Jefferson Lab Data Management Resources and Responsibilities

Modified: December 19, 2024

1 Document Links

- This document.[1]
- JLab Data Management Plans.[2]

2 Overview

Jefferson Lab (JLab) requires that valuable data generated in connection with the Lab's research program be managed in a way that allows future and outside researchers to work with the data, either to validate a result or to conduct additional studies on the same data. The scope of this mandate includes:

- Preservation of the data
- Documentation of the data format
- Preservation of associated data such as run conditions and calibration databases
- Preservation of software used to read and process the data

Collaborations and researchers are required to provide an Experiment Data Management Plan (EDMP) to JLab that identifies these materials, documents how they will be maintained and remain accessible throughout their lifecycle, and is aligned with the requirements of this document. This EDMP will be submitted to the division in which the research takes place. Guidance on requirements and responsibilities and how to fulfill them is outlined below.

In addition, research funding agencies such as the Department of Energy (DOE) and the National Science Foundation (NSF) and others require Data Management Plans (DMP) as part of research proposals. Information in this document can be used to inform both the EDMP submitted to JLab and the DMP submitted to funding organizations.

[1] https://data.jlab.org/data-management-plans/Data_Management_Plan_JLAB-2024.pdf
[2] https://data.jlab.org/

3 Required Data and Metadata

Researchers and collaborations conducting publicly funded research at Jefferson Lab are responsible for ensuring their data are preserved and able to be re-analyzed in the future. The laboratory provides various tools and capabilities to support this responsibility.

The EDMP submitted to JLab must address the following:

1. Raw data
2. Processed data (where significant computing resources were used or data is much more accessible for further investigations, e.g., first-pass event reconstruction which includes tracks, energy deposition, etc.)
3. Run conditions (e.g., machine energy, polarization, beam intensity, and target details, etc.)
4. Electronic logbooks containing information necessary for subsequent data analysis (e.g., identifying periods of poor data quality, detector system modifications, target and beam changes, etc.)
5. Calibration databases
6. Geometry databases
7. Critical simulation and analysis software, including build scripts, documentation for building and using the software, and the same for all software upon which the top-level software depends

Whenever possible, locations (URIs, pathnames, etc.) for the these data and software assets should be clearly listed in the document. Storing these data assets within the JLab infrastructure is strongly encouraged. If they are to be stored offsite, then archival preservation and long-term access provisions must be clearly documented.

4 Data Preservation and Accessibility

All valuable raw and processed data should be stored in the tape library within one week of acquisition. Remaining items should be archived at a regular period (ie. quarterly) using tools provided by the IT Division and a consistent file or directory naming system or other cross-referencing database to facilitate future retrieval.

Databases storing calibrations, run characteristics, or other metadata must remain accessible for the dataset's lifespan. Periodic database snapshots must be archived to ensure they are recoverable and usable on future hardware and operating systems. For example:

- Document software and analysis workflow snapshots regularly
- Ensure that software packages critical for data analysis can be rebuilt and run at least every two years
- Ensure that external databases used by workflows remain available and compatible with archived software

JLab supports these efforts with a software repository (https://code.jlab.org) which includes a software container registry and CI/CD pipelines for automated builds and testing.

Researchers are encouraged to use this service for archiving software and workflows to ensure data reproducibility over time.

5 JLab CST and Accelerator Division Data Management Systems

Capabilities provided by Jefferson Lab include:

- A robotic tape library and associated software and servers for writing and reading files to and from tape
  - *Raw* data (e.g. data coming directly from the Hall DAQ systems) is automatically replicated across two independent tape volumes for added resilience.
- An archival tape storage room for duplicate copies of high-value files
- Disk server cluster for staging files to and from tape storage
- An archival logbook system (https://logbooks.jlab.org) for Accelerator and Hall operations, including production data taking
  - While supplemental logbook frameworks (MIDAS Elog, wikis, DocDB, etc.) may also be maintained in-part by the JLab CST division, they may require non-standard support to provide archival services. Such systems and archive planning must be explicitly called out in the collaboration/experimental data management plan (EDMP).
- An archival database for supplementary accelerator and Hall EPICS data (MYA)
- Software repository and container registry with CI/CD features for documenting, archiving, and automated testing of analysis workflows

Jefferson Lab also provides:

- Automatic duplication of raw experimental data to the JLab data center for archival storage and additional processing
- Automated backup of supplemental directories, files, and selected databases for preserving data, metadata, and data provenance
- A process of tape duplication migrating data to newer generations to maintain long-term file accessibility
- Periodic VM snapshots of the JLab Farm analysis environment to support archived software. These snapshots will be in a format that does not require proprietary software to execute

JLab's intention is to preserve these data indefinitely. Changes to this policy will be advertised in advance and an effort will be made to reach out to points of contact for impacted groups.

6 Experimental Nuclear Physics Plans

This document, along with Experimental Nuclear Physics Plans for each Hall, are available at https://data.jlab.org.

7 Quality Assurance

The Data Management processes are overseen by Jefferson Lab, which will conduct or convene periodic assessments to evaluate compliance and effectiveness.

8 Contact Information

For additional information, please contact:

- Brad Sawatzky, Nuclear Physics Computing Lead (brads@jlab.org)
- Laura Biven, Chief Data Officer (biven@jlab.org)