Response to Industry Canada Consultation

Developing a Digital Research Infrastructure Strategy

September 14, 2015

CYBERA
Calgary Office: Suite 200, 3512 - 33 St NW, Calgary, AB T2N 2A6 T: 403-210-5333
Edmonton Office: 3-43, Computing Science Centre, University of Alberta, Edmonton, AB T6G 2E8
@cybera info@cybera.ca www.cybera.ca

Introduction

Thank you for the opportunity to provide input into Industry Canada's consultation for developing a digital research infrastructure strategy. We applaud the recognition to shape and evolve the digital research system to meet the needs of the Canadian research community.

Cybera is Alberta's National Research and Education Network partner organization, and has been advancing digital infrastructure since the early days of the internet. In addition to managing Alberta's research and education network, Cybera is home to some of the world's top cloud computing experts, and offers cloud infrastructure, data storage, and advanced network solutions to Alberta's public and education sectors. In addition to our provincial perspective, Cybera supports national advancements in knowledge and innovation infrastructure. We strongly endorse CANARIE's submission on this consultation

In developing our response, Cybera consulted with the offices of the Vice Presidents of Research at the three largest research institutions in Alberta: the Universities of Alberta, Calgary and Lethbridge. We felt it important to incorporate the thoughts and concerns of the local research community, where appropriate, throughout our submission. This response is however the opinion of Cybera and does not represent a direct input from the institutions.

Question 1: How can Digital Research Infrastructure be realistically transformed, strengthened and supported over the next five years?

Governments worldwide are investing in digital infrastructure as the foundation for innovation, economic growth and market diversification. Canada is at a critical technological threshold when it comes to the future of Digital Research Infrastructure (DRI). The development of a pan-Canadian framework that aligns both national and provincial digital infrastructure policies needs to be an immediate priority for Canada if a strategy is to take shape over the next five years.

The DRI system is constantly evolving and changing as new technologies come online. This changing landscape is difficult to predict long-term but necessary to plan for, as it supports world-class research and drives innovation. Canada needs to take stock of existing digital infrastructure, leverage what is currently available, and invest in infrastructure that will strengthen and expand the DRI system in an efficient, sustainable and distributed way.

It is important to emphasize the foundational role that networks play in providing connectivity for science and technology research. High-speed and large capacity networks are what carry the increasingly larger data sets used by researchers, students and scientists to develop new knowledge, understanding, and business and commercial applications, and should be considered a critical component of the DRI landscape.

But the biggest and most important step in building a strong DRI foundation moving forward is sustainable and consistent funding mechanisms that equally support all elements of DRI: from networks to compute and data. A coordinated and complementary approach to the management and policy development of DRI is also required. There is strong support from all organizations to build and move forward with a clear and organized digital research infrastructure framework. While this coordination largely sits with Canada's federal organizations, provincial involvement will be instrumental in strengthening acceptance and local participation.

Beyond networks, compute and data, the complexities of the DRI system extends to the game-changing nature of big data. This includes complex software systems, the collection, management and archival of massive data sets, and the analytics expertise to bring it all together. Over the next five years, we need to build the network, ubiquitous compute, and the data repositories necessary to enable open data sharing, collaborative research, and data re-use and discovery. Programs such as CANARIE's Research Software program — which funds the development of tools, software and platforms that aid researchers — should continue to be fostered, as should all programs that provide technology resources to post-secondary institutions.

Question 2: What are the biggest challenges limiting the effectiveness of the Digital Research Infrastructure ecosystem? What opportunities are there to more efficiently deploy the human, technical and financial resources currently being devoted to Digital Research Infrastructure? How, and in what priority, should they be addressed?

As noted in our response to question 1, the biggest challenge affecting Canada's digital research infrastructure ecosystem is the lack of predictable, stable, and long-term operational funding. This severely limits the competitiveness of Canadian researchers, especially in fields such as genetics, biocomputing, and physics for example. This lack of funding also curbs Canada's ability to attract top talent in these fields. Longer-term funding horizons would allow Industry Canada, together with Canadian universities, to facilitate inter-organizational collaboration, stage infrastructure investments across multiple sites, and better meet the DRI needs of the research community.

In the past, institutions and individual researchers were required to apply for digital infrastructure funding or make capital expenditures on an ad-hoc basis. This funding model is unable to accommodate for rapid innovation in the field of HPC, leading to inefficient boom and bust cycles wherein compute-intensive projects may be over-provisioned at the outset, but under-provisioned by the project end-date. As a result, within three years of project completion, the infrastructure often faces obsolescence. Unfortunately, current funding mechanisms continue to force institutions to treat digital infrastructure as an investment or capital expenditure instead of a longer-term, recurring operational expenditure. An effective DRI strategy should incorporate

forecasts for changing compute, data storage, and continuous maintenance / upgrades of the infrastructure

The next challenge is orchestrating a coordinated approach to planning for digital research infrastructure. The current approach is often siloed into three parts: compute, storage, and network. Compute power is required for calculations or data-intensive research activities, including modelling and analysis of large datasets. Storage is required for the short and long term safe-keeping of this research data. Robust networks are required to enable the high speed transfer of research data. It is important that these three components of cyberinfrastructure, and the funded agencies responsible for them, are effectively coordinated to more efficiently and economically satisfy the requirements of Canadian researchers.

Question 3. What do you see as the biggest challenges to effective data management and the development of data standards in Canada? What could be done to promote a more rigorous and coordinated data management system that supports research excellence and maximizes the benefits generated by our investments?

The biggest challenges to effective data management are the lack of comprehensive data management standards and facilities for researchers to make data public and accessible. Furthermore, it would helpful if data sharing was mandated for publicly funded projects. The current system does not incentivize researchers to make their data publicly available and it is often the researchers themselves that are left to establish their own respective discipline-specific databases. Moreover, guidance and rules today often come from journals that publish researchers' studies, but requirements and monitoring for compliance varies greatly from one publisher to another. As a result, data is frequently hidden and siloed rather than openly distributed.

Cybera supports the Tri-Council Plus' recommendation[1] to establish a culture of data stewardship. We recommend the development of policies that require researchers to plan for the effective management of data developed with the funding received. While developing standards and policies, emphasis should be placed on data reuse and secondary uses of research data. Currently, funding decisions are often based on the number of publications a researcher has and the impact factors of the journals they publish in. Additional metrics should be developed to measure research influence, including adherence of research data to data management standards and data reuse or secondary use. Further consideration should be given to supporting mechanisms surrounding the data. Whenever possible, tools (such as the code used to analyze project data) should be archived and made available along with the data, in order to lower the barrier of entry for other researchers to work with the data and expand its utility.

[1] http://www.sshrc-crsh.gc.ca/about-au_sujet/publications/digital_scholarship_consultation_e.pdf accessed 14 September 2015.

Another challenge to effective research data management is a lack of infrastructure to store and archive data in a standardized way. Trusted digital repositories (TDRs) would greatly assist this need by offering researchers a familiar storage environment that also meet standards and are optimized for data reuse and discoverability. TDRs may also facilitate innovation by providing access to the commercial sector, which would help data flow from research to the market — or from campus to commerce. This can be seen in the example of meteorological data. By making public meteorological data openly accessible, an entire industry evolved around it — from TV and radio networks dedicated to weather, to value-added services on top of applications, such as Google maps.

Research Data Canada (RDC) could play a key role in facilitating more rigorous data management practices. CANARIE and Compute Canada represent the network and compute components of digital research infrastructure, respectively, but there is currently no organization responsible for the data management component. RDC could fulfill this role and interface with universities, stakeholders, and, in particular, libraries, which are well set up to manage the TDRs available to researchers. As pointed out in Cybera's *State of Alberta Digital Infrastructure Report*:

> Academic libraries employ some of the leading experts in research data management (RDM) in Canada, particularly in the area of digital preservation and curation. Libraries maintain local services that support research data management and participate internationally in the development of standards internationally. They are also developing collaborative services around RDM which will enable cost efficiencies and national coverage. In March 2014, the Canadian Association of Research Libraries (CARL) initiated Project ARC. The project has two key goals: to implement a centre of expertise for the management of research data in Canada that will support researchers across the country; and to pilot a national preservation service for research data that will evolve and expand over time. The latter activity is being done in collaboration with Compute Canada, CANARIE, and key domain repositories under the coordination of Research Data Canada.

This foundation established through CARL and RDC should be leveraged and further built upon in order to coordinate and create a national research data management strategy and guidelines that will promote responsible data stewardship.

Question 4: What is the current capacity within post-secondary institutions to support research data curation?

According to Cybera's inquiries, the current capacity of post-secondary institutions in Alberta to support research data curation is currently limited. As far as we are aware, there is little to no capacity for future data curation or provisions for data management plans, although key expertise is available at the University of Alberta. It is our position that leadership and coordination in this area are needed. To quickly move forward, this may best be managed and coordinated through Research Data Canada, in concert with institutional libraries as they continue to pursue the digitization of resources and the curation of digitally sourced data.

Question 5: What are the biggest strengths of the Digital Research Infrastructure ecosystem? How will these strengths be affected and prioritized by a transformation of Digital Research Infrastructure in Canada?

The biggest strength in Canada's DRI ecosystem is the willingness and appetite to develop a strategic vision. Canada has a strong research community that is ready to collaborate, build, support and nurture a comprehensive DRI system. Previous infrastructure investments by the Canadian Foundation for Innovation (CFI) have built a solid foundation for compute across the country that is advanced, but due to a lack of resources and coordination often fall short of, world-class benchmarks. A shift toward predictable, stable funding in digital research infrastructure funding will serve to buttress CFI and Compute Canada's contributions to the ecosystem.

Canadian researchers at major institutions also enjoy a strong core of connectivity through CANARIE and the regional research and education networks. However, as data curation and repositories become higher priorities in future comprehensive digital research infrastructure strategies, the required improvements in connectivity to the smaller institutions will become necessary.

Question 6: What is the role of the private sector in supporting a strong Digital Research Infrastructure ecosystem in Canada?

The private sector should supply services and/or infrastructure that support the digital research infrastructure strategy, where it is capable of providing superior services and it is economical to do so. Such a strategy may, for example, incorporate periodic RFPs for data storage through the overarching body that is responsible for data curation. In the case of data storage, it should not matter where in Canada the data is stored, so long as it is secure, the appropriate access paths are defined, and efficient networks are in place to facilitate access.

It is important, however, to avoid contractual vendor lock-in and to ensure any research data entrusted to the private sector is moveable. The institutions and publicly funded bodies that exist to support researchers should retain some capability to supply services and provision infrastructure to ensure that a viable option to private sector services is available to fulfill such

needs.

Question 7: Do you have any other comments or suggestions to support the development of the Digital Research Infrastructure strategy? (750 words maximum)

The development of a national DRI vision would require establishing a federal oversight committee comprised of key stakeholders that represent the compute (Compute Canada), network (CANARIE), data management (Research Data Canada) and research (VPRs of the largest research institutions) components. This committee would establish and coordinate the DRI vision with provincial and territorial counterparts to ensure effective execution and distribution of the digital resources across Canada. It is imperative that this coordination take place in a timely manner to ensure that Canada does not fall further behind other countries with respect to DRI and innovation. For instance, in 2015, the Conference Board of Canada ranked Canada 9th out of 16 peer countries and earned a "C" with respect to innovation, which it defines as "a process through which economic or social value is extracted from knowledge … to produce new or improved products, services and processes."[2]

An integrated and coordinated DRI strategy will empower Canadian researchers to effectively utilize compute and network capacity as needed, and without uncertainties as to what will be available. The availability and distribution of DRI to researchers in Canada in a sustainable manner will enable them to produce world-class research data and results, foster digital literacy, while also helping retain highly qualified personnel in the country. A national DRI strategy that strategically invests in compute, networks and research data management will also enable Canadian institutions to attract top international talent and keep Canada's innovation pipeline competitive. This requires high-speed and high-capacity networks connecting appropriate and accessible compute, supported by trusted digital repositories.

Longer-term and sustainable funding horizons will create a DRI strategy and ecosystem where researchers are not spending valuable research time or funds on—or competing for—computing resources. In addition, investments in research software, such as through the CANARIE research software program and the CFI Cyberinfrastructure initiative, should be coordinated to prevent duplication and to encourage reuse and sharing of existing software code. A unified vision to properly archive, preserve and maintain code is needed to ensure that this valuable commodity does not remain siloed within individual research programs, but is readily identifiable and distributable.

[2] http://www.conferenceboard.ca/hcp/provincial/innovation.aspx accessed 14 September 2015.