Data Management Plan (DMP) for Particle Physics Experiments prepared for the 2025-2029 Consolidated Grants Round (submitted Feb 2024).

Prepared by GridPP and representatives of all experiments
Coordinating contact for comments or questions: peter.clarke@ed.ac.uk

[Note: It has been agreed in advance with STFC (Jane Long) that this may exceed the 2-page limit]

The *Particle Physics Experiment Consolidated Grant* proposals now being submitted are for the exploitation of experiments which will produce large amounts of data. This document provides a summary of the Data Management Plans pertaining to these data, with the necessary references to the DMPs of the experiments themselves, and the means to implement those policies. This document first covers LHC experiments and then supplementary paragraphs for several non-LHC experiments.

1. The LHC Experiments (ATLAS, CMS, LHCb)

Each of the LHC experiments, in conjunction with CERN, implements a Computing Model and publishes an open data policy.

1.1 Data Management and Preservation

The data management and preservation processes of the LHC experiments are part of the Computing Models (CMs) of each of the experiments. They have been developed over the last decades and have operated successfully at Run-1, Run-2 and Run-3. These CMs ensure that (i) multiple copies of all raw data are stored at distinct sites around the world, (ii) resilient metadata catalogues are maintained, (iii) experimental conditions databases are maintained, (iv) and software versions are stored and indexed. Since all data can in principle be regenerated from these raw data, then these models meet the fundamental STFC requirements of resilient data preservation. In addition, multiple copies of derived data are also stored as well as copies of simulated data to facilitate data analysis.

The most recent descriptions of the Computing Models can be found at:

- ATLAS: https://cds.cern.ch/record/2802918/files/LHCC-G-182.pdf
- CMS: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookComputingModel
- LHCb: https://cds.cern.ch/record/2319756?ln=en

Broadly speaking, particle physics experiments produce four “levels” of data.

- **Level-4: Raw data.** These are the raw data produced by the experiments after selection by the online triggers. *Level-4 data is fundamental and must be preserved as all other data may, in principle, be derived from it by re-running the reconstruction.*
- **Level-3: Reconstructed data**. These data are derived from the Raw data by applying calibrations and pattern finding algorithms. Typical content includes “hits”, “track”, “clusters” and particle candidates**.** It is these data that are used by physicists for research. *Selected Level-3 data is also preserved. This is done for efficiency and economy since the process to re-derive it may take significant computing resources, and in order to easily facilitate re-analysis, re-use and verification of results.*
- **Level-2: Data to be used for outreach and education and other non-LHC science**. Several activities have been developed whereby subsets of the derived data are made available for outreach and education and. *Level-2 data has no unique preservation requirement*
- **Level-1: Published analysis results**. These are the final results of the research and are generally published in journals and conference proceedings. *Level-1 data is preserved in the open access journals, and additional data is made available through recognised repositories such as CERN CDS and HEPDATA*

- **Monte Carlo Data:** In addition, the experiments produce simulated "Monte Carlo" data (referred to as MC data) in the same four levels. MC data undergoes the equivalent reconstruction processes as for real data. *MC data can in principle always be regenerated provided the software and the associated transforms have been preserved. However out of prudence some MC data is also preserved along with associated real data.*

The data recorded by the LHC are stored within WLCG. This provides the physical means for each experiment to implement its data management processes. By the end of 2024 it provided ~1.4M logical-CPU cores and ~1.5 exabyte of data storage. The WLCG is comprised of the CERN Tier-0 sites plus 14 Tier-1 and over 80 Tier-2 federations throughout Europe, the US and Asia. These are federated together, and provide common authentication, authorisation and accounting systems, and common workload management and data management systems. They have a well-developed management and communications process, development and deployment planning, fault reporting and ticketing systems, and a security incident response and escalation process. The UK component of the WLCG is GridPP, which provides approximately 10% of WLCG, namely (in 2024) ~100,000 logical cores, ~100 PB of disk and ~140 PB of tape storage. The Tier-1 is linked to CERN via a 200Gb/s optical private network (OPN) and is attached to JANET by dual 200Gb/s links. Main Tier-2 sites are connected to JANET by 40-100 Gbit/s links.

The preservation of Level-3 and Level-4 data is guaranteed by the data management processes of the LHC experiments. The exact details of processes are different for each experiment, but broadly speaking they are as follows:

- The Raw data (level 4) is passed from the experimental areas in near real time to the CERN Tier-0 data centre where it is immediately stored onto tape.
- At least a second tape copy of the Raw data is made shortly afterwards. This second copy is stored at Tier-1 sites remote to CERN. The details and number of copies depend upon the detailed CM of each experiment, but the result is always resilient copies of the Raw (Level-4) Data spread around the world.
- The CERN and remote data centres have custodial obligations for the Raw Data and guarantee to manage them indefinitely, including migration to new technologies.
- Level-3 data is derived by running reconstruction programs. Level-3 data is also split up into separate streams optimised for different physics research areas. These data are mostly kept on nearline disk, which is replicated to several remote sites according to experiment replication policies that take account of popularity. One or more copies of this derived data will also be stored on tape.

In summary several copies of the Level 4 and Level 3 data are maintained in physically remote locations, at sites with custodial responsibilities. This therefore ensures the physical data preservation requirements of the STFC policy are met.

Software is equally important in the LHC context. The knowledge needed to read and reconstruct Raw Data, and to subsequently read and analyse the derived data is embedded in large software suites and in databases which record conditions and calibration constants. All such software and databases are versioned and stored in relevant version management systems. Currently GIT is used. All experiments store information required to link specific software versions to specific analyses. All software required to read and interpret open data will be made available upon request according to the policies of the experiments.

1.2 Open data access

The CERN statement on Open Data Access can be found at:
http://opendata.cern.ch/docs/cern-open-data-policy-for-lhc-experiments
This describes clearly the approach adopted for each Level of data described above.
Each experiment has produced policies with respect to open data preservation & access. These can be found at:
http://opendata.cern.ch/search?page=1&size=20&q=policy

In summary Level 1 data is automatically available through Open Access publications, and Level 2 data is available by construction. Level 4 data cannot practically be made systematically available. Level 3 data is made openly available according to this statement:
The LHC experiments will release calibrated reconstructed data with the level of detail useful for algorithmic, performance and physics studies. The release of these data will be accompanied by provenance metadata, and by a concurrent release of appropriate simulated data samples, software, reproducible example analysis workflows, and documentation. Virtual computing environments that are compatible with the data and software will be made available. The information provided will be sufficient to allow high-quality analysis of the data including, where practical, application of the main correction factors and corresponding systematic uncertainties related to calibrations, detector reconstruction and identification. A limited level of support for users of the Level 3 Open Data will be provided on a best-effort basis by the collaborations.

The proprietary period varies in detail and is typically 50% by 5 years and 100% by 10 years, with full release at the end of the Collaboration. These data will be released through the CERN Open Data Portal: http://opendata.cern.ch/. This portal contains links to data, metadata, educational packages, software and virtual machine images.

At this time both CMS and LHCb have already released their entire Run-1 data sets to the public. Announcements can be found at this page:
http://opendata.cern.ch/search?page=1&size=20&q=policy

2. Non-LHC experiments

2.1 DUNE

DUNE will still be in the construction phase during this CG and will not have any primary scientific data to be managed. However, the ProtoDUNE detector at CERN has taken data and will do so again during this CG period. DUNE uses the same Grid infrastructure and similar processes to the LHC experiments, and makes use of capacity in the WLCG and the OSG. The CM has been published in a Conceptual Design Report:
https://lss.fnal.gov/archive/design/fermilab-design-2022-01.pdf
A DUNE DMP can be viewed at:
https://publicdocs.fnal.gov/cgi-bin/ShowDocument?docid=23
DUNE will follow the FNAL Policies (see Appendix)

2.2 NA62

The NA62 experiment uses the same Grid infrastructure as the LHC experiments. All of the earlier description of physical resources and processes for ensuring data preservation pertains. The data management processes follow the normal scheme of making secure copies of raw data at remote sites and keeping the metadata in a resilient catalogue. In particular the UK is archiving a 2nd copy of all NA62 raw data at RAL. The NA62 Data Management and Open Data Access Policy is at:
http://na62.web.cern.ch/NA62/Collaboration/EditorialBoardDataPreservation.html

2.3 LUX-ZEPLIN

Data generated by the LZ experiment is stored at two mirrored sites, one in the US and one in the UK (Imperial College). An additional independent database archive is maintained in the US. The data are preserved at the mirrored sites. Software is managed through a centralised repository on gitlab. Raw data is regarded as proprietary for the duration of the active experiment and a further period beyond the project end to allow for validation. LZ also adheres to the DOE and NSF data policies. The most recent DMP documents can be found at:
https://lz.lbl.gov/wp-content/uploads/sites/6/2021/09/LZ-DMP-Operations.pdf

2.5 T2K
The T2K experiment also uses the WLCG Grid infrastructure. All of the earlier description of physical resources and processes for ensuring data preservation pertains. T2K data management implements multiple copies of the data, metadata, and essential software on three continents. T2K data will continue to be analysed for many years past the point that operation ceases, which will guarantee that it will be preserved and ported to new storage technology for the foreseeable future.T2K offers access to selected samples of the key reduced data appearing in our publications: http://t2k-experiment.org/results/. The T2K DMP is available at: http://t2k-experiment.org/for-physicists/data-management-plan/

2.6 Hyper-Kamiokande
Hyper-Kamiokande will transition from construction into the operational phase during the period of this CG. Raw data will be stored close to the detector sites in Japan, at KEKCC and Kamioka observatory,University of Tokyo, on non-Grid infrastructures. Raw, processed and simulated data, and the associated metadata, will be distributed and preserved longterm on the WLCG. KEKCC, The University of Tokyo, GridPP, IN2P3, INFN and The Digital Research Alliance of Canada will provide the main infrastructure for the Hyper-K computing model. Data arising from any publications will be published in suitable repositories (such as the DURHAM-HEP database or the CERN open data portal). The experiment will also use the experiments http://www.hyperk.org to publish in an open manner articles of interest to the community and beyond.

2.7 COMET
The COMET experiment will use the WLCG (in addition to non-Grid infrastructures). Thus, all of the earlier description of physical resources and processes for ensuring data preservation pertains. The data management processes are currently being developed and will follow the normal scheme of making secure copies of raw data at remote sites (Grid and otherwise), and keeping the metadata in a resilient catalogue. The international COMET collaboration is currently finalising a DMP and Open Access policy, which are expected to be similar to those of other experiments at J-PARC/KEK such as the T2K experiment.

2.8 SNO+
SNO+ uses similar Grid infrastructure to the LHC experiments. It is based upon WestGrid Compute Canada and GridPP (and in particular RAL). All of the earlier description of physical resources and processes for ensuring data preservation pertains. As physics results are released by the collaboration, relevant data will be released online alongside collaboration papers. As physics results are released by the collaboration, relevant data are made available through the collaboration's public website alongside papers. The SNO+ data management plan is available at https://snoplus.phy.queensu.ca/files/snoplus_datamanagement.pdf

2.9 IceCube/IceCube-Gen2
Raw data written to disks at the South Pole are transported to the University of Wisconsin (UW) data center for archival storage. The raw data is also filtered at the South Pole. The filtered dataset is transferred to UW via satellite. A second copy of the filtered data will be copied to DESY. NSF policies and guidance promote efforts to make data and software available to other researchers. In addition, the Parties to the Antarctic Treaty agree that scientific results from Antarctica should be exchanged and made freely available. IceCube data are released after the main analyses are completed and results are published. The guidance of data management and data use policy are found within the governance document at this site https://icecube.wisc.edu/collaboration/governance

2.10 LEGEND
The first stage of the experiment (LEGEND-200) started operation at the Gran Sasso National Laboratory (LNGS, Italy) in March 2023 and will continue taking data for at least 5 years. LNGS will

act as a first data production site. Data will be copied to a European data distribution centre, which will be mirrored to a second distribution centre in the US. Data will be backed up on tape in all the three sites. The LEGEND software is hosted, and version controlled by git-hub. LEGEND adopts an open-software policy for all projects not containing sensitive information. All LEGEND data analysis and simulation is performed within software containers to guarantee portability and long-term data preservation. LEGEND adheres to the DOE and NSF data policies.

2.4 SuperNEMO
SuperNEMO data processing and management processes are similar to those of the LHC experiments. CC-IN2P3, Lyon, (an LHC Tier 1 site) will act as the hub for SuperNEMO where the data will be initially stored; data is protected by a tape backup system. Resources provided by the collaborating countries of the SuperNEMO collaboration will be used to meet the data, metadata and software preservation requirements, which will be managed using IN2P3's standard tools; this also incorporatess a database server. Software is stored on the SuperNEMO GitHub or in IN2P3's GitLab instance, as appropriate. A long-term data management plan consistent with the requirements of open data access is being developed.

2.11 Other FNAL experiments:
These include CHiPS, g-2, MicroBoone, MINOS/MINOS+, Mu2e, NoVA, SBND.
All of the FNAL experiments implement data management and preservation using disk and tape storage resources at FNAL. They follow the FNAL policy on data management practices and policies: see Appendix.

2.12 PSI experiments Mu3e, MuEDM:
All PSI experiments implement data management and preservation at a long-term storage facility at PSI. Copies of the derived datasets will also be hosted at PSI and collaborating institutions. Software is managed through a central repository on Bitbucket (Mu3E) and gitlab (MuEDM). According to PSI policy, data can be made open-access after a nominal embargo period of 3 years, during which only collaborators have access. The PSI DMP is available here https://www.psi.ch/en/science/psi-data-policy, and collaborations are finalising specific DMPs.

2.13 FASER:
The FASER experiment, approved in 2019 and in operation since the beginning of Run 3, follows closely the ATLAS computing modelling, as depicted in Section 1.10 of https://arxiv.org/abs/2207.11427, and the open data policy in https://cds.cern.ch/record/2745133

Appendix

The Fermilab policy page is currently protected and requires a Fermilab credential. They have been asked to make this public. In the meantime, the contents are copied here.

Data Management Practices and Policies for Fermilab Experiments

Overview: Fermilab is the lead laboratory for many particle physics experiments. Some of those experiments leverage the lab's accelerator facility while many do not. No matter whether part of the Fermilab facility or not, these experiments all go through multiple phases in their lifecycle, from conceptual design to prototype detectors in test beams, to full blown operating experiments and ultimately final data analysis, data archival and knowledge preservation. While the needs for a data management plan vary in detail over the experiment lifecycle, they all have common themes with respect to digital data and how it is treated.

A key deliverable of each experiment is a digital record of the data representing selected physics of interest whether that be a cosmic ray passing through a detector or a digital snapshot of the sky. Fermilab provides the experiments in which it is the lead laboratory the means to store, manage, access and share the raw data, simulation data as well as all of the research dependent reconstruction and calibration data.

Policy: It is the policy of the facility to provide long-term data storage and data access to all of the experiments and scientific programs associated with the facility in order to ensure the integrity, availability and safe keeping of the data products, associated conditions data, as well as relevant simulation data. How long the data is stored and subsequently archived is typically negotiated on a case-by-case basis, depending on the needs and uniqueness of the experimental data being captured. However, the default is that Fermilab will keep the data active for a minimum of 5 years after data taking ceases. All experiments in which Fermilab is the lead laboratory will use Fermilab as its primary repository for the data. Other copies of the data may exist throughout the world depending on the experiments' unique needs.

Resources Available: See Computing and Data Resources at Fermilab for an overall description of available resources. Each experiment will be provided a minimal baseline level of support for their data needs to get each experiment started. A yearly Portfolio Review is held and resource allocations for each experiment are made taking into account the facilities total demand for computing needs, budgets, and scientific priorities. Exceptional needs may require that additional funding be secured in order to satisfy them.

Data Validation: It is expected that each experiment will take ownership of the operation of its experiment though will receive ample support from the laboratory. It is the experiments' responsibility for the integrity of all its public scientific research signed by the collaboration. A vigorous internal review process prior to its public presentation handles that validation.

Data availability and sharing: Fermilab provides the means for experiments' researchers to access data from anywhere in the world and to share data. There are a variety of systems available to meet wide ranging needs; the appropriate technology choice will be made consistent with a particular experiment's requirements.

Researcher Responsibilities: The decision as to who may have access to the data is the responsibility of the experiment to define. The default will be to only allow members of the collaboration to access it. Fermilab provides tools to implement the access decisions once it is informed.

The Experiment must, in conjunction with the laboratory, define the long-term retention period for its data products consistent with the policy described in Computing and Data Resources at Fermilab. The overall retention length will be determined on a case-by-case basis depending on the scientific relevance of the data and the cost required to maintain it. Individual agreements will be established between each experiment and the facility to document these requirements.

It is expected that any experiment that wants something other than the default parameters must contact the head of Scientific Computing Division and CIO to establish a special Service Level Agreement (SLA) that all parties agree to.

This plan complies with: https://www.energy.gov/datamanagement/doe-policy-digital-research-data-management