MicroBooNE Digital Data Management Plan (v0.1)

Mike Kirby and Brett Viren

September 15, 2016

1 Introduction

The MicroBooNE collaboration agrees with the principles affirmed by the DOE Office of Science related to the management of digital research data. This document describes the plan for managing the collaboration digital data satisfying the requirements expressed in the Statement on Digital Data Management and following the Additional Requirements and Guidance for Digital Data Management.

2 MicroBooNE Data

The data are conceptually categorized into different "tiers" based on the volume, their source, the required processing and selection criteria, and the expertise required to consume or reproduce the data. Tiers are listed in order of derivation.

Each tier is described in the subsections below. The nominal policy on sharing and preservation of data in the tier is given and how these enable validation. Data sharing is considered to either be among members of the collaboration or between the collaboration and non-members. Preservation only considers copies of data shared within the collaboration. Requests for any expansion beyond the nominal policies described may be considered by the collaboration on a case-by-case basis.

2.1 Raw Data

The Raw Data Tier includes all files produced directly from experiment devices (e.g., detector DAQ, environment and beam monitors) and files holding the information used to configure these devices. The bulk of this tier's data comes from the MicroBooNE detector itself and consists of digitized signals from TPC wires, optical detectors, and muon detectors in a custom packed binary format and require special software to be read. Also included in this tier is information about the beam and environment held in relational databases.

Raw Data is only shared among collaborators. The volume, infrastructure and expertise required to produce and consume this data makes sharing outside the collaboration impractical. In principle, all collaborators have access to this data but in practice only a few are expected to access a small portion of it. The primary consumer of this data is the official collaboration production processing.

All data in this tier is archived to tape storage at Fermilab for the lifetime of the experiment and at least 5 years after data taking ceases as per the official Fermilab Data Management Plan

Plans for permanent preservation will be made at the time when data taking ceases in order to utilize appropriate technological choices.

Validation of this data tier is largely done by validating the proper operation of the devices that acquired it. This is done through detector commissioning and ongoing special-purpose studies as well as continual monitoring of the data acquisition by human shift operators.

2.2 Simulation Data

The Simulation Data Tier consists of data generated by software which emulates the interaction and detector physics along with readout electronics response. It produces data which is functionally identical to what the real detector produces and includes additional information about intermediate states of the simulation. The data are stored in files formatted as art framework/ROOT format and require substantial software infrastructure to read.

Simulated Data is only shared among collaboration. This policy is chosen for the same reasons as the Raw Data Tier sharing.

Data in this tier is preserved to disk and tape at least until it is superseded by newer simulations. Simulation data are produced in campaigns targeted for a specific detector run periods (e.g. Commissioning Run, pre-Summer 2016) and data for the current and previous two campaigns are maintained for approximately 18 months. As it can be reproduced by rerunning the software long term preservation is not cost effective.

Validation of this data is performed by the collaboration by comparing it to the Raw Data and to published results from other experiments and theory.

2.3 Reconstruction Data

The Reconstruction Data Tier consists of files derived from Raw and Simulated Data. It consists of intermediate results from processes such as noise reduction, signal extraction, imaging, pattern recognition, vertex and particle identification as well as derived calibrations. The volume of data in this tier is at least as large as Raw and Simulation.

Data in this tier is shared following the same policy as the Raw Data Tier.

Data in this tier is preserved to disk and tape at least until it is superseded by newer processing and is no longer actively utilized for measurements. The experiment plans to reprocess all detector data twice each year and retire older processed data after 18 months. As it can be reproduced by rerunning the software long term preservation is not cost effective.

Validation of this data is performed by comparisons between its similar derivations from Raw and Simulation data.

2.4 Analysis Data

The Analysis Data Tier consists of a down sampling of the Reconstruction Data. Selection criteria are applied that reduce which quantities and triggers are kept. The format of this data is such that only common community software tools (ie ROOT) are needed to read them. Expertise not generally available is needed to interpret the data in this Tier.

Data in this tier is shared following the same policy as the Raw Data Tier.

Processes producing this data are relatively simple and validation is done by collaborators to assure the selection criteria perform as expected.

2.5 Published Data

The Published Data Tier consists of files directly used to produce the tables and figures used in published documents. The volume of data is relatively small and is in formats which are readable with common tools (including ROOT) and hold quantities which may be properly interpreted by any individual with general understanding of the field. These data files are maintained in the collaboration "official plots database" along with the derived graphics used in the publications and descriptions of their contents.

Files of Published Data are made available at the time of publication along with the digital document, through references given in the document or by request to the collaboration. The distribution of the Published Data will utilize either INSPIRE or HEPData based upon which is deemed appropriate by the collaboration.

Published Data files will be preserved by the collaboration for the lifetime of the collaboration. Preservation policy of files shared through an external service will be determined by that service.

Validation of these files is done through the collaboration publication policy and procedures.

3 Data Management Resources and Facilities

MicroBooNE data management is in accordance with an agreement between the Fermilab Scientific Computing Division and the MicroBooNE collaboration as detailed in the Technical Scope of Work (TSW), MicroBooNE-DocDB-3537 . The Fermilab resources provide a means to store, manage, access and share the raw data and simulation data, as well as all of the research dependent reconstruction and calibration data. All of the Fermilab Data Management services are utilized within normal Fermilab SCD Service Level Agreements (SLAs) and the details of each SLA are listed in the TSW.

4 Information Protection

MicroBooNE data do not contain any personal identifiable or other confidential information, and therefore the information covered by this plan is not encumbered with any of the qualifiers listed in requirement four of the Statement on Digital Data Management. The exception being that the data is consider proprietary to the MicroBooNE Collaboration, and therefore the collaboration has a responsibility to assure sufficient expertise is employed in interpreting all but the Published Data. For this and practical reasons, access to non-published data is limited to collaboration members.