Helix Nebula – The Science Cloud

Deliverable Title: Data Management Plan

Partner Responsible: CERN

Work Package: 1

Submission Due Date: 26 January 2016

Actual Submission Date: 23 March 2016

Distribution: Public

Nature: Report

Abstract: This document describes the initial Data Management Plan (DMP) for the HNSciCloud project. It is based on the Guidelines on Data Management in Horizon 2020 document version 2.0 of 30 October 2015.
"Participating projects will be required to develop a Data Management Plan (DMP), in which they will specify what data will be open."

Document Information Summary	
Deliverable number:	*1.1*
Deliverable title:	*Initial Data Management Plan*
Editor:	*Jamie Shiers, CERN*
Contributing Authors:	*Bob Jones, Rachida Amsaghrou, CERN*
Reviewer(s):	David Foster, CERN
Work package no.:	*WP1*
Work package title:	*Consortium Management*
Work package leader:	*Bob Jones, CERN*
Work package participants:	*Rachida Amsaghrou, David Foster, Helge Meinhard, Jamie Shiers*
Distribution:	*Public*
Nature:	*Report*
Version/Revision:	*1.0*
Draft/Final:	*Final*
Keywords:	*Data Management Plan, Data Preservation*

Disclaimer

Helix Nebula – The Science Cloud (HNSciCloud) with Grant Agreement number 687614 is a Pre-Commercial Procurement Action funded by the EU Framework Programme for Research and Innovation Horizon 2020.

Grant Agreement Number: 687614
Start Date: 01 January 2016
Duration: 30 Months

Log Table

Issue	Date	Description	Author/Partner
V0.1	04/03/2016	First draft	Jamie Shiers, Bob Jones, Rachida Amsaghrou (CERN)
V0.2	07/03/2016	Clarification in section 3 of the introduction that this first version of the Data Management Plan includes first examples from the scientific use cases, and that others will be added during subsequent document revisions	David Foster, Bob Jones (CERN)
V1.0		Final Version	Rachida Amsaghrou (CERN)

Document Approval

Issue	Date	Name
V0.1	04/03/2016	First draft for circulation within the collaboration
V0.2	23/03/2016	Final draft revised by the HNSicCloud project office

Executive Summary

This document describes the initial Data Management Plan (DMP) for the HNSciCloud project. It addresses Project administration data collected as part of the execution and management of the Pre-Commercial Procurement (PCP) process within the project as well as data used as part of the scientific use cases to be deployed on the cloud services.

The DMP for the scientific use cases is based on the Guidelines on Data Management in the Horizon 2020 document version 2.0 of 30 October 2015.

These guidelines state:

"Participating projects will be required to develop a Data Management Plan (DMP), in which they will specify what data will be open."

As such, this DMP focuses primarily on data sharing and re-use and in particular on the following issues (Annex 1 of the H2020 DMP Guidelines):

- *What types of data will the project generate/collect?*
- *What standards will be used?*
- *How will this data be exploited and/or shared/made accessible for verification and re-use? If data cannot be made available, explain why.*
- *How will this data be curated and preserved?*

Regular updates to this data management plan will be made according to the following draft schedule:

Date	**Milestone**	**Issue(s) to be addressed in revision**
May 2016	Tender publication	Were all the use-cases retained?
April 2017	Start of prototype phase	Could all the use-cases be satisfied?
January 2018	Start of pilot phase	Could all the use-cases be deployed?
June 2018	End of Project	Are the plans for preservation beyond the end of the project still valid?

Table of Contents

1. Introduction

This Data Management Plan (DMP) addresses two distinct sets of data to be managed by the HNSciCloud project:

- **Project administration data**: data collected as part of the execution and management of the Pre-Commercial Procurement (PCP) process within the project
- **Scientific use cases**: data managed by some of the use cases that will be supported during the pilot phase by the cloud services to be developed and procured by the PCP process.

These two sets of data are treated separately by the project and in this document.

This is the initial version of the data management plan. It explains the provisions for project administration data and the general approach for scientific use cases with first examples from high energy physics use cases. The data management plan will be reviewed and updated at major milestones in the project, once the final set of use cases has been refined, and with the following tentative schedule:

Date	Milestone	Issue(s) to be addressed in revision
May 2016	Tender publication	Were all the use-cases retained?
April 2017	Start of prototype phase	Could all the use-cases be satisfied?
January 2018	Start of pilot phase	Could all the use-cases be deployed?
June 2018	End of Project	Are the plans for preservation beyond the end of the project still valid?

Table 1 - Tentative Data Management Plan Update Schedule

2. Project administration data

This section describes the plan for the data to be managed as part of the execution and management of the Pre-Commercial Procurement (PCP) process within the project.

The project office (WP1) will gather and store contact details (name, email address, role within company, company name, and company postal address) provided by individuals representing companies and organisations interested or participating in the PCP process.

Such details will be used by the consortium as part of the communication plan (WP7) and procurement process (WP2). The contact details will also be used as the basis for statistics summarizing the level and scope of engagement in the procurement process and will be reported (anonymously) in the mandatory deliverable reports. Beyond the summary statistics reported in the project deliverables, such information will be restricted to the consortium members.
The data will be maintained beyond the lifetime of the project as email lists and entries in the supplier database of the lead procurer, which may be used for subsequent procurement activities in the cloud services domain.

3. Scientific use cases

This section describes the plan for the data to be managed as part of the scientific use cases that will be deployed during the pilot phase of the pre-commercial procurement process.

As a general rule, data that is created in the Cloud will be copied back to institutional and/or discipline-specific repositories for long-term preservation, curation and sharing. Many of these repositories (e.g. the WLCG Tier0 and Tier1 sites) are in the process of self-certification according to ISO 16363:

ISO 16363:2012 defines a recommended practice for assessing the trustworthiness of digital repositories. It is applicable to the entire range of digital repositories. ISO 16363:2012 can be used as a basis for certification.

As such, the procured services will not be the mechanism by which data preservation for the use-cases presented below will be ensured. This matches the hybrid public-commercial cloud model where the publicly operated data centres provide data preservation facilities.

Once the PCP model has been proven to work, we may begin entrusting it with the procurement of data preservation services, possibly initially on ISO 16363-certified sites. However, this will be outside the scope of the HNSciCloud PCP project.

A number of the experiments that form part of the High Energy Physics (HEP) Use Case, namely the Belle II experiment at KEK and the four main LHC experiments at CERN (ALICE, ATLAS, CMS and LHCb) collaborate through both the Worldwide LHC Computing Grid (WLCG) project (for data processing, distribution and analysis) as well as DPHEP (Data Preservation of Long-term Analysis in HEP). DPHEP maintains a portal through which information on these and other experiments can be obtained, the status of their data

preservation activities and, in some cases, access to data released through Open Data policies. There is significant overlap between the data preservation plans and the H2020 DMP guidelines. For these experiments, we present the current status / plans in the agreed DPHEP format whereas for the LHC experiments it is also presented according to the H2020 guidelines (Annex 1). (The DPHEP format is quite detailed but emphasizes that these "plans" are backed by "implementation" and that there is clear evidence of data sharing and re-use. Furthermore, the data preservation services work at a scale of 100TB – 100+PB, with an outlook to perhaps 10EB and a duration of several decades).

This version of the Data Management Plan includes initial examples from the scientific use cases, others will be added during subsequent document revisions.

3.1 H2020 Data Management Plans for the LHC Experiments

These Data Management Plans were elaborated at a Data Preservation (DPHEP) workshop in Lisbon in February 2016 with the assistance of two of the co-chairs of the Research Data Alliance (RDA) Interest Group on Advanced Data Management Plans (ADMP). Representatives from the LHC experiments as well as HNSciCloud / WLCG Tier0/1 sites contributed to the debate.

H2020 Annex 1 Guidelines		
Guideline	Guidance	Statement
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". In terms of Data Preservation, the software, its environment and associated documentation must also be preserved (see below). Further details can be found at the DPHEP portal site, with entries for each of the above experiments: • http://hep-project-dphep-portal.web.cern.ch/content/alice • http://hep-project-dphep-portal.web.cern.ch/content/atlas • http://hep-project-dphep-portal.web.cern.ch/content/cms • http://hep-project-dphep-portal.web.cern.ch/content/lhcb
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,*	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 the official catalogue of all CERN

	and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.	experiments that is maintained by the CERN Research Board). Hundreds of scientific publications are produced annually. The data is either collected by the massive detectors of the above experiments (the raw data), is derived from it, or is the result of the simulation of physics processes according to theoretical models and the simulated response of the detector to these models. 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. (See also the DPHEP portal for further information: http://hep-project-dphep-portal.web.cern.ch/) The data behind plots in publications has been 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 standards of the discipline. If these do not exist, an outline on how and what metadata will be created.*	The 4 main LHC experiments work closely together through the WLCG 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.).* *In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal*	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 the results of these experiments. All materials are also assigned a persistent identifier and come with citation recommendations.

	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.

		Any data generated on external resources, e.g. Clouds, is copied back for long-term storage to the Tier0 or Tier1 sites. The eventual long-term storage / preservation of data in the Cloud would require not only that such services are cost effective but also that they are certified according to agreed standards, such as ISO 16363. 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). Additional services are required for the long-term preservation of documentation (digital libraries), the software to process and/or analyse the data, as well as the environment needed to run these software packages. Such services will be the subject of the on-going self-certification.

3.2 H2020 Data Management Plans for CTA

H2020 Annex 1 Guidelines		
Guideline	Guidance	Statement
Data set reference and name	*Identifier for the data set to be produced.*	The CTA project is an initiative to build the next generation ground-based very high energy gamma-ray instrument. It will serve as an open observatory to a wide astrophysics community and will provide a deep insight into the non-thermal high-energy universe.
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.*	
Standards and metadata	*Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.*	
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.). *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.*	

3.3 Data Management Plans / Status for HEP Experiments

These plans are presented in tabular form as on the DPHEP portal site: http://hep-project-dphep-portal.web.cern.ch/. These plans will be revised at regular intervals and the site should be consulted for the most up-to-date information.
They were prepared as a result of a DPHEP workshop[1] held at CERN in June 2015 and form part of the DPHEP Status Report (DOI: 10.5281/zenodo.34591).
The workshop aimed to:

1. Establish the motivation for long-term data preservation in HEP in terms of succinct Use Cases
    - Are there a common set of Use Cases, such as those that were recently agreed for the 4 main LHC experiments but in a more global scope?
2. Review the existing areas of "Common Projects"
    - Can these be extended (similarly) from their current scope - often LHC - to become more global?
3. Perform a site-experiment round-table to capture the current situation HEP-wide
    - These are summarized in the Status Report and in the tables shown below.

Re-use covers re-use by the Collaboration that initially acquired and analysed the data, by theorists (e.g. to check their models), by the wider scientific community, by the general public and for education / outreach purposes.

3.4 Belle and Belle II

Preservation Aspect	Status (Belle II)
Bit Preservation	Preamble: The central computing system at KEK is replaced every four years. The main user must be Belle II until the data-taking ends (in 2024). Belle: mDST (necessary for physics analysis) is stored on the disk as well as the tape library. The data is still frequently read by active analysis users. All data will be preserved by migrating to the next system. We experienced data loss in the previous data migration. Main causes of this trouble were the short migration period, miscommunication between researchers and operators and the lack of validation scheme after the migration. We will improve the process of the future migration.

[1] See https://indico.cern.ch/event/377026/ for details.

Data	Belle: raw data (1PB) and other format (incl. simulation, ~1.5PB) are stored at the KEK central computing system. This data will be migrated to the next system, at least (data will be preserved until 2020). However, there is no plan thereafter, because the data will be superseded by Belle II. And a full set of mDST was copied at PNNL in USA. Belle II: data taking has not yet started. But raw data will be stored at KEK and another set will be copied in some places outside Japan. Also, the replicas of the mDST will be distributed to the world-wide collaborated computing sites.
Documentation	Belle: all documentation is stored in the local web server and INDICO system. They are still active and accessible, but not well catalogued at all. Belle II: Using twiki, invenio, svn and INDICO system.
Software	Belle: software has been fixed since 2009 except for some patches. The baseline of the OS is still SL5, but it was migrated to SL6. In parallel, the Belle data I/O tool is developed and integrated in the Belle II software. Thanks to this, the Belle data can be analysed under the Belle II software environment. Other Belle handmade analysis tools are being integrated as well. Software version is maintained with SVN. Belle II: basic features witch are necessary for the coming data taking have been implemented. But need more tuning and improvement. The software version is maintained by SVN. SL5/6 32/64-bits, Ubuntu 14.02 LTS are supported
Uses Case(s)	Continued analysis by Belle.
Target Community(ies)	Belle and Belle II
Value	Quantitative measures (# papers, PhDs etc) exist Belle: During the data taking period (1999-2010), averaged number of journal publications is ~30 papers/year and the number of PhDs is ~12/year. After the data-taking, a moderate decreasing tendency can be seen, but the analysis is still active. (~20 publications/year and ~7 PhDs/year).

Uniqueness	Belle: Comparing with the data from Hadron colliders, the Belle data has the advantage of analysing the physics modes with missing energy and neutral particles. Until the Belle II starts, these data are unique as well as *BABAR*'s data (see http://hep-project-dphep-portal.web.cern.ch/content/babar). Belle II: Belle data will be superseded by 2020. After that, the data must be unique samples.
Resources	Belle / Belle II: at some stage, the Belle data must be treated as a part of the Belle II data, and resources for the Belle data will be included in the Belle II computing/human resources.
Status	Construction of the Belle II detector/SuperKEKB accelerator as well as of the Belle II distributed computing system.
Issues	A couple of items have to be implemented in the Belle II software framework to analyse the Belle data. Further check for the performance and reproducibility is also necessary.
Outlook	Expect to be able to analyse the Belle data within the Belle II software framework. It provides us with less human resource to maintain the Belle software, and longer lifetime for the Belle data analysis.

3.5 ALICE

Preservation Aspect	Status (ALICE)
Bit Preservation	On tape: data integrity check during each access request On disk: periodically integrity checks
Data	7.2 PB of raw data were acquired between 2010 and 2013 which is stored on tape and disk in 2 replicas.
Documentation	ALICE analysis train system & bookkeeping in Monalisa DB: for the last 3-4 years Short introduction along with the analysis tools on Opendata
Software	The software package "AliRoot" is published on CVMFS

	For the Open Access the data and code packages are available on Opendata (http://opendata.cern.ch/)
Uses Case(s)	Educational purposes like the CERN Master Classes Outreach activities
Target Community(ies)	Re-use of data within the collaboration(s), sharing with the wider scientific community, Open Access releases
Value	Analysis, publications and PhDs continue to be produced
Uniqueness	Unique data sets from the LHC in pp and HI Similar data can only be collected by the other LHC experiments
Resources	Since the experiment is still working, budget and FTEs are shared with the operation of computing centre
Status	First data from 2010 has been released to the public (8 TB ≈ 10% of data) Some analysis tools are available on Opendata for the CERN Master class program
Issues	Improve user interface The interaction with the open-access portal is very slow due to long communication times. E.g. the uploading of data is done by some people in the IT department. The interaction via an automated website would be faster.
Outlook	Ongoing analysis within the collaboration Making realistic analysis available on the open-access portal Deployment of more data

3.6 ATLAS

Preservation Aspect	**Status (ATLAS)**

Bit Preservation	Non-Reproducible data exist in two or more geographically disparate copies across the WLCG. The site bit preservation commitments are defined in the WLCG Memorandum of Understanding[2]. All data to be reprocessed with most recent software to ensure longevity.
Data	Non-reproducible: RAW physics data, calibration, metadata, documentation and transformations (jobs). Derived data: formats for physics analysis in collaboration, formats distributed for education and outreach. Greatly improved by common derived data production framework in run 2. Published results in journals and HEPDATA. Sometimes with analysis published in Rivet and RECAST. Format lifetimes are hard to predict, but on current experience are 5-10 years, and changes are likely to coincide with the gaps between major running periods.
Documentation	Software provenance of derived data stored in Panda database. Numerous twikis available describing central and analysis level software. Interfaces such as AMI and COMA contain metadata. The publications themselves are produced via the physics result approval procedures set out in ATL-GEN-INT-2015-001 held in CDS; this sets out in detail the expected documentation within papers and the supporting documentation required.
Software	Compiled libraries and executable of the "Athena" framework are published on CVMFS. Software versioning is maintained on the CERN subversion server.
Uses Case(s)	Main usage of data: future analysis within the collaboration Further usage: review in collaboration and potential for outreach
Target Community(ies)	Re-use of data (new analyses) within the collaboration, open access sharing of curated data

[2] WLCG MOU: http://wlcg.web.cern.ch/collaboration/mou

Value	Publications by the collaboration. Training of PhDs
Uniqueness	Unique data sets (both pp and HI) being acquired between now and 2035. Similar data only acquired by other LHC experiments
Resources	The active collaboration shares the operational costs with the WLCG computing centres.
Status	ATLAS replicates the non-reproducible data across the WLCG and maintains database of software provenance to reproduce derived data. Plans to bring run 1 data to run 2 status. Master-classes exercises available on CERN Open Data Portal, expansion considered. Some analyses published on Rivet/RECAST.
Issues	Person-power within the experiment is hard to find. Validation of future software releases against former processing crucial. No current plans beyond the lifetime of the experiment.
Outlook	On-going development of RECAST with Rivet and collaboration with CERN IT and the other LHC experiments via the CERN Analysis Portal as solution to problem of analysis preservation.

3.7 CMS

Preservation Aspect	Status (CMS)
Bit Preservation	Follow WLCG procedures and practices Check checksum in any file transfer
Data	RAW data stored at two different T0 1. 0.35 PB 2010 2. 0.56 PB 2011 3. 2.2 PB 2012 4. 0.8 PB heavy-ion 2010-2013 Legacy reconstructed data (AOD): • 60 TB 2010 data reprocessed in 2011 with CMSSW42 (no corresponding MC) • 200 TB 2011 and 800 TB 2012 reprocessed in 2013 with CMSSW53 (with partial corresponding MC for 2011, and full MC for 2012)

	Several reconstruction reprocessing's The current plan: keep a complete AOD reprocessing (in addition to 2×RAW) • no reconstructed collision data have yet been deleted, but deletion campaigns are planned. • most Run 2 analyses will use miniAOD's which are significantly smaller in size Open data: 28 TB of 2010 collision data released in 2014, and 130 TB of 2011 collision data to be released in 2015 available in CERN Open Data Portal (CODP) Further public releases will follow.
Documentation	Data provenance included in data files and further information collected in CMS Data Aggregation System (DAS) Analysis approval procedure followed in CADI Notes and drafts stored in CDS Presentations in Indico User documentation in Twiki serves mainly the current operation and usage Basic documentation and examples provided for open data users in CODP Set of benchmark analyses reproducing published results with open data in preparation, to be added to CODP
Software	CMSSW open source and available in github and in CVFMS Open data: VM image (CERNVM), which builds the appropriate environment from CVFMS, available in COPD
Uses Case(s)	Main usage: analysis within the collaboration Open data: education, outreach, analysis by external users
Target Community(ies)	Main target: collaboration members Open data: easy access to old data for collaboration members and external users
Value	Data-taking and analysis is on-going, more than 400 publications by CMS

	Open data: educational and scientific value, societal impact
Uniqueness	Unique, only LHC can provide such data in any foreseeable time-scale
Resources	Storage within the current computing resources Open data: storage for the 2010-2011 open data provided by CERN IT, further requests to be allocated through RRB
Status	Bit preservation guaranteed in medium term within the CMS computing model and agreements with computing tiers, but the long-term preservation beyond the life-time of the experiment not yet addressed (storage, agreements, responsibilities), Open data release has resulted in • data and software access independent from the experiment specific resources • a timely capture of the basic documentation, which, although limited and incomplete, makes data reuse in long term possible common solutions and services.
Issues	Competing with already scarce resources needed by an active experiment. Knowledge preservation, lack of persistent information of the intermediate analysis steps to be addressed by the CERN Analysis Preservation framework (CAP) • CMS has provided input for the data model and user interface design, and defining pipelines for automated ingestion from CMS services. • The CAP use-cases are well acknowledged by CMS. • CAP will be a valuable tool to start data preservation while the analysis is active. Long-term reusability: freezing environment (VM) vs evolving data: both approaches will be followed and CMS tries to address the complexity of the CMS data format
Outlook	The impact of the open data release was very positive: • Well received by the public and the funding

	agencies; • No unexpected additional workload to the collaboration; • The data are in use. Excellent collaboration with CERN services developing data preservation and open access services and with DASPOS • Common projects are essential for long-term preservation • Benefit from expertise in digital archiving and library services • Fruitful discussion with other experiments. Long-term vision and planning is difficult for ongoing experiments: • DPHEP offers a unique viewpoint. Next steps for CMS: • Stress-test CERN Open Data Portal with the new data release • Develop and deploy the CMS-specific interface to CERN Analysis Preservation framework.

3.8 LHCb

Preservation aspect	Status (LHCb)
Bit preservation	Data and MC samples are stored on tape and on disk. Two copies of raw data on tape; 1 copy on tape of full reconstructed data (FULL.DST, which contains also raw data); 4 copies of stripped data (DST) on disk for the last (N) reprocessing. Two copies for the N-1 reprocessing. One archive replica on tape.
Data	For the long term future, LHCb plans to preserve only a legacy version of data and MC samples. Run 1 legacy data: 1.5 PB (raw), 4 PB FULL.DST, and 1.5 stripped DST. Run 1 legacy MC: 0.8 PB DST. Open data: LHCb plans to make 50% of analysis level data (DST) public after 5 years, 100% public 10 years after it was taken. The data will be made public via the Open Data portal (http://opendata.cern.ch/) Samples for educational purposes are already public for the International Masterclass Program and

	accessible also via the Open Data portal (For Education area).
Documentation	Data: dedicated webpages for data and MC samples, with details about all processing steps. Software: twiki pages with software tutorials, mailing-lists. Documentation to access and analyse masterclasses samples is available on LHCb webpage and on the OpenData portal.
Software	Software is organised as hierarchy of projects containing packages, each of which contains some c++ or python code. Three projects for the framework (Gaudi, LHCb, Phys), several "component" projects for algorithms (e.g. Lbcom, Rec, Hlt, Analysis), one project per application containing the application configuration (e.g. Brunel, Moore, DaVinci). Software repository: SVN. Open access: once data will be made public, software to work with DST samples will be released with the necessary documentation. A virtual machine image of LHCb computing environment allows to access and analyse the public samples available on the Open Data portal
Use cases	New analysis on legacy data; analysis reproduction; outreach and education.
Targeted communities	LHCb collaboration; physicists outside the collaboration; general public.
Value	LHCb complementary to other LHC experiments.
Uniqueness	Unique samples of pp an HI collisions collected in the forward region.
Resources	Dedicated working group within LHCb computing group.
Status	Legacy software and data releases defined. Development of a long-term future validation framework ongoing. Masterclasses samples and analysis software available via the Open Data portal. Collaboration with CERN IT and other LHC experiments for the development of an analysis preservation framework.
Issues	Main issue is manpower.
Outlook	Collaboration with CERN IT and other LHC experiments on the Open Data portal and the analysis

	preservation framewok. Enrich the Open Data portal with additional masterclasses exercise and real LHCb analysis. Exploit VM technology to distribute LHC computing environment.

3.9 Summary

We have presented Data Management Plans for some of the main communities that are candidates to use procured services through HNSciCloud. Several of these communities have elaborated DMPs prior to the start of the project, with a focus on data preservation, sharing, re-use and verification of their results. Whilst these plans may be more detailed than is required by the H2020 guidelines, they nevertheless reflect the concrete work in these areas and provide a solid basis on which data management-related work in the project can be evaluated.