

# Canadian Neutron Beam Laboratory

## Data Management Plan

### Responsibility

The CNBL is a national research facility whose users and staff produce scientific data. The proper and secure handling, documentation, storage, sharing, processing, and publication of data are of great importance to CNBL.

Since the CNBL is part of McMaster University, it follows the policies and guidelines set up by the university and relies on IT resources that the university provides. The CNBL data management plan outlines the specific steps this facility takes to ensure data is handled according to best practices. University guidelines on data management, which serve as a basis for this document, can be found here: <https://rdm.mcmaster.ca/>

Integrity, security and archival of research data are important to ensure that CNBL-enabled research follows good scientific practice and can be used to extract the maximum amount of information which in turn can be processed into publishable knowledge. Data at CNBL is handled according to FAIR Data Principles (Findable, Accessible, Interoperable, and Reusable).

In addition to public domain research, the CNBL will also support proprietary research by commercial clients, which raises additional security requirements associated with the data. As part of its annual business planning process, CNBL management will assess the state of its data management and cybersecurity capabilities and practices relative to the expectations outlined in this plan. It will identify actions to be taken and investments to be made to close identified gaps.

### Data Collection

At the CNBL, the following types of data are distinguished:

- **Scientific Data:** These data may be generated by CNBL staff, as part of jobs, or independently by CNBL users. Scientific data can take the following forms:
  - **Raw Data:** A form of scientific data as-collected by the instrument, often using vendor-specific proprietary formats.
  - **Meta Data:** A form of scientific data such as experimental conditions or instrument parameters used to acquire the datasets, typically affiliated with another form of scientific data (e.g. raw data or processed data).
  - **Method Data:** A form of scientific data that is used to describe the inputs, techniques and/or manner by which staff or users acquire, analyze, or process other scientific data.

- Processed Data: Datasets that have been processed, filtered, edited, or annotated in some form, typically to extract knowledge and to be included in manuscripts for publication.
- Administrative Data: These data are required for facility operations. Administrative data can take the following forms:
  - Employee Data: This includes all employee emails and other digitally tracked staff, vendor, facility and user interactions.
  - Business Data: This includes all data necessary for operating the facility. Examples are user records (e.g. contact info, proposals, # of experiments, demographic data), billing information, statistics for reporting to funding agencies, and user data.

CNBL data locations are defined as:

- Individual Workstation: CNBL-supplied computers given to individual staff to conduct their daily tasks. This includes job information or notes; communication with users, service engineers, or other staff; report writing; or data processing.
- Instrument Computers: Computers provided by instrument vendors or supplied by CNBL for the purposes of operating the instruments, and the acquisition and temporary storage of raw scientific data. Instrument computers are typically located in the same room as the instrument or in a nearby operator room.
- CNBL Data Server: A temporary online data storage solution available to users and staff for the purposes of a) safely moving data from instrument computers to a user's personal storage device, b) working storage for data being processed using other CNBL compute resources, c) centralizing user data affiliated with CNBL jobs and sharing with those users.
- The CNBL Data Server is presently McMaster's SharePoint, a privately hosted, secure, cloud storage solution with two-factor authentication.

Data is collected at CNBL through the following means:

- Raw Data is collected on instruments, often using vendor-supplied proprietary software.
- Meta Data is collected through the headers of the datasets.
- Method Data are collected by staff and users, often using software external to the instrument (e.g. digital lab book or note-taking software).
- Employee Data is collected through day-to-day communication with users and staff.
- Business Data is collected by staff and users using both day-to-day communications and formal processes, such as beam time applications.

Sensitivity:

- Some scientific data will be proprietary.
- Some experiments will generate data for sensitive technology areas (e.g. energy storage).
- Employee and some business data will contain personal and administrative records.

Data sizes:

- Data collected for each experiment will be approximately 10 Mb or less.

#### Processed Data:

- Data processing at CNBL may take in the data defined above and generate Processed data. Processing of data is performed on individual workstations or Instrument Computers.
- Users can also opt to utilize their own capabilities or use cloud-based services at commercial third parties.

## Documentation

Scientific data should be annotated with metadata (instrument parameters, experiment description, sample origin, etc.).

Processed data should be thoroughly documented for reproducibility such that other staff and users may use it.

## Active Data Management

The McMaster Research Ethics Board (MREB) Data Storage & Security Guide of McMaster guidelines must be followed. All data are treated as high-risk by default. This ensures maximum security of all data and simplifies data management on the storage systems.

Scientific data are temporarily stored on the instrument computer and moved to central storage immediately after the instrument session. In the central storage, a separate folder for each experiment will be created with no cross-access. Access to storage on the CNBL Data Server is set up so only specified authorized users and staff can access data relevant to them. To facilitate collaboration, users can generate share links to specific people.

Unprocessed raw data are stored whenever feasible, to maintain scientific integrity and reproducibility of experiments.

CNBL instruments produce a variety of data from its custom-designed data acquisition and instrument control system. These systems, as well as any analysis software produced by the CNBL will be designed for interoperability and long-term storage in open formats (e.g., comma-separated values (CSV), tagged image file (TIF) for images). If data is collected with proprietary software, it should be exported to open formats or documented formats when possible.

Storage of data on instrument computers and virtual workstations is only permitted short term, and must be moved as soon as possible, ideally at the end of the instrument session. The CNBL is aware that immediate removal is not always possible, but given proper training, CNBL staff may purge any leftover data on these systems at their discretion, to ensure no unprotected data is exposed to other users.

Data relevant to the user jobs are stored by CNBL staff using libraries on the data server and kept for a storage term of **five years**. This data storage lifecycle is to be acknowledged by the user in a facilities user agreement.

Administrative data are stored indefinitely across various locations depending on the data type. Employee data is regularly maintained through University Technology Services (UTS) storage systems and is property of McMaster University.

## Long-Term Data Management

Users are expected to transfer their Scientific data off-site for permanent storage, if applicable. The CNBL will offer the appropriate encrypted communication channels to move data to the user's preferred storage device. Once the data is transferred off-site, the user acknowledges that CNBL is no longer liable for any lost or stolen data.

CNBL staff and McMaster users are expected use the [McMaster Dataverse](#) as the permanent repository for their Scientific data. McMaster Dataverse is McMaster's institutional open-source repository which provides secure access controls, licensing, version management, data citation, DOI minting, and more. Scientific data will be stored for the long-term storage in open formats, e.g., comma-separated values (CSV), tagged image file (TIF) for images, when possible.

McMaster Dataverse enables options to make the content available publicly, to specific individuals, or to keep it private. Access to data in sensitive technology areas will be released to specific individuals only following an appropriate research security screening.

Proprietary scientific data must be kept confidential and will not be shared without the approval of the CNBL Manager.