# ANNIE Data Management Plan

ANNIE (FNAL-T1063) will in conjunction with the Fermilab Scientific Computing Division established the following policies regarding the retention, archiving and dissemination of data for this experiment. These policies have been realized through the adoption and integration of the data management and storage infrastructure provided by the Fermilab Computing Sector.

This document breaks out these data management policies by the tiers to which the data belong and for which different levels of archival data integrity, proprietary and public accessibility and retention are required.

## Raw Data Tier

ANNIE will acquire data in a proprietary format tuned to the performance characteristics of the experiment's custom readout hardware and data acquisition (DAQ) systems. These data are considered irreplaceable due to their nature as the lowest level of readout information available to the experiment and their temporal correlation with the Fermilab Booster beam complex or with other natural phenomena to which the detectors are sensitive.

The data considered to belong to the raw data tier for the experiment are:

- All data files acquired and constructed by the ANNIE DAQ system in the ANNIE raw data format
- All data recorded by the Intensity Frontier Beam systems, which represent the accelerator event timings, parameters and measurements made during the extraction of beam to the NuMI target station or to the Booster Neutrino (BooNE) target station
- All data recorded by the detector monitoring and environmental monitoring sensors in the ANNIE detector halls, which represent the physical environment and operational parameters of the detectors.
- All parameter data used configure readout hardware and configuration and advance the DAQ systems into a running state.
- All logging and status information generated by the DAQ systems during a the acquisition of physics data.

Data categorized in the raw data tier will be cataloged and archived according to the following policies:

- All raw data that is acquired or generated as digital *files* in a machine readable format will be cataloged using the ANNIE instance of the SAM data catalog.
- All cataloged files will be described in the catalog with a set of meta information which logically describes the data and includes at a minimum a unique filename identifier , the date and time of generation of the data file, the size of the file, an Adler CRC32 style checksum generated and matched to the checksum used by the Enstore mass storage systems, the type or classification of the data being cataloged, and the original registrar

of the data. Cataloged data may contain additional meta information describing the contents of the data or the conditions under which it was generated/acquired.

- All cataloged raw data will be stored in the Fermilab data archive facilities using the Enstore mass storage system. Data files stored in this facility will maintain a minimum of two replicas of the data and each replica will be stored on a physically distinct and independent storage element (i.e. two different tape cartridges). The exception to this policy is that *log* data, which does not contain information directly included in analysis results, will maintain a minimum of one replica stored on a physically distinct and independent storage element from those that hold data files used directly in analysis (i.e. log information and raw data files are not stored on the same tapes).

- All raw data that is generated or acquired as individual digital *records* in a machine readable format will be stored in a relational database system.

- Raw data records will be maintained through a minimum of two replicated database systems hosted on physically different hardware systems.

All data in the raw tier is considered proprietary and precious. General [read] access to the raw data is limited to members of the ANNIE collaboration. Specific access controls are implemented on the raw tier to limit full access only to authorized personnel within the collaboration and to members of the Fermilab staff who provide support for the data management and storage system. These access controls are designed to further protect the data against accidental erasure or other forms of data loss.

Data belonging to the raw data tier of the ANNIE experiment will be hosted primarily by the Fermilab computing and archive facilities. Replicas of any portions of the raw data tier can be disseminated to collaborating institutions via the standard replication tools provided by the Fermilab scientific computing division. This replication can be initiated by members of the ANNIE virtual organization (VO) or by request to the Fermilab computing division staff.

Dissemination of data from the raw data tier to institutions outside of the ANNIE collaboration and VO is provided on a technical level through the standardized replication of both the data catalog corresponding to the data set being published and the corresponding data that constitutes the data set . By this means the ANNIE raw data tier can be duplicated, hosted and disseminated using the tools provided by the Fermilab Scientific Computing Division, which are fully compatible with the Open Science Grid (OSG) analysis infrastructure as well as other common grid computing infrastructures which would be required to analyze and interpret the ANNIE data.

Dissemination of data from the raw tier to non-ANNIE collaboration parties will require approval of the ANNIE collaboration and due to the proprietary nature of the data may require dissemination of additional software, computing infrastructure, or intellectual property to be properly interpreted.

All data belonging to the raw data tier for the ANNIE experiment shall be retained and supported for the active life of the experimental collaboration. Data shall be retained past the dissolution of the ANNIE collaboration at to at least a minimum level corresponding to of an archival form of the raw data along with the ability to restore both the data and associated tools at a level that complies with DoE Directives.

**Analysis Data Tier**

As part of the data analysis process the the ANNIE experiment converts information from the raw data tier into expanded data collections which extract or refined the raw data in such a way as to enable the their examination for sophisticated scientific analysis. These data are considered to be a derived product of the raw data tier and the specific analysis algorithms, modeling and simulation systems that are used to process the raw data. As such data in this tier is considered non-precious as it can be re-constituted by the re-processing or re-analysis of the raw data with the same algorithms. This allows data in this tier to be retained at a reduced redundancy/replication factor and overall reduction in cost for long term retention.

The analysis data tier is considered to be an intermediate tier, that requires the highly specialized knowledge of the ANNIE collaboration to work effectively with. These data are considered highly proprietary and may represent the intermediate work and intellectual property of the ANNIE collaboration and its members. The data in this tier does not represent final physics results, but is often the direct inputs that are used to perform the final, more general scientific analysis which form the basis of publications.

The data considered to belong to the analysis data tier for the ANNIE experiment are:

- All data files derived from data in the raw data tier by application of documented algorithms or analytic processing.
- All data generated by modeling or simulation systems, which can be deterministically regenerated at a later time (i.e. Monte Carlo simulations which known configurations and seed values).
- All data records which represent calibration constants, derived detector response functions or other record based information which is reconstruct-able from information in the raw data tier through application of documented algorithms or procedures.

Data categorized in the analysis data tier will be cataloged and archived according to the following policies:

- All analysis data that is acquired or generated as digital *files* in a machine readable format will be cataloged using the ANNIE instance of the SAM data catalog.
- All cataloged files will be described in the catalog with a set of meta information which includes all information required by the raw data tier and which additionally includes the provenance information regarding both the parentage information describing the chain from which the data was produced and meta information which describes the procedures or algorithms which were used in its generation. The meta information must be sufficient to permit the regeneration of the data.
- All cataloged analysis data will be stored on a data storage element supported by the SAM data management system and the tools provided by the Fermilab Scientific Computing Division for data retrieval and management. These systems may include the Fermilab central disk systems (commonly referred to as the Bluearc NAS), the dCache based storage pools which are part of the Fermilab archive facility, the Enstore

tape library system, or other storage systems which are part of the Fermilab computing infrastructure. The data may also reside on non-Fermilab hosted data storage systems such as specific university disk arrays, cloud storage systems such as the Amazon Web Services S3 facilities, or other academic or commercial systems that have been integrated with the SAM platform. Data from this tier will maintain a minimum of one replica of the data.

- All cataloged analysis data that is used to produce published physics results will maintain at least one replica of the data in the Fermilab data archive facilities using the Enstore mass storage system.

- All analysis data that is generated as individual digital *records* in a machine readable format will be stored in a relational database system or as appropriate a noSQL database system.

- Raw data records will be maintained through a minimum of one replicated database system for which regular backups or snapshots are performed.

- The primary database systems in which the analysis data records are stored will be hosted by Fermilab and maintained and supported by the Fermilab Computing Sector.

Data belonging to the analysis data tier of the ANNIE experiment will be hosted primarily by the Fermilab computing and archive facilities. Replicas of any portions of the raw data tier can be disseminated to collaborating institutions via the standard replication tools provided by the Fermilab scientific computing division. This replication can be initiated by members of the ANNIE virtual organization (VO) or by request to the Fermilab computing division staff.

Dissemination of data from the analysis data tier to institutions outside of the ANNIE collaboration and VO is provided through the standardized replication of both the data catalog corresponding to the data set being published and the corresponding data that constitutes the data set in a manner identical to the way in which data from the raw tier is disseminated. By this means the ANNIE raw data tier can be duplicated, hosted and disseminated using the tools provided by the Fermilab Scientific Computing Division, which are fully compatible with the Open Science Grid (OSG) analysis infrastructure as well as other common grid computing infrastructures which would be required to analyze and interpret the ANNIE data.

Dissemination of data from the analysis data tier to non-ANNIE collaboration parties will require approval of the ANNIE collaboration and due to the proprietary nature of the analyzes being performed may require additional dissemination of software, computing infrastructure, or intellectual property to be properly interpreted.

Data belonging to the analysis data tier for the ANNIE experiment which is used directly or indirectly as the inputs to a published scientific or technical result, shall be retained and supported for the active life of the experimental collaboration. Analysis data which is directly or indirectly used as inputs to a published scientific or technical result, will be retained past the dissolution of the ANNIE collaboration at a level corresponding to at least an archival form of the data along with the configurations and additional procedural information that

would be required to restore the data and associated analysis tools to a level where the data could be used to replicate any resulting publications or published results.