14 September 2015

University of Ontario Institute of Technology Response to Industry Canada Consultation on a Digital Research Strategy

Authors: I. Tamblyn, C. McGregor, and H. de Haan

University of Ontario
Institute of Technology
2000 Simcoe Str. North
Oshawa ON Canada
L1H 7K4

1) *How can DRI be realistically transformed, strengthened and supported over the next five years?*

Strengthening DRI over the next five years will require a combination of directed spending (money), establishment of long-term national priorities together with their supporting procedures, and developing nationally recognized training and certification for users and support staff.

Given the stated time horizon of 5 years, in practice this will mean working with existing institutions on their current roadmaps. For example, Compute Canada has recently announced 4 new national data centres to be rolled out over that time frame. CANARIE (and the corresponding provincial equivalents) have already identified sites for network connectivity upgrades. Supporting these existing initiatives should be a priority.

Strengthening DRI during the next 5 years means ensuring that these activities go as planned, and working on a transition plan to a longer term plan that ensures there is an overarching strategy guiding the priority areas moving forward for Compute Canada and CANARIE (and the corresponding provincial equivalents).

Infrastructure broadly refers to core computing, data, network infrastructure, centre support staff, web based training, available software on core and database management systems, visualization and virtual reality. Strategic approaches across all of these areas are required.
The DRI must support all stages of the innovation pipeline. Different levels of quality of service and support are required through each of these stages.

Technology gets better and less expensive every year. Steady funding for regular upgrades is more efficient than large, rare spending events.

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

A lack of proper personnel training across all strata of the DRI ecosystem significantly impedes the efficiency in the use of DRI. This is felt acutely by researchers who take on undergraduate students (from 1st to 4th year) to work on computation based research projects via work-study programs, summer employment, or upper year thesis projects. The time required to impart a basic skill set such that the student can conduct research can easily end up consuming a great portion of the allotted time. The same problem arises at the graduate level. Although it may not be as extreme in these cases, the impact is even larger as the output of research labs relies on competent graduate students to perform the work. Taking time out of a two year Masters degree to train a student in basic programming severely limits the amount of research that the student is able to complete and thus significantly impacts the productivity of the lab. In fact, this is an issue for personnel in general and not just students.

The lack of available training and strategic guidance for local IT personnel, specific to each institution, ultimately limits the possibilities for research initiatives at those institutions. For example, staff that are only competent in administering Microsoft Windows based machines will be unable to assist a large sector of the research community where systems are almost solely Unix/Linux based. Researchers developing, deploying, and utilizing software solutions to solve research problems are hampered by this lack of basic support. Strategic guidance and training for local IT personnel is still lacking. A national platform for developing core competencies is required immediately.

Currently, infrastructure is maintained and deployed by a large variety of institutions and organizations, each with different mandates and priorities. For example, a researcher sitting at her desk in Halifax running CPU intensive tasks on a cluster in Victoria uses multiple networks and multiple hardware platforms that are paid for, and maintained by, a large group of organizations. Within a university, providing all users with data intensive connections is not necessarily a priority. The same is true for maintaining a research “software stack” on local machines.

Finally, the lack of basic computational skills in the primary investigators themselves hampers the effective use of the DRI and can in fact preclude its use entirely. Computation can assist the vast majority of research currently being done in Canada. A lack of computational skills in the project leader can cause available computational solutions to be neglected or to be used ineffectively, if they are known about at all. This extends all the way from basic skills, such as accessing a remote machine, to advanced topics such as the use of parallelization and GPU based acceleration to significantly increase the efficiency of computation. In the current model, acquiring such skills is largely done in an self-taught manner on an individual basis with perhaps some instruction from graduate level courses or online resources offered from external sources such as Coursera (Computer Science is a possible exception to this).

To address this issue, ultimately programming and computational skills need to become part of a national vision for the education and training of the workforce. This can include specific programs such as Microsoft Excel, but the focus should be placed on a more fundamental skill set such as programming basics (reading/writing data files, "for" loops, etc.) and operating within a Unix/Linux environment at the command line level. While it would be best if these goals were first addressed at the secondary and primary education levels, national programs for post-secondary training would address the immediate need and could serve as a catalyst for further program development. To be precise, the establishment of a national program of certificates in specific skill areas (e.g. basic Linux commands, basic programming (in Python or C++), advanced programming) would provide a consistent and unified method for developing these skills and provide documentation that the student's accomplishment. The latter point provides incentive for the student and a useful tool for researchers in evaluating potential lab members. The development of such programs across a variety of topics and skill levels would also begin to address the issue of non-student personnel being under-trained by providing a clear and high quality route to bolster their skills and thus help increase the efficiency of researchers at their respective institutions. Taking the obvious route of online training, there would seem to be no significant barriers to the immediate development and rapid implementation of such programs. This should begin at the basic level such that different certificate programs can feed into more advanced programs.

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?

Some of the key challenges in creating an effective data management policy and data management systems are in the areas of data ownership, data sharing, and data retention. Data ownership is a particular challenge in any research where data is collected from, or in relation to humans. Currently it is not clear if this data should be owned by the individual or the organization performing the research. A recent public opinion survey run in both Australia and Canada notes that the public have an increasing interest in being engaged with research using their data. New models are required that engage participants for research involving humans. A popular view for research data is that source data is owned by the participant (for example, electrocardiogram tracings) but that data derived from it using the researchers' techniques should be owned by the researchers. The issues of data ownership exist in other domains in varying degrees and a policy for an approach to data ownership is required.

There is a need for policies and procedures to enable data sharing. There has been increasing concern in the research community about the repeatability of research. Likewise, there is a need to ensure that the same datasets are used to determine whether new approaches to solving a problem are more effective than prior approaches. In the current model, there is much more incentive against sharing data than there is for sharing data. To address this, the recent Tri-

Council policy on open access publication is a good model: if you want sharing to take place, require it.

To support new research discoveries in emerging areas, such as the analysis of Big Data, it is not always clear what data should be retained and hence policies in the area of data retention are required. As a result, usually the request is to store all data collected. While the cost of storage has been falling exponentially, the rate of data production has been increasing and there can be significant expense (hardware and time) associated with needless data storage. Storing unnecessary data makes the relevant information more challenging to access and use in the future (needle in the haystack).

In addition, there has been limited research on what granularity of data and what retention term is the most effective to balance research integrity, research efficiency, and effective storage. Such issues can be a disincentive to collecting data in the first place.

4) What is the current capacity within post−secondary institutions to support research data curation?

The recent trend in Canada has been a move away from institutional computing to national platforms (e.g. Compute Canada). It is not clear whether or not individual institutions should be responsible for research data curation (or at least this would seem to go against the philosophy of the past few years).

Where ethics approval is required for the completion of research involving human or animal subjects, statements of data retention intent are required. There is currently a lack of policies and procedures for data curation which incorporate automatic management of data retention and ensuring that data is destroyed at the time that was required by the research ethics approval. File systems where a future delete date can be defined upon file creation or transfer can mitigate this issue.

Overall, policies, procedures and processes are lacking. These should be handled at a national level, rather than requiring individual institutions (between which many principal investigators move) to develop their own specific infrastructure.

5) What are the biggest strengths of the DRI ecosystem? How will these strengths be affected and prioritized by a transformation of DRI in Canada?

One of the biggest strengths in the current DRI ecosystem is the centralization of computation to dedicated data centres (and away from smaller institutional and departmental clusters). This has significantly reduced infrastructure costs (including labour for support) and increased the availability, uptime, and scale of computation that is available to researchers in Canada.

The high-speed network which links institutions to each other and compute resources is another strength. Although this network should be upgraded in the years to come, over the past decade it has performed extremely well and made the centralization possible.

6) What is the role of the private sector in supporting a strong DRI ecosystem in Canada?

The private sector can support DRI by providing consistent feedback on the types of tools and skills that are necessary within their industry. For example, if a company is interested in working in the area of large scale data analysis (using Hadoop, for example), conveying this information to universities can be extremely helpful when designing courses and selecting research tools. Universities can be much more effective at providing a highly trained talent pool when this information is available. The private sector can also support DRI by working with local educational institutions in the development of training and certification materials.

The private sector should also make a point of consulting with local expertise within the academy when making long term technical decisions. Universities are in many ways an under-utilized source of expertise that private companies could make more effective use of. Rather than relying solely on vendors for advice about technology solutions (who obviously have a strong bias and will preferentially promote their own products), private companies should contact local research groups working within the area. This form of digital consultancy can not only reduce costs for private companies, but also strengthen the feedback loop between University curriculum and private sector activity.

One major issue that the private sector faces in Canada is the extremely high price of wireless and wired internet. Many commercial applications which appear feasible within a research environment are impossible in Canada due to the low quality of public internet. Although this may be beyond the scope of this consultation, improving the cost and availability of reliable high-speed internet connections in Canada would have immediate and significant impacts for private industry.

The private sector should be encouraged to make their own software and tools available to academic researchers. This would allow a broader group of students to access and learn how to use their software (and take that familiarity to their future jobs in industry).

7) Do you have any other comments or suggestions to support the development of the DRI strategy?

During the course of our discussions, many ideas were generated. Given the space constraints of this consultation paper, here we list in point form some of the ideas that do not fit into the consultation questions but could help to improve DRI in Canada. For further information on these points, please contact isaac.tamblyn@uoit.ca.

1) Create and maintain a national website pointing to open data sets in the scientific literature

2) Change funding metrics to incentivize open data publication

3) Establish a national cross platform cryptographic standard for research data (e.g. PGP)

4) Fund a national economic simulator (similar to weather forecasting)

5) Incentivize targeted data collection in the national interest (e.g. arctic sovereignty, resource exploration, Maritime intelligence)

6) Provide researchers with free credits for CANSIM data which meet federal priorities

7) Create a Canadian-specific version of "Kaggle" (https://www.kaggle.com). Use data challenges to connect public researchers with private institutions

8) Collect and release "Main Street" business data (which businesses operate where, and for how long). This will enable more accurate economic, business, and policy modelling relevant to Canadian communities.

9) Modernize the Social Policy Simulation Database and Model (SPSDM). Make it available through a web portal.

10) Provide open and real time data sets on Maritime traffic, and commercial aerospace.

11) Collect and make available data from auto emissions testing (huge amounts of data are measured from vehicles during safety inspections with very little of it being used elsewhere).

13) Create and fund faculty sabbatical training programs for non-traditional DR disciplines (digital boot camp).

14) Fund training and certification (DRI) for existing IT workers in universities, colleges, etc. Many staff do not have the expertise required to make use of DRI.

15) Incentivize data sharing by utilities (e.g. water, power, etc).

16) Create data challenge grants (where the objective is to produce a tool or solve a specific problem).

17) Incentivize cities to invest in open tools for data collection, storage, and monitoring.

18) Enforce the use of the Canadian Common CV for all public sector hiring (this will provide much more accurate and detailed labour force data).

19) Ensure “last mile connectivity” to research labs and PI offices. A research specific tool like speedtest.net would allow researchers to confirm that their connection is configured properly.

20) Provide funds for IPV6 switchover and 10 GigE connectivity within research. Outward facing IP addresses should be the standard, not the exception.

21) Produced standardized training for support staff and end users (MOOC style) including web based training for common software tools. Provide students with an on-ramp to practical data analysis approaches. Establish national certificate programs for proficiency.