Gaps and Opportunities for NDRIO Support of Research Data Management

Insights from the Global Water Futures Data Managers

K. Dukacz[1], B. Persaud[2] , A. Peterson[3], L Moradi[3] , G. Saha[4]

[1] McMaster University, Hamilton, ON, Canada

[2] University of Waterloo, Waterloo, ON, Canada

[3] University of Saskatchewan, Saskatoon, SK, Canada

[4] Wilfrid Laurier University, Waterloo, ON, Canada

December 14, 2020

Gaps and Opportunities for NDRIO in Research Data Management

Global Water Futures: Solutions to water threats in an era of global change is a Canada First Research Excellence Fund (CFREF) program led by the University of Saskatchewan in partnership with University of Waterloo, McMaster University and Wilfrid Laurier University. The Global Water Futures (GWF) program is the largest, academic led freshwater research program in the world with partnerships across 18 Canadian universities, various levels of government, dozens of Indigenous communities, and various organizations in Canada and abroad. The program currently supports 58 projects including 6 indigenous-led projects and several affiliate projects. GWF focuses on transdisciplinary collaboration involving almost 200 university faculty and 965 highly qualified personnel (HQP) researching the gamut of water issues from observing climate change impacts, modelling future climate and developing new sensors, to increasing our understanding of the role of water in perceptions of health, and engaging citizen science. Within the GWF program there is a dedicated interdisciplinary core data management team consisting of computer scientists, hydrologists, academic librarians and data managers working to put systems in place to support the implementation of the GWF Data Policy. Despite these significant resources it remains a challenge to provide a clear and supported pathway for researchers to implement the best practices of Research Data Management (RDM). The researchers that we interact with use a range of tools including but not limited to those outlined in Appendix 1.

The breadth of disciplines, institutions and individuals represented by GWF have highlighted many challenges and opportunities for the strengthening of Digital Research Infrastructure (DRI) in Canada. While a comprehensive review and extensive consultation with researchers was not possible due to time constraints, the data managers of the core GWF Data Management team prepared this paper to highlight issues that have been identified in working on data management with researchers within GWF and at our home institutions over the past three years, as well as opportunistic discussions targeted at addressing this call. The value of our perspective is that it is defined by the experience of researchers across disciplines, faculties, and institutions; the main focus of GWF is water science but the common focus of our input is on supporting research excellence across the board as *"data underpins quality research in every discipline"* (Government of Canada. Digital Research Infrastructure). Systems need to be put in place to transcend the heterogeneity of infrastructure across disciplines, academic institutions, and time to create a stable foundation for innovation and discovery in Canada.

DRI is no longer the sole domain of computer scientists or those with high computational demands, it is critical to the work of all researchers. There is currently a disconnect between the resources that are provided to researchers and the expectations surrounding the collection, safe storage and sharing of data coming from ethics boards, funders, journals, stakeholders and the public. In short, easy to access tools need to be provided for storage, collaboration, sharing and preservation on trusted and enduring platforms to create a stable foundation for research in Canada. Our observations and recommendations follow the framework of the key elements and supports of DRI identified by the Government of Canada (https://www.ic.gc.ca/eic/site/136.nsf/eng/home).

Element 1: Digital network for research and education*, allowing researchers to share data and collaborate across Canada and around the world*

In a recent discussion, researchers identified the following as requirements of a research system: Support for collaboration; version control, storage and logs to describe changes to data. A digital network for research and education is critical to promote cross-institutional, transdisciplinary research. Funders and institutions value these synergies, but the current infrastructure does not facilitate them. Institutional storage is typically quite siloed with significant barriers to file sharing and collaboration with external investigators and stakeholders. Favouring expediency and the path of least resistance, researchers turn to low-barrier, easy to use alternatives such as Dropbox, Google Drive, Google Docs and MS Teams. While many researchers would prefer a more secure platform in terms of increased access control, server location and provider (many viewing Canadian-based, government or institution servers as preferable and more trusted), these external alternatives are deemed necessary for research to progress at a reasonable pace.

Compute Canada provides many resources that could help researchers the tools, but they are not easily accessible and are piecemeal solutions. For instance, Compute Canada offers secured share groups that could enable researchers across institutions to share data, but few people are aware of these resources. Training is required to implement the storage solution and it addresses just one part of the toolkit needed to collaborate effectively. As well, Compute Canada can provide the basic building blocks with which researchers could build their own system (virtual machines, IP addresses, software etc.), but there is a knowledge gap and a high barrier to use. Not all researchers have the time or inclination to, for example, develop the skills to set up and manage their own remote server hosting a Nextcloud Hub.

To address this challenge, NDRIO should provide a basic collaborative platform with low barriers to use, both in terms of usability and cost. When selecting a platform, it will be critical to look not only at the tools but also how the tool is managed. For example, the PURR system at Purdue University, based on HubZero, may be one model to take into consideration. PURR offers free access for all researchers through a tiered resourcing model. All researchers receive the base level resources at no cost. Researchers with grant funded projects or other supports receive additional resources without added fees and beyond those allocations, storage can be expanded at a reasonable cost to the researcher.

Element 2: Data management (DM), *allowing researchers to find and access data*

Data is at the heart of the research findings presented in journals. Increasingly, journals request that data be provided when articles are published so that findings can be verified. As well, governments are also recognizing the importance of sharing data to promote collaboration, build knowledge and spur new research. This is demonstrated through recent initiatives around Open data, as well as through strengthening requirements for the sharing of research data collected through public grants. As data collection is very expensive, using a lot of resources (time, personnel, money), it is critical that data be viewed as durable assets. For data to endure over time it must be well managed throughout its life cycle. Funding that is tied to making data public is becoming increasingly a common but the tools to help researchers manage the path from proposal to publication are missing.

The DMP Assistant, in concert with the many resources provided by the Portage Network, gives researchers the tools to develop a solid Data Management Plan (DMP). While these resources offer guidance on best practices in areas such as back-up, metadata development, and preservation, it relies on the individual researcher to put the digital tools in place to operationalize the plan. Researchers also receive input on how they need to manage their data from ethics boards which may stipulate, for example, secured storage, secure communications, auditable access control, data anonymization or data destruction. It is necessary for the digital infrastructure to be in place to facilitate researchers' compliance with these requirements.

Special attention needs to be paid to DRI to support the active research phase, characterized by data collection, metadata and collaboration as well as creating a clear path from start to end. This means having infrastructure, tools and training to keep data safe. In addition to the safety of data, it is crucial that data be well documented as it is developed to ensure its usefulness over time. Data and document version control and change log records are key components of this. NDRIO should provide additional active Research Data Management (RDM) tools that bridge data management from the DMP to a repository such as the Federated Research Data Repository. Additional guidance should be developed to support the use of existing established repositories (e.g., Polar Data Catalogue, The Gordon Foundation DataStream etc.) and new repositories to ensure that they meet specific standards, such as the CoreTrustSeal and ISO 16363:2012

Space data and information transfer systems–Audit and certification of trustworthy digital repository, would also help define the RDM path for researchers.

In addition to the collaboration platform described above, additional resources are needed, including:

Cloud Storage: Easy access to cloud storage for researchers **-** these resources are available to people if they know how to install and run their own storage application. Cloud storage should be offered as a hosted solution to make access as easy as Dropbox, Google Drive, or Microsoft Azure, AWS so that researchers have a safe, reliable place to store data that allows cross institutional, external access in a controlled way.

Sensitive data storage: Provide storage solutions configured to appropriately handle sensitive data so that researchers can easily enact the requirements of ethics boards and meet the needs of partners and stakeholders. Let researchers focus on what they do best without having to become cybersecurity experts.

Storage to support First Nations data sovereignty: Engage with First Nations organizations such as the First Nations Information Governance Centre (FNIGC) to develop storage to support First Nations data sovereignty and which respects the OCAP (Ownership, Control, Access, Possession) Principles.

Backup capacity: While Compute Canada again can provide the infrastructure to set up an incremental backup service (e.g., open source UrBackup on a Linux server) very few researchers or labs have the capacity to institute this type of system. Providing a platform and support for data backup would be very helpful. Ensuring that data is stored in accordance with recommended 3-2-1 redundancy rules deposited with the project is increasingly important as fewer student researchers work solely (or at all) on institutionally owned computers. A service such as UrBackup can push changes to the server whenever a computer is connected to a network providing added protection in the active data management phase and prevent data loss that can occur if data transfer is left to the end of the project.

Element 3: Research software (RS)*, enabling researchers to access and use data*

There is a significant push by many organizations and entities to make data more accessible such as the GoC Open Government portal, Conservation Ontario, The Gordon Foundation Datastream and the FRDR. It would be helpful for NDRIO to provide additional leadership in this area to build an inventory of data portals available and harvested by FRDR, and additional metadata guidelines. FRDR is poised to add great value in this respect by enabling users to query many databases through one search tool. To promote data deposit by streamlining the decision-making process for researchers, a list of established and trusted repositories should be curated to make it easier for researchers to quickly identify a target deposit location for their data. Additional tools for researchers include the provision of software such as MATLAB and geospatial tools. As well, application development that is funded through NDRIO should require an assessment of how the tool could be utilized by the broader research group. Often tools developed have a specific discipline focus which may limit the return on investment but sometimes it is necessary as some discipline do need domain specific tools.

Element 4: Advanced research computing (ARC)*, involving super computers that allow researchers to analyze massive amounts of data*

The ARC resources that a recurrently provided through Compute Canada are seen to have a steep learning curve. This is particularly true for graduate students who have a limited time to ramp up and get on with their research. A need for better support and training at the grassroots level was identified. Additionally, researchers noted that the Resource Allocation Competitions (RAC) applications are quite cumbersome as they need to be frequently renewed and can miss supporting programs that require allocations to meet the needs of a group of projects rather than a single project. Further, some researchers have concerns about the sustainability of the allocated resources given the finite financial support of Compute Canada.

Support Element 1: Highly qualified personnel, *skilled people with the expertise to support the DRI system and help researchers make the most of cutting-edge tools*

It is clear that there is a great deal of expertise in the organizations that are now under the NDRIO umbrella. As described, there is a need for additional direct support of researchers tailored to their specific level of expertise. There is an opportunity to streamline services and make more efficient use of resources by taking advantage of the individuals within the system. Additional human resources would help propel researchers in their engagement of DRI tools. One specific resource that was identified was that of a "Liaison": A dedicated contact person to support researchers along the path from DMP inception to project closure. Such a resource could assist researchers in making their DMP more actionable by advising on active data management, providing guidance on potential repositories, managing embargo, metadata needs and obtaining a digital object identifier for publications. This resource should have an in-depth knowledge of the offerings of NDRIO and assist researchers in identifying appropriate resources to meet their needs. They should provide an actionable roadmap for the researcher aligned with their data management plan.

Support Element 2: Cybersecurity, *to ensure that the knowledge that is created is protected*

Cybersecurity is an extremely important aspect of DRI and one that should not be left up to individual researchers to contend with. As described above, there is a need for cybersecurity experts to manage the infrastructure that Canadian researchers will rely on. Again, there are resources currently available that would allow individual researchers to build their own server and run their research applications on it. This type of resource requires monitoring for security updates and breaches which exceed the capacity of most individual researchers. Offering managed platforms will allow researchers to confidently focus on their research.

While cybersecurity is vital to maintaining the integrity of DRI systems, researchers would like access to a wider range of tools when accessing sensitive data through access points such as those hosted at UofT, ICES or through Statistics Canada Data Centres.

Additional Considerations:

Communication and Guidance Role

NDRIO has the potential to be the research and data nexus of government, academia, industry, indigenous nations and non-profit organizations. Leadership is needed to bring together the many research focused initiatives and align the efforts. Increasing alignment between organizations will provide efficiencies in resource use and create a more agile and responsive research ecosystem. NDRIO should look to become the authoritative resource for research in Canada. Through this role it should work to organize and amplify the diverse efforts around data management, research and innovation across the various sectors. There is a need for an authoritative source for information. Further, with the integration of the various organizations, there is tremendous potential for offering education and hands-on practical training in data management, reproducible research techniques and tools through various channels.

Lastly, NDRIO is in a unique position to provide leadership in RDM policy accountability by developing systems that support and monitor the meeting of requirements. As policies are developed and enacted there will be a need for the overall system to provide incentives and provide checks and balances associated with data sharing.

Final Thoughts: NDRIO is well positioned to play an instrumental role in strengthening the research landscape in Canada. Seamless support for researchers from DMP to Publication is needed to meet demand for tools for planning, collaborating, publishing and preserving research data. Tools need to be provided to replace the multitude of easy-to-use external DRI that researchers rely on due to the high barriers associated with the current infrastructure. As stated previously, DRI is a requirement for all researchers and must be provided to the masses. Sustainability is also a concern so DRI should be prioritized to place it above the budgeting churn of governmental change. In addition, training and support should be readily available to researchers and their HQP to provide clear direction on implementing RDM and accessing the DRI tools. Having a strong DRI foundation will help build researcher readiness to share data and build trust in the system. While much work needs to be done to create a solid RDM DRI, we are excited to see the momentum that is building with NDRIO and hope that this momentum will scale up across all levels of expertise and discipline at an exponential rate.

Gaps and Opportunities for NDRIO in Research Data Management

Table 1: Some typical DRI resources that are being used by researchers within our community

DRI Resource	**Brief description of how it being used by our researchers**	**Link to resource**
Portage DMP Assistant and associated resources	Data management planning, and training resources	https://portagenetwork.ca/
Compute Canada	For cloud computing, high performance processing, and storage (*but typically users tend to have strong computer skills*)	https://www.computecanada.ca/
Federated Research Data Repository (FRDR)	Access to data and deposit data	https://www.frdr-dfdr.ca
Dataverse/Scholars Portal	Access to data and deposit data	https://dataverse.scholarsportal.info/
Globus	Facilitates transfer of files	https://www.globus.org/
The Gordon Foundation DataStream	Access to data and deposit only water quality data	https://gordonfoundation.ca/initiatives/datastream/
Polar Data Catalogue	Access to data and deposit metadata and data for cold regions	https://www.polardata.ca/
Canadian Surface Prediction Archive (CaSPAr)	Access to numerical weather predictions data issued by Environment and Climate Change Canada	https://caspar-data.ca/
The Cuizinart	Cloud base platform that is used to *slice and* dice large, gridded datasets	www.cuizinart.io
WISKI (Water Information Services Kisters)	Storage, quality control, and access management of quantitative hydrological and meteorological timeseries data in the active research phase	https://www.kisters.net/NA/
Github	Code management	https://github.com/
Zenodo	Deposit data and codes	https://zenodo.org/
Ameriflux	Access to data and deposit only flux data	https://ameriflux.lbl.gov/