# The Needs of Canada's Future Digital Research Infrastructure Ecosystem

Submitted on behalf of the University of Ottawa research community by:

**Guy Levesque |** glevesq4@uOttawa.ca
Associate Vice-President, Research Support and Infrastructure
Office of the Vice-President Research | University of Ottawa

This document was drafted by:

**Sean Geddes, MSc, MBA**
Executive Advisor
Office of the Vice-President Research | University of Ottawa

**Une version de ce document en français est disponible sur demande**

uOttawa

## Scope

Canada's digital research infrastructure (DRI) is a key component in supporting the defined strategic areas of research at the University of Ottawa (uOttawa). Here we outline current challenges, needs, and opportunities presented by our research community with respect to DRI tools, services, and support, the ideal future state of DRI in Canada, as well as how this state can be achieved with the support and guidance of NDRIO. Taking into consideration the diverse requirements and uses of DRI across all research disciplines, uOttawa has incorporated information obtained through consultations across multiple disciplines, including medicine, artificial intelligence, engineering, psychology, the social sciences, the health sciences, information technologies, and library research.

## The University of Ottawa's perspective

The University of Ottawa employs a number of DRI tools, services, and resources to complement and achieve goals linked to our research programs. These DRI-related components can largely be divided into two groups: resources provided internally and resources provided by external third parties.

Internal resources, which are provided and generally supported by uOttawa, include: rack-mounted servers/clusters located in the uOttawa Research Data Centre; limited database storage for our research groups on a local network-attached storage (NAS); virtually unlimited cloud-based storage; virtual machine infrastructure for supporting specialized low-intensity tasks that do not warrant a dedicated server; a limited locally hosted web-based computer application development and deployment platform, and; limited in-house local high-performance computing infrastructure.

Actively used resources that are provided by external third parties include: the data stewardship resources provided through the Canadian Association of Research Libraries' (CARL) Portage Network (e.g. Data Management Planning Assistant, Data Management Plans, Sensitive Data Tools); Compute Canada resources (e.g. the Rapid Access Services for access to high-performance computing, virtual machine clusters); the CANARIE national backbone network; population-based data repositories (e.g. Statistics Canada, Institute for Clinical Evaluative Sciences (ICES), etc.); and the SOSCIP platform, which has been particularly valuable for academia-industry collaboration.

## Current issues

As is the case in other Canadian academic institutions whose increasing research activities leverage DRI resources, the University of Ottawa is sensitive to a number of challenges faced by Canadian research groups in terms of the accessibility and availability of specific DRI tools, resources, and capabilities.

Compute Canada and CANARIE resources and tools, which are widely leveraged by the uOttawa research community, have been instrumental in developing and executing DRI-dependent research programs. In using these resources, our research groups have encountered some challenges and concerns, including the requirement for more education and training, greater awareness of available resources, increased availability of remote servers, and better use of the linked scheduling platform.

Research groups that actively use Compute Canada and CANARIE resources have identified training and educational resources as a consistent concern and challenge. More specifically, these challenges arise due to a lack of awareness of, and limited training on: the capabilities of these tools; information on selecting the appropriate tool for a given task; and appropriate training to fully capitalize on the excellent, but limited, resources provided by Compute Canada and support from CANARIE. It is important to note that in some

instances, the appropriate training and educational resources may, in fact, be available to the research community, but the community is simply not aware of their availability. This has raised the additional, and certainly related, concern that there is limited awareness of the availability of resources linked to Compute Canada and CANARIE. Indeed, this has highlighted the need for further dedicated activity and resources to contribute to increasing awareness and use of these resources and tools by our research community.

Compute Canada server availability has also been raised as a periodic challenge: there are instances when the servers have limited capacity due to priority initiatives taking precedence (e.g. computing resources are demanded from funding agencies). In line with resource availability, there are also concerns regarding the available capacity of graphics processing units (GPUs) to meet the needs of DRI-intensive research programs. In the past, to circumvent limited GPU support to our research activities, we have invested in some internal GPU servers, but it would be an asset for Compute Canada to expand this capacity. Finally, in terms of scheduling and accessing these resources, we have encountered certain challenges in navigating and using the Slurm scheduler. Superficially, this appears to stem from inexperience with this type of Linux-based platform, but this certainly stresses the need for resources and tools to use it properly.

In disciplines such as healthcare, life sciences, and human behavioral psychology, a clear reoccurring challenge encountered by our researchers has been their access to resources that manage participant- and patient-confidential information within the DRI. As one might expect, researchers must be extremely careful with personal information related to subjects and patients since this information often contains sensitive data that can be connected back to individual participants. Additionally, in storing and transferring data, there are strict requirements for compliance with provincial and national legislation, such as the *Personal Health Information Protection Act* (PHIPA) and the *Personal Information Protection and Electronic Documents Act* (PIPEDA). Correctly managing and securely storing this data has made data anonymization a valuable asset and a necessity for researchers working with patient and participant data. However, there are often inconsistent methodologies used to accomplish this task, and this can sometimes present a risk of increased personal information vulnerability. Access to resources to securely store this type of data would be a valuable asset to the research community.

Beyond the correct management of patient and participant information discussed above, there is also a challenge in having access to resources, tools, and services to effectively transfer this type of data to collaborating institutions while maintaining its security. A significant challenge has been the ability for some of our research groups to transfer confidential data to colleagues externally to accomplish research objectives. Access to resources and tools to enable the sharing of this type of data would also be considered an asset to the research community.

Finally, resources for the development and deployment of web-based applications have been an area of interest within the uOttawa research community. Locally, some limited resources have been provided to assist our researchers with accomplishing these research objectives; however, there is a strong need for external support as challenges present themselves while working through this type of activity. Indeed, our research groups have leveraged the Compute Canada virtual machine resources to host APIs for mobile applications but some of our interdisciplinary groups have expressed interest in receiving further support in the actual development and refinement of the applications themselves. Our researchers would find it quite valuable to have access to such resources.

In summary, uOttawa research groups have collectively experienced immense value in leveraging available DRI resources to help drive their research programs. However, clear identifiable challenges have arisen, along with requirements that need to be addressed to further support and contribute to research excellence in Canada. Some of the challenges in using the Compute Canada and CANARIE resources include: access

to, or availability of, training and educational tools; awareness of the resources available; availability of remote servers; and problems in using the scheduling platform. Our research groups have further expressed a need for resources such as: web-based or mobile application development resources; resources to effectively and securely manage participant data; and resources to share this data safely and collaborate with research groups at external institutions. Further access to these resources would certainly promote Canadian research excellence through the development of the national DRI.

## Future DRI state

Over the years, the Canadian DRI ecosystem has certainly become more cohesive. For example, Compute Canada data centres/clusters have continued to converge, bringing together common software stacks and promoting the use of the same software by the majority of clusters. This increasingly cohesive nature is also evident in the creation of a single interface to request accounts that works on all clusters. Collectively, this has contributed to a more unified view of the Canadian DRI ecosystem.

The University of Ottawa's vision for the further development of a cohesive Canadian DRI ecosystem includes the availability of resources to actively and safely share confidential data amongst national collaborators; enhanced communication and task-transfer between clusters; continual development of Research Data Management (RDM) resources; overall improved data literacy; continuous and evolving training and educational resources; and overall homogeneous integration of the available resources in the DRI space.

One challenge previously identified was linked to the perceived limited capacity of the Compute Canada data centres. A potentially viable solution would be the ability of schedulers to remain agile and distribute scheduled jobs to less active clusters at that time, thereby readily distributing tasks across the network. Furthermore, it would be beneficial to see an increase in the overall capacity of the Compute Canada servers (e.g. GPUs).

When developing a vision for the future DRI ecosystem, special consideration needs to be given to the educational and training resources accompanying the rapidly developing DRI framework. It is essential that the appropriate training be readily available to members of the research community to support the growth of the national DRI. Furthermore, agile educational programs need to be consistently implemented to run simultaneously with the integration of new tools. The agile nature of the training paradigms being incorporated would allow them to adjust to this rapidly evolving field.

Given the rapid expansion of national DRI resources, further consideration and development efforts should be directed towards Research Data Management (RDM) resources, training, and guidance. RDM best practices help ensure the accessibility and security of data, thereby enhancing the level of "data literacy" within our research community. Resources such as these will be essential in developing consistent and reproducible practices in the research community. In addition, there is a future need for NDRIO to support projects to build on some foundational work related to data management policies (DMPs) because without a nationally developed standard to reference, there is an increased risk of inconsistencies and unsatisfactory research data management.

Collaboration is a term at the heart of academia and the research ecosystem. The future DRI state is envisioned to be one that fosters more collaborative activities by leveraging DRI to help safely share data across multiple institutions nationally (and even internationally), while maintaining sufficient security in line with national policies. This could enable researchers, especially ones working with sensitive participant information, to more actively collaborate and build their research programs (e.g. permitting activity such as multi-site testing and development).

Finally, leading into this future state of Canadian DRI, it is anticipated that with further DRI developments, academic institutions will gradually expect that the Canadian DRI will accommodate access to tensor processing units (TPUs). Indeed, it is recognized that the Graham cluster currently has 144 of the NVIDIA T4 tensor core GPUs. Access to these artificial intelligence enabling resources is certainly a relevant future consideration in the development and growth of DRI in Canada.

## How to bridge the gap

Addressing the current gaps in the national DRI ecosystem will certainly be a gradual and national effort requiring buy-in and collaboration among institutions and organizations, which will effectively build on the past and recent successes of our local and national networks.

Continued commitments to universities in order to support experts in leveraging current DRI resources will be an essential aspect of growing the DRI in Canada. This type of support (e.g. CANARIE) has been invaluable in supporting staff and students, not only in reaching their research objects, but also in leveraging the full capabilities of available DRI. In the past, this type of support has been instrumental in ensuring that researchers are appropriately trained in the tools and resources available. These resources have proved to be especially beneficial to multidisciplinary researchers who may have less knowledge of DRI tools.

Communication will also be essential to bridging the gaps within the Canadian DRI ecosystem. This will need to consist of a number of different aspects, including regular feedback channels as the DRI resources grow (e.g. regular meetings such as an annual stakeholder forums), regular communication with academic institutions on developments with DRI to promote resources within institutions (thereby ensuring efficient change management practices in the DRI ecosystem), and regular training on resources to enable researchers to fully capitalize on the capabilities of the DRI.

Security will be an important factor in managing research data as NDRIO integrates additional features into the national DRI. It will be essential to implement controls and tools to help safeguard research data and prevent the possibility of security incidents. One tool that could help enhance this security effort is the implementation of multi-factor authentication (MFA) for users. To date, it is certainly recognized that some activities have begun to address these security concerns, but this will absolutely need to be a key consideration at the forefront of future initiatives.

As NDRIO begins to address the gaps in DRI, it will also be important to continuously support defined research projects through programs such as the Resource Allocations Competition (RAC). Through research activities, these programs will directly enhance knowledge of DRI within the community and support DRI ecosystem growth.

Finally, in light of the rapid growth of the DRI, institutional support is a key aspect of executing our research programs. Indeed, one challenge faced by institutions is how to determine the right support structures and levels based on their data management needs. Given that individual institutions have different requirements, sizes, and research intensities, the most efficient way to determine this required support is through tools such as research data management (RDM) maturity frameworks. Indeed, there are a number of available RDM maturity frameworks for evaluating needs; however, there is gap when considering the lack of consistency in the absence of a unified RDM maturity framework. As NDRIO services mature, a RDM maturity evaluation framework should be formalized at the national level to help institutions evaluate the current RDM support they provide and identify areas of focus for future development, paying particular attention to the diversity of institutions in Canada.

## Conclusion

Through its research-related activities, the University of Ottawa research community frequently leverages and capitalizes on Canada's digital research infrastructure. Through these activities, we have become sensitive to the DRI's current challenges and opportunities, and we believe that this submission has provided some insights that will contribute to the development of NDRIO's Strategic Plan to produce "a single and unified vision of the highest DRI priorities for 2021-2024." These insights have highlighted the need for further, and continuous, training and education linked to DRI, improvements in data literacy, increases in DRI capacity and investments, re-evaluation of research data management frameworks, consideration of institutional diversity and inclusiveness, resources to facilitate inter-institutional collaboration and data transfer, security frameworks and training, and finally, increased awareness of the resources available to our research communities.

The University of Ottawa research community is grateful for the opportunity to share our thoughts with the NDRIO and remains committed to continued productive collaboration and partnership.