Data Management Plan (DMP) for Particle Physics Experiments prepared for the 2018 Consolidated Grants Round.

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

The *Particle Physics Experiment Consolidated Grant* proposals now being submitted are all for the support of experiments which will produce very significant amounts of data. As such STFC requires a Data Management Plan (DMP) to be provided. This document provides a summary with the necessary references to the DMPs of particle physics experiments and responses to the questions set out in the STFC guidelines.

This document covers firstly LHC experiments and then non-LHC experiments.

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

1.1 Computing Models

The data management and data processing processes of the LHC experiments are part of the Computing Models of each of the experiments. They have been developed over the last decades and have operated successfully at Run-I and Run-II.

The Computing Models 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 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 computing models of the LHC experiments were most recently updated in 2014:

- *"Update of the Computing Models of the WLCG and the LHC Experiments"* http://cds.cern.ch/record/1695401?ln=en

More recently, with a view to Run-III and beyond, the HEP Software Foundation (HSF) was formed to address the future of software and computing for particle physics, which includes the means and processes for data management. The draft roadmap addresses many items pertinent to future computing models and data management, and can be found at:

- http://hepsoftwarefoundation.org/activities/cwp.html

1.2 Types of data

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-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.

- **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-1: Published analysis results**. These are the final results of the research and are generally published in journals and conference proceedings.

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.

1.3 Data to be preserved

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.

Some 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. More recently "NTUPLE" level data is also preserved.

Level-2 data has no unique preservation requirement

Level-1 data is preserved in the journals, and additional data is made available through recognised repositories such as CERN CDS and HEPDATA

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.

1.4 Data Preservation

1.4.1 Data centres

The data recorded by the LHC is stored within WLCG, the UK part of which is provided by GridPP. This provides the physical means for each experiment to implement its data management plans.
The WLCG is comprised of the CERN Tier-0 sites plus 12 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.
In 2017 WLCG provides 375 PB of disk storage and 340 PB of tape storage. As noted earlier all data is stored with multiple copies at physically separated sites, thus guaranteeing resilience against any foreseeable disaster. The UK component is provided through GridPP and provides (in 2017) approximately 10% of WLCG, i,e. 30 PB of disk storage and 55 PB of tape storage at the Tier-1 and 20 PB of disk at the Tier-2 centres. The Tier-1 is linked to CERN via a resilient 30Gb/s optical private network (OPN) and is attached to JANET by resilient 40Gb/s links. Most Tier-2 sites are connected by 10 Gbit/s links.

1.4.2 Processes

The preservation of Level-3 and Level-4 data is guaranteed by the data management processes of the LHC experiments. The LHC experiments use the Worldwide LHC Computing Grid (WLCG) to implement those processes. The exact details are different for each experiment, but broadly speaking the process is as follows:

- The Raw data is passed from the experimental areas in near real time to the CERN Tier-0 data centre where it is immediately stored onto tape.
- CERN has a remote Tier-0 centre in Hungary (Wigner Centre), which provides resilience.
- At least a second tape copy of the Raw data is made shortly afterwards. This second copy is stored at other sites remote to CERN, typically the Tier-1 data centres. The details and number of copies depend upon the detailed computing model of each experiment but the result is resilient copies of the Raw 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 Raw 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.

1.5 Software

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 SVN and GIT are 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.

1.6 Data preservation period

There is currently no upper time limit foreseen for the retention of LHC raw data. Naturally the ability to do so depends upon the continuation of CERN and the remote data centres, and of available expertise, and staff and hardware resources.

1.7 Analysis preservation

An analysis preservation system has been developed at CERN. This allows completed analyses to be uploaded and made available for future reference. This includes files, notes, NTUPLE type data sets, and software extracted from SVN or GIT. This is being used by the experiments to deposit completed analyses. This can be viewed at https://analysispreservation.cern.ch/welcome and http://cernanalysispreservation.readthedocs.io/en/latest/(although as it pertains to internal analysis preservation a valid credential is required).

1.8 Open data access

Each experiment has produced policies with respect to open data preservation & access. These are the result of agreement between the full set of international partners of each experiment. These can be found at:
http://opendata.cern.ch/search?page=1&size=20&collections=Data-Policies

In respect of the STFC guideline questions these cover:
Which data is valuable to others: In general Raw (level 4) data could not be interpreted by third parties without them having a very detailed knowledge of the experimental detectors and reconstruction software (such data is rarely used directly by physicists with the collaborations). Derived data (level 3) may be more easily usable by third parties. Level-1 data is already openly available.
The proprietary period: Experiments specify a fraction of the data that will be made available after a given reserved period. This period ranges up to several years reflecting the very large amount of effort expended by scientists in construction and operation of the experiments over many decades, and in part following the running cycle that defines large coherent blocks of data. For details of periods and amounts of data release please see the individual experiment policies.
How will data be shared: Data will be made available in format as specified by the normal experiment operations. This may be exactly the same format in which the data are made available to members of the collaborations themselves. The software required to read the data is also available on a similar basis, along with appropriate documentation.

CERN, in collaboration with the experiments, has developed an Open Data Portal. This allows experiments to publish data for open access for research and for education. The portal offers several high-level tools such as an interactive event display and histogram plotting. The CERN Open Data platform also preserves the software tools used to analyse the data. It offers the download of Virtual Machine images and preserves examples of user analysis code. The portal can be found at http://opendata.cern.ch/
In some cases individual experiments have also taken the initiative to develop or engage with value added open data services using resources obtained in participating countries. In ATLAS Recast and RIVET are the recommended means for reinterpretation of the data by third parties. They have also developed light-weight packages for exploring self-describing data formats intended mainly for education and outreach.

1.9 Resources required
All experiments carry out their basic data preservation activities as a natural result of scientific good practice. However, experiments in general do not have any specific resources available (either staff of hardware) for carrying out pro-active open data access activities over and above those described above.

2. Non-LHC experiments
2.1 NA62
The NA62 experiment uses the same Grid infrastructure as for the LHC experiments. 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, and keeping the metadata in a resilient catalogue. The NA62 Data Management Policy is at:
http://na62.web.cern.ch/NA62/Collaboration/EditorialBoardDataPreservation.html

2,2 FNAL experiments: CHiPS, LBNF/DUNE, g-2, MicroBoone, MINOS/MINOS+, Mu2e, NoVA, SBND.

All the FNAL experiments adhere to the FNAL policy on data management practices and policies, which is available at [1].

This policy is implemented within the experiments through dedicated members of the collaborations who are also members of FNAL's Computing Division, which oversees the implementation of the policy. The data retention policy of the experiments will follow the template established by the Tevatron experiments, which is underpinned by CernVM-FS and is detailed in documents from the "ICFA Study Group on Data Preservation and Long Term Analysis in High Energy Physics" [2] as part of the wider DASPOS group [3].

Fermilab supports the usage of electronic publishing methods and the principles of Open Access Publishing, which includes granting free access to publications to

all. Results are disseminated to the public through scientific publications, technical reports, presentations at scientific conferences and the e-print service arXiv which is free of charge to all readers. These are also all available on the public Fermilab library web-server.

The codes for processing, accessing, and extracting data are well-documented and maintained centrally for all the members of both collaborations to ensure reproducibility and backward compatibility. Irreproducible raw data is replicated to multiple locations in the US and Europe, using the Office of Science and Worldwide LHC Computing Grid respectively.

[1] http://computingint.fnal.gov/xms/Science_&_Computing/Policies_and_Publications/Data_Management_Practices_&_Policies
[2] http://www.dphep.org
[3] https://daspos.crc.nd.edu

2.3 LUX and LUX-Zeplin

Data generated by the LUX experiment is stored at two mirrored sites, and LUX-Zeplin will adopt the same strategy with one in the US and one in the UK. An additional independent database archive is maintained in the US. The typical data volume to be generated will be ~1 PB per year, stored on RAID arrays at the mirrored sites and subsequently committed for permanent storage. Version control and traceability of software developed for processing and analysis is performed through a centralised repository at LBNL. The two data centres control and monitor freely distributed raw and processed data within the collaboration. Results intended for publication are internally verified by the collaboration and relevant data and information made available to the public once published. 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. As with LUX, LZ also adheres to the data policies of the DOE and NSF. The most recent DMP documents can be found at:
http://lz.ac.uk/wp-content/uploads/2017/08/LZUKDataManagementPlan_v2.0.pdf

2.4 SuperNEMO

SuperNEMO is a small experiment by HEP standards, but the data types, processing and management issues are essentially the same as those of the LHC experiments given in section 1 above. The main difference from the LHC experiments is that CC-IN2P3, Lyon, will act as the hub for SuperNEMO rather than CERN. CC-IN2P3 is one of the main WLCG Tier-1 sites, providing extensive data storage and processing capabilities via Grid and non-Grid interfaces. (It was used for the NEMO-3 experiment.) The data from the experiment at the LSM (Laboratoire Souterrain Modane) will be transferred to CC-IN2P3. Resources and infrastructure provided by the collaborating institutes and countries of the SuperNEMO collaboration will be used to meet the data, metadata and software preservation requirements. A data management plan consistent with the requirements of open data access is being developed; it is planned to make use of data management processes and tools developed by WLCG, DPHEP etc.

2.5 T2K

The T2K experiment also uses the WLCG Grid infrastructure. Thus all of the earlier description of physical resources and processes for ensuring data preservation pertains. T2K has been running for eight years, so has a well-exercised data distribution network which features multiple archives of the data, metadata, and essential software, preserving them in multiple copies on three continents and leaving any serious loss of data extremely unlikely. It is planned that T2K will transition fairly seamlessly into Hyper-Kamiokande, and thus the T2K data will continue to be analysed for many years, which will guarantee that it will be preserved and ported to new storage technology for the foreseeable future. This transition will include the upgrading of the ND280 near detector with new, more granular detectors, which will somewhat increase its storage and data-processing needs. T2K has been considering what type of Open Access is appropriate for the T2K data, and as a first step the public web pages of T2K already offer access to selected samples of the key reduced data appearing in our publications (see http://t2k-experiment.org/results/), which enables other researchers to directly use the results of our research (for instance our measured neutrino cross sections as a function of energy can be downloaded). As part of the work on T2K and Hyper-K, the UK is also involved in Super-Kamiokande, the DMP for which is available at http://t2k-experiment.org/for-physicists/data-management-plan/

2.6 Hyper-Kamiokande

The Hyper-Kamiokande experiment will be in pre-construction phase, but anticipates transitioning to the construction phase during this CG.
During this period the experiment expects to produce simulation data and test-beam data which will use the WLCG infrastructure to process and store the data. As Hyper-K will make use of existing T2K infrastructure it will follow, as a basis, the T2K data management plan. Data arising from publications will be published in suitable repositories (such as the DURHAM-HEP database or the CERN open data portal) to ensure data are accessible and reusable by the community and beyond. The experiment will also use the experiments website http://www.hyperk.org to publish in an open manner articles of interest to the community and beyond. The experiment makes use of version control systems (GIT and CVS) for all code and published results will cite the versions of software used. Beyond published results the experiment will develop a policy during the construction phase on access to further levels of data.

2.6 DUNE

See general statement above for all FNAL experiments.

There are no UK-specific data management requirements for this Pre-

construction Phase proposal. The DMP statement most recently accepted by the PPRP is:
*"The simulated data to support the scientific studies of **WP1** will be stored centrally at Fermilab and will be processed using the existing LArSoft-based software framework. The repositories for the reconstruction code developed in **WP2** will be hosted in the central LArSoft repository. Software development work will be carried out on local machines in Cambridge, Lancaster and Warwick using data stored at Fermilab.The DAQ development in **WP3** is a programme of prototyping high-speed FPGA-based processing and there are no data management/storage requirements. **WP4** focuses on commissioning and detector characterisation. Data will be stored at CERN and Fermilab. The processing of this data will use the standard DUNE software and will primarily take place at Fermilab. At this stage in the project we do not foresee the usage of significant UK grid resources. The pre-construction activities in **WP5** focus on production processes and there are no data storage/management requirements*."

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 Data Management Policy that will document Data Preservation and Open Access policies, which are expected to be similar to those of other experiments at J-PARC/KEK such as the T2K experiment.

2.8 SNO+

SNO+ data is managed centrally by the experiment. Raw data (the direct output of the DAQ) and processed data (after reconstruction has been applied) are both stored on grid enabled storage elements located in both Canada and the UK using the same infrastructure as LHC experiments. This plan provides for redundancy in data storage. Raw data and select processed data will be maintained on these sites indefinitely. These data are accessed via grid certificate with membership of the snoplus.ca VO, which is managed by collaboration members. As physics results are released by the collaboration, relevant data will be released online alongside collaboration papers. The international SNO+ collaboration is currently drafting a Data Management Policy to document Data Preservation and Open Access policies in line with the requirements of SNO+ funding agencies. The Data Management Policy will be included in the collaboration statutes.

2.9 IceCube/PINGU

Data from the IceCube / PINGU project is stored at the University of Wisconsin, Madison. The Madison cluster also provides the primary resources used to process the data and generate Monte Carlo. At The University of Manchester, a GPU cluster is being used to generate PINGU Monte Carlo. During this

Consolidated Grant round, we also intend to begin using the GridPP resources for the reconstruction of both IceCube data and IceCube / PINGU Monte Carlo.