# UBC Dynamic Brain Circits Cluster Whitepaper Documentation

***Release 2019***

**Ashutosh Bhudia, Glaynel Alejo**

**Nov 22, 2022**

# Research Curtailment and Remote Work

# CHAPTER 1

# Overview

The Dynamic Brain Circuits and Connections in Health and Disease research cluster 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, lSearning, and disease. This white paper aims to provide information, recommendations, and best practices in data management to aid labs in the Brain Circuits cluster, and the greater DMCBH community, to develop and maintain Data Management Plans (DMPs). This includes considerations to data storage, data sharing, research data workflows, and data stewardship throughout the research life cycle.

We are addressing a global shift towards Open Science and the pressing need within the cluster for secure data storage throughout the research process, from collection to long-term preservation. There is an understandable reluctance to invest time, money, and effort into data sharing and there are also potential risks involved, including misuse and misinterpretation of data, and inappropriate or absence of proper assignment of scientific credit[1]. We therefore encourage cluster labs to recognize that the potential and demonstrated benefits of being scientifically open can greatly outweigh its drawbacks. 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[2]. Sharing data can also boost citation count[3],[4] and a proven record of open science can positively impact careers of both new and established neuroscientists.

We present a number of possible solutions to data management, the selection of which was guided by UBC standards and legal and ethics policies. Since some of the services outlined in the paper are in development, we created this website to update the information presented in the document. This website contains example use cases, demonstrations, and documentation to aid in training and to speed up adoption. The primary purpose of the white paper and is to reduce the barriers that deter or hinder the implementation of DMPs and data sharing within the cluster.

In pursuit of this goal, we have also provisioned resources for the use of the cluster, including:

- Dataverse

[1] Gardner D, Toga AW, Ascoli GA, et al. Towards effective and rewarding data sharing. Neuroinformatics. 2003;1(3):289-295. doi:10.1385/NI:1:3:289

[2] Spires-Jones TL, Poirazi P, Grubb MS. Opening up: Open access publishing, data sharing, and how they can influence your neuroscience career. Eur J Neurosci. 2016;43(11):1413-1419. doi:10.1111/ejn.13234

[3] Piwowar HA, Day RS, Fridsma DB. Sharing Detailed Research Data Is Associated with Increased Citation Rate. Ioannidis J, ed. PLoS One. 2007;2(3):e308. doi:10.1371/journal.pone.0000308

[4] Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. PeerJ. 2013;1:e175. doi:10.7717/peerj.175

- Federated Research Data Repository collection
- JupyterHub
- GitHub Team Repository

CHAPTER 2

# Contents

## 2.1 Research Curtailment and Remote Work

### 2.1.1 Introduction

COVID-19 and UBC's research curtailment are impacting all aspects of teaching and scholarship and changing the way we approach these activities. The Dynamic Brian Circuits (DBC) cluster will present approaches and tools which have become useful (or more useful) to enable collaboration and remote work during the COVID-19 pandemic. This meeting will be hosted by Dr. Tim Murphy (lead Dynamic Brain Circuits cluster) and consist of 5, 10 min talks: Slack Workspaces (Peter Hogg, DBC Neurodata tutor, Haas lab), Open Science Framework (Annika Wevers, DBC co-op student), remote access to computers (Jeff LeDue, DMCBH), and remote data analysis with Jupyterhub (Patrick Coleman, Haas lab).

### 2.1.2 Slack Workspaces

Slack is a platform that facilitates communication during remote work. Slack workspaces can be organized into channels for specific needs, and members can communicate in a group or privately. Some uses of the DBC Slack include data analysis, project collaboration, and coordinating meetings.

Join the DMCBH Slack

Join the Dynamic Brain Circuits Slack

### 2.1.3 OSF

Open Science Framework (OSF) is a free and open source project management tool. OSF enables collaboration on projects and streamlines the workflow process by integrating already existing platforms such as Dropbox and GitHub. UBC is an OSF Institution, which means researchers can affiliate their OSF account and projects with UBC and login to OSF is possible using their university credentials.

### 2.1.4 Remote Access

Remote Access to computers can be achieved in many different ways. Computers in the DMCBH NeuroImaging and NeuroComputation Centre are available for remote access with teamviewer or Microsoft Remote Desktop. Each of these computers has 6-cores and 64GB RAM and run windows and linux. Join us on the DBC slack for help with access.

### 2.1.5 Remote data analysis with JupyterHub

JupyterHub is a notebook server that allows multiple users to access a pool of resources more powerful than a single workstation for interactive analysis.

## 2.2 Federal Initiatives on Open Science

Although directed towards government scientists, we present the following federal initiatives on Open Science to show that data sharing is a task of enough importance to be addressed on a national scale.

### 2.2.1 Canada's Commitment to Open Science

In an open letter, the Treasury Board of Canada Secretariat discusses the concern that federal scientists funded by the federal government cannot discuss their work openly, based on the outcome of the survey on scientific integrity by the Professional Institute of the Public Service of Canada (PIPSC).

To address this, they have undertaken the following key initiatives:

- Created a new Directive on the Management of Communications (2016) aimed at "fostering greater openness, transparency and accountability, and clearly stating that subject-matter experts, including scientists, may speak publicly on their own areas of expertise and need not be explicitly designated to do so."
- In August 2016, letters went out to all Ministers and departments to reiterate the commitment to ensure government scientists are allowed to speak publicly about their work.
- The Government has also recently announced significant additional investments to make basic science a priority, including $540 million for the National Research Council and significant investment for new and renewed federal research infrastructures.

### 2.2.2 Canada's 2018-2020 National Action Plan on Open Government

In section 5 of Canada's 2018-2020 National Action Plan on Open Government titled "Open Science", the plan indicated the following concerns:

- Scientific research is often not open, accessible by, or appropriately communicated with the general public.
- Members of the public are not aware about how they can get in touch with scientists conducting research on issues that are relevant to them.

To address this, the government aims to:

- make improvements to open.canada.ca
- help Canadians learn more about the Government of Canada's work on open government
- improve the quality of open data available through open.canada.ca
- expand the Open by Default pilot project

- provide tools for government and citizens to work together
- develop open data privacy guidelines

Other relevant documents:

- Progress made on the milestones
- End-of-Term Self-Assessment Report on Canada's Third Biennial Plan to the Open Government Partnership 2016-2018 - Commitment 14: Increase openness of federal science activities (Open Science)

## 2.3 Tri-Agency Policies

The Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC) are three federal research funding agencies collectively known as the Tri-Agency.

### 2.3.1 Tri-Agency Open Access Policy on Publications

The objective of the Tri-Agency Open Access Policy on Publications (last modified: 2016/12/21) is to increase the accessibility of the results of Agency-funded research. As of January 1, 2008, research funded by the CIHR must satisfy two requirements:

1. Deposit bioinformatics, atomic, and molecular coordinate data into an appropriate database upon publication. (Examples of research outputs and corresponding publicly accessible archive, repository or database)
2. Preserve original datasets, both published and unpublished, for at least five years after the end of the grant.

### 2.3.2 Tri-Agency Research Data Management Policy For Consultation

The Tri-Agency Research Data Management Policy for Consultation (last modified: 2018/05/25) promotes best practices in research data management.

- Section 3.2 "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.
- As indicated in section 3.3 "Data Deposit", the policy requires grant recipients to deposit all digital research data, metadata, and code into a repository. This applies to all data that directly support the research conclusions in journal publications, pre-prints, and other research output from agency-funded research.

Note that as of 2019, this document remains a draft and the policy will be implemented incrementally.

### 2.3.3 Tri-Agency Statement of Principles on Digital Data Management

The Tri-Agency Statement of Principles on Digital Data Management (last modified: 2016/12/21) specifies the expectations of the agencies regarding research data management, which encompasses:

1. Data Management Planning
2. Constraints and Obligations
3. Adherence to Standards
4. Collection and Storage
5. Metadata

6. Preservation, Retention and Sharing
7. Timeliness
8. Acknowledgement and Citation
9. Efficient and Cost Effective

The statement also defines the responsibilities of the researchers, research communities, research institutions, and research funders that must be fulfilled in order to meet these expectations.

From the statement:

> Responsibilities of researchers include:
>
> - incorporating data management best practices into their research;
> - developing data management plans to guide the responsible collection, formatting, preservation and sharing of their data throughout the entire lifecycle of a research project and beyond;
> - following the requirements of applicable institutional and/or funding agency policies and professional or disciplinary standards;
> - acknowledging and citing datasets that contribute to their research; and
> - staying abreast of standards and expectations of their disciplinary community.

## 2.4 UBC Information Security Standards

UBC's Chief Information Officer has published Information Security Standards, which each lab must carefully consider and adhere to when choosing data sharing and storage services. The documents most pertinent to the purposes of this white paper are listed below and are outlined in the following sections.

1. Policy 104, Acceptable Use and Security of UBC Electronic Information and Systems
2. Information Security Standard #01: Security Classification of UBC Electronic Information
3. Information Security Standard #03: Transmission and Sharing of UBC Electronic Information

### 2.4.1 Obtaining Informed Consent

Please see Sensitive Data Toolkit for Researchers Part 3: Research Data Management Language for Informed Consent

### 2.4.2 Security Classification of UBC Electronic Information

**UBC Electronic Information** is "electronic information needed to conduct University Business" as defined in Policy 104, Acceptable Use and Security of UBC Electronic Information and Systems.

The relevant precautions and standards depend on the nature of the data and is outlined in Information Security Standard #01: Security Classification of UBC Electronic Information. It is therefore crucial for data to first be classified before actions are taken to store and/or share it. The Information Security Classification Model has four levels: *Low Risk*, *Medium Risk*, *High Risk*, and *Very High Risk*. Research data that is non-personal and non-proprietary is considered Low Risk, while non-personal and proprietary is Medium Risk. Employee IDs and home addresses fall under Personal Information and are therefore considered High Risk. Very High Risk UBC Electronic Information include biometric data, date of birth, and personally identifiable genetic data.

**Note:** The classification of data may change over time, hence the method of data sharing and/or storage being used can also be changed as other options become permissible or more desirable.

### 2.4.3 Transmission and Sharing of UBC Electronic Information

There are two sections of note in this standard. The table under Section 9 provides the method of transmission(s) appropriate for each information security classification. To ensure compliance, Table 2 categorizes the major services presented in this paper by method of transmission, however note that the classifications made here have not been approved by the Office of the Chief Information Officer.

Section 11 must also be heeded for decisions regarding data storage and sharing. It is as follows:

> Subject to section 9, if the User is using personal accounts or other information sharing tools to share UBC Electronic Information, they are responsible for ensuring that a copy of this information is stored on UBC Systems, in addition to any desktop computers and mobile devices, at all times.

**UBC Systems** include but is not limited to Compute Canada, Teamshare, Educloud, FRDR, Dataverse, and servers and computer systems in UBC. Hence, it is recommended that data is stored and/or shared on UBC Systems first if other services are to be used.

Please be advised that sharing of Very High, High, and Medium Risk UBC Electronic Information through personal email is not permitted under Policy 104, Acceptable Use and Security of UBC Electronic Information and Systems. Contact information of UBC faculty and staff is considered low risk information and is not recommended for sharing through personal email. It is therefore highly recommended for members of the lab to secure and use a UBC email account for University Business.

| Solution | Method of Transmission | *Permitted* Information Security Classification | Considered As |
|---|---|---|---|
| Teamshare | UBC-endorsed File Sharing Collaboration & Messaging Tools | recommended for all levels of risk | Full Backup, UBC copy |
| On-Campus Network Attached Storage, RAID Arrays, removable external storage media (flash drives, SSDs, hard drives) | Mobile Storage Devices/ Media (e.g. USB drives, CDs/DVDs, tapes) | acceptable for Low Risk, encryption strongly recommended for Medium Risk, encryption required for Very High Risk | On-site, UBC copy |
| Compute Canada clusters | Other Internet Transmissions (e.g. SSH, FTPS, SFTP) | All risk levels with authentication and encrypted connections (insecure internet transmissions e.g. Telnet, FTP are not permitted) | Off-site, UBC copy |
| Compute Canada cloud, Educloud | Other Internet Transmissions (e.g. SSH, FTPS, SFTP), Websites Hosted Within Canada | All risk levels with authentication and encrypted connections (insecure internet transmissions e.g. Telnet, FTP are not permitted) | Off-site |
| Cloud Object Storage (AWS, GCP, Azure, IBM Cloud) | Other Internet Transmissions (e.g. SSH, FTPS, SFTP), Websites Hosted Within Canada (if server is located in Canada) | All risk levels with authentication and encrypted connections (insecure internet transmissions e.g. Telnet, FTP are not permitted) | Off-site |
| Scholars Portal Dataverse, FRDR, OSF | Websites Hosted Within Canada (must specify storage location as Canada) | Very High Risk, High Risk, Medium Risk permitted with authentication and HTTPS (encrypted) connections. HTTPS (encrypted) strongly recommended for Low Risk. | Off-site |

*Major data sharing and storage platforms classified according to Information Security Standard #03: Transmission and Sharing of UBC Electronic Information*

## 2.5 Data Management Plans (DMPs)

### 2.5.1 Introduction

From the UBC library website:

> A Data Management Plan (DMP) is a document you create that sets out how you will organize, store and share your research data at each stage in your project. A DMP is a living document that can be modified to accommodate changes in the course of your research.

DMPs are increasingly becoming a part of grant applications. Granting agencies such as the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), the Social Sciences and Humanities Research Council of Canada (SSHRC) and the Canadian Foundation for Innovation (CFI) require that the research they fund is conducted using best practices in data management for the following reasons:

- To ensure that data management practices align with ethical and legal requirements. This includes animal and patient ethics, university policy, granting agency and government requirements.

- To ensure that research data is accessible, reusable, and stored properly. This includes data standardisation, backup, documentation, and metadata.
- To foster open science practices to enhance research productivity and quality.

### 2.5.2 Developing a DMP

The white paper is intended to aid PIs in developing DMPs. In addition to this, the following resources are available:

#### UBC Library

The UBC library provides resources for Research Data Management through its website. Additionally, research data librarians can be contacted at research.data@ubc.ca

#### DMP Assistant (Portage)

Portage was launched by the Canadian Association of Research Libraries and works with libraries to enhance shared stewardship of data. This includes availing expertise such as research librarians and services and technologies such as the Federated Research Data Repository and the Portage DMP Assistant, a step-by-step tool for preparing a data management plan. Under *Data Management Planning* of section 3, the *Tri-Agency Statement of Principles on Digital Data Management* states that DMPs should be developed using standardized tools, where available. It is therefore highly recommended that researchers use the Portage DMP Assistant for Canadian funders[1].

### 2.5.3 Additional Resources

#### Metadata

Metadata is data that describes data. Ideally, all datasets should be accompanied by, at minimum, metadata that fully describes the dataset such that lab members, external collaborators and preferably, any user, can reproduce or extend the study.

Resources on research data metadata and metadata standards:

- UBC Library - Document and Describe
- University of Western Australia - Research Data Management Toolkit: Documentation
- Cornell University - Research Data Management Service Group: Metadata and describing data
- FairSharing is a curated, informative, and educational resource on data and metadata standards, inter-related to databases and data policies.

#### File Naming Conventions

File naming conventions should be interoperable between different computer systems (length, special characters, and case sensitivity), eliminate ambiguity, support versioning, be concise, and be conscious of directory structure and hierarchy.

Resources on file naming conventions:

- UBC Library - Organize

[1] For US funders, such as the NIH and NSF, use the analogous DMP Tool.

#### Best Practices from Exemplar Labs

- The **3-2-1 Backup Rule**: At the very least, research data should have 3 copies, with 2 copies on different media and 1 copy off-site. This includes having 2 UBC copies.
- As part of your DMP, set up infrastructure, services, and training to ensure that at worst, **only one day of work can be lost in the event of an emergency**.
- **All data is accompanied by a metadata file**. Secondary data is accompanied by a wiki article describing how the data was processed and analysed.
- **Automated metadata entry** for experimental data.
  - Example 1: Murphy Lab dataset on Zenodo
  - Example 2: Alyx
- **Backup snapshots** of all workstations and the central data storage servers.

## 2.6 Data Storage Platforms

To facilitate the construction of data management plans, we make a distinction between data in an ongoing research investigation and final data that has been processed and/or analyzed. Thus, we divide data storage services into two sections: Live Storage and Data Archival.

For both sections, we provide overviews and brief evaluations of various storage platforms and technologies.

### 2.6.1 Live Storage

Live storage is appropriate for data that is actively changing and accessed frequently. This typically requires a high performance file system with low wait times and high read-and-write speeds to support demanding research workflows. They are normally characterised by mid to high storage costs.

#### TeamShare

TeamShare is a UBC service that allows Faculty and Staff to store and share files with their teams.

#### Cost

As of 7 August 2019, the UBC teamshare service charges a flat rate of CAD 0.15 per gigabyte per year. A minimum of 20 GB must be purchased. Charges will be made monthly to the speed chart provided when making the application.

#### Signing Up

To sign up for TeamShare, the following requirements have to be met:

- An Enterprise Active Directory (EAD) account, which in turn requires a CWL account.
- All UBC faculty, staff and students have a CWL account. People external to UBC can be sponsored a CWL account to gain access.
- The department must be onboarded onto EAD
- A billing contact and speed chart

Once these requirements are met, department administrators can order TeamShare Storage Service by submitting a Systems Service Request.

### Access

Teamshare can only be accessed through the UBC network or through the UBC VPN.

CWL is used to log into the VPN, which can be accessed on a variety of systems, of which Windows and MacOS are officially supported.

### Backups

From the TeamShare FAQs:

> Backup snapshots are taken:
>
> - Every 2 hours between 6:00 am and 6:00 pm and kept for 5 days
> - Daily at midnight and kept for 30 days

Teamshare is fully compliant with the 3-2-1 Backup Rule with an offsite backup in Kamloops. This means that it can be a one-stop shop option for data storage and backup, especially for labs that produce small disk volumes of data.

### Compute Canada (CC)

Compute Canada is a provider of Advanced Research Computing infrastructure, including systems, storage, and software. Their regional partner is Westgrid which provides additional support.

Compute Canada provides heterogeneous, general purpose clusters and clouds that allow researchers to access resources such as CPU and GPU time, software, as well as different storage systems. A list of the available national systems can be found here.

### Cost

Compute Canada and Westgrid resources can be used at no financial cost to researchers.

### Signing up

You can apply for a Compute Canada account following the documentation.

- If you are a PI, please create an account as other lab members must be sponsored under your account.
- If you are a lab member, contact your PI for their Compute Canada Role Identifier (CCRI) so that you can complete your application. Your PI must then confirm your role for your account to be created.

### Storage Types

**Home** Small fixed quota that cannot be changed. Does not offer high performance read and write speeds. Unique to each user. It is not clear whether this space is backed up. Best location to store smaller files like source code, scripts and configs.

**Project** Larger quota than Home. Resource shared by PI and all users registered under the PI. Details on backups can be found here. Best location to share files from and for live storage.

**Scratch** Large quota for each user, ~20TB. For high performance read and write operations. This space is not backed up and is purged at 60 day intervals. Files should only be in stored in Scratch when running jobs.

**Nearline** A file system virtualized onto tape. Files copied to nearline are transferred to tape. Access speeds are slow as files have to be copied onto disk from tape. Resource shared by PI and all users registered under the PI. Data is not backed up. Best location for data archival.

### Security and Status

Compute Canada is considered to be off-site UBC storage as per *UBC's Information Security Standard #03: Transmission and Sharing of UBC Electronic Information*.

### Resource Allocations

The following information pertains to the national heterogeneous clusters as of 8th August 2019.

### Default Resource Allocations

- CPU/GPU time is for opportunistic use (no priority is given on the queue). If you require priority, apply to the Resource Allocation Competitions (RAC)
- No Cloud Allocation
- Default storage allocations[1]:

| | Home (user) | Project (project) | Scratch (user) | Nearline (project) |
|---|---|---|---|---|
| Beluga | 50 GB, 0.5 M files | 1 TB, 0.5 M files | 20 TB, 1 M files | 1 TB, 0.5 M files |
| Cedar | 50 GB, 0.5 M files | 1TB, 5 M files | 20 TB, 1 M files | 2 TB, 5 K files |
| Graham | 53 GB, 0.5 M files | 1 TB, 0.5 M files | 20 TB, 1 M files | 1 TB, 0.5 M files |

### Rapid Access Service (RAS)

From the CC website:

> Rapid Access Service (RAS) allows Principal Investigators (PIs) to request a modest amount of storage and cloud resources without having to apply to the Resource Allocation Competitions (RAC).

- CPU/GPU time is for opportunistic use (no priority is given on the queue). If you require priority, apply to the Resource Allocation Competitions (RAC)
- Cloud allocations available

PIs are encouraged to apply for storage through RAS. Details can be found on their website, linked above.

[1] Obtained usign the diskusage_report command-line utility of the Compute Canada clusters.

### Resource Allocation Competitions (RAC)

RACs are held annually in the fall and are awarded the following April. They enable researchers to request resources beyond what they can apply for through RAS.

- CPU/GPU time by priority
- larger storage allocations
- larger cloud allocations

### Recommended Usage Scenario for Live Storage

The project file system can be used as a backup, made at regular intervals such as daily or bi-hourly. This can be automated as a cron job using shell scripting tools like Globus, rsync, scp, or sftp.

Documentation on these tools:

- Transferring data from the Compute Canada documentation
- Data Transfer and Backup on Remote Computers from the white paper website

Archeion or the Globus API/SDK can be used to script Globus transfers.

### Network Attached Storage (NAS)

A **NAS** is a storage server that allows clients to access a centralised filesystem over a network. Redundant Array of Inexpensive Disks (RAID) can be incorporated into the setup to prevent data loss due to hard drive failure.

NAS is a great solution for centralising and sharing large volumes of data, but plans must be made for making backups of the data and for archival towards the end of the research project life cycle.

We cautiously recommend NAS because all cluster labs currently using NAS have problems with backing up their data as they accumulate large amounts of data without clear plans on how to manage the growth or volume.

### Storage

NAS can be scaled up to tens of terabytes on consumer grade options and hundreds of terabytes on enterprise grade offering. Usually, NAS modules have to be purchased separately from the hard drives, which have to be slotted in. For use cases needing high throughput, more expensive Solid State Drive (SSD) options can be considered.

Labs considering a NAS solution should think about factors like cost, Input/Output speeds, RAID levels, sophistication, expandability, support, security, and storage needs.

### Cloud Object Storage

Cloud storage is increasingly being used in research institutions to store data, especially if they also use the cloud platform for analytics.

Due to the large costs associated with this option, as well as the fact that it is not considered a UBC System, we do not recommend this option for purposes other than data archival. However, such cloud platforms are the best option for on-demand storage and analysis resources and have large flexibility for scaling.

Due to the transaction-based pricing model employed by cloud platforms, it is challenging to determine the costs of their services and therefore labs considering using this solution should consult with experts such as UBC ARC.

### 2.6.2 Data Archival

Research data that does not change or is not accessed often (< 1 time a month) should be archived for preservation to reduce storage costs and effectively manage live storage resources. Archival and long-term availability of research data is also a requirement of some journals and research grants.

There are solutions available that are designed for archival, with common characteristics including low storage costs, high retrieval times, and high retrieval costs.

#### TeamShare

Consult the introduction to TeamShare in the Live Storage section for more information.

Teamshare is cost-effective approximately until 15 TB of usage after which cloud storage becomes the better strategy. It has some benefits over the other solutions, including access on demand, fast access speeds, no extra cost of retrieval and being a fully compliant backup.

#### Compute Canada Nearline

Nearline is a file system virtualized onto tape. Files copied to nearline are transferred to tape. Access speeds are slow as files have to be copied onto disk from tape. Resource shared by PI and all users registered under the PI. Data on nearline not backed up. It is the best location for data archival on Compute Canada.

Read more about Compute Canada file systems *here*.

#### Cloud Object Archival

The bare minimum criterion for off-site storage is that the data should be stored in Canada. Providers of cloud-based object archival have transaction-based pricing models. This can make it challenging to determine the costs of their services and therefore labs considering using this solution should consult with experts, such as UBC ARC.

As of 7th August 2019, the following cloud providers have data centers in Canada

| Cloud Provider | Data Center(s) |
|---|---|
| Amazon S3 Glacier | Canada Central |
| Google Cloud Platform | Montréal |
| Microsoft Azure | Canada Central, Montréal |
| IBM Cloud | Toronto, Montréal |

*Cloud providers and the location of their data centers in Canada*

#### Amazon S3 Glacier

Pricing information is available here.

Usage is billed monthly. Transactions billed include storage, retrieval requests, upload requests[1], data retrieval, and outbound internet data transfers. All data transfers into Glacier are free.

Glacier has the highest retrieval costs compared to other storage options. It is therefore probably not ideal for an emergency recovery situation where massive amounts of data must be retrieved.

[1] LISTVAULTS, GETJOBOUTPUT, DELETE and all other requests are free.

### Google Cloud Platform

Pricing information is available here.

Usage is billed monthly. Transactions billed include storage, data retrieval, data operations and requests, and outbound internet data transfers. All data transfers into GCP are free.

**Storage class definitions**

**Regional storage** store data more cheaply at the expense of data being stored at one location instead of having geographic redundancy.

**Nearline Storage** 30-day minimum storage, ideal for data accessed once a month at most.

**Coldline Storage** 90-day minimum storage, ideal for data being accessed once a year at most.

The minimum storage durations for Regional and Coldline storage are 30 and 90 days respectively. If data is removed early, you will be charged the storage cost for the fraction of the time remaining. For instance, if 1,000 GB is deleted from coldline after 60 days, you will be charged 1,000 GB * $0.007 USD/GB/Month * 1 Month = $7 USD.

### Microsoft Azure

Pricing information is available here.

Usage is billed monthly. Billed transactions include storage, data retrieval, data operations and requests, and outbound internet data transfers. All transfers into Azure are free.

#### Definitions of storage classes

**Hot** Optimized for storing data that is accessed frequently.

**Cool** Optimized for storing data that is infrequently accessed. Data must be stored for at least 30 days.

**Archive** Optimized for storing data that is rarely accessed. Data must be stored for at least 180 days with flexible latency requirements (on the order of hours).

#### Definitions of redundancy options

**Locally Redundant Storage (LRS)** Keep multiple copies in one data center.

**Zone Redundant Storage (ZRS)** Keep multiple copies across multiple data centers or across regions.

**Geographically Redundant Storage (GRS)** Keep multiple copies of data in one region while asynchronously replicating in another region.

**Read-access Geographically Redundant Storage (RA GRS)** Allow read access from the second region used for GRS.

### IBM Cloud

Pricing information is available here.

Usage is billed monthly. Billed transactions include storage, retrieval, data operations and requests, and outbound internet transfers. All transfers into IBM cloud are free.

**Definitions of storage classes**

**Standard** Optimized for storing data that is accessed frequently (many times in a month).

**Vault** Optimized for storing data that is infrequently accessed. Data must be stored for at least 30 days.

**Cold Vault** Optimized for storing data that is rarely accessed. Data must be stored for at least 90 days.

**Flex** Dynamic movement between storage classes on a per month basis.

**Definitions of redundancy options**

**Single Data Center** Keep multiple copies on different devices in one data center.

**Regional** Keep multiple copies of your data in different data centers in one region.

**Cross Region** Keep copies across three regions.

| Solution | Considered as part of UBC | Located in Canada | Cost | Recommended Use Case |
|---|---|---|---|---|
| Teamshare | Yes | Yes | $ 0.015 per GB per year [] | < 15 TB volume |
| Compute Canada Nearline | Yes | Yes | Free | Highly recommended. Quotas can be increased through RAS and RAC |
| Cloud Object Based Archival (AWS, GCP, Azure) | No | Yes, with international storage locations also available | Transaction based pricing. See below for more details | For additional off site backup of high data volumes > 15 TB |

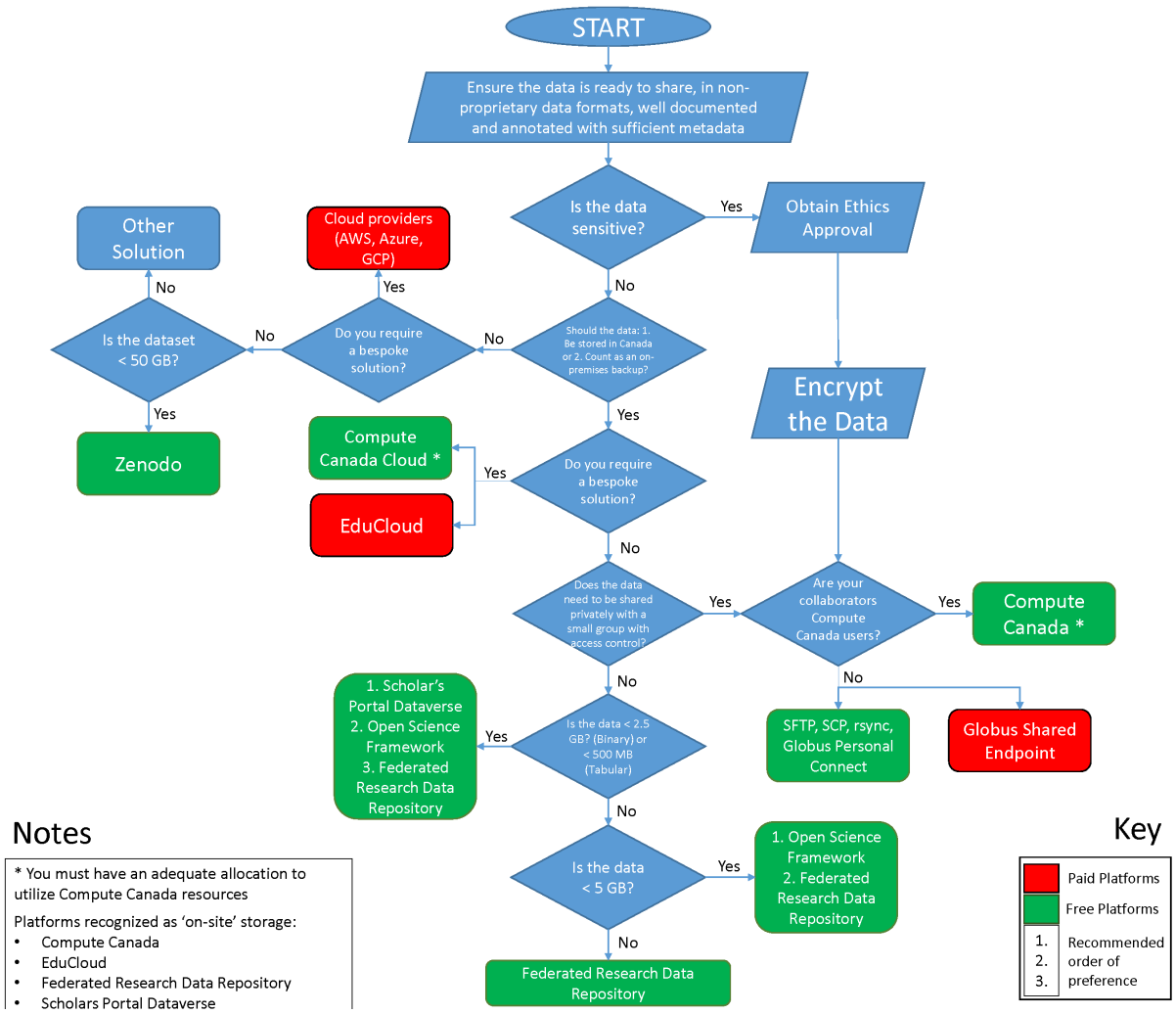
*Summary of data archival solutions*

## 2.7 Data Sharing Platforms

It is highly recommended for researchers to share data associated with the findings and results of publications as this promotes the cluster's aim to further engage in open science. In addition to providing a data archival solution whenever appropriate, data sharing platforms have other benefits such as:

- Reduced overall cost of storage - the use most of these platforms are free of charge
- Backups are created by the platforms and data integrity is highly assured
- Meeting data management requirements, such as number of backups, backup location and mode, and long-term preservation
- Embargo features allows fresh data to be backed up and released publicly after publication

The use of APIs provided by these platforms can be used to automate the upload of large quantities of data to these platforms during the initial process of archival and standardised in-lab procedures can then be set up to make archival on such platforms a frequent activity.




*Data sharing flowchart*

### 2.7.1 Recommended Data Sharing Platforms

| Platform | Cost | Total Storage Quotas | Dataset Size Limits | Curation | Shareability | Maintains File Hierarchy | Backed up |
|---|---|---|---|---|---|---|---|
| Teamshare | CAD 0.15 per gigabyte year | Total Storage Purchased | None | Internal | UBC only | Yes | Yes |
| Compute Canada | Free | On default allocation, 50 Gb per user and 1 TB per group per Compute Server. More available through Resource Allocation Competitions. | None | Internal | Compute Canada Users Only. Full access control possible with a Globus Subscription | Yes | Yes |
| Scholars Portal Dataverse | Free | None within reasonable usage | 2.5 GB for binary files, 500 MB for tabular data formats | Internal | Full access control possible | Only if double zipped | Yes |
| Federated Research Data Repository | Free | None within reasonable usage | None within reasonable usage | External | Open | Yes | Yes |
| Open Science Framework | Free (with free and paid add-ons available) | None within reasonable usage | 5.1 GB on OSF storage | Internal | Full access control possible | Must be done manually on OSF dashboard | Yes |

#### Borealis: The Canadian Dataverse Repository

Borealis is a publicly accessible and secure Canadian data repository provided by Scholars Portal on behalf of the Ontario Council of University Libraries (OCUL) and other participating institutions. Borealis can be used by affiliated researchers to deposit, share, and archive research data.

**Tip:** Consult the Borealis User Guide for more information.

**Attention:** We are pleased to announced that we have set up a dataverse for the cluster within the UBC dataverse, under which we can set up dataverses for individual labs.

Contact Jeffrey LeDue to set up a dataverse for your lab.

#### Storage and Backup

#### Location

Borealis is hosted in Canada at the University of Toronto Libraries. Although institutional Datavarses can be set up, the data is still stored at this data center.

### Size Limits

> **Attention:** The limit for an individual file upload on Borealis is **2.5 GB**.
>
> There is no set limit on the overall storage per dataverse.

There are also other considerations to be made when making uploads:

1. When .zip or .tar archives are uploaded, they can be:
   - unpacked, with loss of hierarchy and file organisation. They recommend having fewer than 500 files in each archive due to processing and user experience concerns.
   - packed, maintaining hierarchy and file organisation. To achieve, this, the archive must be zipped or tarballed.
2. Tabular data files (Excel, SPSS, R Data, CSV) should be less than 500MB in size as the processing that takes place at upload time will make the process slow.

For larger datasets, dataverse support can be contacted at dataverse@scholarsportal.info. However, the recommended next step is to use FRDR.

### User Permissions

Users can be granted the following tiered access and permissions:

**Admin** A person who has all permissions for dataverses, datasets, and files.

**Contributor** For datasets, a person who can edit License + Terms, and then submit them for review.

**Curator** For datasets, a person who can edit License + Terms, edit Permissions, and publish datasets.

**Dataset Creator** A person who can add datasets within a dataverse.

**Dataverse + Dataset Creator** A person who can add subdataverses and datasets within a dataverse.

**Dataverse Creator** A person who can add subdataverses within a dataverse.

**File Downloader** A person who can download a published file.

**Member** A person who can view both unpublished dataverses and datasets.

| Role | Dataverse | | | | | | |
|---|---|---|---|---|---|---|---|
| | Add | Edit | View Unpublished | Manage Permissions | Publish | Delete | Download File |
| Admin | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Contributor | | | | | | | ✓ |
| Curator | ✓ | | ✓ | | | | ✓ |
| Dataset Creator | | | | | | | |
| Dataverse + Dataset Creator | ✓ | | | | | | |
| Dataverse Creator | ✓ | | | | | | |
| File Downloader | | | | | | | ✓ |
| Member | | | ✓ | | | | ✓ |
| **Role** | **Dataset** | | | | | | |
| Admin | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Contributor | | ✓ | ✓ | | | ✓ | ✓ |
| Curator | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Dataset Creator | ✓ | | | | | | |
| Dataverse + Dataset Creator | ✓ | | | | | | |
| Dataverse Creator | | | | | | | |
| File Downloader | | | | | | | ✓ |
| Member | | | ✓ | | | | ✓ |

*Dataverse roles*

**Note:** It is recommended that curators are clearly identified in your DMP, as well as guidelines and procedures they should abide by to prevent misuse of permissions.

### Additional Features

***Findability and Reusability*** Automatically generated DOI and citation (**F** A I **R**).

***Version Control*** Built-in versioning with ability to view differences.

***Metrics*** Track number of downloads and collect data about users who download files using the Guestbook feature.

***Dataset Template*** Dataverse provides the following dataset templates:

1. CC Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)
2. CC Attribution-Non-Commercial 4.0 International (CC BY-NC 4.0)
3. CC Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
4. CC Attribution 4.0 International (CC BY 4.0)

The templates contain metadata fields that are prepopulated based on the chosen license.

It is possible to create custom templates so that dataset creators will not have to enter values for metadata fields that do not change from dataset to dataset. While it is not possible to create custom fields, there are large selections of metadata fields available in the template creator and the ability to create keywords for other particulars.

Labs are encouraged to create standardized templates for their datasets to ensure all required metadata are captured.

### Federated Research Data Repository (FRDR)

Portage, Compute Canada (CC) and the Canadian Association of Research Libraries (CARL) have collaborated to produce the Federated Research Data Repository.

**Attention:** We are pleased to announce that we have set up a FRDR collection for the cluster, on which all cluster-associated datasets will appear.

To deposit a dataset into the cluster's FRDR collection, refer to the *Deposit* section.

### Storage and Backup

#### Location

From FRDR's About page:

> Data submitted to FRDR is housed on Compute Canada managed infrastructure at the University of Victoria, BC or at the University of Waterloo, ON. Research data submitted to FRDR does not leave Canada. The metadata related to datasets is housed in a database the University of Victoria. Most of that metadata is shared with Globus, running on Amazon Web Services services in the USA, to be indexed and made available for discovering datasets. Certain metadata fields, for example, submitter contact information, are not shared with Globus.

Larger uploads are made using Globus. While Globus is hosted in the USA on AWS, only the public metadata is stored there; the datasets themselves are transferred securely between the endpoints so the data does not leave Canada.

### Size limits

**Attention:** There is a theoretical dataset limit of **4TB** due to the limitation of the Archivematica data preservation system. There has not been further comment on whether they will impose their own upload size limit.

However, they do have limited resources so they may impose restrictions during curation if datasets of unreasonable magnitude are uploaded.

### Changes

Only authorised curators can make changes to submitted data. A request must be made via email in case any changes need to be made.

### Curation

Curators perform tasks such as:

- Checking metadata record for completeness
- Linking DOIs for related publications
- Validation of data, for instance, flagging tabular null data with no explanation, corrupt files
- Checking documentation
- Checking for copyright and ethical violations

They typically take up to 48 hours to complete this process, after which it takes 15 minutes to complete DOI registration, and a further 2 hours for the dataset to appear for download in the FRDR portal.

### Data Sharing and Collaboration

FRDR includes search functionality for its own datasets and datasets that it harvests from other sources such as the Scholars Portal Dataverse. Users can search for/deposit datasets using the online web interface or by using an API. Note that while depositing data requires authorization, anybody can search for datasets. FRDR is also format agnostic, and allows users to manage the dataset file hierarchy. It also support embargos. They also issue DOIs for all deposited datasets. Other features include data integrity checks using checksums, curation and upload authentication. Through the use of Globus, FRDR also enables secure transfer of large datasets with the ability to make asynchronous and resumable transfers that are automatically managed.

### Deposit and Download

### Globus Connect Personal

To download or deposit a dataset, Globus Connect Personal must first be installed then your computer must be set up as an Endpoint.

1. Log into FRDR using one of these accounts: Orcid, Compute Canada, Globus ID, Google. Detailed log-in instructions can be found here.
2. Click on *Endpoints* from the toolbar on the left-hand side of the page, then click on *Create new endpoint* at the top right corner.

3. Select "Globus Connect Personal".
4. Follow the three steps on the page.

The endpoint you created will show up in the "Endpoints" page. You are now able to download datasets from FRDR and if you are already an approved depositor, you can also submit datasets.

### Deposit

Any lab in the cluster can deposit datasets into the "UBC Brain Circuits" special storage group in FRDR. PIs can become depositors by sending an email to Jeffrey LeDue with the email address or account that they used to log into FRDR. It can be one of the following (copied from FRDR, Getting Authorization To Submit):

- username@computecanada.ca
- 0000-0002-1342-3550@ORCID.org
- username@globusid.org
- username@gmail.com

You will receive an email from noreply@globusonline.org with the subject "You are invited to join FRDR Depositors - Déposants DFDR". Once you accept the invitation, "FRDR UBC Brain Circuits Depositors" should appear when you click on *Groups* in the toolbar. You are now eligible to deposit datasets into the cluster's storage group.

FRDR provides information and instructions on

- Before Depositing
- Depositing (includes Using Globus to Upload Dataset)
- After Depositing

**Attention:** To submit a dataset, log into FRDR and click on "Deposit Dataset" in the toolbar at the top of the page. Click on "Submit a New Dataset" in the "New Submission" box. This will take you to a page titled *Submit: Select Storage Group*.

Make sure you select "UBC Brain Circuits" under *Special Storage Groups*!

### Download

To download a dataset from FRDR,

1. Navigate to the page of the desired dataset. Click on *Download Dataset*, which is located near the bottom. This will take you to the *File Manager* page but if you're not already logged into FRDR, it will first prompt you to log in.
2. You should see two columns: the left one contains the dataset under a Collection that is similar to "FRDR-Prod-2". Select that files you want to download or click *select all* in the blue toolbar if you want to download the entire dataset.
3. Click *Transfer or Sync to* located in the middle of the two columns.
4. Click on *-select a collection-*. This is where the dataset will be downloaded.
5. Choose the Endpoint corresponding to the computer you wish the download to occur in. If you don't already have an Endpoint set up on the computer you are using, click on *Install Globus Connect Personal* and follow the instructions.

6. Once you've selected the endpoint, the *File Manager* page reappears. Make sure you check out the "Transfer & Sync Options". To start the download, click on the blue *Start* button underneath the left column with an arrow pointing towards the column corresponding to the Endpoint you've chosen.
7. A green banner should appear with the message: "Transfer request submitted successfully" followed by the task ID. You can track the transfer's progress in the Activity page.

You will receive an email from noreply@globusonline.org with the subject: "SUCCEEDED" followed by the task id once the transfer is complete.

### Sharing on Compute Canada

On Compute Canada clusters, Home spaces are unique to individual users while the Project space is shared by a research group. Both are backed up and are not purged. Quotas are exclusive to each cluster and do not extend to the other clusters.

| Cluster | Home Space | Project Space |
|---|---|---|
| Cedar | 50 GB and 0.5 M files per user | 1 TB and 5 M files per group |
| Graham | 50 GB and 0.5 M files per user | 1 TB and 0.5 M files per group |
| Béluga | 50 GB and 0.5 M files per user | 1 TB and 0.5 M files per group |

*Storage Quotas*

This means they can be used to share files to Compute Canada users. If a Globus subsription is purchased, files can be shared with non Compute Canada users as well. Learn more here.

This section contains quick references. For detailed documentation and instructions, refer to these docs.

---

**Note:**

- You must know the Compute Canada username of the person you want to share with.
- You must also own the files/directories that you intend on sharing. Use chown to change file owners and groups.
- Learn more about UNIX file permsissons here.

---

### Internal Sharing

To share files within your own research group, refer to these docs.

### External Sharing

To share files with other Compute Canada users, refer to to these docs

### Quick Reference

### Sharing a Single file

If you are logged in as `<user>` on compute cluster `<server>` and want to share a generic file, e.g. `<data.zip>` with another user `<other_user>`,

```
[<user>@<server>]$ setfacl -m u:<other_user>:rx <data.zip>

# substitute the variables, but don't include the <>'s
#                              `u` to identify other users to share with
#                                                  'rx' - give the user the rights to:
#                                                  `r` - read
#                                                   'x' - execute
```

### Sharing a folder

For most use cases, it is more practical to share one or more folders and give different groups access to them using Access Control Lists (ACL).

---

**Note:**

- You must have a group created in CCDB
- You must own the directory you want to share
- All parent directories of the directory you want to share must have public execute properties

If you do not meet these requirements, consult the docs.

---

If you are logged in as `<user>` on compute cluster `<server>` and want to share a folder `<shared_data>` in some project or home space to a group `<group_name>`

```
# make any file inside the folder inherit the same ACL rules (for new files)
[<user>@<server>]$ setfacl -d -m g:<group_name>:rwx /home/<user>/projects/<def/rrg-PI>
↪/<shared_data>

# make any file inside the folder inherit the same ACL rules (for existing files)
[<user>@<server>]$ setfacl -R -m g:<group_name>:rwx /home/<user>/projects/<def/rrg-PI>
↪/<shared_data>
```

Members can be added to the a group through CCDB.

### Open Science Framework

Open Science Framework (OSF) is a free and open source cloud-based data management system developed by the Center for Open Science (COS).

The following introduction is based on the OSF Guides.

**Attention:** osfclient is a Python module and command-line program for executing commands in OSF, such as uploading and downloading files and folders.

A tutorial on osfclient created for and by the Brain Circuits cluster is available as a Jupyter notebook. View a browser-friendly version of the notebook here.

### OSF Institutions

UBC is an OSF Institution, which means researchers can affiliate their OSF account and projects with UBC and log into OSF is possible using their university credentials.

Public projects affiliated with UBC appear in the institutional portal: https://osf.openscience.ubc.ca/.

To find out how to connect your OSF account to UBC, visit the OSF FAQ page of the Open Science @ UBC website.

### Storage and Backup

OSF uses Google Cloud for active and archival storage and Amazon Glacier as a backup location.

---

**Note:** OSF is maintained and developed by the Center for Open Science, which has established a $250,000 preservation fund in the event that COS closes down. This fund can preserve and maintain read access to hosted data for more than 50 years.

OSF is free due to the support of COS sponsors.

---

### Location

The U.S. is OSF's default storage location. A variety of storage locations are available, including Canada (Montreal). The global (default) storage location can be changed and will be applied to new projects and components. Each project and component can also have its own storage location.

### Size limits

**Attention:** There is no limit on storage per project and no cap on the amount of OSF Storage per user. Direct upload of individual files to OSF Storage has a **5GB** limit.

Third-party add-ons like Dropbox can be used to integrate different services already in use and connect them to OSF to allow access to existing data/materials. There is no limit for the amount of storage used across add-ons.

### Add-ons

### Citation add-ons

There are two reference managers supported by OSF:

1. Mendeley
2. Zotero

### Storage add-ons

If data needed for a project already exists in one of the services below, it can be connected to OSF rather than transferring it. This feature is also useful if a file exceeds the 5GB limit for upload.

This page provides a list of available storage add-ons in OSF, which includes platforms like Amazon S3, ownCloud, Dataverse, GitHub, Bitbucket and GitLab.

> **Warning:** OSF does not store or back up content within an add-on. Ensure that each individual add-on being used complies with *UBC's Information Security Standard #03: Transmission and Sharing og UBC Electronic Information*.

## Collaboration

### Within a lab

#### Contributors

OSF projects and components are private by default. Members of the lab can be added as "Contributors" to projects. There are two types of contributors:

1. Bibliographic (displayed in contributor list and in project citations)
2. Non-bibliographic

Each component within a project can have its own set of contributors.

#### User Permissions

A contributor can be given three levels of permissions:

1. Administrator
2. Read + Write
3. Read only

Each contributor can be granted different levels of permissions across components.

#### Checkouts

OSF has a checkout feature that prevents multiple contributors from editing the same file at once.

### Across labs

#### Access requests

Researchers can request access to both private and public projects, which means they can be added as a contributor and given a level of permission.

#### View-only (Projects)

For peer review, presentations, or data sharing, a view-only link can be created for a project. The contributor list can be anonymized for blind peer review and only selected components will be visible via the link.

#### Quick Files

A single file can be shared independently from a project using the Quick File feature, which allows files to be publicly searchable and accessible on the profile page.

### Public Access

**DOIs**

DOIs can be created for projects but OSF does not currently support DOI versioning.

**Citations**

OSF generates citations (APA, MLA, Chicago, or custom) automatically for every project and component.

**Licensing**

A license can be added to a project either by choosing from a list provided by OSF or uploading your own. Components will have the same license as the top-level project by default but they can also be licensed individually.

### Version Control

OSF has built-in version control and provides access to previous versions of files, including those stored on add-ons.

### Registration

A registration is a time-stamped copy of an OSF project that cannot be edited or deleted. This feature is useful for archiving and to capture and preserve significant moments in the research process (i.e. before submission for peer review, etc.).

A registration can be withdrawn, which means the project contents will be removed but its basic metadata will be maintained. All registrations will be made public, which can be done immediately or embargoed for up to 4 years.

### Bespoke solutions

In some cases, a custom data sharing solution is required, for instance, due to the need to host a persistent database, a website or a highly specialised analysis environment.

### Compute Canada Cloud

To get a Cloud project, the PI must have an active cloud resource allocation as part of their Resource Allocation Competition (RAC) allocation. Users sponsored under such a PI can request access to a Cloud project, and have to be granted access by their PI. Read the documentation here.

### EduCloud

EduCloud can be accessed on-demand and may be faster and easier to set up than on Compute Canada Cloud, but costs money. You must apply for an EduCloud account before you can use the service.

**From the UBC IT Website :** EduCloud Server is a self-managed, private higher education cloud server service that provides simple and secure virtual data centre access to provision, manage and utilize servers at a fraction of the cost of implementing physical servers. EduCloud Server is designed to provide functionality which allows self-management and self-deployment. EduCloud Server is intended to provide the same convenience and cost effectiveness as other cloud services, while meeting all provincial privacy requirements under the FIPPA legislation. EduCloud Server provides the ability to easily self-deploy virtual machines from templates. With this service, you can create and delete virtual machines, and change resource allocation all within minutes. Managed using a web portal, you have the flexibility of managing your own environment at any time, from anywhere.

The costs of deploying the virtual machines can be found here. Find documentation and other information here.

### Canadian Open Neuroscience Platform (CONP)

From Brain Canada:

> The Canadian Open Neuroscience Platform (CONP) is a national platform for open sharing of neuroscience research data and brings together many of the country's leading scientists in basic and clinical neuroscience to form an interactive network of collaborations in brain research, interdisciplinary partnership, clinical translation and Open Publishing. The goal of the platform is to improve the accessibility and re-usability of neuroscience data and, by increasing awareness of ongoing and past research efforts, it will reduce unnecessary duplication and overlap, resulting in a more efficient use of funding support. The CONP will also engage young investigators across the country in order to develop the next generation of "open" scientists.

The platform is currently under development, and is slated to be available to researchers in 2021. Researchers are encouraged to become early adopters and work with CONP during this beta phase. Their website is available here, where they publish updates and other information.

The platform provides the following technologies:

### CONP Portal

The portal aggregates metadata from different sources to serve as a central repository for Canadian neuroscience, giving researchers the ability to discover and access past and ongoing research data to reduce unnecessary duplication and overlap and foster collaboration. While they do not host the data, they point to different data sources, allowing researchers the flexibility of choosing how and where their data is kept. They are also developing a method of tiered access to datasets for controlled sharing of data.

### NeuroLibre

NeuroLibre is a curated repository of interactive publications. It uses Binder to allow researchers to publish Jupyter Books with interactive inline code, media and widgets. The dedicated Binder server is hosted on Compute Canada servers and enables users to execute, modify and manipulate code interactively.

This allows:

- researchers to produce a clear and reproducible publication
- reviewers to gain further insight and develop a better understanding of the data
- other researchers to contribute to a research project as they can run their own experiments and analysis on the data in situ
- researchers to communicate complex ideas and research to stakeholders without a neuroscience background, such as the media.

NeuroLibre is seen as the next step in open Neuroscience publication. CONP has not placed any restrictions on what can be published through NeuroLibre, and have encouraged researchers to submit Jupyter Notebooks/Books developed at any stage of the research life cycle.

### 2.7.2 Other Data Sharing Platforms

| Platform | Cost | Total Storage Quotas | Dataset Size Limits | Curation | Shareability | Maintains File Hierarchy | Backed up |
|---|---|---|---|---|---|---|---|
| Zenodo | Free | None | 50 GB | Internal (Changes have to be requested) | Open | Yes | Yes |
| DataLad | Self-Hosted (Dependent on system configuration) | None | None | Internal | Open | Yes | Optional |

#### Zenodo

Zenodo is an open access data repository operated by CERN . It is hosted on the high-performance infrastructure operated for the needs of high-energy physics.

#### Location

Zenodo is hosted by CERN, which has an experimental programme defined for at least the next 20 years.

The technical infrastructure is located on CERN's premises on CERN's EOS service in an 18 PB disk cluster, with two copies kept of each file on difference disk servers.

Checksums are used to detect changes and enable automatic corruption detection and recovery.

#### Size Limits

**Attention:** The maximum size that Zenodo accepts per dataset is **50 GB**, with the ability to upload an unlimited amount of datasets.

If larger files need to be uploaded, they can be contacted to discuss further.

#### Changes

If changes need to be made to the dataset, due to reasons such as typos or accidental omissions, and the upload was made within a week, a contact request can be made with Zenodo to make the modification. Once the record has been published, you cannot change the files on record.

#### Collaboration and Data Sharing

You can create collections and control what uploads can be made into them. There are no limitations to what type of data is uploaded any file format and any results can be submitted. Zenodo also assigns all publicly accessible records a DOI to make the upload more accessible and citable. It provides DOI versioning, such that uploads have a DOI representing a specific version of the dataset or a DOI representing all the versions. This allows for specific citations to be made. Github integration is also included, which allows users to track each release. There are also over 400 licenses available to choose from to assign to your datasets.

#### DataLad

DataLad is a solution that provides decentralised access to data from open data annexes. It provides a command line interface to interact with the technologies it is built on, such as git and Git-Annex to provide functionality such as dataset searches, dataset nesting, dataset collections and git annex file management. This makes it a good candidate for further extensibility using a web interface.

DataLad also comes with support for metadata, with a search command that enables metadata queries via a flexible query language.

#### Installation: Linux

**Note:** sudo priviledges are required

Install the Git-Annex from NeuroDebian. Install DataLad

```
$ sudo apt-get install datalad
```

### 2.7.3 Discovering Open Data

The CONP portal promises to deliver a central repository aggregating metadata from multiple platforms, allowing researchers to discover open data. As the portal is still under development, we have compiled a list of resources that can be used for this purpose:

- Registry of Research Data Repositories - Discover open data repositories across the globe
- bioCADDIE - Biomedical and healthcare data discovery and indexing ecosystem
- Federated Research Data Repository - FRDR crawls through multiple other Canadian data repositories including dataverses, FRDR's own data and government open data (view a full list here), allowing users to search for and discover data from multiple sources
- CRCNS - Collaborative Research in Computational Neuroscience. Hosts data from projects funded by their funding program.
- Center for Open Neuroscience - open software frameworks, platforms, data and methodologies for neuroscience
- OpenNeuro - free and open platform for sharing MRI, MEG, EEG, iEEG, and ECoG data

Lists that have been compiled by other sources:

- Awesomedata - A list of open neuroscience data platforms
- Opening up: open access publishing, data sharing, and how they can influence your neuroscience career - Table 2. Popular data sharing repositories

## 2.8 Version Control

### 2.8.1 Overview

Version control is a system that record changes made to one or more files. In this section I will discuss distributed version control systems e.g. Git+GitHub

We are pleased to announce that we have set up a GitHub team account for the cluster [github.com/ubcbraincircuits], under which members can set up their repositories.

---

**Note:** Go to the Onboarding section below for instructions on becoming a member

**Resources**

- Learn Git and GitHub

---

### 2.8.2 Not Using Version Control?

#### Why Use Version Control?

- Revert your files to a previous version
- Compare changes
- Provides a record of what changed and Why
- Lets you experiment
- Recover files if you make a wrong turn or delete them
- Distruibuted clones of a repository act as full backups
- Collaborate in an organised manner
- Clear attribution of contributions

#### What Can I use Version Control for?

- Code
- Papers
- Theses
- Journals
- Critical data files

#### Version Control Systems (VCS)

#### Git

Git was created to manage the development of the Linux kernel. It is the most popular distributed version control system. You can read more about it here. Git is complex and requires training to use safely and appropriately. This complexity gives Git greater flexibility and more functionality for power users.

#### Mercurial

Mercurial is simpler and easier to use than Git. It makes it more difficult for users to cause inintentional damage. You can read more about it here.

**Tip:** Git and Mercurial are available on Windows, MacOS and Linux. There are also several graphical user interface options for both options for those who do not prefer to use the command line interface.

#### Hosting

The following platforms can be used to host your repositories remotely:

| | GitHub | GitLab | Bitbucket |
|---|---|---|---|
| VCS Support | Git Only | Git Only | Git and Mercurial |

None of these platforms store data exclusively on Canadian servers, so sensitive data should not be pushed to the remote. If this is absolutely necessary, there is the option of setting up your own server as a remote. This is possible through:

1. Enterprise offerings from these platforms that may not benefit from academic pricing
2. Setting using an open-source offfering at the expense of losing the rich interfaces of the hosting platforms. One such example is **'gogs.io'_**.

### 2.8.3 Using Version Control?

#### Moving existing repositories to central cluster repository

It is highly recommended that all existing respositores be transferred to the main cluster GitHub repository. Please refer to the `Setup Instructions` section below for instructions.

The main aims of doing this are:

- Reducing fragmentation by maintaining a centralised code repository under the administration of PIs for easier management
- Access to research repositories after personell leave labs
- Making GitHub Pro and GitHub Team available to all cluster members, providing access to more powerful tools

#### Spread the Word!

Encourage colleagues who have not adopted VCS to do so. Incorporate VCS training during the onboarding process when brining new people into the lab.

### 2.8.4 Benefits of GitHub Education

With the GitHub Education subscription granted to the cluster, members enjoy the benefits like:

#### Wikis and Websites

Free wikis for private and public repositories. Use them to: write down documentation, write lab notes, create a whiteboard, create meeting agendas. Find out more here

You can also create websites using GitHub Pages

#### Team Collaboration Tools

PIs and Researchers can create teams and control who can edit certain repositories. There are also spaces for team discussions

### 2.8.5 Onboarding

#### 1) Get a GitHub account

If you do not already have a GitHub account, you can sign up for one here. Skip this step if you have an existing account.

#### 2) Become a Member of the Brain Circuits Repository

Contact Jeffrey LeDue and provide your GitHub account name or email to be added as a member. You will receive an invitation via the email account you used to register for GitHub.

#### 3) Move Existing Repositories to the Central Repository

If you have exisitng respositories that you desire to move to `github.com/ubcbraincircuits`, the transfer procedure can be found here. When prompted, set the `New Owner's Github username or organisation name` field to `UBCBrainCircuits`.

---

**Note:** The main aims of doing this are:

- reduce fragmentation by maintaining a centralised code repository under the administration of PIs for efficient management. This also enhances discoverability and the ability to collaborate across labs.
- enable labs to access repositories after members move on
- give cluster members free access to GitHub Team and its benefits

---

#### 4) Create or Join a Group

If your lab has not yet creted a group, you should create one by going to `github.com/ubcbraincircuits` clicking on the `Teams`, then on `New Group`. Temas can be made private or visible to all member of the cluster.

If your lab has a group, ask an existing member to add you to it.

Visible groups can be nested, and are useful for coordination. All members are able to create other groups as they wish.

## 2.9 Continuous Integration, Delivery and Deployment

### 2.9.1 Continuous Integration

Continuous Integration (CI) involves automating the integration of code from different developers of a centralised codebase managed using a version control system. This includes automated testing of correctness and compatibility of code across different platforms and environments to ensure that changes do not cause failures.

CI allows teams to enforce the quality and correctness of code quickly and reliably.

### 2.9.2 Continuous Delivery and Continuous Deployment

Continuous Delivery (CD) builds upon CI and involves the automation of software delivery by making new releases available automatically upon approval.

Continuous Deployment builds upon CD and involves the automation of the approval process. This allows end users to obtain bleeding edge updates as soon as possible instead of having to wait for a major release and enhances the feedback loop. It also further enforces the quality and reliability of the software.

### 2.9.3 CI/CD Solutions

Below is a comparison of some popular solutions, using information from the providers' websites as at the 1st of August, 2019.

For factors like full control, compliance with hosting policies and privacy of source code, readers may need to consider what solution CI/CD solution is deployed and where. They are also encouraged to conduct their own research into picking the best solutions for their projects.

| CI/CD solution | On-Premises | Hosted | Free for Open Source |
|---|---|---|---|
| Travis | Yes | Yes | Hosted |
| GitHub Actions | Coming Soon | Yes | Hosted |
| Jenkins | Yes | No | Yes |
| Buddy | Yes | Yes | Hosted |
| TeamCity | Yes | No | No |
| Bamboo | Yes | No | No |
| CodeShip | Yes | No | Free for 100 builds/month |

## 2.10 Pipelines

### 2.10.1 Data Pipelines

A data pipeline is an automated system that can be used to retrieve data from multiple sources and manage dependencies in data processing that would otherwise take time and effort to do manually. Various forms of automation can be integrated into a pipeline to create powerful and robust systems.

Pipelines are useful for researchers who need to:

- Produce or use large volumes of data from different sources
- Require real time output
- Perform analytics with a complex logical flow

Pipelines, when implemented correctly, have benefits like:

- Enhanced productivity due to automation
- Greater reproducibility

- Quality assurance through automated testing and checks

Frameworks have been developed to build data pipelines, and have been tried and tested by multiple organisations and research institutions, such as:

- Python - Luigi, Airflow, Pinball, DataJoint
- R - Drake
- MATLAB - DataJoint

#### DataJoint

**Best Use**

Ideal for MATLAB/Python-proficient users.

DataJoint is a free and open toolbox for building scientific data platform using Python, MATLAB, or both. It was developed by neuroscientists in Andreas Tolias' lab at the Baylor College of Medicine. DataJoint provides data integrity and accessibility with intuitive language for defining, querying, and visualizing data pipelines. Through DataJoint's relational data model, users can automate data entry, processing, and analysis by simply storing data in tables. Relationships between tables enforces data integrity while efficient querying is achieved by performing operations on tables.

A GUI such as Heidi SQL can be used to enter and view data but data pipelines must be defined in a Python or MATLAB environment. DataJoint is therefore ideal for labs with members that are already familiar with these languages, otherwise its use will require extensive training. DataJoint has both documentation and tutorials, which are easy to follow and conveniently uses neuroscience examples.

DataJoint enables multi-user access and the ability to grant privileges hence collaboration among researchers is simplified, particularly through features that prevent duplicated, missing, and otherwise erroneous data. Note that this platform entails resource considerations. DataJoint relies on a database server which can be self- or cloud-hosted. Depending on the scale of its use and the number of users, significant IT support may be needed to establish a database server and hardware may also be required to accommodate DataJoint. The complexity of the setup and configuration depend entirely on the needs of the lab.

### 2.10.2 Assisted Metadata Entry

Researchers need to be proviced with an easy to use lab notebook for metadata entry and experiment registration with enforced entry requirements. Preferrably, this should operate with relational database backend. This would ensure that at least a minimum amount of metadata and documentation are available for each experiment, that the data are well catalogued, searchable and shareable.

This could also lead into automated data pipelines for automatic processing of raw data and data preservation/storage, which would be a valuable tool for experiments with routine workflows and should reduce the time it takes to organise and curate experimental data manually.

#### Alyx

The International Brain Lab at UCL has created a tool built using python on a Django web framework with a PostregeSQL database system, which when deployed provides users with a web interface that they can use to log the metadata for their experiments. Features include templated input forms, user account control, search, date range filtering and data administration. The version of Alyx currently on Github is adapted for mouse brain imaging experiments, but a system like this could be implemented with different templates for different experiment groups provided

based on this framework. A major benefit of this kind of architecture is the ability to deploy it on a web server to allow centrallised remote access.

## 2.11 Data Standardization

A critical but often overlooked aspect of data sharing is data standardization. Currently, there is likely variability in data organization even between neuroscientists from the same lab. However, data is shareable only if it is understandable and interoperable, hence feasible data pooling requires universal standards. Data must be structured and organized in a way that is intuitive and readily usable by researchers within the lab and in the greater neuroscience community. When choosing a system or format, it is best to focus on the following:

- Simplicity: Does the format reflect the contents of my files, my lab practices?
- Accessibility: Is it widely used and recognized by researchers in my field?
- Adoptability: Can members of my lab learn it quickly? Can data be handed off easily?

### 2.11.1 Minimum Information about a Neuroscience Investigation (MINI)

Best use: electrophysiology data

Minimum Information about a Neuroscience Investigation (MINI) is a minimum information standard for reporting electrophysiological studies. MINI is registered with Minimum Information for Biological and Biomedical Investigations (MIBBI), a collection of guidelines for reporting bioscience data organized by method of collection. MINI is a metadata standard; it specifies recommended metadata fields pertaining to an electrophysiology dataset to ensure it is accurately interpreted, analyzed, and corroborated by the neuroscience community. The MINI reporting requirements has seven subdivisions: General features, study subject, task, stimulus, behavioral event, recording, and time series data. The original paper can be consulted for definitions.

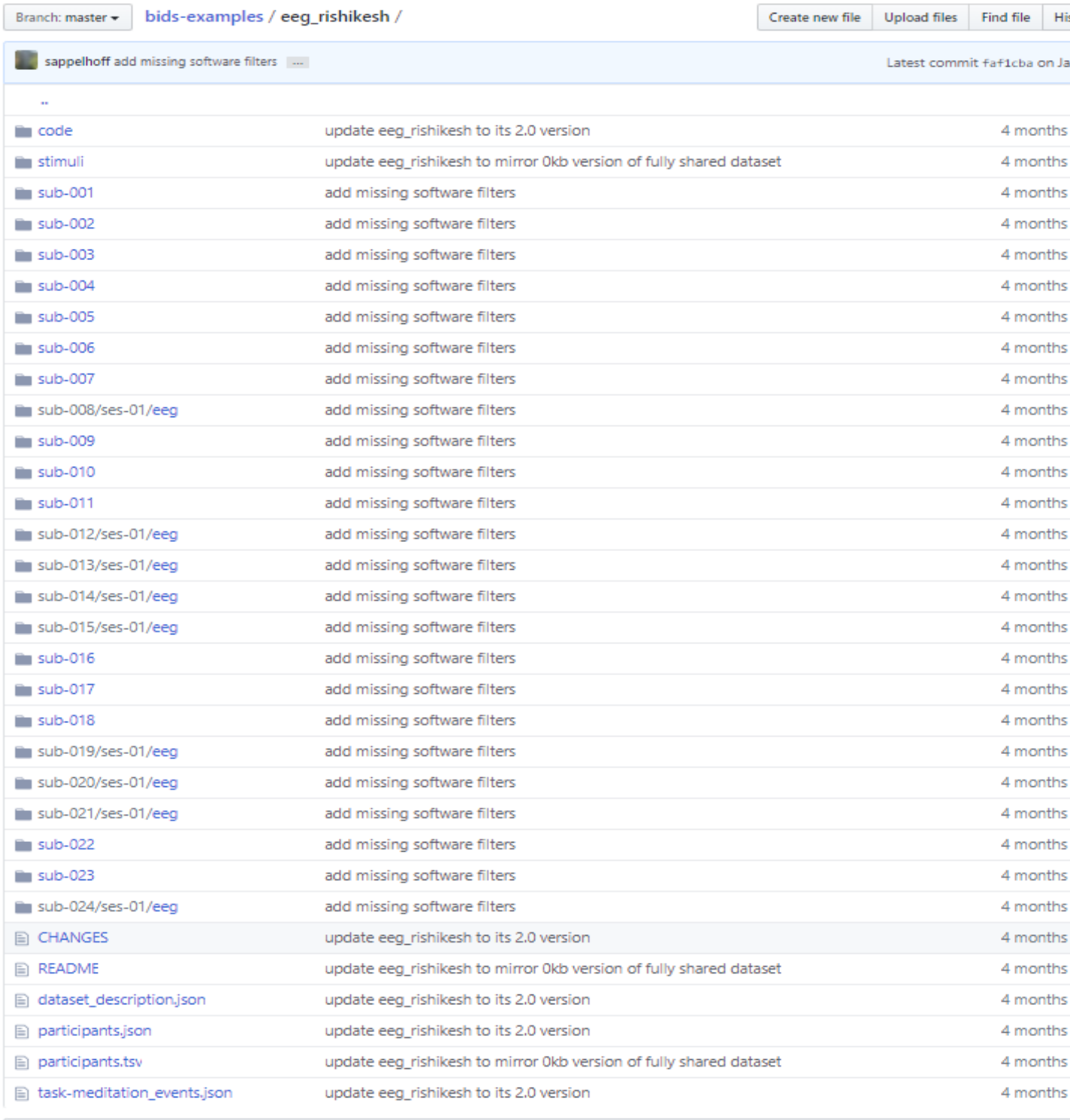
### 2.11.2 Brain Imaging Data Structure (BIDS)

#### Use Case Demo: EEG

This demo aims to familiarize members of the cluster with the BIDS format and its documentation, most importantly the BIDS Specification which this demo is heavily based on.

The following example uses a BIDS-compliant dataset eeg_rishikesh. It may be helpful to click around the dataset as you go through the demo. Note that the raw data files are empty but they have been shared by the authors on Zenodo.

The corresponding paper is "Reduced mind wandering in experienced meditators and associated EEG correlates". *Brandmeyer T, Delorme A. Reduced mind wandering in experienced meditators and associated EEG correlates. Exp Brain Res. 2018;236(9):2519-2528. doi:10.1007/s00221-016-4811-5.*



Branch: master ▾ bids-examples / eeg_rishikesh / Create new file Upload files Find file Hi

sappelhoff add missing software filters ... Latest commit faf1cba on Ja

| | | |
|---|---|---|
| .. | | |
| code | update eeg_rishikesh to its 2.0 version | 4 months |
| stimuli | update eeg_rishikesh to mirror 0kb version of fully shared dataset | 4 months |
| sub-001 | add missing software filters | 4 months |
| sub-002 | add missing software filters | 4 months |
| sub-003 | add missing software filters | 4 months |
| sub-004 | add missing software filters | 4 months |
| sub-005 | add missing software filters | 4 months |
| sub-006 | add missing software filters | 4 months |
| sub-007 | add missing software filters | 4 months |
| sub-008/ses-01/eeg | add missing software filters | 4 months |
| sub-009 | add missing software filters | 4 months |
| sub-010 | add missing software filters | 4 months |
| sub-011 | add missing software filters | 4 months |
| sub-012/ses-01/eeg | add missing software filters | 4 months |
| sub-013/ses-01/eeg | add missing software filters | 4 months |
| sub-014/ses-01/eeg | add missing software filters | 4 months |
| sub-015/ses-01/eeg | add missing software filters | 4 months |
| sub-016 | add missing software filters | 4 months |
| sub-017 | add missing software filters | 4 months |
| sub-018 | add missing software filters | 4 months |
| sub-019/ses-01/eeg | add missing software filters | 4 months |
| sub-020/ses-01/eeg | add missing software filters | 4 months |
| sub-021/ses-01/eeg | add missing software filters | 4 months |
| sub-022 | add missing software filters | 4 months |
| sub-023 | add missing software filters | 4 months |
| sub-024/ses-01/eeg | add missing software filters | 4 months |
| CHANGES | update eeg_rishikesh to its 2.0 version | 4 months |
| README | update eeg_rishikesh to mirror 0kb version of fully shared dataset | 4 months |
| dataset_description.json | update eeg_rishikesh to its 2.0 version | 4 months |
| participants.json | update eeg_rishikesh to its 2.0 version | 4 months |
| participants.tsv | update eeg_rishikesh to mirror 0kb version of fully shared dataset | 4 months |
| task-meditation_events.json | update eeg_rishikesh to its 2.0 version | 4 months |

First, let's check out the metadata `dataset_description.json` which is required for all datasets.

### Dataset description


```
{
  "Name": "Meditation study",
  "ReferencesAndLinks": [
    "https://www.ncbi.nlm.nih.gov/pubmed/27815577"
  ],
  "License": "CC0",
  "BIDSVersion": "1.1.1"
}
```


Of the four fields, only "Name" and "BIDSVersion" are required. A full list of fields for dataset_description.json is available in the BIDS Specification.

### README

> "This meditation experiment contains 24 subjects. Subjects were meditating and were interupted about every 2 minutes to indicate their level of concentration and mind wandering. The scientific article (see Reference) contains all methodological details
>
> -Arnaud Delorme (October 17, 2018)"

A README file is a description of the dataset. It is recommended by BIDS but must be in ASCII or UTF-8 encoding (text file).

### CHANGES

> "Revision history for meditation dataset
>
> **version 1.0 beta - 17 Oct 2018**
>
> - Initial release
>
> **version 2.0 - 9 Jan 2019**
>
> - Fixing event field names and various minor issues"

This is an optional text file detailing any changes, updates, and corrections made to the dataset, i.e. a version history. Note that BIDS requires adherence to the CPAN Changelog convention.

### Participants file

This is an optional document that contains a table of participants and their properties as a Tab-Separated Values (TSV) file.

You may notice that a JSON file of the same name also exists, which simply contains descriptions of each column.


```
{
  "gender": {
    "Description": "sex of the participant",
    "Levels": {
      "M": "male",
      "F": "female"
    }
  },
```


(continues on next page)

(continued from previous page)

```
  "participant_id": {
    "Description": "unique participant identifier"
  },
  "age": {
    "Description": "age of the participant",
    "Units": "years"
  },
  "group": {
    "Description": "group, expert or novice meditators",
    "Levels": {
      "expert": "expert meditator",
      "novice": "novice meditator"
    }
  }
}
```

### Code

The folder `code/` contains a MATLAB script called `run_mw_experiment6.m`. In general, this folder should contain scripts used on the dataset, such as deface.py and other anonymization tools. Identifiable information should be eliminated.

### BIDS Folder Hierarchy

Before examining the data, let's take a look at the four levels of the BIDS folder hierarchy.

### Project

In this case, it is `eeg_rishikesh/`.

### Subject

It must have the following structure:

```
sub-<participant label>
```

with *participant label* typically being a padded number.

### Session

It must have the following structure:

```
ses-<session label>
```

with *session label* typically being a padded number.

According to the BIDS specification, a session is

> "a logical grouping of neuroimaging and behavioral data consistent across subjects. Session can (but doesn't have to) be synonymous to a visit in a longitudinal study. In general, subjects will stay in the scanner during one session. However, for example, if a subject has to leave the scanner room and then be re-positioned on the scanner bed, the set of MRI acquisitions will still be considered as a session and match sessions acquired in other subjects. Similarly, in situations where different data types are obtained over several visits (for example fMRI on one day followed by DWI the day after) those can be grouped in one session. Defining multiple sessions is appropriate when several identical or similar data acquisitions are planned and performed on all -or most- subjects, often in the case of some intervention between sessions (e.g., training)."

### Acquisition

This can be one of eight types:

```
func (task based & resting state functional MRI)
dwi (diffusion weighted imaging)
fmap (field inhomogeneity mapping data such as field maps)
anat (structural imaging such as T1, T2, etc)
meg (magnetoencephalography)
beh (behavioural)
eeg (electroencephalography)
ieeg (intracranial electroencephalography)
```

In this case, it is `eeg/`.

The location of a raw data file in BIDS format will therefore have the following structure:

```
<project>/sub-<participant label>/ses-<session label>/<acquisition>/<data>
```

### Modality Specific File: EEG

Notice the 24 subject folders; one for each of the 24 participants. The folders are almost identical in content so let's focus on one of them.

The folder `sub-017/` contains two sessions:

```
ses-01
ses-02
```

There is only one acquisition per session, `eeg/`. Let's take a look inside `ses-01/eeg`.

```
sub-017_ses-01_task-meditation_channels.tsv
sub-017_ses-01_task-meditation_eeg.bdf
sub-017_ses-01_task-meditation_eeg.json
sub-017_ses-01_task-meditation_events.tsv
```

Note the structure of the file names. The template for EEG data is given in the BIDS Specification and re-stated below.

```
sub-<label>/
    [ses-<label>]/
      eeg/
        sub-<label>[_ses-<label>]_task-<label>[_acq-<label>][_run-<index>]_eeg.
↪<manufacturer_specific_extension>
        sub-<label>[_ses-<label>]_task-<label>[_acq-<label>][_run-<index>]_eeg.json
```

### Sidecar JSON

The metadata file `sub-017_ses-01_task-meditation_eeg.json` contains the following information describing the corresponding raw data file of the same name.


```
{
  "InstitutionAddress": "Centre de Recherche Cerveau et Cognition, Place du Docteur␣
↪Baylac, Pavillon Baudot, 31059 Toulouse, France",
  "InstitutionName": "Paul Sabatier University",
  "InstitutionalDepartmentName": "Centre de Recherche Cerveau et Cognition",
  "PowerLineFrequency": 50,
  "ManufacturersModelName": "ActiveTwo",
  "TaskName": "meditation",
  "EEGReference": "CMS/DRL",
  "Manufacturer": "BIOSEMI",
  "EEGChannelCount": 64,
  "MiscChannelCount": 15,
  "RecordingType": "continuous",
  "RecordingDuration": 2787,
  "SamplingFrequency": 256,
  "EOGChannelCount": 0,
  "ECGChannelCount": 0,
  "EMGChannelCount": 0,
  "SoftwareFilters": "n/a"
}
```


"TaskName" is a required generic field, while "EEGReference", "SamplingFrequency", "PowerLineFrequency", and "SoftwareFilters" are required specific EEG fields. The rest are recommended fields, either generic or specific to EEG. The full list is in the BIDS Specification.

### Raw data

The file `sub-017_ses-01_task-meditation_eeg.bdf` stores the EEG data. BIDS allows four EEG formats:

- Europena data format (.edf)
- BrainVision data format (.vhdr, .vmrk, .eeg)
- EEGlab, a MATLAB toolbox (.set, .fdt)
- Biosemi data format (.bdf)

If the original data is not in one of the formats above, then it must be made available in a separate folder called `/sourcedata` (more on this at the end of the demo).

### Channels description

Let's check out `sub-017_ses-01_task-meditation_channels.tsv`.

This table has all three of the required columns in the proper order: name, type, and units. Other recommended columns as well as a list of types is in the BIDS Specification.

### Task Events

BIDS defines an event as "a stimulus or subject response recorded during a task" (see the Common Principles section in the specification for more definitions). Since this study is on mind wandering during meditation, it involves stimuli and responses which is why there exists a file called `sub-017_ses-01_task-meditation_events.tsv`.

Note that Task Events are classified separately from EEG.

The first two columns, onset and duration, are required whereas trial_type, response_time, sample, and value are optional.

One might ask, *'Where is the accompanying JSON file?'*. Now is a good time to introduce the Inheritance Principle.

### The Inheritance Principle

The Inheritance Principle states that values from a metadata file defined at the top directory level are inherited by all lower levels unless overriden by a file at a lower level.

Return to the project folder `eeg_rishikesh/`. Observe that the JSON file `task-meditation_events.json` for the events file discussed above lives here. This file applies to all participants therefore it is defined at the top level and its values are then *inherited* by all subfolders.

```
{
  "onset": {
    "Description": "Event onset",
    "Units": "second"
  },
  "duration": {
    "Description": "Event duration",
    "Units": "second"
  },
  "trial_type": {
    "Description": "Type of event (different from EEGLAB convention)",
    "Levels": {
      "stimulus": "Onset of first question",
      "response": "Response to question 1, 2 or 3"
    }
  },
  "response_time": {
    "Description": "Response time column not use for this data"
  },
  "sample": {
    "Description": "Event sample starting at 0 (Matlab convention starting at 1)"
  },
  "value": {
    "Description": "Value of event (numerical)",
    "Levels": {
      "2": "Response 1 (this may be a response to question 1, 2 or 3)",
      "4": "Response 2 (this may be a response to question 1, 2 or 3)",
      "8": "Response 3 (this may be a response to question 1, 2 or 3)",
      "16": "Indicate involuntary response",
      "128": "First question onset (most important marker)"
    }
  }
}
```

As expected, it describes each column of `sub-017_ses-01_task-meditation_events.tsv`.

A more detailed explanation of the Inheritance Principle is in the Common Principles. section of the BIDS specification.

#### Stimuli

Last but not least, the audio files used as stimuli are stored in `eeg_rishikesh/stimuli`.

#### Source and Derived Data

Other datasets in the GitHub repository bids-examples have `sourcedata/` and `derivatives/` folders. The BIDS format segregates **raw** (unprocessed), **source** (pre-file format conversion or pre-harmonization), and **derived** (processed) data into separate folders to prevent accidental changes to the raw data.

These folders preserve the same folder and file name structure as the the folders containing the raw data.

#### References

Original paper: Gorgolewski KJ, Auer T, Calhoun VD, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci Data. 2016;3. doi:10.1038/sdata.2016.44

BIDS Extension Proposal: Pernet CR, Appelhoff S, Flandin G, Phillips C, Delorme A, Oostenveld R. BIDS-EEG: an extension to the Brain Imaging Data Structure (BIDS) Specification for electroencephalography. PsyArXiv. 2018. doi:10.31234/osf.io/63a4y

## 2.12 HPC Working Environments

Instructions and recommendations on setting up your working environments and workflows on High Performance Computing (HPC) systems such as:

- Digital Research Alliance of Canada (formerly Compute Canada and Westgrid clusters)
- UBC ARC Chinook
- UBC ARC Sockeye

These systems use similar software configurations. Where procedures differ between systems, this will be indicated. Otherwise, a generalized approach will be used in the documentation that follows.

This documentation was written to accommodate users on a variety of operating systems, including Windows, MacOS and Linux distributions. Corrections, further documentation, feedback and other contributions are highly welcomed.

### 2.12.1 UBC Advanced Research Computing Resources

#### UBC ARC Chinook

UBC ARC Chinook is an object storage platform, available to UBC researchers by application. With an initial 5 PB of storage, Chinook is an integral part of UBC's efforts to significantly increase storage capacity to meet the immediate needs of UBC researchers and supplement the storage resources available through the national platforms.

#### UBC ARC Sockeye

UBC ARC Sockeye ("Sockeye") is a high-performance computing platform available to UBC researchers accross all disciplines by application. With nearly 16,000 CPU cores and 200 GPUs, Sockeye is designed to significantly increase UBC's computing capacity and supplement the national platform for digital research infrastructure (DRI) in order to meet the immediate needs of UBC researchers.

#### Signing Up

Applications for Sockeye can be found under the Sockeye Resource Allocation section on UBC's ARC Sockeye page. Applications for Chinook can be found under the Who Can Apply - Chinook Resource Allocations section on UBC's ARC Chinook page.

### 2.12.2 Digital Research Alliance of Canada

The Digital Research Alliance of Canada (formerly Compute Canada and Westgrid) is a provider of Advanced Research Computing infrastructure, including systems, storage and software. The Alliance provides heterogeneous, general purpose clusters and Clouds that allow researchers to access resources such as CPU and GPU time, software, as well as different storage systems.

#### Signing Up

You can apply for a Compute Canada account following this documentation

---

**Note:**

- If you are a PI, please create an account as other lab members must be sponsored under your account.
- If you are a lab member, contact your PI for their Compute Canada Role Identifier (CCRI) so that you can complete your application. Your PI must then confirm your role for your account to be created.

---

---

**Tip:** There is a wealth of documentation provided by Compute Canada and WestGrid that should serve to fill in any gaps left by these docs.

---

### 2.12.3 Logging in remotely via SSH

#### Linux and MacOS

Test availability of ssh by entering the following command:

```
$ ssh -V
```

You should see something like

```
OpenSSH_7.9p1, OpenSSL 1.1.1a  20 Nov 2018
```

Log into the remote computer

```
$ ssh <username>@<remote computer>
```

With `<username>` replaced with your username on `<remote computer>` Input your password and press enter.

### Examples

### Compute Canada Clusters

```
# cedar (SFU)
$ ssh <username>@cedar.computecanada.ca

# graham (U Waterloo)
$ ssh <username>@graham.computecanada.ca

# beluga (ETS Montreal)
$ ssh <username>@beluga.computecanada.ca
```

### Windows

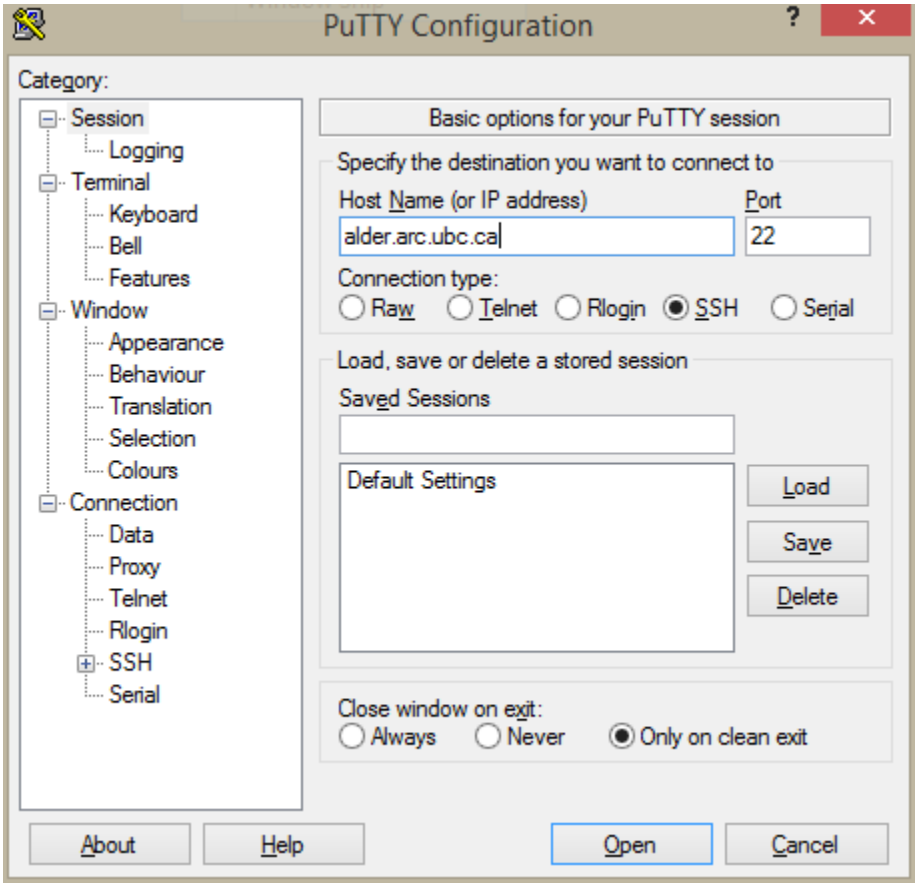
### A. Using Git Bash

Download and install Git Bash (Highly Recommended). Once you have it set up, you can use the same process as in the previous section.

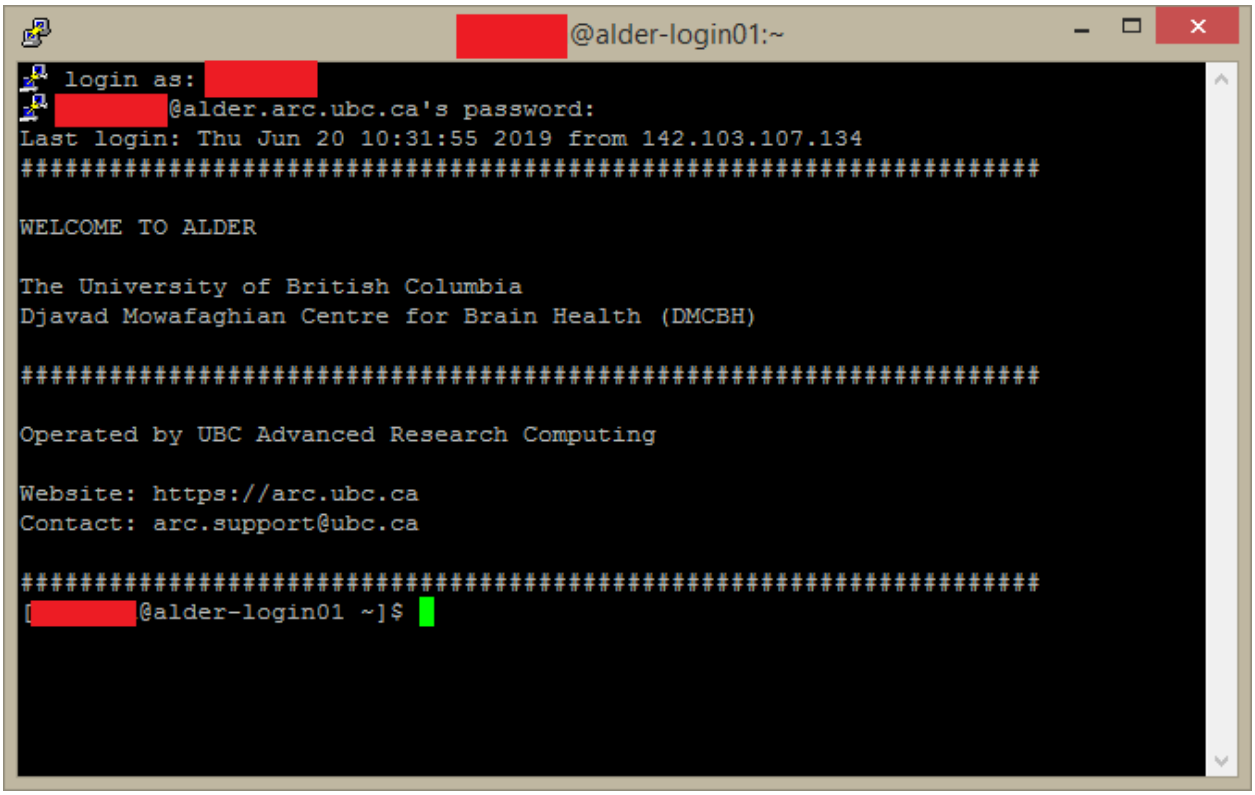
### B. Using PuTTY

Download and install the latest version of PuTTY from here

Open up PuTTY. Under the **Session** section, set the **Host Name** to the appropriate server e.g. `cedar.computecanada.ca`, and so on.




Click on **Open**. This should open up a terminal. Input your login credentials. If the login is successful, you should see something like this:




### 2.12.4 Set Up an SSH Key

SSH keys allow you to conveniently authenticate to remote computers by allowing you to connect to them without entering your passsword each time.

This procedure is meant to be done on your local machine.

---

**Note:** This procedure should work for Windows, MacOS and Linux users. If you are on Windows and the command below does not work in PowerShell, you must install Git Bash.

---

#### 1) Generate the Key

Open the git bash command line and enter:

```
$ ssh-keygen
```

Follow the prompts. It is highly recommended that you use a secure password for your key. It is recommended that you use the default location suggested to stopre the key. You should see something like this:

```
Generating public/private rsa key pair.
Enter file in which to save the key (/c/Users/user/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /c/Users/user/.ssh/id_rsa.
Your public key has been saved in /c/Users/user/.ssh/id_rsa.pub.
The key fingerprint is:
```

(continues on next page)

(continued from previous page)

```
SHA256:2mrihQwvnsIx8gDp8pO01EmCT4PgnTSVgv8Xj1vKNOg user@May2015b
The key's randomart image is:
+---[RSA 2048]----+
|  . ...          |
|.. + .           |
|o++ +            |
|=.++.  .         |
|oo.=... S        |
|++++o+ B o       |
|+*+o= * *        |
|.o*o.E.=         |
| .oo.o.          |
+----[SHA256]-----+
```

We can check the contents of the `.ssh` directory with

```
$ ls ~/.ssh
id_rsa  id_rsa.pub
```

`id_rsa` is the private key and `id_rsa.pub` is the public key

### 2) Add key to `ssh-agent`

Start the agent by running

```
$ eval "ssh-agent"
```

Add the key

```
# Windows and Linux
$ ssh-add ~/.ssh/id_rsa # or wherever your private key is stored
```

```
# MacOS
$ ssh-add -K ~/.ssh/id_rsa # or wherever your private key is stored
```

For MacOS to remember your private key password, create a file called `~/.ssh/config` and input the following:

```
Host *
UseKeychain yes
```

### 3) SSH agent forwarding

If a `~/.ssh/config` does not already exist, create it. Add to the file the following:

```
Host cedar
  Hostname  cedar.computecanada.ca
  User  username
  ForwardAgent  yes

Host graham
  Hostname  graham.computecanada.ca
  User  username
  ForwardAgent  yes
```

(continues on next page)

(continued from previous page)

```
Host beluga
  Hostname  beluga.computecanada.ca
  User  username
  ForwardAgent  yes
```

Where `username` is your username on the remote computer. You can add other blocks like these for other remote computers if you wish.

Now, you should be able to log in to a remote machine using only `ssh <host>` instead of `ssh <username@host.address.com>` e.g.

```
$ ssh cedar
```

Instead of

```
$ ssh user@cedar.arc.ubc.ca
```

#### 4) Install your ssh public key on the remote machines

Copy your public key to each of the remote machines in your `~/.ssh/config` file, for instance:

```
$ ssh-copy-id -i $HOME/.ssh/id_rsa cedar
```

You will be prompted for your password on the remote machine and the key will be installed.

Once your key is installed, you should be able to run commands like `ssh`, `scp`, `sftp` and `rsync` without having to enter your password.

**Note:** You may be prompted for the password to your key when you first log into the remote server via `SSH`

#### 5) (Optional) Add SSH key to GitHub Account

You can also add your ssh key to your GitHub account to avoid entering your password each time you push to the cloud. Instructions are available here.

### 2.12.5 Loading Software

#### Working with Modules

On Compute Canada Systems, a list of all available modules can be obtained by running:

```
$ module spider
```

To see detailed information about a module, run:

```
$ module whatis <module name>
```

To load a module, for instance a module called `python`, run:

```
$ module load python
```

To see a list of modules you have loaded, run:

```
$ module list
```

See also: List of software available on Compute Canada

#### Bash Configuration

If you are working in Bash, you can configure your account to automatically load the modules you need each time you log in.

#### 1) Loading modules

Add to your `$HOME/.bashrc` file (on the remote computer) bash commands you want to execute when you log in. For loading some modules, for instance, there is the following example:

```
module load python
module load machinelearning
module load hdf5
```

#### 2) Ensure `.bashrc` is run when you log in

Create a file called `$HOME/.bash_profile` (on the remote computer) the following:

```
if [ -f ~/.bashrc ]; then
    . ~/.bashrc;
fi
```

### 2.12.6 Setting up a Python Environment

**Tip:** You can also use these instructions to set up python environments on your personal computer

#### Setting up a python environment using Anaconda

The Anaconda distribution gives you access to many commonly used libraries and is the fastest way to set up a ready to use python environment. It is also the recommended way of installing jupyter.

Begin by going to the Anaconda website and copying the link to the lastest Linux '64-Bit (x86) Installer.

Download it using `wget "<link to installer file>"` e.g.

```
$ wget "https://repo.anaconda.com/archive/Anaconda3-2019.03-Linux-x86_64.sh"
```

Install Anaconda using `bash <name of the file you just downloaded>`, then following the installer's instructions. It is highly recommended to use the default settings provided by the installer.

```
$ bash Anaconda3-2019.03-Linux-x86_64.sh
```

### Installing conda only

The links to the Miniconda installer can be found here, and can be installed using the same procedure as in the previous section.

Miniconda only includes python and the conda package manager.

### Setting up a custom python environment

#### 1. Using conda

Conda combines pip and venv and offers a better user experience. This excellent tutorial from UoA e-research gives concise instructions on setting up virtual environments using conda.

#### 2. Using venv

venv comes into play if you prefer not to work with conda.

Replace 'myvenv' with whatever path you want to keep you virtual environment in.

```
# create the venv
$ python3 -m venv myvenv
# activate the venv
$ source myvenv/bin/activate
# you should now see the name of our venv in the terminal prefix
(myvenv) [user@<cluster> ~]$
```

Upgrade pip

```
$ pip install --upgrade pip
```

You can view the packages you have install into your virtual environment using `pip --list`. You should see something like:

```
Package    Version
---------- -------
pip        19.1.1
setuptools 28.8.0
```

Now we can install some packages

```
# install a package
$ pip install numpy
# install many packages at once
$ pip install scipy matplotlib
# install a particular version of a package
$ pip install h5py==2.9.0
```

You can deactivate the virtual environment with

```
$ deactivate
```

Whenever you want to use the environment again, activate it using `source <path to venv activate>`, e.g.

```
$ source myvenv/bin/activate
```

If you want to save your environment configuration, use `pip freeze`

```
$ pip freeze > requirements.txt
```

You can also use a requirements file to build or modify your virtual environment

```
$ pip install -r requirements.txt
```

### 2.12.7 Run Jupyter Remotely

In some instances, you may need to run Jupyter (notebook or lab) on a remote computer to use its resources but open the web interface on your local machine.

#### Syzygy

Syzygy is an interactive computing environment that allows access to Jupyter notebooks that combine rich text with code in a languages such as R and Python. You can log into Syzygy by using your university credentials. To do so, navigate to the top right drop down menu labelled "LAUNCH", and select your university.

## 2.13 File Transfer and Backups

This section provides:

### 2.13.1 Backup Snapshots

To prevent the loss of work and data, as well as to keep it secure and enable versioning, the use of backup software is recommended for all live storage in a lab, including data storage drives, servers and workstations.

Backup snapshots create images of your filesystems at configurable intervals, for instance, storing hourly snapshots for 7 days, storing weekly snapshots for 6 months and monthly snapshots indefinitely. File duplication is minimised as unchanged files are not backed up again, and features such as checksum verification exist to maintain integrity. In addition, the files can be encrypted to maintain high levels of security. This enables researchers to easily revert their filesystem to its state in an earlier snapshot or retrieve an earlier version of a file. The number of redundancies can also be configured, enabling automatic backups across different backup locations and media.

There are also free open source software solutions available, including Duplicati, Restic and Borg. Various paid software solutions are also available, such as Crashplan and Acronis True Image. Labs considering such an automated backup solution should consider their needs and resource availability when making their decision.

### 2.13.2 Tarballing Files

Tarballing is the process of bundling multiple files into one file. It is Compute Canada's recommendation that a collection of multiple small files be tarballed into large files. It is also the recommended way to store files in Nearline since smaller files are not written to tape. Systems like Cedar, which have a conservative number of files allowed in Nearline (5K) also warrant tarballing to keep the file count low.

A tarball can also be compressed using the gunzip command. This has its limitations, such as not being able to add files to an existing gunzipped tarball. It is also unlikely to reduce the size of certain file formats like videos and images.

### Working with Tarballs (CLI)

**Note:** You can also use a GUI of your choice. The CLI is often the only choice for working over `SSH`.

### Tarballing/Compressing Files

To make a tarball

```
# tarball
$ tar -cvf <.tar file> /path/to/folder/to/tarball/
# e.g. $ tar -cvf example.tar videos/

# gunzipped tarball
$ tar -cvzf <.tar.gz file> /path/to/folder/to/gunzip/
# e.g. $ tar -cvzf example.tar.gz videos/
```

Adding files to existing archive (only `.tar` files)

### Extracting files

To extract a tarball

```
$ tar -xvf <.tar file> -C </directory/to/extract/to/>

# or to extract in current directory
$ tar -xvf <.tar file>
```

To extract a gunzipped tarball

```
$ tar -xvzf <.tar.gz file> -C </directory/to/extract/to/>
# e.g. $ tar -xvzf

# or to extract in current directory
$ tar -xvzf <.tar.gz file>
```

Extract particular file/folder from tarball/gunzipped tarball

```
$ tar --extract <.tar/.tar.gz file> <path/to/file>
# e.g. $ tar --extract example.tar example.txt
```

Extract multiple files

```
# tarball
$ tar -xvf <.tar file> "<path/to/file1>" "<path/to/file2>"

# gunzipped tarball
$ tar -xvzf <.tar.gz file> "<path/to/file1>" "<path/to/file2>"
```

Extract files by wildcard

```
# tarball
$ tar -xvf <.tar file> --wildcards "*.mp4"

# gunzipped tarball
$ tar -xvzf <.tar.gz file> --wildcards "*.mp4"
```

#### Viewing Contents of Archive

To view contents of a tarball/gunzipped tarball

```
$ tar -tvf <.tar/.tar.gz file>
```

### 2.13.3 Data Transfer and Backup on Remote Computers

This section covers the use of tools like `scp`, `Rsync`, `sftp` and Globus to make transfers and backups. While the examples place an emphasis on Compute Canada, they can be generalised to apply to any remote computer.

#### `rsync-time-backup`

`rsync-time-backup` is a utility available here that builds on top of `Rsync`.

It allows backups over SSH, backup resumes, uses hardlinks to avoid duplication and provides functionality to create full backups. In short, it can be used to create full snapshots, the frequency of which can be configured e.g. keep each hour of the last 24 hours, each day of the previous month and each month in the previous year.

You may find it useful to combine several techinques in the examples below and in the docmentation of this script.

#### Example: Backing up to Compute Canada Project Space

The project space is a great place for:

- Frequent backups
- Internal and external sharing

You may want to tarball collections of small files (multiple files < 100 MB) to keep the file count down and to keep the diretory uncluttered.

---

**Note:** This works if you set up the SSH key on the Compute Canada server

---

In this example, we are backing up a drive mounted to our system. In this instance, it is a NAS drive, althogh it could be a hard drive or any other external media.

```
$ rsync_tmbackup.sh /media/user/data cedar:/home/user/projects/def-pi/data/
```

You can replace `cedar` with `graham` or `beluga` as required

### Example: Backing up to Remote Computer

Suppose you have an `ssh` machine `computer.domain.com`

```
$ rsync_tmbackup.sh /media/user/data user@computer.domain.com:/data/
```

### Example: Backing up using an exclusion list

Suppose you do not want to backup certain directories, files or file types. You must create an exclusion list `exclude-file.txt`, for instance:

```
secrets.txt
folder1/*
folder2
folder3/scratch/
*.mat*
```

This will exclude the file `secrets.txt`, copy the folder `folder1` but not its contents, will not copy the folder `folder2` or its contents, will not copy the subdirectory `folder3/scratch/` or its contents, and will exclude all files with the `.mat` extension.

Local machine to external drive

```
$ rsync_tmbackup.sh /home/user /media/user/data/ excluded_patterns.txt
```

Compute Canada

```
$ rsync_tmbackup.sh /media/user/data cedar:/home/user/projects/def-pi/data/ excluded_
↪patterns.txt
```

### Example: Change the Expiration Strategy

From the README:

```
The default strategy is ``1:1 30:7 365:30``, which means:

* After 1 day, keep one backup every 1 day (1:1).
* After 30 days, keep one backup every 7 days (30:7).
* After 365 days, keep one backup every 30 days (365:30).
```

To change the strategy to: After 30 Days, keep one backup every 14 days,

#### STFP

From the Compute Canada Documentation:

```
SFTP (Secure File Transfer Protocol) uses the SSH protocol to transfer files between␣
↪machines which encrypts data being transferred.
```

Unlike `SCP`, `SFTP` omes with an interactive prompt.

### Dropping into the SFTP Prompt

```
$ sftp user@remote_hostname_or_ip_address
```

For instance,

```
$ sftp john@cedar.arc.ubc.ca
```

If you set up your SSH key on the remote computer, you won't even need a password.

```
$ sftp cedar
```

If it worked, you should be in the prompt e.g.

```
$ sftp cedar
Connected to cedar.
sftp>
```

### Exiting the SFTP Prompt

```
sftp> exit

or

sftp> bye
```

### Help

```
sftp> help

or

sftp> ?
```

### Navigating the Remote and Local Machines

### Current working directory

Remote Machine

```
sftp> pwd
```

Local Machine

```
sftp> lpwd
```

### List directory contents

Remote Machine

```
sftp> ls
```

Local Machine

```
sftp> lls
```

### Change working directory

Remote Machine

```
sftp> cd <remote dir>
```

Local Machine

```
sftp> lcd <local dir>
```

### Make new directories

Remote Machine

```
sftp> mkdir <remote dir>
```

Local Machine

```
sftp> lmkdir <local dir>
```

### Transferring File to Remote

```
sftp> put <local file or directory> <new name on remote [OPTIONAL]>
```

e.g.

```
sftp> put data.hdf5
Uploading data.hdf5 to /project/6006382/user/data.hdf5
data.hdf5                                     100%   11GB  100.3MB/s   01:50
```

```
sftp> put data.hdf5 data_20181012.hdf5
Uploading data.hdf5 to /project/6006382/user/data_20181012.hdf5
data.hdf5                                     100%   11GB  100.3MB/s   01:50
```

### Transferring File from Remote

```
sftp> get <remote file or directory> <new name on local [OPTIONAL]>
```

e.g.

```
sftp> get data.hdf5
Fetching /project/6006382/user/data.hdf5 to data.hdf5
/project/6006382/user/data.hdf5               100%   11GB  100.1MB/s   01:50
```

```
sftp> get data.hdf5 data_20181012.hdf5
Fetching /project/6006382/user/data.hdf5 to data_20181012.hdf5
/project/6006382/user/data.hdf5                100%   11GB  100.1MB/s   01:50
```

#### SCP

From the Compute Canada Documentation:

```
SCP stands for "Secure Copy". Like SFTP it uses the SSH protocol to encrypt data␣
↪being transferred. It does not support synchronization like Globus or rsync. Some␣
↪examples of SCP use are shown here.

SCP supports an option, -r, to recursively transfer a set of directories and files.␣
↪We recommend against using scp -r to transfer data into /project because the setgid␣
↪bit is turned off in the created directories, which may lead to Disk quota exceeded␣
↪or similar errors if files are later created there.
```

### Basic Usage

```
$ scp <location/file to copy from> <location/file to copy to>
```

### Transferring Files

Suppose a folder in your current local working directory is as follows:

```
package/
├── package
│   ├── conf.py
│   ├── __init__.py
│   ├── models.py
├── LICENSE
├── README.md
├── setup.py
└── tests
    ├── test_interface
    |   ├── tests.py
    ├── test_models
    ├── run_tests.py
```

Running this will only copy `LICENSE`, `README.md` and `setup.py`, and nothing in the other folders or subdirectories

```
$ scp package cedar:/home/user
```

Running this will copy everything

```
$ scp -r package cedar:/home/user
```

**Note:** The above examples will only work if you set up an ssh key on the remote computer

If using the full address of the remote computer, the equivalent examples are:

```
$ scp package username@cedar.computecanada.ca:/home/user
```

```
$ scp -r package username@cedar.computecanada.ca:/home/user
```

Transferring between two remote Computers

```
$ scp graham:/home/user cedar:/home/user
```

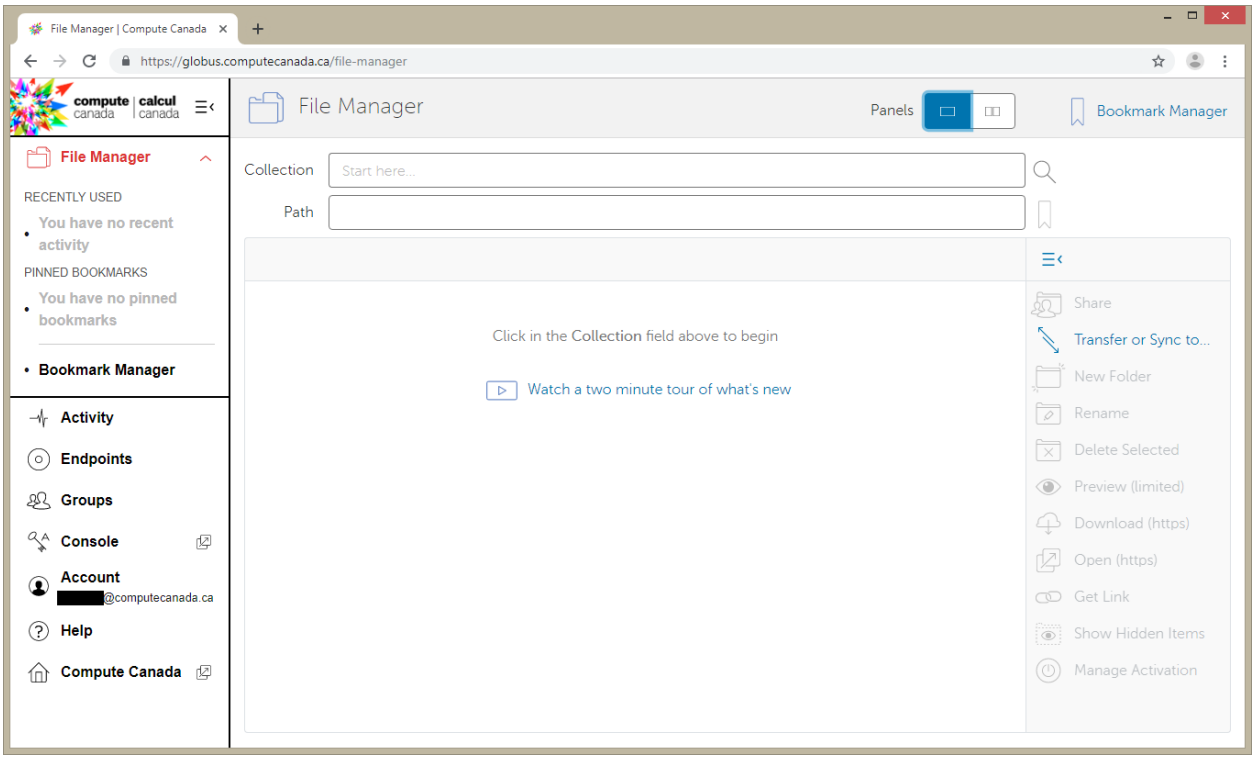
```
$ scp username@graham.computecanada.ca:/home/user username@cedar.computecanada.ca:/
↪home/user
```

## Globus

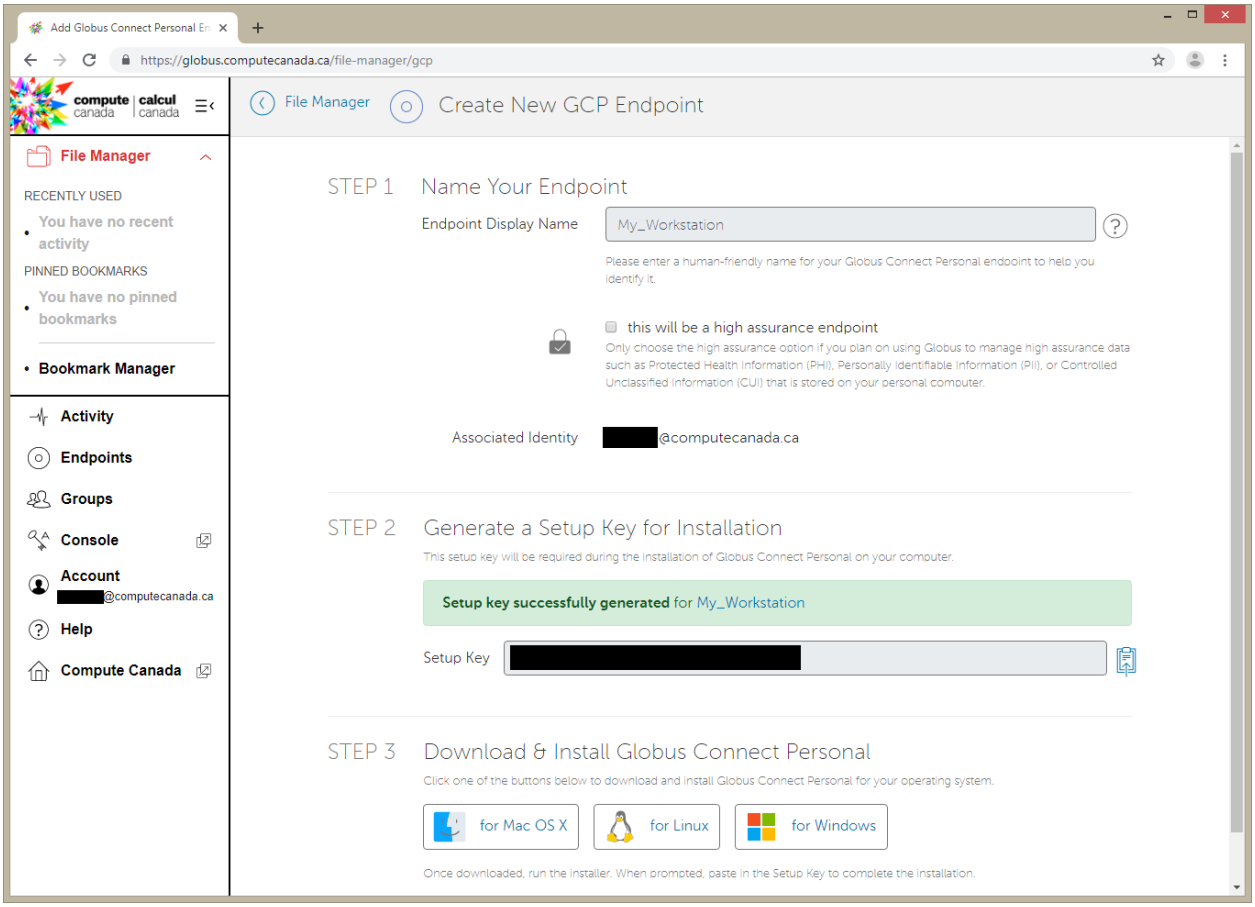
### Option 1: Globus GUI

**Tip:** This can be extended to file transfers between personal endpoints and between Compute Canada servers

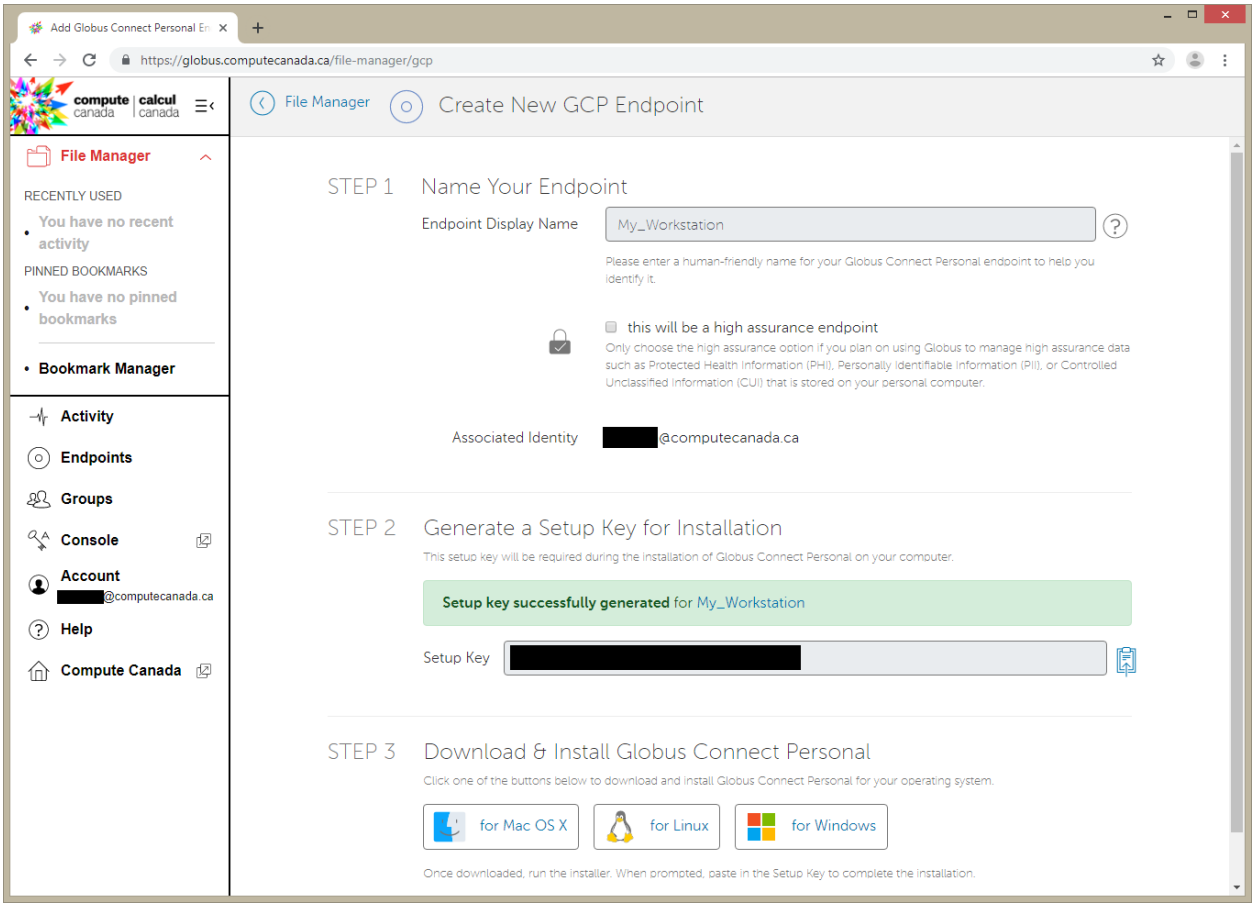
1. Log in



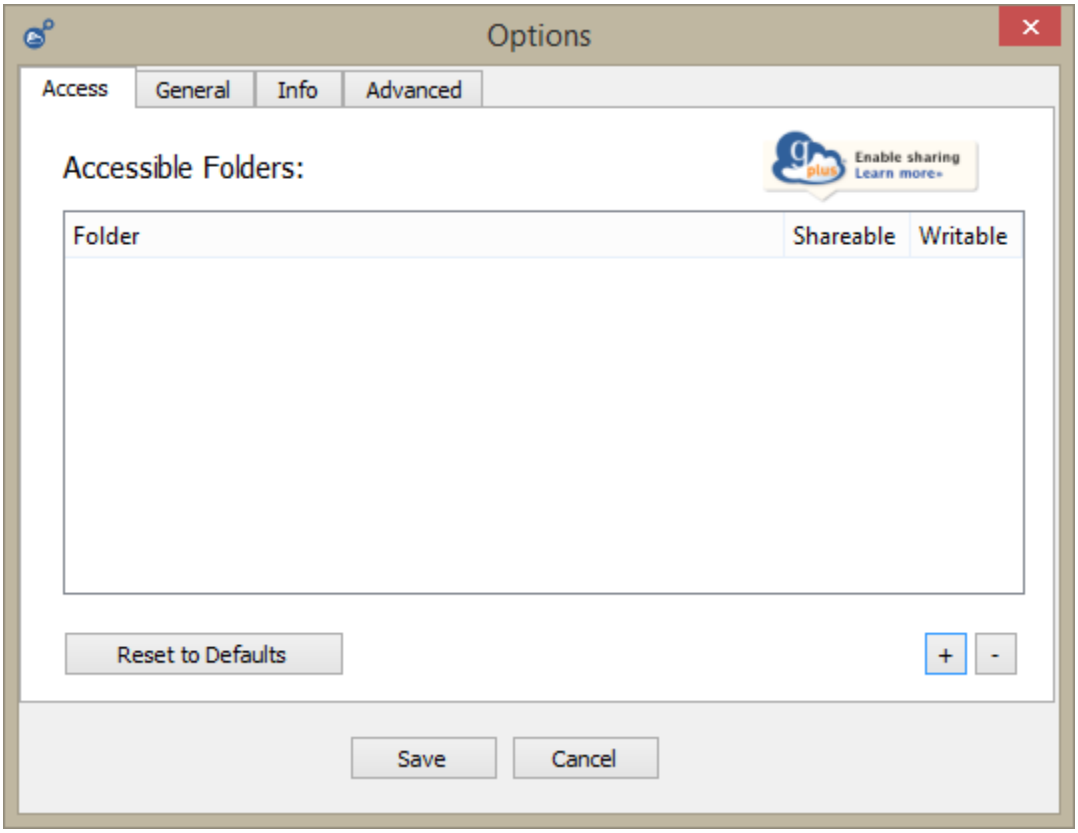

2. Click on `Endpoints`
3. Click on `Add an Endpoint` then `Globus Connect Personal`
4. Enter a name for your endpoint, e.g. `My_workstation` in this case. Check the `This will be a high assurance endpoint` box if dealing with highly sensitve data




5. Generate the Setup Key and copy it to your clipboard.
6. Using the links at the bottom of the page, install the Globus Connect Personal Client on your machine and follow the on screen instructions
7. Run the client and paste your Setup Key, then click `OK`



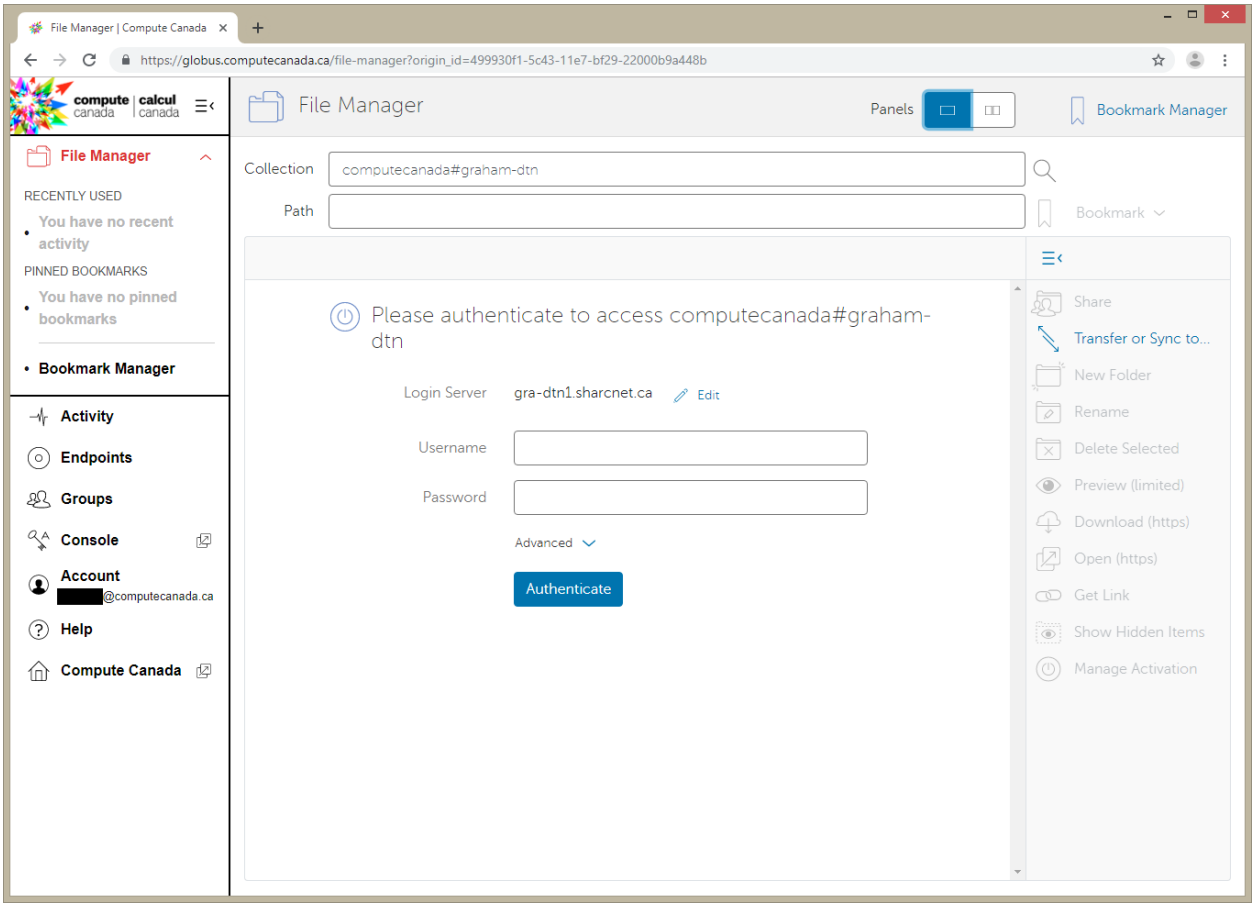

8. Add location(s) of data you want globus to be able to access by clicking on +.



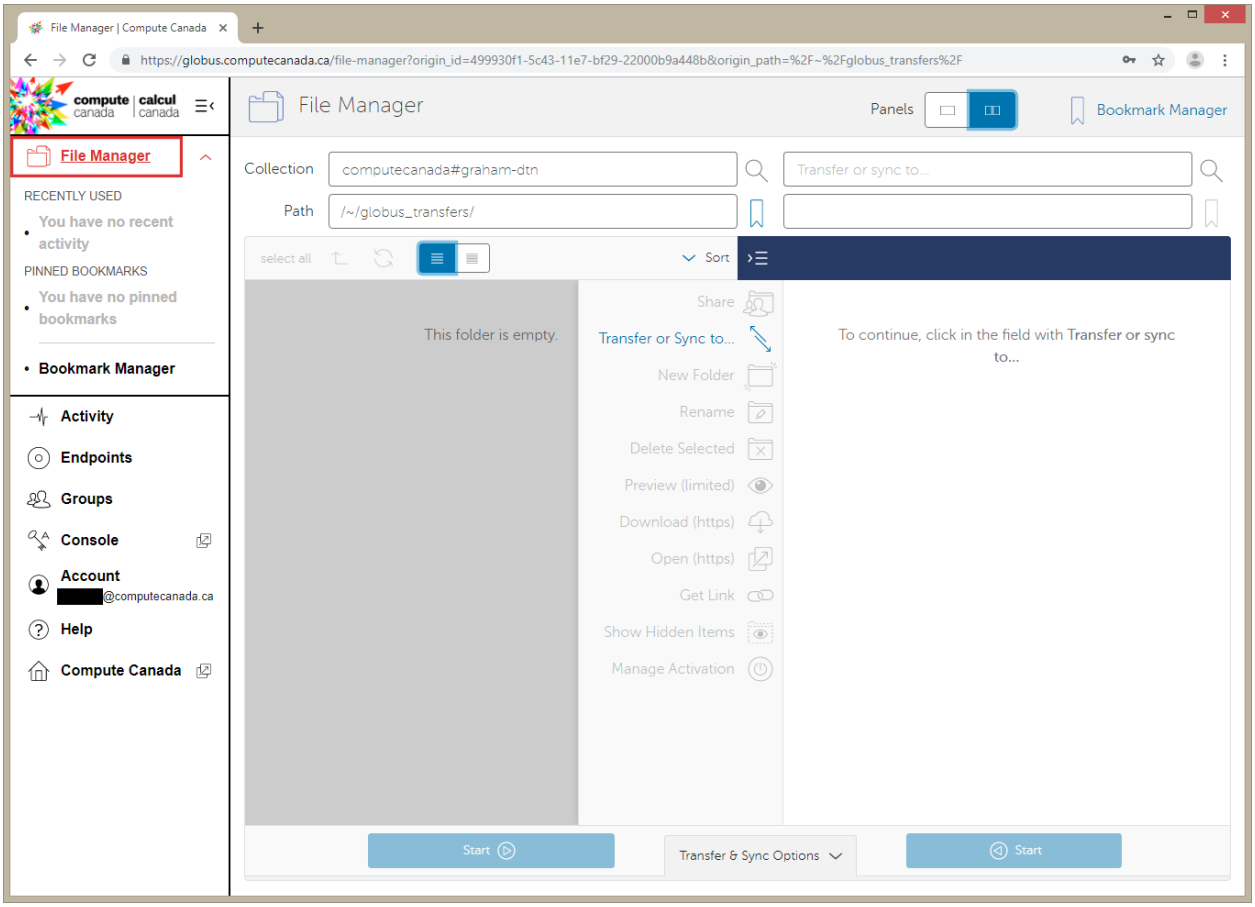

**Tip:**

- Ticking *Shareable* will allow outbound transfers
- Ticking *Writable* will allow inbound transfers

9. Click `Save` when done
10. Go back to your browser and click on the `File Manager` Tab
11. Search for the Compute Canada server you want ot upoad the files to. In this example, we are using Graham.

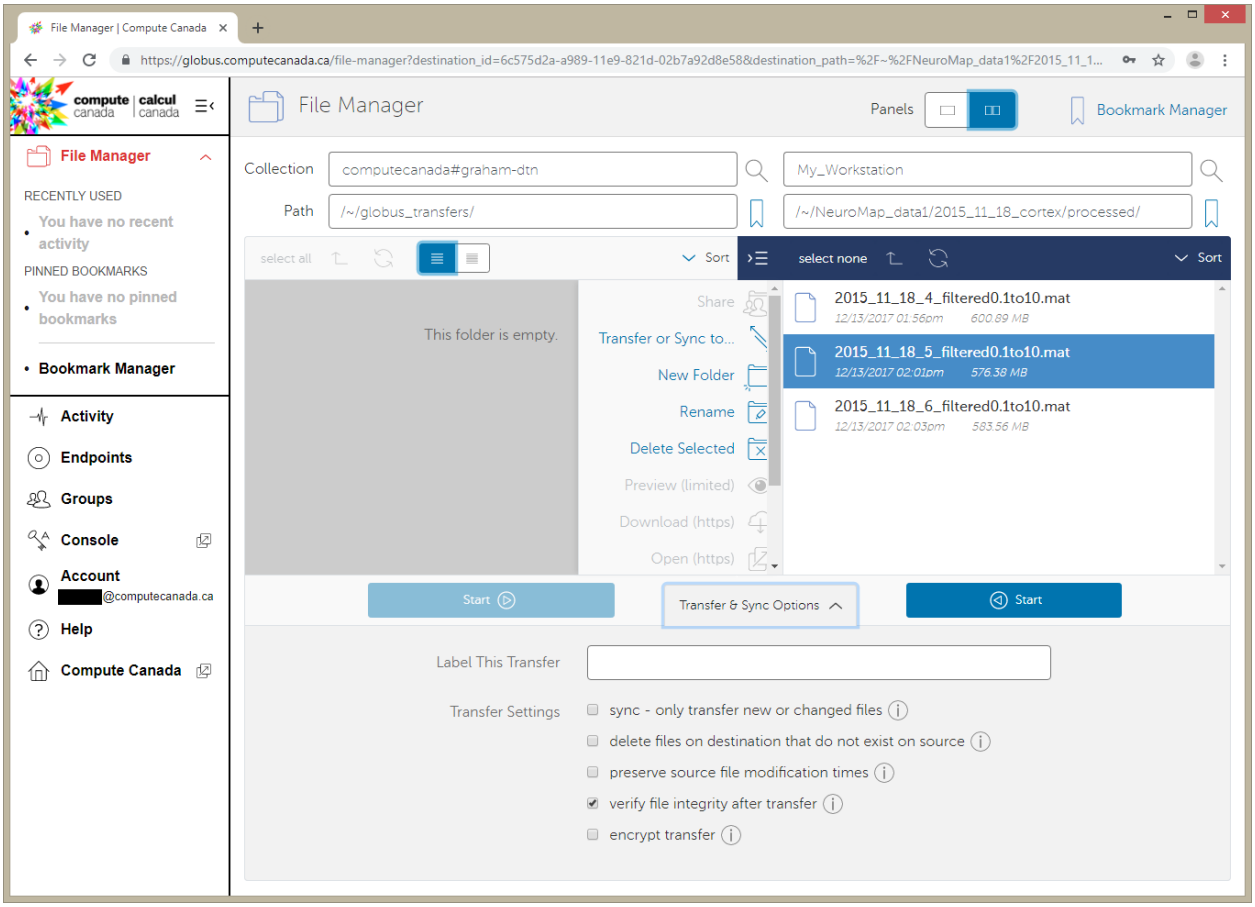



12. Log in with you compute canada credentials and click on `Authenticate`
13. Your Home directory should now be displayed. Navigate to the folder you want to keep the data in. In this example I will use the `globus_transfers` directory in my home directory.
14. Click on `Transfer or Sync to...` on the right side menu

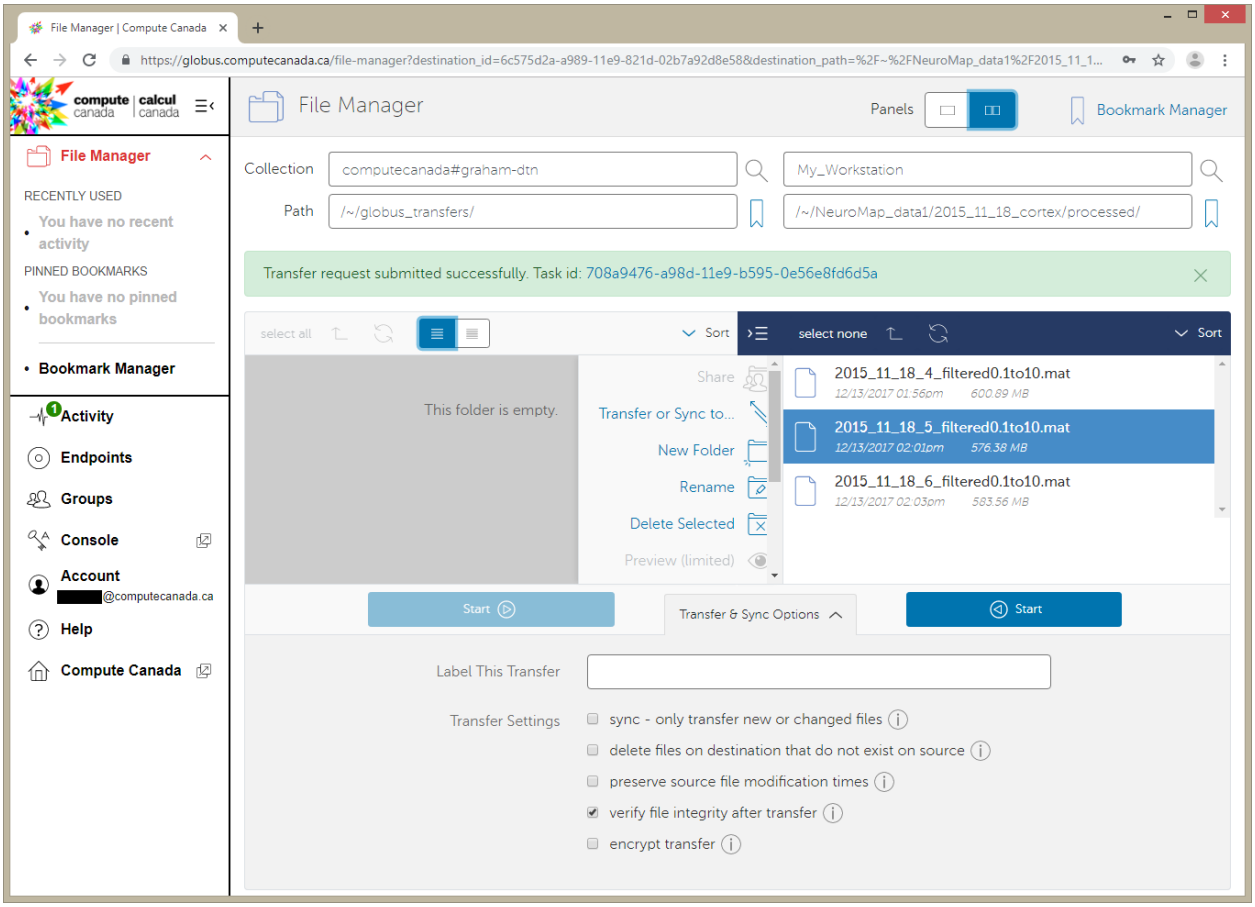



15. Click on `Transfer or sync to...` box. Click on `Your Collections` then on your desired endpoint's name.

16. Navigate the directory structure on either endpoint and select the folder(s) or file(s) you want to transfer/sync to the other endpoint. Clicking on Transfer and Sync Options below, you can select a multitude of options for managing the content on the destination endpoint. Click on `Start` when done.




17. You should see a message like: `Transfer request submitted successfully. Task id: <TASK_ID>` where <TASK_ID> is a system generated hash for your task.




18. The client on our endpoint will handle your transfer and send you an email when it is done. You can view the status of the transfer in the *Activity* tab 19. Looking at the filesystem on Graham upon completion, we can see that the file is indeed there:

```
[<user>@gra-login2 globus_transfers]$ ls -lha
total 555M
drwxr-x--- 2 <user> <user>     3 Jul 18 14:54 .
drwx------ 6 <user> <user>    13 Jul 18 14:38 ..
-rw-r--r-- 1 <user> <user> 550M Jul 18 15:02 2015_11_18_5_filtered0.1to10.mat
```

### Option 2: Globus CLI

Globus offers a command line interface, which is useful for its convenience and for automating transfers and backups. Its documentation is available here.

### Option 3: Archeion

Archeion can be downloaded here. Requirement: Globus personal must be set up on personal endpoints. It can be used to script transfers using python and provides functionality that handles authentication and transfer management.

### Git-Annex

Git-Annex uses git to create an annex, which presents files to the user in a single directory structure, even though the individual files are distributed across multiple locations. It can also be confiogured to create a number of copies of a file distributed across different annexes. This enables users to remove a local copy while ensuring redundancies are available on other storage locations. It is also able to synchronise files across redundancies. File versions can be uniquely tracked and referenced using git changeset hashes.

It also comes with a webapp interface.

---

**Note:** Git-Annex is a powerful tool but requires knowledge of git, UNIX command line and careful scripting to use effectively.

---

### Use Case Demo: Syncing files with Git-Annex using Linux CLI

The following demo was tested on Ubuntu 18.04 LTS

---

**Note:** sudo priviledges are required to install git-annex

---

Install the Git-Annex from NeuroDebian.

Create a repository in a location of your preference

```
$ mkdir annex
$ cd annex
$ git init
Initialized empty Git repository in /home/user/annex/.git/
$ git annex init
init ok
```

Add file to the repository

```
$ cd /home/user/annex/
$ cp ~/Pictures/121406.jpg ./
$ git annex add .
add 121406.jpg ok
(recording state in git...)
$ git commit -a -m added
[master (root-commit) 1259e1c] added
1 file changed, 1 insertion(+)
create mode 120000 121406.jpg
```

Add a remote, in this case, an external hard drive called ‘My Passport’

```
$ cd /media/user/USB\ DISK/
$ git clone /home/user/annex/
Cloning into 'annex'...
done.
$ cd annex
# get the desktop and the hard drive to get to know each other
$ git annex init
$ git remote add desktop /home/user/annex
$ cd /home/user/annex
$ git remote add harddrive1 /media/user/My\ Passport/annex
```

Now let's add a bunch of files to the Desktop's Annex

And now let's add a bunch of other files to the Hard Drive's Annex

Looking at the contents of the Desktop annex, we see the following:

```
$ ls
121406.jpg  121419.jpg  121421.jpg
```

Looking at the contents of the Hard drive annex, we see the following:

```
$ ls
121406.jpg  121415.jpg  121420.jpg
```

Now we need to sync the files and make sure our annexes have the same contents

Now, looking at the contents of the Desktop annex, we see the following:

```
/home/user/annex$ ls
121406.jpg  121415.jpg  121419.jpg  121420.jpg  121421.jpg
```

And also when looking at the contents of the Hard drive annex, we see the following:

```
/media/user/My Passport/annex$ ls
121406.jpg  121415.jpg  121419.jpg  121420.jpg  121421.jpg
```

This can be automated as a cron job that syncs your files with your backups in regular intervals

Refer to the documentation to learn more about setting up ssh remotes, removing and transferring files and troubleshooting.

### Downloading files from data respositories

#### FRDR

FRDR offers the option to download files using globus. Refer to the Globus GUI section above for instrutions on how to download files using Globus. Globus alkso provides the option of downloading the file(s) using a direct download link.

#### Direct Downloads

To download files using a direct downlaod link, for instance, via Dataverse, use `wget` or `curl`.

#### Example

To download the **'Neurophotonics tutorial on making connectivity diagrams from Channelrhodopsin-2 stimulated data <>'_** dataset from Dataverse, using `wget`

```
$ wget https://dataverse.scholarsportal.info/api/access/datafile/77286?gbrecs=true
```

Alternatively, you can use `curl`

```
$ curl https://dataverse.scholarsportal.info/api/access/datafile/77286?gbrecs=true --
↪output download.zip # or whatever you want to call the file you download. Keep in␣
↪mind the file format.
```

## 2.14 Data Encryption

Data encryption is a means of securing data by encoding the information such that it can only be decrypted using the appropriate encryption key. Data encryption is a requirement when storing or transmitting sensitive data, especially human data. Nevertheless, it is highly recommended to encrypt all research data. There are several ways of encrypting data:

- UBC offers full device encryption
- Using backup software with built in encryption capabilities [see previous section]. This is recommended for automation and ease of use.
- Encryption solutions such as eCryptfs

## 2.15 JupyterHub

### 2.15.1 Introduction

JupyterHub is a multi-user notebook server, allowing multiple users to access a pool of resources for interactive analysis.

Python 3.7.4 (list of packages) and R 3.6.1 (list of packages) kernels are currently installed, with multiple packages already available.

Benefits and features:

- Shared filesystem for quick collaboration
- Work anywhere from any device
- Access more powerful compute resources than a single workstation
- All standard Jupyter Notebook features, including a Text editor, file manager and Command Line Interface

---

**Note:** To use the JupyterHub, you must

1. use your Compute Canada username and password to login

To install packages, contact Jeffrey LeDue with your requirements.

---

#### Resource Allocation

Currently, when users log in to a notebook session, they can specify a certain number of cores and memory.

#### Future Directions for Resource Allocation

1) Uniform resource allocation for each user, much like the current configuration, for demonstrations and workshops. User accounts will be generated on an as needed basis.
2) Users will be given a list of configurations to choose from so as to be allocated the resources they need.
3) A schedule for resource allocation requests, whereby researchers can request dedicated resources and compute time to ensure they have sessions and resource availability when needed. This would mean that certain nodes will not be accessible through the dynamic scheduler in 2) and will only be available to a user/group of users on a manually administered queue.

**Tip:** PIs are encouraged to advocate for additional resources for cluster upgrades. This includes GPUs, storage backup, and other computational resources.

## 2.16 Coming Soon

### 2.16.1 NextCloud - Advanced Research Computing (ARC)

From the ARC website:

> We are a free service supporting the high-performance computing and data management needs of UBC researchers. Our experienced, dedicated team provides consultation, expertise, and access to cyberinfrastructure.

ARC has indicated that they plan to set up a NextCloud server service. This will be free to use for UBC researchers and will provide cloud storage, with an interface similar to offerings like Dropbox. You can get a free demo of the interface here. There have not been further comments on storage quotas or a timeline, but we are confident that we will not see this being implemented in the near future.

### 2.16.2 TeamShare for Research Data

There was a meeting planned by TeamShare management to discuss a new service, which would provide segment Teamshare into two services, with one dedicated to research data and the other for smaller files and documents. This probably would be done to cater to the different use cases more appropriately and to lower costs. However, the meeting was postponed indefinitely and we have not yet heard back from them since. Contact Jeffrey LeDue if you wish to be kept updated on the status of this proposed service.

### 2.16.3 Dataverse

While the Dataverse core development team at Harvard has added a file hierarchy preservation feature to their latest version, the team at Scholar's Portal have yet to incorporate this update into their deployment, which they will probably do in the later future.

### 2.16.4 Open Science Framework CWL Integration

UBC will soon be an OSF Institution, which will enable CWL authentication into OSF. The university will also have its own unique OSF URL as well as a landing page to increase visibility of public projects associated with UBC.

### 2.16.5 Cloud Innovation Center (CIC)

AWS has announced the launch of a CIC at UBC, which is slated to open in early 2020. Its theme is "Community Health and Wellbeing". From their website:

> [CIC] provides UBC students, staff, and faculty access to cloud technology to advance projects, along with employing Amazon's innovation processes.

Read more about it here.

## 2.17 Terms and Definitions

**Metadata** [+] "data about data". Research data file metadata includes elements such as title, file format, language, creator, and date

**Data format** [+] a particular way of encoding information within a computer file so that it can be recognized by an application

**Data Management Plan (DMP)** [+] a document you create that sets out how you will organize, store and share your research data at each stage in your project

**Data pipeline** [-] a collection of processes and systems for organizing the data, computations, and workflows used by a research group as they jointly perform complex sequences of data acquisition, processing, and analysis

**(Scientific) Workflow** [@] the description of a process for accomplishing a scientific objective, usually expressed in terms of tasks and their dependencies. Typically, scientific workflow tasks are computational steps for scientific simulations or data analysis steps.

**Raw data** [*] data directly obtained from the instrument, simulation, or survey

**Processed data** [*] result from some manipulation of the raw data in order to eliminate errors or outliers, to prepare the data for analysis, to derive new variables, or to de-identify the human participants

**Analyzed data** [*] results of qualitative, statistical, or mathematical analysis of the processed data. They can be presented as graphs, charts, or statistical tables.

**Final data** [*] processed data that have, if needed, been converted into a preservation-friendly format

**Embargo** [x] a formal request by an author to restrict access to documents or data for a specified period of time where the data contains:

- Sensitive information and/or names that cannot be released at the time of publication
- Cases that could be identified, even if anonymised
- Confidential government statistics
- Information relevant to current court cases
- Information subject to copyright or other intellectual property restrictions

[+] As defined by the UBC Library

[-] As defined by DataJoint

[@] As defined in the document Scientific Workflows by the UC Davis Genome Center

[*] As defined in the University of British Columbia Generic Template within the Portage DMP Assistant

[x] As defined in the Research Data Management: Publication page of the Curtin University Library website

## 2.18 Abbreviations

**API** Application Programming Interface

**ARC** Advanced Research Computing

**AWS** Amazon Web Services

**BIDS** Brain Imaging Data Structure

**CARL** Canadian Association of Research Libraries

**CC** Compute Canada *or* Creative Commons

**CD** Continuous Delivery

**CI** Continuous Integration

**CIHR** Canadian Institutes of Health Research

**CLI** Command Line Interface

**CONP** Canadian Open Neuroscience Platform

**COS** Center for Open Science

**CWL** Campus Wide Login

**DMCBH** Djavad Mowafaghian Centre for Brain Health

**DMP** Data Management Plan

**DOI** Digital Object Identifier

**EAD** Enterprise Active Directory

**FRDR** Federated Research Data Repository

**GCP** Google Cloud Platform

**GUI** Graphical User Interface

**MINI** Minimum Information about a Neuroscience Investigation

**NAS** Network Attached Storage

**NSERC** Natural Sciences and Engineering Research Council of Canada

**OSF** Open Science Framework

**PI** Principal Investigator

**RAC** Resource Allocation Competitions

**RAS** Rapid Access Service

**SCP** Secure Copy Protocol

**SDK** Software Development Kit

**SFTP / FTP** Secure File Transfer Protocol / File Transfer Protocol

**SSD** Solid State Drive

**SSH** Secure SHell

**SSHRC** Social Sciences and Humanities Research Council of Canada

**VCS** Version Control System

**VPN** Virtual Private Network

CHAPTER 3

# Indices and tables

- genindex
- modindex
- search

# Index