Jefferson Lab Data Management Plan*

Modified Aug 26, 2013

http://scicomp.jlab.org/DataManagementPlan.pdf

Summary: Jefferson Lab 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 be able 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 the preservation of the data, documentation of the data format, the preservation of associated data such as run conditions and calibration databases, and the preservation of software used to read and process the data.

Responsibilities

Researchers and collaborations conducting publically funded research at Jefferson Lab are responsible for having and following a data management plan to preserve their data and the ability at some future date to re-analyze that data. The laboratory makes available to the researchers a number of capabilities and tools to facilitate fulfilling this responsibility.

The following data and metadata must be addressed by the researchers' data management plan:

1. raw data
2. processed data, where the processing involved significant computing resources, or where the processed data is much more accessible for additional investigations (example: first pass event reconstruction, where the processed data includes tracks, energy deposition, etc.)
3. run conditions (machine energy, polarization and intensity, target, etc.)
4. electronic log books containing pertinent data for subsequent data analysis (e.g. periods of time for which the data is known to be of poor quality)
5. calibration database(s)
6. geometry database(s)
7. analysis software sources, make files or build scripts, documentation for building and using the software, and the same for all software upon which the analysis software depends

All valuable raw and processed data should be stored in the tape library in a timely fashion, typically within one week of acquisition. All remaining items should be stored at a reasonable frequency (e.g. quarterly), using tools provided by the IT Division and using a file naming system or directory naming system that allows any of these snapshots to be recovered in the future.

Snapshots of databases must be stored in an externalized human readable form, such as mysqldump for MySQL databases, and the plan must list the commands used for

* This Jefferson Lab Data Management Plan (DMP) is distinct from the DMP which will be required by DOE for all new research proposals submitted in response to solicitations issued on or after 10/01/2013, but can be referenced as part of writing a research DMP.

taking the snapshots and for their recovery. This externalization must also be performed for all metadata. Binary versions may also be kept, but cannot be the sole backup for any item except event data.

Collaborations may develop a plan that covers all or most elements of a data management plan, such that a researcher may reference that plan and specify any necessary additional items needed to make the plan complete for his or her research. The collaboration plans may reference this document to cover those aspects of a complete plan that are being provided by the IT Division (listed below).

Data Management Plans must be submitted to the division in which the research takes places.

Major long lived collaborations are also responsible for testing their ability to go back and rebuild software and run a standard analysis job at least once every two years. Every decade this process must touch a snapshot from every five year period for which snapshots exist. If third party software becomes completely unavailable, then alternatives will be investigated for running an older operating system as a virtual machine, and then keeping an older machine operational for some period of time.

The laboratory will conduct additional random tests of the snapshot system to ensure that it is functioning properly.

IT Division Data Management Systems

The following capabilities can be used by research groups and collaborations in building a data management process and plan:

1. a robotic tape library and associated software and servers for writing and reading files to and from tape
2. an archival tape storage room to hold duplicate copies of high value files for which this is appropriate
3. a disk server cluster and distributed file system for staging files to and from tape

In addition, Jefferson Lab manages in an automated and checked fashion the following processes:

1. automatic duplication of raw data from an experiment (triggered by writing the files into one of a specified number of raw data directories)
2. automatic duplication of additional directories for the purposes of preserving both data, metadata and data provenance
3. an ongoing process of duplicating tapes from older generations to newer generations so that the ability to read files is preserved

User tools documentation:

1. Raw data duplication: http://scicomp.jlab.org/doc/<tbs>
2. Metadata, databases and source code snapshots (TBS)
3. Tape library commands: (TBS)
4. File & disk system commands: (TBS)
5. Additional general scientific computing usage information available at http://scicomp.jlab.org/doc/

Quality Assurance

The Data Management processes will be overseen by the Deputy Director for Science. Under his direction, the laboratory will conduct periodic self-assessments of its Data Management processes. A readiness review of Data Management will be conducted prior to the start of 12 GeV research.

Point of contact for additional information: Chip Watson watson@jlab.org Head of Scientific Computing.