Policies on the Retention, Archiving and Dissemination of Data
for the NOvA Experiment (FNAL-E929)

Introduction

The NOvA experiment (FNAL-E929) has, in conjunction with the Fermilab Scientific Computing Division, established the following policies regarding the retention, archiving and dissemination of data for the NOvA 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

The NOvA experiment acquires data in a custom 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 NuMI 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 NOvA experiment are:

- All data files acquired and constructed by the NOvA DAQ system in the NOvA 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 NOvA detector halls, which represent the physical environment and operational parameters of the detectors.
- All parameter data used to configure the readout hardware and advance the DAQ systems into a running state.
- All logging and status information generated by the DAQ systems during 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 NOvA 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[1] , 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 are 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.
- The primary database systems in which the raw data records are stored will be hosted by computing systems located at the NOvA far detector site and supported by the NOvA collaboration through their data acquisition group. The secondary (replica) database system used to host raw data records will be hosted at Fermilab and the Fermilab Computing Sector will provide support for the database infrastructure.

[1] Uniqueness is imposed across the entire NOvA data catalog namespace

All data in the raw tier are considered proprietary and precious. General read access to the raw data is limited to members of the NOvA 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.

Retrieval of data files from the raw data tier is provided and controlled by the SAM data management tools provided and supported by the Fermilab Scientific Computing Division. These tools provide both optimized retrieval of the data files from the mass storage system, and permit the creation of replicas of the data to additional storage elements either at Fermilab or at other collaborating NOvA institutions. The mass retrieval or restoration of data from the raw tier (e.g. restoration of petabyte scale data sets), requires special considerations and will be performed only by authorized NOvA collaborators in conjunction with Fermilab staff supporting the data management and storage systems.

Data records from the raw tier are made accessible to the NOvA collaboration through database servers and replica servers that provide authenticated connections through a web (http protocol) based application layer. Full access to the raw data records and direct connections to the database servers are protected through strong authentication mechanisms that permit access only by authorized NOvA personnel and Fermilab staff members who support the database systems.

Data Dissemination

Data belonging to the raw data tier of the NOvA 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 NOvA virtual organization (VO) or by request to the Fermilab computing division staff.

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

Dissemination of data from the raw data tier to institutions outside of the NOvA 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[2]. By this means, the NOvA 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 NOvA data.

Data Retention

All data belonging to the raw data tier for the NOvA experiment shall be retained and supported for the active life of the experimental collaboration. Data shall be retained past the dissolution of the NOvA 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 Statement on Digital Data Management[3].

Analysis Data Tier

As part of the data analysis process the NOvA experiment converts information from the raw data tier into expanded data collections that extracted or refined the raw data to enable 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 are 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 NOvA collaboration to work effectively with. These data are considered highly proprietary and may represent the intermediate work and intellectual property of the NOvA collaboration and its members. The data in this tier do not represent final physics results, but are often the direct inputs that are used to perform the final, more general scientific analyses that form the basis of publications.

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

- All data files derived from data in the raw data tier by application of documented algorithms or analytic processing.

---

[2] Replication includes the instantiation of a public SAM data catalog for the published data sets which is independent from the primary SAM data catalog that is used by the NOvA collaboration and imposes strict access controls on its use.
[3] http://science.energy.gov/funding-opportunities/digital-data-management/

- All data generated by modeling or simulation systems that 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 reconstructable 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 NOvA 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 are 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 are 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 Dissemination

Data belonging to the analysis data tier of the NOvA 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 NOvA virtual organization (VO) or by request to the Fermilab computing division staff.

Dissemination of data from the analysis data tier to non-NOvA collaboration parties will require approval of the NOvA collaboration and, due to the highly customized nature of the data format, may require additional dissemination of software, computing infrastructure, or intellectual property in order to be properly interpreted.

Dissemination of data from the analysis data tier to institutions outside of the NOvA 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 NOvA analysis 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 NOvA data.

Data Retention

Data belonging to the analysis data tier for the NOvA 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 NOvA 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.

Scientific Results Tier

When forming physics analysis results, the NOvA experiment produces event summary information, summary ntuples, plots/fits and other data that describe the different parts of the final analysis chain. Many of these samples hide the highly specialized knowledge/characterization of the detectors and instead provide quantities relating to the underlying physics that can be directly used and compared

to theory and other experiments. The highest level of these data are often formally published as part of peer-reviewed articles or made available as public datasets. As a result, these data and the methods and methodologies used to create them require special retention policies.

The data considered to belong to the scientific results tier for the NOvA experiment are:

- All data presented in the form of histograms, plots, graphs or other graphical representations that are included in publications or public presentations.
- All data presented in the form of tables or other tabular formats which are included in publications or public presentations.
- All summary data which is used as direct input to plots, graphs or tables which are included in publications or public presentations (i.e. summary ntuples which are selected against when making a final plot)

Data categorized in the scientific results tier will be cataloged and archived according to the following policies:

- All data presented in graphical formats will be stored in their final published form in digital formats corresponding to the actual format(s) that were used to generate the publication(s) (i.e. if the plot was included in an article as a portal network graphic [PNG] then a PNG copy of that file will be retained. If, in addition, the graphic was used in a Portal Document Format [PDF] to generate a publication, then a PDF copy will also be retained.)
- All data presented in graphical formats will also have their contents stored in a self-describing, human readable digital format (i.e. an ASCII or UTF-8 encoded text file). The information included will include the information required to regenerate the graphic. As an example, to reconstruct a published histogram, the bin edges, bin contents, error bars/bands, axis titles and legend/statistics information and any additional information placed on the histogram will be stored in a self describing text file.
- The graphical data and their corresponding text representations will be cataloged together in the NOvA Official plots database together with meta information about the graphic, including a detailed description appropriate for or corresponding to a caption included in a publication.
- All tabular data presented in publications will be stored in both a file containing the formatted source of the table (i.e. a latex formatted table object appropriate for direct inclusion in a publication) and in a human readable digital format (i.e. an ASCII or UTF-8 encoded text file). The human readable digital format will be self-describing and provide all information required to reconstruct the printed tables (i.e. columnar headings, row demarcation's, units etc…) The format will be parse-able in such a way that it could be used or adapted to be the basis of the input to other analysis software.
- All published data, graphs and tables will be stored by the collaboration on archival medium through the Fermilab mass storage system and will

maintain a minimum of two copies in that system. Additional copies may be kept on other storage systems or at other locations.

The intention of this archival policy to ensure that the data are preserved in a form that can be used without reliance on NOvA-specific intellectual property or other proprietary software.

In addition the following items relating to the published data will be cataloged and archived according to the following policies:

- The analysis code, algorithms, scripts, and procedures used to generate any published results will be cataloged and indexed in such a way that it is associated with the particular result or publication that it generated.
- These code bases, along with their supporting infrastructure, will be collected and packaged into a fully redistributable digital form (example: standard Unix tape archive).
- The redistributable form of the code bases will be stored by the collaboration on archival medium through the Fermilab mass storage system and will maintain a minimum of two copies in that system. Additional copies may be kept on other storage systems or at other locations.

The intention of this archival policy for the code base is to ensure that method by which the published data were derived is fully preserved and that the exact methods and algorithms could be used at a later point in time to verify past, present, or future data against the published data.

Data Dissemination

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

Dissemination of data from the scientific results data tier to institutions outside of the NOvA collaboration and VO is provided through an official HTTP based web portal which is hosted at Fermilab through the standard Fermilab central web services. The portal provides access to both the individual data elements (plots/tables) and their metadata descriptions. The portal provides two separate access doors (one authenticated and one unauthenticated) which provide the ability to disseminate the information to the public as well as to specific collaborating institutions which may need expanded access to the NOvA collaboration's digital representations of the data or the underlying analysis code and procedures.

In addition, unpublished documents which support the scientific analysis tier are available for dissemination, with the approval of the NovA collaboration, through the Document Database Service (DocDB) hosted by Fermilab. The DocDB service provides both authenticated and unauthenticated access and permits the NOvA collaboration to provide both fully public access as well as partially restricted access to these supporting documents.

Dissemination of data from the scientific results data tier beyond the published plots and data tables (for example: individual event kinematics or energy distributions from the ν_e appearance analysis) to non-NOvA collaboration parties will require the approval of the NOvA collaboration. Even when approval is granted, these requests may be deferred for a pre-determined waiting period after the initial publication of a NOvA result. Due to the nature of the high-level analyses that can be performed with this data, the intellectual property rights of the NOvA collaboration must be considered. This waiting period is designed to permit the completion of any analysis efforts internal to NOvA that may be pending or planned for publication (example: a joint fit combining NOvA data with data published by another experiment). In these cases the data, after approvals and any waiting period, will be disseminated through the standard services.

Requests for NOvA data from the scientific results tier that do not fall into the previously described categories will be subject to approval by the NOvA collaboration and disseminated through the standard authenticated web services provided by Fermilab.

Data Retention

Data belonging to the scientific results data tier for the NOvA experiment will be retained and supported for the active life of the experimental collaboration. Results data which is the direct output of a published scientific or technical result, will be retained past the dissolution of the NOvA collaboration at a level corresponding to at least an archival form of the data along with the supporting analysis code base and procedures described previously. The core archival copies of the data will resided at Fermilab in the mass storage systems provided by the lab and will be subject to the published data retention and media migration policies that the lab maintains.

Summary

The NOvA data management policy is design to both protect the data collected and analyzed by the NOvA collaboration as well as ensure its availability to the collaboration and the scientific community as a whole. The policy leverages the

strengths of the Fermilab Policy on Data Management[4] and is compliant with the Department of Energy, Office of Science Data Management Guidelines.

[4]http://computing.fnal.gov/xms/Science_Computing/Policies_and_Publications/