of data projects for several years. This effort was taken up for a time by the National Research Council of Canada. At the same time, a number of individual Canadian university libraries attempted to create their own lists of useful data sources. As the open data movement has emerged, a number of international, “crowd sourced”data source registries have also developed.

By 2015, a review undertaken by Research Data Canada showed that different disconnected listings of the data landscape were based on somewhat different criteria. All available lists were out of date. Canadian and international lists contained overlapping, but incomplete records of Canadian data sources. (*Where Does Canada’s Social Science Research Data Live*, Webster, 2018)

Between 2016 and 2018, several agencies coordinated an effort to review all available sources and to create a single inventory listing of Canadian data repositories. This effort was undertaken by Saint Mary’s University, in cooperation with the National Research Council of Canada, and with assistance from several other academic libraries, including the University of Toronto and York University.

The international data source registry Re3data.org was used for the inventory for several reasons. “Re3data covers global research data repositories from different academic disciplines. It includes repositories that enable permanent storage of, and access to data sets. Over 2400 data sources from over 30 countries are included. The registry is funded by the German Research Foundation (DFG), begun in 2012, and came under the auspices of DataCite in 2015. Re3data.org also became a partner with the European OpenAIRE project.

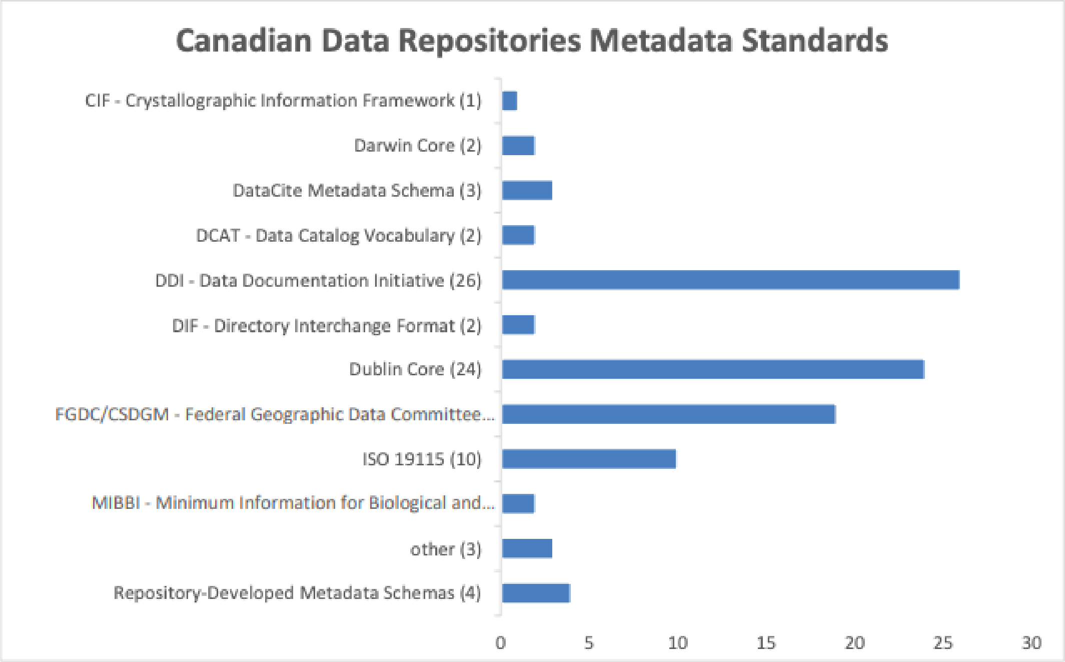
The FAIRsharing database catalogue was also considered as a resource for the Canadian inventory. FAIRsharing is an online registry which lists databases,data repositories, metadata standards, and data management policies. FAIRsharing is a project of the “Data Readiness Group” at Oxford University, began in 2009. Although FAIRsharing provides considerably more detailed metadata than Re3data.org, it is considerably less comprehensive in its coverage of Canadian repositories, and more limited in its discipline subject coverage. FAIRSharing now contains around 80 Canadian listings.

Re3data.org coverage of Canadian repositories was found to be the most complete and to have the most extensive subject coverage. Re3dat.org editors were open to working on the Canadian collection project. At Canadian request, Re3data.org added functionality to their application programming interface (API) to provide easy extraction of national repository metadata sets, allowing analysis of subject coverage, metadata and software being used.

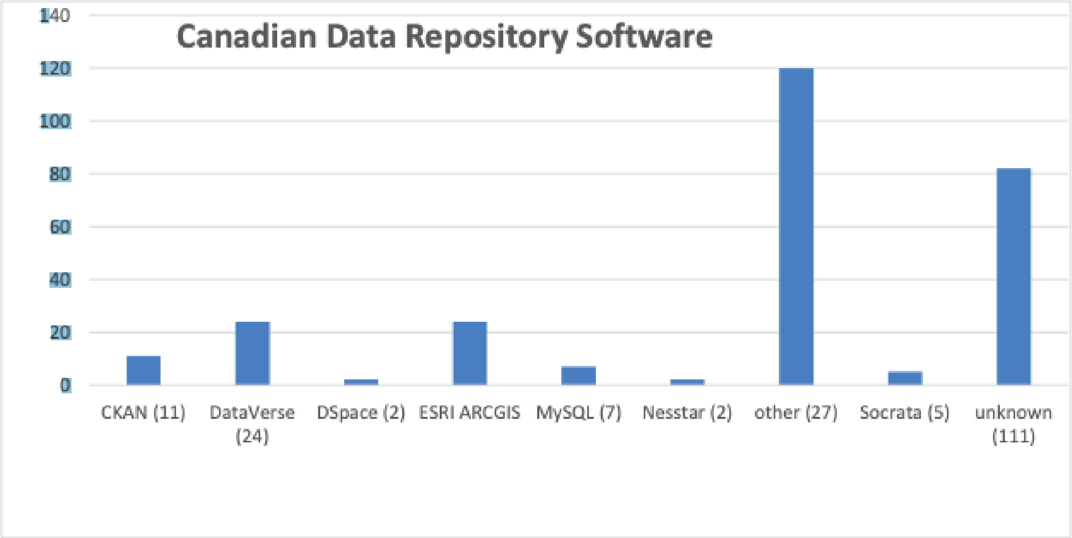
Individual graduate students worked full time over two summers to review all available lists of Canadian data repositories and submit entries to Re3data.org for review. The Canadian inventory was largely completed in 2018.

## Appendix 1. Canadian Data Repository Inventory Analysis

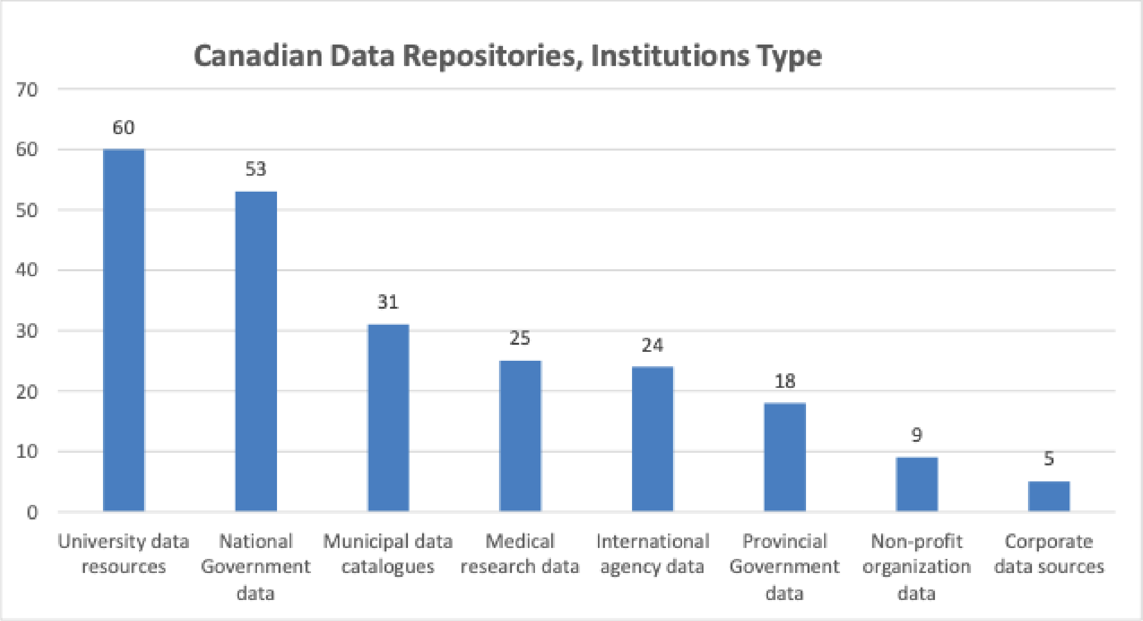
Charts from: Webster, Peter. “Integrating Discovery and Access to Canadian Data Sources. Contributing to Academic Library Data Services by Sharing Data Source Knowledge NationWide,”IFLA World Library and Information Congress 08, 2019



Work on the inventory has made Re3data.org a reliable and comprehensive reference tool for Canadian, as well as other data sources. It provides a wealth of new information about the kinds of data being collected in Canada. It provides additional information about the terms of use, access and sharing arrangements for each data source, as well as information about metadata standards, and software being used.



(Inventory work so far has provided valuable information about data repository software being used, but software for many repositories still needs to be identified.)



## References

- Barbot, L., Moranville, Y., Fischer, F., Petitfils, C., Ďurčo, M., Illmayer, K., Parkoła, T., Wieder, P., & Karampatakis, S. (2019). SSHOC D7.1 System Specification—SSH Open Marketplace. https://doi.org/10.5281/zenodo.3547649

- Ferrari T., Scardaci D., Andreozzi S. (2018) The Open Science Commons for the European Research Area. In: Mathieu PP., Aubrecht C. (eds) Earth Observation Open Science and Innovation. ISSI Scientific Report Series, vol 15. Springer, Cham. https://doi.org/10.1007/978-3-319-65633-5_3

- Webster, Peter. "Where Does Canada's Social Science Research Data Live? An Evaluation of Data Disposition." Presented at the IASSIST 2018: Once Upon A Data Point: Sustaining our data storytellers (IASSIST 2018), Montreal, May 2018.. https://doi.org/10.5281/zenodo.3775742.

- Webster, Peter. "Integrating Discovery and Access to Canadian Data Sources. Contributing to Academic Library Data Services by Sharing Data Source Knowledge Nation Wide,"IFLA World Library and Information Congress 08, 2019. http://library.ifla.org/2514/1/248-webster-en.pdf.

- Webster, Peter. "Toward an open and integrated research data repository landscape in Canada". 2018. Internet Librarians International Conference Presentation, London, August 2018.