

## **Alcator C-Mod Data Management Plan**

The Alcator Data Management Plan (DMP) describes the elements and procedures for storing, securing and sharing data associated with the C-Mod facility.

### **1. Data Covered**

This plan covers the following data:

- Continuously acquired engineering data from the C-Mod supervisory control systems
- Set-up information for C-Mod pulses, including plasma control, fast timing and data acquisition
- Raw data acquired from each C-Mod pulse (shot)
- Data processed automatically between pulses
- Data processed manually and stored in the MDSplus archive
- High-level analyzed C-Mod data, run information and an electronic logbook stored in relational databases

### **2. Data Acquisition, Storage, Archival and Retention Policy**

#### **2.A. Data Acquisition**

Data from the C-Mod tokamak is acquired through the MDSplus system ([www.mdsplus.org](http://www.mdsplus.org)). This system integrates all set-up, raw and processed data into a single, coherent hierarchical structure. Engineering data is acquired from the supervisory control systems, which is based on programmable logic controllers (PLC) monitoring about 750 distinct system values at a typical acquisition rate of 1 per second on a 24/7 basis. These data are transferred to MDSplus every 2 minutes. Most raw data from experiments is acquired by transient recorders and other devices and written into MDSplus immediately after each pulse. During experiments, pulses occur at the rate of ~4 per hour for 8 hours each run day. Between shots, a large number of automated analysis routines create and store processed data. Together, these amount to 10 GB of data per shot for each of approximately 2,000 shots per year. Over the ensuing days, months and years, additional analyzed data is written into the same data hierarchy. Relational databases are used to store a comparatively small quantity of run metadata, highly analyzed data, data summaries and an electronic lab notebook. Together, integrated over all C-Mod operation, these comprise on the order of 2-3 million database records, most entered manually or by user provided software.

#### **2.B. Data Storage**

Primary storage currently consists of a 110 TB RAID6 disk array, that contains all of the C-Mod data taken from its first day of operation. Every night, any new or modified data is copied to a secondary 110 TB disk array, maintained in a separate building. Data and software required for machine operation are stored on separate, much smaller disk volumes. This allows machine operation to continue even if an equipment failure made the main storage unusable. (Restoring the full data set would take more than a week, hence the use of small volumes for this critical data.) For similar reasons, there is enough local storage on the data acquisition server to operate C-Mod for a week without the primary or secondary storage units.

#### **2.C. Data Backup and Archival**

An archive copy of all original data – that is data in the state when it was first acquired – is maintained by nightly transfers to MIT’s Tivoli Storage Manager (TSM), a large enterprise-class automated tape library. In addition a backup copy of the data, in its current state, is also maintained and updated daily on TSM. Older versions of files are maintained on TSM for 30 days, allowing further redundancy and a path for recovery from short-term data integrity problems. The relational databases are backed up nightly, weekly, and monthly; nightly and weekly backups are saved for 8 weeks and the monthly backups are saved permanently. All of these database backups are in turn backed up to TSM on campus.

## **2.D. Long Term Data Retention Policy**

All data ever acquired from C-Mod is stored on magnetic disk and archived as described above, which provides for 4 copies of raw data and 3-4 copies of processed data, spread over 3 separate buildings on campus. While there is no contractual obligation to retain this data beyond the life of the C-Mod cooperative agreement, a recently funded proposal supports data retention and access. Our intention is to keep this data accessible and usable beyond this agreement, limited only by future fiscal constraints, for at least 3-5 years beyond the life of the experiment.

## **3. Data Access and Sharing**

MDSplus provides access to all of the data described above through a simple application program interface (API) adapted for many common programming languages. Data is specified by a globally unique name, decoupling users from details of the underlying storage mechanisms. Remote access is provided by MDSIP, a software-based network layer that allows the API to store or retrieve data using the internet IP protocol. All C-Mod data is available to everyone on the C-Mod team, subject to the use and publication conditions described in the C-Mod collaboration agreement. [http://www.psf.mit.edu/research/alcator/program/collab\\_agree\\_2.pdf](http://www.psf.mit.edu/research/alcator/program/collab_agree_2.pdf)

## **4. Publication of Digital Data**

Digital data in support of publications will be provided in accordance with evolving DOE policy. Based on current guidance from SC and FES, researchers will store the data required, that is, those used in published figures or tables, at the PSFC and provide it through the PSFC web site. Discussions with journal publishers and with MIT libraries for either publication-based or institutional approaches are underway and may supplant this local solution in the future.