UNIVERSITY OF ALBERTA

OFFICE OF THE VICE-PRESIDENT (RESEARCH AND INNOVATION)

2-51 South Academic Building (SAB)
Edmonton, Alberta, Canada T6G 2G7
Tel: 780.492.5353
Fax: 780.492.3189
www.ualberta.ca/research.

University of Alberta response to NDRIO call for CDN DRI needs assessment white papers

https://engagedri.ca/ndrio-call-for-white-papers-on-canadas-future-dri-ecosystem/

9 December 2020

Contact: Walter Dixon, Interim Vice-President (Research and Innovation); walter.dixon@ualberta.ca

Co-Contributors: Scott Delinger, Director of Advanced Research Computing; James Doiron, Academic Director of the Alberta Research Data Centre & RDM Services Coordinator

Canada’s national digital research environment is undergoing significant change. Its stakeholders must retain and bolster existing people and resources, while both addressing these challenges and supporting ongoing digital research of all types, from individual researchers’ projects through participation in international research collaborations and future research including large-scale research like the Square Kilometer Array telescope project. To put the bottom line up front, Canada’s current DRI provisions, although of good quality, are in short supply. As important as increasing the scale of resources is, it is even more important to market DRI for researchers’ awareness and to train and support researchers to make effective and efficient use of all available DRI resources. Appropriate support for the highly qualified personnel involved in delivering and supporting these resources to researchers is also critical, including financial and positional stability as well as training to remain current with new developments in the ARC, RDM, and research software fields.

Current DRI provisions that must be carefully carried across to the new organization:
High-performance computing clusters and high-throughput computing environments with storage are critical DRI for computationally intense research. Much of Canada’s current ARC provision is well-designed to support a variety of research fields and the scales of demand from those fields; the Canadian ARC community has been responding to these needs for decades. For example, Canadian innovation in HPC services include modular software packages available via CVMFS (CERN Virtual Machine file system); a method of software provision that European ARC providers adopted because of its efficiency. For some research projects, the current ARC resources are used as an integral part of a larger project and so coordination with other research elements is an important part of planning a research workflow, and resource limitations can inhibit project progress. For other research projects, the entire project may occur on DRI, requiring digital research workflow design to ensure research team members know what is complete and what remains to be done. Resource limitations can slow or stymie progress altogether. ARC support, from orientation and training to support tickets, data storage, data visualization, and transfer to archives involves very specialized knowledge of large-scale specialized systems and software that may have only a handful of users across the world. Not only must these highly qualified personnel (HQP) have deep knowledge, they must maintain current knowledge of developments in ARC so as to anticipate needs and suggest platform enhancements and replacements as they become available.

In addition to HPC and HTC clusters, cloud computing resources suit many digital research needs. For modest needs beyond the common desktop computer, cloud computing resources are ideal for ensuring thousands of researchers have computational capabilities on demand. Both computational tasks as well as web-based applications, portals, and platforms are supported by these cloud services. Annual evaluation of cost of operations for cloud within NDRIO and in commercial cloud platforms is important to ensure value. Many cloud vendors make adoption easy, but data egress costs can make commercial cloud adoption expensive in the end. Building cloud services atop migrated cloud environments such as Arbutus would facilitate common cloud-based software needs, such as web-based databases with other tools required by the portal, platform, or project. Recent testing of CVMFS-mounted filesystems further leverage this Canadian innovation in software provision; cloud users need only run a script to mount hundreds of software packages on their cloud-based environments.

Research Data Management Planning (RDM) is a critical element in 21st century research. The current Portage DMP Assistant, endorsed by Tri-Council, is more important than ever, given the Canadian funding agencies' desire to increase the efficiency of research and to make results as accessible as possible. Considering data assurance during research projects, as well as appropriate accessibility of results and data for beneficial application and reuse, is an essential part of planning a research project or program at its inception. While tools like the DMP Assistant and exemplar data management plans are useful, RDM experts, resources, and supports are also needed to guide and advise researchers during the development of their data management plans and when specific questions arise during research planning.

Data repositories and discovery platforms that assign persistent identifiers (PIDs) to digital research assets enable Canadian researchers to steward, archive, and curate research data for discovery and reuse, both important aspects of assuring the full value of the research projects' funding is derived by the public. While RDM measures are beginning to be required of researchers, appropriate repositories such as Dataverse and the Federated Research Data Repository (FRDR) need to be instantiated, operated, maintained, and grown according to demand so that the value of the research results can be fully appreciated. Appropriate support from research data managers, data archivists, and metadata experts is crucial for the effective and efficient use of these systems, and IT support for the underlying infrastructure is important as well.

In all of the above cases, HQP are critical for researchers to concentrate on their projects and gain insight and produce results. A graduate student, for instance, should be concentrating on the field, gaining mastery of tools including ARC, data repositories, and discovery platforms; those same students should not be required to become part-time, under-qualified systems administrators just to accomplish their research.

Current gaps in DRI provision:
ARC resources significantly lag behind demand, with allocations meeting only 40% of current demand for CPUs and ~20% of GPUs. A significant issue was the lack of national investment in ARC resources for five years, resulting in the recent investments only replacing the existing resources deployed up to 2012, with no additional supply for increased demand both from existing users and the increased number of users. Research support from HQP in Compute Canada, regional organizations, and local institutions, *primarily from individuals working at all three levels every day*, remains exemplary and coordinated. Supply of hardware resources lags behind other OECD countries' funding and supply significantly, putting Canadian researchers at a disadvantage compared to those in the US, Germany, and even Italy. Support for bursting into commercial cloud providers resources could be a part of ensuring cycles are available, although billing/funding administration is the issue here. Several institutions are currently looking at this, as the resources available nationally are so constrained. While cloud hardware resources are available, cloud services atop that cloud hardware infrastructure lags behind what is available commercially from vendors like Google Cloud Platform and Amazon Web Services and Microsoft Azure. Parity is neither expected nor affordable, but the next step in ARC support on cloud would be the most commonly required services.

Data storage in active research projects and programs, even that supplied only for research with ARC needs as currently supported by Compute Canada, is becoming inadequate as demand increases annually. Storage in both file and object formats needs to be available in larger quantities than currently offered, so that data transfers needn't occur so often to get the data "adjacent" to the computational resources used for the analytical elements of the research projects. In addition, too many researchers install separate copies of commonly used datasets from their research domains, such as ImageNet, sentiment analysis datasets, natural language processing datasets, clinical datasets, and many bioinformatics datasets. Obviously, not every publicly available dataset can be hosted for common use, but when more than one or two groups require the same dataset to be available, having a common provision would save needless duplication and resultant storage waste.

Data management planning gaps include machine-actionable DMPs, versioning support and control of DMPs, and support for development and implementation of discipline- and methodology-specific DMP templates and guidance for use and growth. Portage is working on a DMP repository in order to support the deposition, curation, discovery, and citability of DMPs, but this is currently in initial stages and needs coordinated and sustained effort to facilitate its adoption as a core RDM tool like the DMP Assistant has been to date. Versioning support in DMP Assistant would enable trackable changes to DMPs for large or longitudinal research projects or programs. Training around data management planning will need revision, support, and delivery as DMP integration into DRI platforms matures.

Associated with Research Data Management efforts are strategies for Persistent Identifiers (PID) to enhance reproducibility of research and citability of results. This need for PID strategies is interwoven with the need for Data Deposit and Archives also emphasized in Tri-Council requirements for DPs. Data Management Plans are only as good as the ability of the researchers to follow through on them, and making data available for reuse requires repositories and/or archives. These facilities need to be identified, characterized for any field-specific needs, sized for both initial needs and planning for scalability, trustworthiness, security, and funding sustainability. Commercial models will be proposed to fill any gaps in archival availability, and as seen in the past with commercial academic publishing, reverting to accessible archives not hidden behind paywalls would be an unfortunate situation that can be foreseen and avoided if approached properly now. Also, long-term preservation must be considered if research funding is to be fully leveraged over time with data reuse. Returning to PIDs/DOIs, Canada has significant national-level data available (eg, StatsCan, CIHI) not supported by PIDs and/or DOIs, limiting the discoverability, accessibility, and citability of this data. Support for extending the range of data with persistent identifiers would enhance the value of that data.

As Canada integrates the delivery of all aspects of DRI, keeping Open Science in mind will be important. The Government of Canada released a Roadmap for Open Science in February 2020, and as an element of open scholarship, the DRI environment will need to be designed to align with the principles and guidelines of the Roadmap as integration proceeds. The principles included in the Roadmap emphasize the need for non-commercial archives so that all researchers have access to published data.

While ARC support and documentation is available in French, bilingual support for other elements of DRI may need improvement: services from Portage, StatsCan data access, etc. The bilingual support and documentation available in ARC will, of course, need to be maintained.

Currently used DRI tools:
University of Alberta researchers make use of a wide range of national resources like the Portage DMP Assistant; ARC resources like the Béluga, Cedar, Graham, and Niagara clusters as well as cloud resources like Arbutus and those in the clusters; and digital repositories like Dataverse and the Federated Research Data Repository (FRDR). Other resources include, but are not limited to, RedCap servers, the Library's Education and Research Archive (ERA) and Digital Scholarship Centre (DSC), data access supports such as those available through the Statistics Canada Data Liberation Initiative (DLI) as well as the Canadian Research Data Centre Network (CRDCN), GPU-rich servers and workstations for development and small-scale production computation, and local visualization resources.

DRI challenges include:
The paramount need is researcher education so that researchers know what resources are available, what training is available for these resources, and what support exists for all of the DRI elements. Perfect but unknown services and support would be a waste. Marketing the services is a shared responsibility of communications experts in NDRIO and DRI people involved at the research institutions. Given awareness, providing training on ARC, RDM, and research software provisions is crucial. Regardless of federal funding levels for DRI, there will never

be a surplus of computational resources; training ensures that the funded resources are effectively and efficiently used to best serve the research goals of all research projects and programs and teams.

In summary, much of the DRI provision in Canada to academic researchers is of good quality, if insufficient in quantity. More resources are needed, but most importantly awareness through marketing, bilingual support and training through well-supported highly qualified personnel, and sustainable growth through consistent funding and good management are required to consolidate the disparate DRI elements into a sustainable and useful DRI ecosystem. As new researchers adopt DRI for their research, and DRI intensity within a group increases, the basics of ARC resources, DRM support and research data archives, and research software development and reuse are the expensive elements of a sustainable Canadian DRI ecosystem. Awareness, training, and support are far less expensive than the infrastructure, but even more important for best use of resources.