Towards an Ideal Digital Research Infrastructure (DRI)

Some perspectives from Athabasca University[i] prepared in response to NDRIO's call for discussion papers on Canada's future DRI

Robyn Stobbs, Research Data Management Librarian, Library & Scholarly Resources, Athabasca University, rstobbs@athabascau.ca

Dr. Vivekanandan Kumar, Professor and Associate Dean, Research and Innovation, Faculty of Science and Technology, Athabasca University, vive@athabascau.ca

Introduction

For this brief discussion paper, we chose to highlight perspectives on digital research infrastructure (DRI) through vignettes. Each vignette is based on a real-world example experienced by one of the authors who is affiliated with Athabasca University. The vignettes are intended to give examples of current DRI being used, followed by a brief discussion of how this DRI could be further supported to speak to the "ideal" state of DRI.

We have focused not only on national technical infrastructure, but also on the potential for national infrastructure to include recommendations and guidance that can be adapted to various institutional contexts. Such guidance is important to consider in moving towards compatibility and the potential for integration between institutional, regional, and national DRI.

Our vignettes take place at Athabasca University (AU), a distance education institution and research university that is in Alberta but serves students across the country and internationally. AU is a smaller research university (9,717 FLE; ~1,000 staff)[ii] and a distance education institution, and it is home to renowned research on distance education. As a part of its strategic plan for research, AU has identified developing a "reputation for open scholarship, data, and scholarly achievements" as a strategic goal.[iii] The vignettes we have chosen to highlight here are rooted in this context.

Vignette: Research and/on Data Privacy

The knowledge and infrastructure necessary for proper data management is important not only to support effective and ethical approaches to research, but also as a subject of research. For example, the importance of data privacy and proper handling of research data is a part of the data lifecycle, but data privacy and the creation of tools to support data privacy in multiple realms, such as that of distance education, is also an area of research interest. Recent controversies involving privacy concerns and the use of proctoring services for remote examinations have been highlighted in the media.[iv]

AU researchers have an interest in this area and the creation of tools to support the use of data collected by distance institutions to better the experiences of their students while also protecting the privacy of those students.

AU is a distance institution that uses a variety of processes to proctor exams. Some of these processes use tools to record student data during proctored exams. Most of this dataset is about the exam but it may include some unwarranted data. These datasets are then sent to a centralized server for processing. That is, the required data as well as the unwarranted data are transmitted outward, out of the control of the student who generated them.

AU researchers are exploring ways to block this outward transmission completely thus ensuring that the data stays in the control of students. That is, instead of the outward transmission of data, they explore ways for the inward transmission of the algorithm. Thus, students retain control of their data while availing the proctoring functionality that the algorithms provide. Such an algorithm would evaluate whether a student's exam should be reviewed for misconduct. Only in the event of a misconduct check by a third party, say the Ombud's office, the data needs secure outward transmission. Transmission of algorithms for specific functionalities can address privacy issues in many data-sensitive situations.

Such a reverse-transmission technique can also be used to empower students to formatively monitor the growth of their competence, on their own, without sharing data with anyone else.

Other datasets collected by the institution (such as data from individual courses on access, submissions, and grades) could be accessed through a similar mechanism, one that allows queries to be run on the data without revealing individually identifying data points.

At AU, this data is currently distributed due to the nature of the computing infrastructure, but as researchers and instructors gain access to it and use it for research on educational practices, the release of an open dataset would be ideal.

Ideal DRI

This example demonstrates a need for a computational mechanism that empowers data privacy for end users (i.e., students in distance courses) while also enabling research to improve educational practices.

Increased support for the development of such mechanisms and for curation and preservation of existing data would benefit the research and teaching community. Datasets such as those compiled for AU courses could have value for broader communities (e.g., learning analytics data to examine the effectiveness of distance delivery methods); however, it is currently difficult to curate that data and ensure appropriate privacy to enable such use.

Vignette: Selection of Cloud Computing Resources for Active Data Use

Selecting a proper place to store and access data during active phases of a project is an involved process and knowing which service to choose, and what criteria to use in that selection process, is a crucial decision for all researchers. There are national tools (e.g., Compute Canada, WestGrid), provincial tools (e.g., Cybera[v]), and institutional tools. Recently, AU shifted its operational computing infrastructure to the Amazon Web Services (AWS) cloud. AU is starting to explore AWS as a research space, its appropriateness for storing research data, and accompanying data security considerations.

Knowing which resource is best suited for a project can help with the preparation of project proposals and resource plans. For example, Compute Canada's HPC is meant for computationally-intensive batch processing jobs such as the training of a deep learning model and is not meant for interactive, student-facing services. Compute Canada's Cloud resources, Cybera's Cloud resources (RAC) and West Grid's Cloud resources (Arbutus) are fine for student-facing services.

Ideal DRI

Researchers will require infrastructure training and infrastructure procurement training to understand the characteristics of these research platforms and map their research-specific infrastructure requirements to these platforms. Researchers will also require training in NDRIO-supported DRI tools, along with case studies of use by the community. Such training and sharing of cases should be promoted as inter-institutional collaboration.

Vignette: Open Scholarship and Publishing Research Outputs

Increasingly, researchers are required to make their datasets (and methods of data collection and preparation) readily accessible. For example, the forthcoming Tri-Agency research data management policy[vi] will require that projects funded by them have their data deposited and made available (when ethically and legally possible), and journals in some disciplines are beginning to require data sets be made accessible as a condition of publication. Such practices are a positive step in the support of open scholarship. National DRI and a network of expert guidance on the sharing of data, research designs, analytical models, and research protocols would be a significant benefit to researchers at a variety of institutions (with varying levels of internal support).

AU does not (yet) have an institutional data repository, and as such, researchers have been putting together solutions for themselves (e.g., Kumar & Boulanger (2020) used the open science framework[vii] to meet a publisher's requirements). National DRI and guidance (such as resources produced by the Portage Network) will assist institutions like AU in building greater support for open scholarship.

Ideal DRI

Continued support for endeavours such as the Portage Network (DMP Assistant, published RDM guidance),[viii] Scholars Portal Dataverse,[ix] and the Federated Research Data Repository[x] alongside mechanisms for integration with initiatives of regional networks, such as those of the Council of Prairie and Pacific Libraries (COPPUL),[xi] would be ideal to assist research institutions, like AU, that are in the process of developing further institutional support and access to DRI.

As a part of planning for increased research data management (RDM) support at AU, we have been reviewing and relying on these resources and are hoping to further develop institutional implementation of them.

Vignette: RDM Support and DMPs

AU, as an institution, is working on providing support for researchers related to RDM. However, there is a lack of institutionally based guidance in this area. Dr. Kumar's experience in the previous vignette is an example of a gap researchers at AU are currently experiencing.

As the new research data management librarian, Ms. Stobbs has begun collaborating with the Research Centre at AU towards developing RDM resources and support. In these early conversations, resources developed by the Portage Network have been invaluable. They have been used to demonstrate sample data management plans (DMPs) and the potential for an institutional implementation of DMP assistant, and they have served as effective starting points for discussion of tangible and immediate ways to connect AU researchers with RDM support as well as providing starting points as we consider the structure for future supports. These tools have led to specific conversations about avenues for incorporating RDM into existing grant and ethics application processes.

Our last conversation had an interesting pause, where after discussing potential avenues for connecting researchers with RDM guidance, we stopped and asked: How do we deal with the DMPs once we have them? Do we have the capacity to evaluate them? This question had arisen in earlier conversations the Research Centre has been involved in, and its recurrence exemplifies an on-going need.

Ideal DRI

As AU team members begin to explore ways to provide RDM support for our researchers, a particular need has been expressed: the need for guidance on evaluating DMPs. Continued support for Portage and other national networks could be targeted toward the development of further RDM resources, particularly around advice and training on DMP evaluation and further exemplar DMPs.

Conclusion

In responding to NDRIO's needs assessment, we chose to present vignettes representative of our current experiences with DRI. In doing so, we aimed to emphasize key gaps and ideal support that could be provided through national DRI. These identified gaps and ways of bridging to an ideal DRI include:

- Development of computational mechanisms to empower data privacy, particularly for learning analytics
- Continued and expanded support for data curation and preservation (e.g., in collecting and de-identifying data collected through online learning systems)
- Expanded support for infrastructure - such as Compute Canada, West Grid, Cybera, and AWS - to increase the number of researchers able to access their resources
- Continued support for national RDM guidance and resources, such as Portage's DMP Assistant and training modules and Scholars Portal's Dataverse, and mechanisms for integration of regional and national infrastructure
- Guidance for evaluating data management plans (e.g., through the development of training and guidance documents or training modules)

[i] The views expressed here are of the individuals listed as authors and are not intended to be taken as representative of all researchers and staff at Athabasca University. However, the context of Athabasca University as a distance institution with distributed infrastructure is important to the needs and suggestions presented herein. It is our intention that our vignettes exemplify key aspects of DRI that could benefit all Canadian researchers, particularly those who like researchers at AU, are located at smaller institutions or institutions earlier in the process of developing their own institutional supports for RDM and digital research.

[ii] Statistics from the AU Institutional Data Analysis Department. Full load equivalent (FLE) last updated July 3, 2020. Staffing as reported in the 2019 orientation presentation.

[iii] Athabasca University. Strategic Research Plan: 2018-2022. URL

[iv] Stewart, B. (Dec. 3, 2020). *Online exam monitoring can invade privacy and erode trust in universities*. The Conversation. URL

[v] Cybera is an Alberta-based not-for-profit organization that provides computing resources. More details are available on their website.

[vi] Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, & Social Sciences and Humanities Research Council of Canada. (2018, May 25). DRAFT Tri-Agency research data

management policy for consultation. Government of Canada; Innovation, Science and Economic Development Canada. URL

[vii] Kumar, V., & Boulanger, D. (2020, October 6). Frontiers In Education 2020 - Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value. URL

[viii] Portage Network URL

[ix] Scholars Portal Dataverse URL

[x] Federated Research Repository URL

[xi] For example, the COPPUL Digital Stewardship Network offers opt-in services including WestVault for preservation storage, Dataverse (with Scholars Portal), and Archivematica-as-a-Service). See their website. AU is exploring options for developing a data repository.