Considerations from researchers from Université Laval

Authors: Jacques Corbeil, François Laviolette and Florent Parent.

Existing Landscape and Current Issues

Advanced Research Computing (ARC)

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

We are observing constant growth in demand for ARC resources over the past years, whereas the growth of ARC resources has not followed this demand. This deficiency has a direct impact on researchers in severely curtailing the allocation of sufficient resources. Researchers of the Big Data Research Centre, from the Institute of Intelligence and Data and Engineering faculty, are among the largest users of Compute Canada services at our University. Unfortunately, some researchers need to resort to alternative sources of computing power by accessing old clusters, on-premise ARC, or commercial cloud services. This tendency will continue until more resources are made available, the sooner the better. One can understand the need to do it right, but absolutely no one would argue that NDRIO has the mandate to act immediately. We cannot be competitive without this resource.

Moreover, there is a strong tendency of research-intensive universities (UBC, U. Laval, among many others) to acquire their own digital research infrastructure since Compute Canada could not presently address their needs. This tendency will continue if NDRIO does not step up significantly and do so rapidly this will contribute to furthering the inequity between universities.

The need for computing resources may come in bursts (this is common in AI for health and bioinformatics research projects). Unfortunately, the scheduling assumes a steady use of resources, so jobs are given lower priority by the scheduler for using too much computing (burst) during these times. By its nature, research has uncertainties on experiments and unexpected delays getting access to data, while still needing to meet deadlines. A mechanism to support short-term burst computing requirements would be beneficial to many researchers.

The assistance provided by Compute Canada to researchers working on COVID19 was genuinely proactive, and we believe it was a great success. Compute Canada should be commended for this. NDRIO should institute a specific request mechanism that is more formal than calling your local or regional providers.

In some cases, researchers need tools and services that are either not available (yet), or not feasible or practical on the national ARC. Here are some examples of services required by research projects at Université Laval, and available on campus DRI:
- Secure platform to store and analyze sensitive research data. This is a requirement in many medical research fields, or any research involving population data. Such a platform must meet technical and legal requirements that may require that data be stored on campus or within the provincial jurisdiction.
- Data-intensive science: Scientific instruments in laboratories (MRI, electron microscopes,

genome sequencers, etc.) produce substantial amounts of data. Moving large datasets from the campus to the national DRI is challenging. The campus networks are not typically designed to support long-lived and high-speed data flows over the wide-area networks[1]. Although we have high-speed R&E networks from CANARIE and NRENs, data-intensive science will require bridging the campus to the national DRI.

- Near-real-time data processing: Supporting research where data from the scientific instrument needs to be processed in near-real-time on computing and GPU resources, and providing feedback during experimental procedures in the laboratory.
- Data streaming and sensor networks: Projects under preparation involve data streaming from sensor networks (IoT, drones, vehicles) present on campus or in remote locations with direct connectivity to the campus. Such a project requires a stable data streaming service and computing near the data sources.

These examples illustrate some of the challenges researchers are facing today, and require infrastructure to be deployed and supported on the campus. But there is currently no structure or program in place to assist campus IT in providing this type of infrastructure for research or connect infrastructure already available with researchers who need them, which is a missed opportunity to create an integrated DRI ecosystem.

Research Data Management (RDM) and Research Software (RS)

The distribution of the Compute Canada software stack across clusters using CVMFS is an example that has a positive impact on the research community. This facilitates the reuse of available software and provides a common user experience across our ecosystem. JupyterHub/Lab and Magic Castle are examples of Research Software development that started as local innovation and scaled to the benefit of the national platform and international partners.

But without national support for RS, there is no coordination for the development and integration of scientific gateways, software pipelines and other research software development. Researchers with the means to develop their software often end up with suboptimal solutions since they do not know or understand the underlying ARC architectures, or may miss opportunities to reuse existing software due to the lack of visibility of the research software ecosystem. The sustainability of these solutions is also a problem when there is no national support to integrate and maintain these solutions on the national DRI.

Future DRI State

The future DRI would benefit from an integrated ecosystem that can deliver the DRI to accelerate the time-to-discovery to researchers, no matter where it is happening in Canada.

ARC

The NSF's Blueprint for a Cyberinfrastructure Ecosystem for Science and Engineering in the 21st Century[2] presents NFS's vision for an integrated national Cyberinfrastructure ecosystem for the US, including a blueprint for the future computational ecosystem. NSF's vision and blueprint document

---

[1] https://fasterdata.es.net/science-dmz/motivation/
[2] https://www.nsf.gov/cise/oac/vision/blueprint-2019/

should be used as a guide to shape the future Canadian DRI ecosystem.

Increased computing and storage resources in order to meet the increasing demand from researchers is obviously a desirable state for our future DRI. In addition, NDRIO should take the opportunity to better bridge the national DRI with the research campuses (How to bridge the gap), and leverage capabilities and capacities provided by commercial cloud services.

NDRIO should also actively support exploratory pilots and prototypes to promote innovation for new capabilities and services. We would benefit in engaging with existing NSF-funded pilots and initiatives (e.g. Open Storage Network[3], Global Research Platform[4], FABRIC[5]). Projects for pilots and prototypes should not be limited to hosting sites.

Research Data Management

Supporting researchers in Research Data Management requires staff who will take the role of "Data Champions"[6] to guide and provide assistance on implementing research data management practices from the beginning of a research project. These Data Champions will engage with various research communities, promote good and established practices in RDM and encourage the support of FAIR[7] principles. Mentoring and training will be an important role of the Data Champion, and proximity to the researcher is important to facilitate outreaching.

In a future DRI, NDRIO should foster a community of Data Champions across institutions to share experience and encourage collaboration across teams. The community should be open to all RDM stakeholders in Canada. NDRIO should also support research institutions to meet the Tri-Agency research data management policy (platforms that securely preserve, curate and provide continued access to research data), and with the Roadmap for Open Science of the Chief Science Advisor of Canada[8].

Research Software

Highly qualified personnel is the key element to support research software development. Through a funding program from CANARIE[9], Université Laval has a dedicated research software team that will help in the development and integration of software tools to facilitate good practices in research data and to democratize the use of digital research infrastructures.

In a future DRI, NDRIO should have a program to pursue the creation and support of research software teams to provide local support to researchers. NDRIO should also foster a research software community that brings together institutional teams on the national level. This community will promote best practices in software development, encourage software reuse and help in bringing innovative software from teams at the scale of the national platform.

---

[3] https://www.openstoragenetwork.org

[4] http://www.theglobalresearchplatform.net

[5] https://fabric-testbed.net

[6] https://www.data.cam.ac.uk/intro-data-champions

[7] https://www.go-fair.org/fair-principles/

[8] https://www.ic.gc.ca/eic/site/063.nsf/eng/h_97992.html

[9]https://www.canarie.ca/to-equip-canadian-researchers-with-powerful-software-tools-and-expertise-canarie-awards-3-6 m-to-six-institutions-to-fund-local-research-software-development-teams/

How do you see NDRIO's role in addressing current gaps in the national DRI ecosystem?

Even though the research campuses are an important stakeholder in the national DRI ecosystem, an important gap exists that needs to be bridged. There is no program in support of connecting the research on campus to the national DRI.

The NSF Office of Advanced Cyberinfrastructure (OAC) identified this gap back in 2012 and launched the NSF Campus Cyberinfrastructure program (CC*) to invest in coordinated campus-level networking and cyberinfrastructure improvements, driven by science requirements. (https://www.nsf.gov/pubs/2021/nsf21528/nsf21528.htm). A common theme of the CC* program is the importance of the partnership among the campus-level cyberinfrastructure experts, including the campus IT. Indeed, the proposals are evaluated on the strength of institutional partnerships since they are expected to play a central role in developing and implementing the network and data infrastructure upgrades.

The NSF CC* program should be used as a guide to establishing a similar program in Canada. This would have a direct and positive impact for the researchers on campus and would be an enabler for data-intensive research. Such a program would require collaboration with all stakeholders in the DRI ecosystem. NDRIO and CANARIE could lead the initiative in concert with the NREN. The campus CIO and VPR would be the point of contact for the campus.

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

We need a more agile structure that responds to all types of researchers from different domains, particularly the social sciences.

Moreover, we have some research teams that are massive users of the infrastructure. They must be supported in their endeavours but not at the expense of all others. There should be a process to make them more accountable. The excess computing resources necessary for their program should be supported by their university or by providing part of the additional resources (through CFI request or other means of funding) to be integrated into the NDRIO infrastructure. This mechanism would be more equitable to all.

A buy-in from everyone involved is essential and a mechanism, possibly through the researchers' council, is the best process to connect to the NDRIO executive team directly. This immediate contact should not flow through the Board and strengthen the link with the community. The previous structure was fairly dysfunctional.

The prior model of a federated partnership (with regional partners) was not very responsive to the researchers' needs, had difficulties speaking with one voice to make adequate representation to government, and was not directly accountable and ultimately fomented divisions. The model should be completely reassessed, and it appears that this is where NDRIO is heading in that direction at present.

Research is primarily university-based in Canada, and the human resources, the most critical key to success, are within universities. The central decision-making process must remain within NDRIO to benefit all Canadians. Moreover, the consultation process should be a continuous one, and a formal mechanism to receive suggestions should be put in place.

Speed prevails, and seizing opportunities is crucial in the race for advanced research computing and software development. We cannot be competitive without this resource. NDRIO should invest as soon as possible in additional infrastructure to compensate for the lack of resources and delays. The ISED extension announcement more than 18 months ago ($50M provided) and fund provisions are still not fully operational despite the requirement that the resources should be made available without delays. The community urgently needs additional computing resources to meet this dire need.

We suggest making agreements with US entities (National centers such as Argonne National Laboratory, for example) primarily on the R&D associated with optimizing supercomputer infrastructures and service delivery. We definitely should collaborate with for-profit corporations in the field for both cloud management and networking solutions. Some of these (AWS/Google/Microsoft) already support researchers individually. We are essential partners for training HQPs and their future employees. It is reasonable for them to contribute to the process. These collaborative agreements would be best handled at the National level (NDRIO) to gain leverage.

With respect to the upcoming project proposals, they should be aimed at large national projects with industries and healthcare. NDRIO would be a natural partner in bringing digital health care to the whole country, including remote regions. Investing and partnering with satellite internet providers would bring the national DRI in all regions, including indigenous groups, within our community. This is simply a question of equity and inclusion.

Conclusions

The DRI ecosystem encompasses Advanced Research Computing, Research Software, Research Data Management, Networking, and Support and Training. Although this document aims to provide useful feedback to NDRIO, we understand that some points also fall under the leadership of CANARIE and NRENs. An integrated, agile and sustainable DRI ecosystem in Canada must incorporate all DRI stakeholders: NDRIO, CANARIE, NRENs and research campuses.

Université Laval is fully engaged in supporting DRI for researchers, and has made important investments in recent years:

Data centers

> A new on-campus tier-3 datacenter inaugurated in Sept 2019 (Centre de Valorisation des Données), increasing the campus hosting capacity in energy-efficient data centers providing high availability;

VALERIA

> A data-science platform providing researchers with the tools to collect, store, process, analyze and share data in a secure platform (https://valeria.science). Today, VALERIA offers a secure and sustainable campus DRI platform to support researchers with sensitive data (e.g. health), store and process data from scientific instruments, and the capacity for near-real-time data processing and streaming.

VALERIA was designed to facilitate moving data and workloads with Compute Canada systems. CVMFS is used to provide a common set of software and module environments. A campus DMZ provides dedicated data transfer nodes (DTN) using Globus for data movement. VALERIA and the campus DMZ are built on a new networking infrastructure dedicated to research, offering a high-speed data path to RISQ and the R&E networks[10].

We have a long history of constructive collaboration with Compute Canada to support the national DRI and demonstrated that innovation accomplished locally has a global impact on the research community. We are committed to continue on this path and provide additional value to the Canadian research community. Université Laval's experience on implementing a campus DRI by the campus IT in partnership with campus ARC experts and NREN can be used as a stepping-stone towards building an integrated DRI ecosystem.

---

[10]https://canheit-arc2019.exordo.com/files/papers/198/presentation_files/1/CANHEIT2019-Bridging-Campus_National_ARC.pdf