This report is part of the project ’**Insights’** that has received funding from the European Union’s Horizon 2020 research and innovation programme H2020-MSCA-ITN-2017, under grant agreement n. 765710.

Data Management Plan

in fulfilment of INSIGHTS Deliverable 5.7 (D41)

Deliverable	D5.7
Deliverable No	D41
WP	5 (Management)
Type	Report
Dissemination	Public
Report Date	21 November 2019
Author/Institution	Glen Cowan
Version	1

Contents

1. Data Summary

The INSIGHTS ITN includes a variety data collection and analysis activities. Foremost amongst these are the two large experimental facilities at the Large Hadron Collider (LHC), ATLAS and CMS. When in data-taking mode, each experiment collects roughly 1 petabyte of data per year on high-energy proton-proton collisions. The data is used to explore interactions between elementary particles and thus investigate the fundamental laws of particle interactions. The summary below of the data management policies for these two experiments is adapted from material prepared for the UK Science and Technology Research Council by Prof Peter Clarke, University of Edinburgh.

The data management and data processing processes of the LHC experiments are part of the Computing Models of each of the ATLAS and CMS collaborations. They have been developed over the last decade, have operated successfully for the first two data-taking periods (run I and run II) of the LHC.

Broadly speaking the LHC experiments produce four "levels" of data.

- Level-4: Raw data. These are the raw data produced by the experiments at the LHC interaction points 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. Several activities have been developed whereby subsets of the derived data are made available for outreach and education.
- 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.

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 updated by the experiments and the Worldwide LHC Computing Grid (WLCG). The description of these updated models can be found at:

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

2. FAIR data

2. 1. Making data findable, including provisions for metadata

In accordance with the computing models described above, the data are discoverable with metadata and they are identifiable. The naming conventions and other details are specific to each experiment.

The data are naturally divided into data sets (either by a physics theme, or a run period). These are catalogued in the experimental cataloguing systems running across the WLCG. These systems index physical file locations, logical file names, and the associated metadata which describes each data set.

In the ATLAS case, Rucio will handle replicating files to multiple sites. For the web front end there are multiple servers (stateless). Behind that, resilience and redundancy is provided by Oracle with the usual RAC configuration and a Data Guard copy to another set of machines.

The CMS Dataset Bookkeeping Service (DBS) holds details of data sets, including metadata, and includes mappings from physics abstractions to file blocks. Multiple instances of DBS are hosted at CERN. The transfer of data files amongst the many CMS sites is managed by PhEDEx (Physics Experiment Data EXport), which has its own Transfer Management Database (TMDB), hosted on resilient Oracle instances at CERN. Logical filenames are translated to physical filenames at individual sites in the Trivial File Catalogue (TFC).

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.

2.2. Making data openly accessible

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?cc=Data-Policies

Data preservation and open data access issues have also been developed through pan experimental initiatives at CERN which include:

- Open data access: http://opendata.cern.ch/
- DPHEP Community study group: http://www.dphep.org/

2.3. Making data interoperable

CERN and the experiments have taken open data access very seriously, and all of the experiments have developed policies, referenced above, which contain further details. These specify:

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. CMS is already about to use this for data release according to its policy. 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.

2.4. Increase data re-use (through clarifying licences)

The CERN Open Data Portal products are shared under open licenses. Further details can be found at http://opendata.cern.ch/ .

Use of CERN Open Portal data in subsequent publications is allowed in accordance with the FORCE 11 Joint Declaration of Data Citation Principles.

3. Allocation of resources

Both the ATLAS and CMS experiments carry out their data preservation activities as a natural result of scientific good practice. This leads to marginal extra staff costs over and above those operating the WLCG, and some additional storage costs. Naturally these activities rely upon the continuation of CERN and the remote sites.

Experiments in general do not have any specific resources for carrying out active open data access activities over and above those described above.

The additional cost of storage for data preservation is starting to be specified in the annual resource projections of each experiment which go to the CERN RRB and which are scrutinised and subsequently approved.

4. Data security

The preservation of data follows the following basic principles:

- 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.
- 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 (see later). However out of prudence some MC data is also preserved along with associated real data.

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 which 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.

CERN has developed an analysis preservation system. 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 already being used by the experiments to deposit completed analyses. This can be viewed at http://data.cern.ch/ (although as it pertains to internal analysis preservation a valid credential is required).

The wider HEP community has been working together in a collaboration to develop data preservation methods under the DPHEP study group (Data Preservation for HEP). The objectives of the group includes (i) to review and document the physics objectives of the data persistency in HEP (ii) to exchange information concerning the analysis model: abstraction, software, documentation etc. and identify coherence points, (iii) to address the hardware and software persistency status (iv) to review possible funding programs and other related international initiatives and (v) to converge to a common set of specifications in a document that will constitute the basis for future collaborations. More on DPHEP can be found at http://www.dphep.org/.

5. Ethical aspects

LHC data have no connection to individuals and thus have no relation to personal data or privacy issues.

Ethics issues related to the project of ESR8 at Pangea Formazione have been described in the deliverables 8.1 (D53) and 8.3 (D55) (cf. Sec. 6 below).

6. Other issues

The INSIGHTS ITN includes several smaller projects not described above: the AWAKE experiment at CERN, the NEWSdm experiment at Gran Sasso, Italy, and a project related to traffic flow carried out at the Italian SME Pangea Formazione.

As a CERN experiment, AWAKE adheres to that organisation's general rules relating to data management. For the INSIGHTS ESR working on this experiment, special data sets were recorded that are only expected to be of value for the ESR and to another student working on the AWAKE project; there is thus no significant need to make the raw data public.

The data management of NEWSdm is bound to the policies of the Italian National Institute for Nuclear Physics (INFN). The experiment uses the CNAF site in Bologna (the INFN national centre for data processing and computing technology, https://www.cnaf.infn.it/en/) for data storage and centralized computing.

The traffic data used by Pangea Formazione is based on material hosted on public github servers and released either in the public domain or under permissive licenses like Creative Commons (CC0, CC-BY, CC-BY-SA), BSD or MIT. Ethics issues related to these data (privacy and potential dual use) were addressed in deliverables 8.1 (D53) and 8.3 (D55).