WLCG Draft Data Management Plan – Annex 1

This draft data management plan has been prepared according to the H2020 (and other) guidelines for input to the DPHEP Workshop in Lisbon in February 2016: https://indico.cern.ch/event/444264/.

It will also be used as the basis of the HEP component of the HNISciCloud Data Management Plan (Deliverable due in Project Month 2 – i.e. February 2016).

H2020 Annex 1 Guidelines		
Guideline	Guidance	
Data set reference and name	*Identifier for the data set to be produced.*	This Data Management Plan (DMP) refers to the data set generated by the 4 main experiments (also know as "Collaborations") currently taking data at CERN's Large Hadron Collider (LHC). These experiments are ALICE, ATLAS, CMS and LHCb. For the purpose of this plan, we refer to this data set as "The LHC Data".
Data set description	*Description of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.*	The 4 experiments referenced above have clear scientific goals as described in their Technical Proposals and via their Websites (see https://greybook.cern.ch/greybook/ for a catalogue of all CERN experiments). Hundreds of scientific publications are produced annually. Similar data – but at lower energies – have been produced by previous experiments and comparisons of results from past, present and indeed future experiments is routine. The data behind plots in publications is made available since many decades via an online database: http://hepdata.cedar.ac.uk/. Re-use of the data is made by theorists, by the collaborations themselves, by scientists in the wider context as well as for Education and Outreach.
Standards and metadata	*Reference to existing suitable*	The 4 main LHC experiments collaborate through the WLCG

	standards of the discipline. If these do not exist, an outline on how and what metadata will be created.	Collaboration on data management (and other) tools and applications. At least a number of these have found use outside the HEP community but their initial development has largely been driven by the scale and timeline of the above. The ROOT framework, in particular, is used as "I/O library" (and much more) but all LHC experiments and is a *de-facto* standard within HEP, also across numerous other laboratories. The meta-data catalogues are typically experiment-specific although globally similar. The "open data release" policies foresee the available of the necessary metadata and other "knowledge" to make the data usable (see below).
Data sharing	*Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).*	The 4 LHC experiments have policies for making data available, including reasonable embargo periods, together with the provision of the necessary software, documentation and other tools for re-use. Data releases through the CERN Open Data Portal (http://opendata.cern.ch/) are published with accompanying software and documentation. A dedicated education section provides access to tailored datasets for self-supported study or use in classrooms. All materials are shared with Open Science licenses (e.g. CC0 or CC-BY) to enable others to build on CERN's results. All materials are also assigned a persistent identifier and come with citation recommendations.

	In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).	
Archiving and preservation (including storage and backup)	*Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.*	The long-term preservation of LHC data is the responsibility of the Tier0 and Tier1 sites that form part of the WLCG Collaboration. A Memorandum of Understanding (MoU) outlines the responsibilities of sites that form part of this collaboration (Tier0, Tier1s and Tier2s). In the case of the Tier0 and Tier1s, this includes "curation" of the data with at least two copies of the data maintained worldwide (typically 1 copy at CERN and at least 1 other copy distributed over the Tier1 sites for that experiment). The costs for data storage and "bit preservation" form part of the resource requests that are made regularly to the funding agencies. A simply cost model shows that the annual storage costs – even including the anticipated growth – go down with time and remain within the funding envelope foreseen. (The integrated costs of course rise). Personnel from the Tier0 and Tier1 sites have followed training in ISO 16363 certification – A Standard for Trusted Digital Repositories – and self-certification of these sites is underway.

		The data themselves should be preserved for a number of decades – at least during the active data taking and analysis period of the LHC machine and preferably until such a time as a future machine is operational and results from it have been compared with those from the LHC. The total data volume – currently of the order of 100PB – is expected to eventually reach 5-10 EB (in circa 2035 – 2040).