Data Management Plan : Research Data

The CMS experiment is dedicated to timely dissemination of its data and procedures, in addition to documentation of results in publications and journal articles. The LHC experiments (including CMS) are world leaders in grid computing and cloud-like data storage, and the solutions that have been developed have robustly handled the many petabytes of data that have been collected. There are several "tiers" of data, which are designated by how widely they are deployed throughout the Open Science Grid (OSG) and the other LHC sites. The data "tiers" are :

- **RAW** : The raw data, collected in terms of simple detector readouts and event information. While rarely used for analysis, this is always available for all of the collected data at CMS indefinitely.
- **Reconstruction (RECO)** : The reconstructed data, which utilizes the RAW data tier and computes relevant information in higher-level objects to be used by analyzers.
- **Analysis Object Data (AOD)** : In Run 1, the RECO tier was too large to transmit throughout the OSG. Instead, a subset of the RECO was stored, the AOD tier. This was the primary tier for analysis usage, although it was more transient in nature. In Run 2, the computing model for this stage is still under development, although the functionality of the AOD tier will always be present.

These are stored at the various OSG sites of CMS as follows.

- **Tier 0** : The long-term data collection and storage facilities are located at CERN, where the experiments are, including CMS. Data are collected and stored in RAW format at the Tier 0 site.
- **Tier 1** : Subsequent to data-collection at the Tier-0 facility, the data are shipped in RAW format to several sites worldwide (including to FNAL in the US) at facilities where the reconstruction software is run. Data are processed and stored in the RECO format at the Tier-1 sites.
- **Tier 2** : For simulation of data, and for analysis, many smaller clusters are utilized throughout the world. These are typically stored locally in the AOD format. Since the simulated data are not stored locally at the Tier 0 site, there is one "custodial" Tier 2 which stores each generated sample centrally for the long term.

In addition to these "tiers" of the actual data collected, there are also software and documentation schema for the LHC data and analysis. The software for the reconstruction is stored and versioned locally at CERN and duplicated at FNAL, and is visible to the public [1]. This has also now been fully migrated to `github` [2]. While the data collected are initially private to CMS, there are now mechanisms in place to make the entirety of the data completely public, although this may take several years to fully realize and deploy. In the meantime, there are well-defined approval procedures to ensure that the data collected by CMS are made public via documentation in webpages and journal publications, or by data-sharing projects such as `HEPDATA` [3].

[1] https://cmssdt.cern.ch/SDT/lxr/
[2] https://github.com/cms-sw/cmssw
[3] http://hepdata.cedar.ac.uk