Toward Petabyte Scale Open Neuroscience: UBC Dynamic Brain Circuits Research Excellence Cluster Experience

Timothy H. Murphy[1,2,3,5*], Paul Pavlidis[1,2,3,4], Jeff M. LeDue[1,2,3,5]

1 Dynamic Brain Circuits in Health and Disease Research Excellence Cluster
2 Department of Psychiatry
3 Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z3
4 Michael Smith Laboratories
5 Canadian Neurophotonics Platform; * lead contact

Our research cluster is addressing a global shift towards Open Science and the pressing need within our group for secure data storage throughout the research project lifecycle, from collection to long-term preservation. Increasingly, neuroscience publications rely on large and often complex data sets. To support the conclusions made in research publications, journals and granting agencies are beginning to require ready access to primary and/or processed data. In addition to issues of compliance, data sharing and transparency increases reliability and reproducibility of research findings and promotes collaboration. Many journals, repositories, and funding agencies now require or encourage open data, and several grant agencies now require researchers to outline their data management and sharing plans. Sharing data can also boost citation count and a proven record of open science can positively impact careers of both new and established scientists. **Here we survey the landscape of current terabyte to petabyte quantity storage and offer some recommendations based on our experience and Canadian grant agency data retention and availability requirements.**

Current requirement for funded research: Canadian Tri-Agency Statement of Principles on Digital Data Management: Data should be collected and stored throughout the research project using software and formats that ensure secure storage, and enable preservation of and access to the data well beyond the duration of the research project. **Metadata** All research data should be accompanied by metadata that accord with international and disciplinary best practices to enable future users to access, understand and reuse the data.

Draft policy: The Tri-Agency Research Data Management Policy for Consultation promotes best practices in research data management. "Data Management Plans" states that grant applicants must ensure that proposals submitted to the agencies include methods that represent best practices in research data management. In particular, the creation of data management plans is encouraged by the agencies, and is required by some grants. "Data Deposit", "Grant recipients are required to deposit into a recognized digital repository all digital research data, metadata and code that directly support the research conclusions in journal publications, pre-prints, and other research outputs that arise from agency-supported research. The repository will ensure safe storage, preservation, and curation of the data. The agencies encourage researchers to provide access to the data where ethical, legal, and commercial requirements allow, and in accordance with the standards of their disciplines. Whenever possible, these data, metadata and code should be linked to the publication with a persistent digital identifier."

In neuroscience, current human resting state as well as preclinical animal task and spontaneous activity imaging data sets can easily approach 10 TB per project (Murphy et al. 2020). Automation of preclinical

experiments has increased data volumes and state of the art histology techniques, such as Expansion Microscopy combined with Lattice Light Sheet Imaging, have already produced data sets up to 0.6 PB for a single publication (Gao et al. 2019). Additionally, with cluster member Frangou's V-Brain initiative, the cluster is working to harmonize MRI acquisition parameters creating a database of scans which can easily be combined across patient populations leading to the need to process and store much larger data sets. Currently there are few options for long-term, cost-effective and accessible terabyte to petabyte quantity storage of imaging or other data sets. Our **Dynamic Brain Circuits (DBC) Research Excellence Cluster** (DBC) at UBC's Djavad Mowafaghian Centre for Brain Health has invested a significant amount of time reviewing national and international academic and non-profit providers. Our experience includes: Compute Canada (CC), Scholar's Portal Dataverse, Zenodo, Federated Research Data Repository (FRDR), and Open Science Framework (OSF), as well as local university servers. DBC is composed of researchers across departments and faculties united by their collective pursuit of advancing the study of brain connections and their dynamic changes during development, learning, and disease.

In the broader Canadian environment, open science in neuroscience received a major push with the Brain Canada-funded Canadian Open Neuroscience Platform (CONP). The CONP has addressed, in the neuroscience context, many digital resource infrastructure (DRI) issues and developed solutions that can be readily generalized to the wider Canadian research community. Launched in 2018, the CONP provides infrastructure for the promotion of open-science workflows and the sharing of neuroscience data, both nationally and globally. The CONP is composed of neuroscientists working alongside computer scientists, ethicists and research software developers to build a national ecosystem for open neuroscience. CONP has a distributed data model which largely relies on the availability of data in existing infrastructure. NDRIO can play a major role in the future of the adoption of open neuroscience by contributing mechanisms for secure long-term data storage.

This white paper aims to provide information and recommendations on archiving terabyte and beyond scale data related to research publications.

Of the solutions mentioned above, **only CC, FRDR, and university servers are solutions for terabyte+ data**. In our hands Compute Canada servers, while having capacity, have many limitations. The greatest being a need for a yearly renewal of storage space. This yearly renewal requirement and a "use it or lose it" philosophy may help to weed out those that are not active users, but it also potentially conflicts with the Tri-Agency 5 year data retention requirement from the "Tri-Agency Open Access Policy on Publications". Other limitations include a maximum number of files at 5,000 per user allocation. While, at first glance this may not sound like a roadblock, depending on the organization of data sets, there could be many small files and one is forced to pack or archive the data. Due to hardware constraints the file archives (typically tar format) cannot be larger than 50 GB necessitating the creation of an index and reassembling the data from split archives. An ideal storage mechanism should preserve directory structure and should not require up front archiving of files by researchers. Zipping or compressing files also reduces the ability of users to browse the data which impacts data sharing.

We currently appreciate the data repository FRDR for its capacity and publication supporting features, but realize it has a serious limitation of not being able to anonymously share unpublished data with a

reviewer. Reviewers will need to be able to easily browse submitted data sets to potentially confirm findings using software also deposited by authors. A requirement of FRDR is the use of Globus file transfer for large deposits. Globus, as an upload tool, is useful and powerful in that it verifies transfer integrity, but has a steep learning curve, for one there needs to be better documentation on how to map specific endpoints (i.e. drives).

While FRDR remains in limited production, there are no data volume quotas nor cost recovery. **Cost recovery is indeed an important aspect for sustainability** and the model adopted will greatly impact the rapidity, willingness and capacity of the research community to move toward open science practices and compliance with existing and future Tri-Agency policies. **Here we consider two options: 1. Grant budgeted (researcher pays) 2. Funds are allocated directly to the platform (overhead model)**

Option 1 (researcher pays): It has been suggested that researchers should develop a data management plan (DMP) for each grant and, based on this plan, make a budget line item for long term storage and archiving of the data for their project. At face value, this is a straightforward and reasonable approach. However, it comes with a number of serious drawbacks and DBC would prefer other options such as the overhead model below.

Drawback 1: Variable Computing Proficiency: Crucial to its success as a highly collaborative discipline, Neuroscience includes diverse researchers coming from many backgrounds ranging from physiotherapy and medicine to physics and engineering. This means that comfort with computing varies widely across labs. Without consultation, DMPs may be inaccurate or rest on false assumptions about the scale and availability of resources.

Drawback 2: No/limited accountability: The funds granted for data storage/archiving are not directly tied to these efforts. Researchers may not reserve the funds or sufficient funds for the intended purpose.

Drawback 3: May limit adoption of new techniques: New techniques are consistently more data intensive than existing ones. Situations may arise where researchers would not have a sufficient contingency to pay for the extra data produced during a project should they adopt a new technique. Choosing not to adopt a new technique reduces the impact and insight gained from the research and represents a waste of public resources.

Option 2 (overhead model): As an alternative to Option 1 (researcher pays), DBC proposes a model in which DMPs are reviewed as part of funding decisions and they can be used as a means of allocating aggregate funding for long term storage and archiving. This funding would be routed directly to the platform of interest (FRDR in our example) to ensure support for the eventual deposit of the data from the research project. Given the time lag between funding start and research outputs, this gives the opportunity for capacity to be added. Contingencies for data volumes in excess of what is specified in the DMPs would be the responsibility of the platform with the added advantage that risks are mitigated by being spread across a large number of projects.

Drawback 1: This option relies on a review of DMPs to ensure accuracy in allocation of resources. As noted above there will be a large number of DMPs potentially mitigating any capacity excess or shortfall.

Recommendations: we assemble a wish list of attributes for such a long-term data storage plan that will support the transparent review and presentation of research findings in publications.

1) The data residency must be Canadian and fulfill all research medical ethics and legal requirements around privacy (PIPEDA and FIPPA).
2) Consistency with existing and future Tri-Agency data retention and data deposit policy.
3) Published datasets must have a digital object identifier.
4) The ethics around upload of human datasets should be clearly indicated and include steps to ensure anonymization and de-identification.
5) The storage must comply with FAIR principles for open data: findable, accessible, interoperable, and reusable.
6) At least 10 TB/lab member capacity, set by trainee (lab member) number not linked to the PI. Capacity should renew each year and/or be linked to Tri-Agency funding.
7) Streamline the process of mapping shared drives and resources where data exists.
8) The data storage cost should be ideally free and should have no recurrent cost or potential of long term vulnerability or hidden costs. Ideally these fees should be paid to NDRIO and funneled to providers like FRDR by the Canadian Government and not be collected from individual users.
9) Preserve directory and file structure: have options where directory structure can be preserved (uncompressed data taking this option may come as a penalty with your data quota).
10) Anonymously browsable datasets by reviewers in unpublished forms (not requiring the download and reassembly of complex directory relationships needed to demonstrate software).
11) Published versions of datasets can be compressed or moved to slower storage means if not accessed over a reasonable time.

Acknowledgements:
This White Paper uses some material from a larger Dynamic Brain Circuits in Health and Disease White Paper developed originally by 2 talented UBC Science coop students Ashutosh Bhudia, and Glaynel Alejo that were supervised by Jeff LeDue and Tim Murphy. We thank the CONP for their input.

References:

Cite this work using: Identifier: DOI 10.17605/OSF.IO/C29XQ

See cited hyperlinks to government documents

https://ubcbraincircuits.readthedocs.io/_/downloads/en/latest/pdf/

Gao, Ruixuan, Shoh M. Asano, Srigokul Upadhyayula, Igor Pisarev, Daniel E. Milkie, Tsung-Li Liu, Ved Singh, et al. 2019. “Cortical Column and Whole-Brain Imaging with Molecular Contrast and Nanoscale Resolution.” *Science* 363 (6424). https://doi.org/10.1126/science.aau8302.

Murphy, Timothy H., Nicholas J. Michelson, Jamie D. Boyd, Tony Fong, Luis A. Bolanos, David

Bierbrauer, Teri Siu, et al. 2020. "Automated Task Training and Longitudinal Monitoring of Mouse Mesoscale Cortical Circuits Using Home Cages." *eLife* 9 (May). https://doi.org/10.7554/eLife.55964.