Appendix 8: Intensity Frontier Data Management Plan

The scope of work described in this supplemental proposal does not expand beyond that which is originally described in funded project. The data management plan included here is a restatement of that original data management plan.

The projects described in this supplement proposal for the intensity frontier do not produce any data on their own; they merely make use of data generated by the activities related to the, SBND, and ICARUS experiments. These experiments have a data management policy consistent with the Department of Energy's data management plan (found here http://science.energy.gov/funding-opportunities/digital-data-management/). This policy conforms with that of the Data Preservation in High Energy Physics (DPHEP) study group, which has described a hierarchy for the types of data that particle-physics experiments produce, and given recommendations for how such data should be preserved for future use.

The types of data these experiments produce include the raw data produced by the detectors, the reconstructed version of the raw data, and simulated events. SBND and ICARUS have their own mechanisms for archiving data, including producing both digital and tape storage of raw data, and any data that resides at our institution will merely be a copy of data that is stored permanently by the experiment. Thus, there is no need for this project to separately manage or archive any experimental data. No personally identifiable information is expected to be generated during the execution of the project, and thus no explicit plans to protect confidentiality or personal privacy are used.

The analysis of the experimental data is described in published, peer-reviewed journal articles; summaries of data analyses that are released to the public (often as contributions to conferences); and notes that are circulated internally within SBND and ICARUS as well as public technical notes made available using Fermilab's Technical Publication. The journal articles are archived by the journals themselves, and are also typically available through the arxiv.org e-print archive. The public analysis summaries and internal notes are archived by the SBND and ICARUS experiments and are available through Web interfaces. All collaborations encourage frequent and timely publication of results related to the research described in this proposal and thus data generated by the activities described here are expected to have near annual publication releases (through peer reviewed papers and contribution to conferences).

SNBD and ICARUS data management will be defined in accordance with an agreement between the Fermilab Scientific Computing Division and the SBND and ICARUS collaborations, as will be detailed in a Technical Scope of Work (TSW) document to be drafted and agreed upon before the start of data taking. Fermilab resources will provide a means to store, manage, access, and share the raw data and reconstruction data, as well as all of the research dependent analysis and calibration data.
Precise details of SBND and ICARUS data formats are still to be determined, but the data may be 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. Substantial quantities of simulated data will also be generated as an important ingredient in the analysis of real data collected by the detectors. The policies outlined here apply to both real and simulated forms of data. Each Tier is described in the subsections below. Tiers are listed in order of derivation: Raw Data, Reconstruction Data, Analysis Data, Published Data. The nominal policy on sharing and preservation of data in each Tier and how these data can be validated are given. 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 at any time on a case-by-case basis.

The Raw Data Tier includes all files produced directly from experiment devices (e.g., detector DAQ, environment monitors, beam monitors) and files holding the information used to configure these devices. The bulk of this Tier's data will come from the SBND and ICARUS detectors and consist of digitized signals from TPC wires, optical detectors, and muon detectors in custom packed binary formats that require special software to be read. Also included in the Raw Data 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. Additionally, the ICARUS raw data will be replicated to tape storage at CERN. Plans for permanent preservation will be made at the time when data taking ceases in order to utilize appropriate technological choices. Validation of the Raw Data Tier is largely done by validating the proper operation of the devices that acquired it. This is done through detector commissioning and special-purpose studies as well as continual monitoring of the data acquisition by human shift operators.

The Reconstruction Data Tier consists of files derived from Raw 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 Data. Data in this Tier is shared following the same policies as the Raw Data Tier. Data in the Reconstruction Data Tier is preserved to disk and tape at least until it is superseded by newer processing and is no longer actively utilized for measurements. Typically, long term preservation is not cost effective as Reconstruction Data can be reproduced by rerunning the software. Validation of this data is performed by comparisons between its similar derivations from Raw Data.

An Analysis Data Tier consists of a down-sampling of the Reconstruction Data. Selection criteria are applied that reduce which quantities and triggers are kept. Expertise not generally available to the general public 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.

Published Data consists of files directly used to produce the tables and figures used in published documents. The volume of data is relatively small, typically in formats that are readable with common tools (including ROOT), and hold quantities which may be properly interpreted by an individual with general understanding of the field. Files of Published Data can be made available at the time of publication along with the digital document, either through references given in the document or by request to 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 will be done through the publication policy and procedures of the collaboration.