A PERSPECTIVE ON A CANADIAN DIGITAL RESEARCH INFRASTRUCTURE FROM: MCGILL CENTRE FOR INTEGRATIVE NEUROSCIENCE

INTRODUCTION

The McGill Centre for Integrative Neuroscience (MCIN) at the Neuro has been at the forefront of digital research infrastructure (DRI) development in Canada for the past two decades, providing Canada's first comprehensive neuroinformatics platforms for large multi-centre studies and portals for collaborative distributed high-performance computing (HPC). Two of MCIN's main platforms, CBRAIN and LORIS, have been built as open-source projects over the last 20 years. MCIN leads large cyber-platform efforts for several national projects, e.g. a dementia network across 6 provinces, the Canadian Consortium on Neurodegeneration in Aging, CCNA, and a national open data-sharing platform, the Canadian Open Neuroscience Platform, CONP. MCIN is a core partner to major international projects, e.g. our central data coordinating role in the NIH-funded Infant Brain Imaging Study (IBIS). MCIN has built Open Science data repositories, including the BigBrain high-resolution 3D brain model, the Open MNI iEEG Atlas of normal brain activity and the C-BIG biospecimen repository. MCIN's expertise in research computing, data management and research software engineering has maintained a respected Canadian leadership presence at an international data science organizations, for instance within the Global Brain Consortium and the International Neuroinformatics Coordinating Facility (INCF). These initiatives, supported by CIHR, CFI, CFREF, FRQS, Brain Canada and CANARIE, have established MCIN as a platform technology leader. Here, we present MCIN perspective on the evolution of NDRIO.

1. CURRENT ISSUES

1.1 What are the main DRI tools, services and/or resources you currently use in your research?

MCIN's LORIS and CBRAIN platforms provide a comprehensive open-source data management and processing solution for hundreds of collaborators around the globe.

LORIS is a web-based data management platform enabling researchers to collect, curate, and share data. It has three main functions: (i) data repository, (ii) project management of multi-site data collection and sharing, and (iii) querying portal linking to fully automated analysis pipelines. Its main aim is to make research transparent and reproducible, and to alleviate technical challenges in managing and curating multi-modal (imaging, behaviour, genetics, biospecimen) research data. Initially developed as the data platform for the NIH MRI Study of Normal Brain Development Study, LORIS has since grown to support projects in 22 countries. Recently LORIS has been extended to include epigenetic data, EEG data and a laboratory information management system (LIMS) capability that handles biospecimen data (fluids, tissue samples, iPSC) and laboratory workflows. The platform has an open-source architecture that is customizable for new data modalities and features.

CBRAIN is a web-portal to HPC resources, both from 'supercomputing' facilities, e.g. the Compute Canada network, and from cloud-mediated services. CBRAIN currently interacts transparently with 13 computing clusters and nearly 80 remote data sources across Canada as well as around the world at major supercomputing facilities such as the Texas Advanced Computing Center (TACC), Pittsburgh Supercomputing Center (PSC) and SURF-SARA in the Netherlands. CBRAIN presents a powerful orchestration interface to users, removing barriers and abstracting away the complexities of the Compute Canada ARC systems. The user community leveraging the CBRAIN platform continues to grow annually, supporting over a thousand researchers, including 600 from Canada. Most of these portal users do not have the experience normally required to interact with high-performance supercomputers. Importantly, the CBRAIN platform code itself is generic and **not restricted to neuroscience data**. CBRAIN can accommodate any data type and any scriptable tool, whether in life sciences, astronomy, physics, humanities, finance or engineering

NeuroHub: Our users benefit from an unmatched ability to share analysis and visualization tools, computing resources and data online, entirely within the ease and comfort of a web-browser. This has been further enhanced by the development of the CBRAIN API layer so as to enable flexible interfacing with other tools, to marry ease of use with maximum flexibility. This has the immediate benefit of standardizing tools, methods and traceability, ensuring better experimental result reproducibility across dispersed teams. This broader ecosystem is exemplified with the NeuroHub platform, a core facility of McGill's CFREF-funded Healthy Brains, Healthy Lives initiative. NeuroHub integrates data and tools for research in neuroinformatics, genomics, social sciences, neurobiology and more, linking datasets with computational infrastructure. LORIS and CBRAIN are the central components of NeuroHub, along with other open-source tools, e.g. Jupyter Notebooks, DataLad and Zenodo. NeuroHub represents a core element of DRI in service and support to the neuroscience community and beyond.

Hardware: ARC hardware resources are operated both in-house and through Compute Canada:

In-house: MCIN currently develops, operates and monitors on a 24/7 basis; platforms, visualization services, databases, websites and portals, storage, high performance networking and computing servers serving dozens of research projects for hundreds of scientists. In total, MCIN locally hosts >170 host servers, ~1800 virtualized services, about 1 PB of local disk storage, 3.5 PB at Compute Canada, as well as 350 TB of usable tape storage. Past growth indicates that these needs will triple every two years.

Compute Canada: Our allocation for the period 2020-23 includes 429 processor cores, 1950 TB of disk storage and 1950 TB of tape (nearline) storage on the national platform. However, our work is evolving towards higher memory requirements, making traditional HPC nodes less suitable for our needs. This includes machine learning where very large datasets, e.g. UK Biobank, are transformed into formats suitable for machine learning and processed using specialized hardware such as GPUs and FPGAs.

1.2. Do you have access to all DRI tools & resources needed for your research? What is missing?

As described under section 1.1, the MCIN ecosystem comprises a set of tools covering the full life cycle of a research study: data acquisition, validation, storage, dissemination, processing, analysis and publishing. As multi-disciplinary consortia of researchers build larger and increasingly heterogeneous data cohorts, the need for data sharing across platforms intensifies. Despite much progress in findability and accessibility of data, it remains difficult for researchers to discover and examine specialized datasets, in particular where access to data is governed by specific ethical and legal considerations. The CONP was launched to address these gaps in the greater DRI landscape. This platform provides a centralized web portal for sharing of (i) cohort datasets with open metadata descriptors, and (ii) analysis workflows mounted on CBRAIN. Guided by the FAIR principles, the CONP links resources and tools optimized for interoperability and data re-use, with tools for data publishing, data exploration, distributed storage, privacy and access control, version control and rich metadata layering. The CONP has also undertaken a coordinated effort to develop a national ethics framework to ensure that ethics and privacy concerns are well governed while facilitating research access.

1.3 What are your biggest challenges accessing DRI tools, services and/or resources that exist?

Barriers to interoperability across systems and data standards remain a significant factor. Research datasets are often not well described and come in many data formats and database structures. Most DRI tools are developed for a narrow scope of domain-specific use and generally do little to ensure the reproducibility and accessibility of research outputs. Open access to datasets is another limitation in research domains which are otherwise ready to leverage machine learning applications for scalable knowledge discovery. At a more granular level, analysis tools and pipelines are typically constructed for a specific initiative. Their long-term sustainability and generalizability to other research areas is limited, often due to lack of funding and maintenance. These issues are addressed by the FAIR principles and respecting these principles contributes greatly to sustainability of data assets, infrastructure and tools.

The diversity of research data repositories in the current landscape also present challenges to comprehensive RDM coordination. Data sharing platforms generally require data to be moved to a central location, giving rise to duplication and data versioning issues. Distributed datasets may become unreliably available for a variety of transient reasons - network glitches, host platform downtime or configuration errors. Large federated data repositories may contain impoverished definitions of the data being shared and sparse metadata, whereas well-curated centralized repositories may contain too small samples and specialized data definitions which limit re-usability.

The above challenges faced within our community are of high priority to CONP in its mission to reduce barriers to access and interoperability. Such challenges are not unique to neuroscience. Working together to develop standards and interoperability for data discovery, provenance and reproducibility will continue to contribute to progress in those areas and benefit the national research community.

2. FUTURE DRI STATE

2.1 What is your vision for a cohesive Canadian DRI ecosystem that would fulfill your research needs?

A cohesive Canadian DRI ecosystem consists of not only the various best practices involved in data and software management, but also in ensuring that all these components are linked effectively and resourced properly. Many elements of an effective DRI ecosystem are downstream considerations that are not addressed at the outset. Data acquisition logistics are prioritized but subsequent analysis and dissemination details are often overlooked, leading to problems with data sharing and reproducibility.

Strong standardization of research output is critical to future interoperability and collaboration. This lack of standardization has long hindered data-sharing for established techniques, such as EEG, and is a growing concern for scalable research applications in emerging technologies, e.g. wearable sensors.

Low-cost resources available to all researchers for storage, processing, curation with transparent workflows and abstraction of technical barriers would enhance the accessibility of these assets. A major time-cost for researchers is dealing with tedious, varied and confusing institutional IT protocols that are beyond a typical scientist's expertise, such as i) deciphering how to transfer research data between fire-walled hospital centres, or ii) becoming familiar with batching systems for processing pipelines, or iii) understanding intricate and heterogeneous operating systems. Simplifying these IT barriers and freeing researchers to focus on their science, rather than spending time and precious funds on infrastructure and support services, would hugely benefit modern scientific discovery.

Transparency and sustainability as foundational aspects of a Canadian DRI ecosystem would enable the research community to use, create and take best advantage of the evolving DRI landscape. An infrastructure to support the importance of research outputs beyond the end-product publication will be critical for advancing innovation in compute-intensive disciplines. Recognizing the wide range of factors that promote discovery and its clinical, policy or commercial translation (data, software, analytics, open data/tool dissemination, open publishing) will promote sustainable R&D best practices. Promulgating such best practices within the scientific community will undoubtedly foster a more productive research landscape in Canada. A cohesive pan-Canadian DRI ecosystem will include the following elements:

- Data and tools are integrated without duplication for scalability, and with data integrity checks
- Ingestion of data in infrastructure is feasible by community members
- Data governance remains with the original data stewards or providers
- Datasets and processing tools adhere to FAIR principles
- Open formats and standards adopted to promote sustainability/extensibility across workflows

2.2 What DRI tools, services and/or resources you would like to use, or envision using, in the future ?

An important driver of Canada's future DRI ecosystem will be the global movement toward Open Science, the unfettered sharing of research data and analysis tools, that has found expression in many fields already. Accordingly, MCIN is the main hub of the Canadian Open Neuroscience Platform (CONP). CONP has addressed, in the neuroscience context, many DRI issues and developed solutions that can be readily generalized to the wider Canadian research community. Launched in 2018, the CONP provides an infrastructure for the promotion of open-science workflows and the sharing of neuroscience data, both nationally and globally. The CONP is composed of brain researchers (neurology, psychiatry, cognitive neuroscience, neurogenetics) working alongside computer scientists, ethicists and research software developers to build a national ecosystem for Open Neuroscience. The CONP provides tools and training to facilitate fully open or restricted sharing of research objects (data and processing pipelines). The CONP web portal integrates several open-source technologies to provide: i) structured search capabilities for data and software tools, ii) the ability to federate decentralized data storage platforms, iii) the ability to connect to Canada's HPC infrastructure with relative ease or to run computations locally, iv) cross-disciplinary student training, v) data governance/ethics materials, and vi) dissemination of data and tools to the global community.

Data Publishing is also becoming an increasingly influential feature of modern DRI ecosystems. As the need to make larger amounts of data available for aggregation and 'big data' analysis, such as with machine learning, it becomes imperative for researchers to share their data and techniques. Providing them with mechanisms that facilitate sharing and rewarding those that do share will accelerate this evolution. Several groups have created platforms to enable this new form of publishing, such as the CONP-supported initiative Neurolibre. This project uses Jupyter notebooks for seamless handling and processing of data. Typical steps of processing and analyzing data are structured and automated into seamless workflows that can be vetted and approved, as in traditional reviewing of scientific findings but, within NeuroLibre, this process is also applied to tools and data. Neurolibre is already being used to streamline publication to international publishing platforms such as Aperture.

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

Ongoing challenges for DRI utilization will hinge upon reliability, persistence and open access to services and data assets. While privacy concerns and ethics of data sharing are an increasingly central part of the research discourse, nevertheless ever more research datasets will be shareable and accessible for re-use, in a manner compliant with global data governance ethics. Data collections which can be shared under a controlled(registered) access framework will continue to see limitations in the detail and granularity of metadata that can be shared, particularly in medical research disciplines. Competing data standards and interoperability priorities require solutions to bridge different practices across scientific domains, DRI platforms, institutions and jurisdictions. To accommodate these realities, CONP adopts a decentralized architecture that respects the various governance, ethical, and performance models required by data owners. For instance, some data may not be stored outside the province where they were acquired, some institutions require control of physical data storage, and some projects prefer to remain in control of data access. This is possible in CONP, as data can reside anywhere on the internet.

3 HOW TO BRIDGE THE GAP

3.1 What tools, services and/or resources should NDRIO leverage to achieve your desired future state?

As outlined in section 2.2. the CONP has developed many features that would enhance the NDRIO vision for Canada's DRI and its engagement with the global DRI landscape.

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

NDRIO could provide leadership and support through the following actions:

- Foster explicit, competitive career paths and retention for digital HQP working in universities.
- Sustain and grow existing, successful infrastructures and avoid "reinvention of the wheel".
- Promote FAIR principles within funding programs, training and outreach initiatives.
- Advance data sharing efforts that emphasize reproducibility.
- Prioritize interoperability and standardization for research infrastructure and supported outputs.
- Offer accessible and low-cost storage, processing and curation tools.
- Invest in technical skill development across diverse user backgrounds and academic stages.
- Provide training in best practices for data format standardization and technology usage.

3.3 What other suggestions do you have?

Current Canadian engagement with the Research Software Alliance and models similar to the Software Sustainability Institutes in the UK and US could be leveraged to support the evolving DRI ecosystem.

CLOSING REMARKS

We look forward to working with the NDRIO. We are confident that we can meet the challenges ahead and remain ready to partner with NDRIO in creating and delivering a Canadian DRI solution. If you have questions, please contact MCIN Director, Dr. Alan Evans.

Name: Alan Evans, PhD, FRSC. FCAHS
Title: James McGill Professor of Neurology and Psychiatry, McGill University
Email: alan.evans@mcgill.ca ; Phone: 514-398-8926 (office) or 514-984-6919 (cell)