



International Collaboration for Data Preservation and  
Long Term Analysis in High Energy Physics

# Preparing Data Management Plans for WLCG and HNISciCloud

Jamie.Shiers@cern.ch



HNISciCloud

# What are DMPs about? – Typical DMP Questions

- How will the data be curated and preserved?
  - The strong link to Certification of repositories – another key w/s topic
- How will it be shared / made accessible for verification and re-use?
- How will publications refer to the data (the plots in the paper, the data behind them and so forth)?

# DMP Requirements from Funding Agencies

- To integrate data management planning into the overall research plan, all proposals submitted to the **Office of Science** for research funding are required to include a **Data Management Plan** (DMP) of no more than two pages that describes how data generated through the course of the proposed research will be shared and preserved or explains why data sharing and/or preservation are not possible or scientifically appropriate.
- At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.
- Similar requirements from European FAs and EU (H2020)

From WLCG OB, May 2014. Similar slides shown since at WLCG GDBs

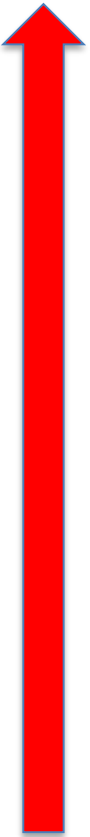
Template	Funder	Funder Links	Sample Plans (if available)
Alfred P. Sloan Foundation 	Alfred P. Sloan Foundation	Alfred P. Sloan Foundation	
Department of Energy: Office of Science 	Department of Energy (DOE)	DOE Statement on digital data management	
Department of Energy: Generic 	Department of Energy (DOE)	Policy for Digital Research Data Management	
Gordon and Betty Moore Foundation 	Gordon and Betty Moore Foundation	Guidelines	GBMF: Sample Plan #1  GBMF: Sample Plan #2  GBMF: Sample Plan #3 
GoMRI Research Consortia DMP Template 2015 	Gulf of Mexico Research Initiative		
Institute of Education Sciences (US Dept of Education) 	Institute of Education Sciences (US Dept of Education)	IES Data Sharing Implementation Guide	
IMLS (2014-): Digital Content 	Institute of Museum and Library Services	Institute of Museum and Library Services (IMLS)	
IMLS (2014-): Research Data 	Institute of Museum and Library Services	Institute of Museum and Library Services (IMLS)	
IMLS (2014-): New Software Tools or Applications 	Institute of Museum and Library Services	Institute of Museum and Library Services (IMLS)	
Joint Fire Science Program 	Joint Fire Science Program		
NEH-ODH: Office of Digital Humanities 	National Endowment for the Humanities	Guidelines	NEH-ODH Sample 
NIH-GEN: Generic 	National Institutes of Health	Guidance	NIH: Sample Plans 
NIH-GDS: Genomic Data Sharing 	National Institutes of Health	Guidance	NIH-GDS: Sample 

# Proposal: WLCG & HNISciCloud

- **For WLCG**, we follow the structure given in **annex 1 and annex 2** of the H2020 guidelines, with a further section covering important points from other key FAs that are not, or not clearly, covered by the above (e.g. linking of publications to data, education and outreach etc - more details below). The guidelines can be found at <https://indico.cern.ch/event/444264/>;
- **For HNISciCloud**, and for each discipline, we follow the structure given in **annex 1 only** of the H2020 guidelines.
  - We may decide to provide additional information in a subsequent update of the DMP.
- **Due as a Deliverable in PM 2 (February 2016!)** with updates in ~PM14 & PM28
- We should **probably** update the DMPs in sync, depending on the schedule in the HNISciCloud DOW.
- HNISciCloud partners: **CERN**, **INFN**, **DESY**, **CNRS**, **KIT**, **SURFSara**, **STFC**, **EMBL**, **IFAE**, (egi, trust-it)
- [DPHEP ISO 16363 training](#): **CERN**, **INFN**, **CNRS**, **KIT**, **STFC**, **IFAE**, dkrz

# H2020: Annex 1 (DMP Template)

The DMP should address the points below...

- 
1. Data set reference and name
    - Identifier for the DS to be produced
  2. Data set description
    - Description; origin; nature & scale; to whom useful; underpins publication? similar data?
  3. Standards and metadata
    - Reference to standards *of the discipline*
  4. Data sharing
    - How will it be shared? Embargo periods? Mechanisms for dissemination, s/w and other tools for re-use, access open to restricted to groups, where is repository? Type of repository?
  5. Archiving and preservation
    - Description of procedures, how long will it be preserved? End volume? Costs? How will these be covered?

# H2020 DMP Guidelines – Annex 2

Are the data:

- Discoverable,
- Accessible,
- Assessable and
- Intelligible,
- Usable beyond their original purpose,
- Interoperable to specific quality standards?





# “Annex 3” – other key questions

- What is the relationship between the data you are collecting and any existing data? **(NSF)**
- Requirement #2: DMPs should provide a plan for making all research data displayed in publications resulting from the proposed research open, machine-readable, and digitally accessible to the public at the time of publication.
  - This includes data that are displayed in charts, figures, images, etc.
  - In addition, the underlying digital research data used to generate the displayed data should be made as accessible as possible to the public in accordance with the principles stated above.
  - This requirement could be met by including the data as supplementary information to the published article, or through other means.
  - The published article should indicate how these data can be accessed. **(DoE)**
- See November 2015 workshop at CERN: <https://indico.cern.ch/event/395374/other-view?view=standard#20151109.detailed>

# Some Words about the Workshop

Current schedule:

- Wednesday pm
  - **DPHEP Status Report: main changes wrt Blueprint (2012)**
  - **Certification of Sites – Motivation**
  - **Overview of Certification Process**
  - **(Active) Data Management Plans – Introduction**
- Thursday
  - **(Self) Certification of WLCG / HNISciCloud Sites (Repositories)**
  - **(Active) Data Management Plans for WLCG & HNISciCloud**

Thursday in particular should be highly interactive

Timing may vary, e.g. Status Report presentations may expand

Vidyo will be available but is a poor substitute...

# DMP Timeline

- HNISciCloud: DMP v1 as Deliverable: February 2016
- (WLCG: draft, but more comprehensive DMP: Q1 2016)
  
- HNISciCloud: DMP v2: + 1 year
- WLCG: DMP v2: + 1 year
  
- **“We” (DPHEP) will draft the first version of the HEP content to be reviewed at the DPHEP w/s in Lisbon**
  
- **First drafts available for comment: Wednesday 27<sup>th</sup> January 2016**

# Strategic Directions (DP + OD)

- ✓ Define a **Data Preservation (Management) policy** for CERN experiments
  - Perhaps linked to the approval process through the Research Board & monitored through reviews
  - See HL-LHC ESFRI Roadmap questions (and answers)
- ✓ At least a “**self-audit**” for the CERN Tier0 and WLCG Tier1 sites in the context of the WLCG project
  - See WLCG/DPHEP Workshop in Lisbon: Certification and Data Management Plans
- ✓ Extension of DPHEP’s activities to consider also those of potential FCCs
  - Presentation of DPHEP BLUE TOO conclusions to FCC in Rome proposed (**but not a lot of traction**)
  
- Further developments in terms of Analysis Capture and Preservation
  - Use Cases & Metrics to validate DP strategy (AKA “DSA++”)
- Further releases of Open Data through the CERN Open Data Portal
- [ Harmonization of similar activities across various laboratories and projects ]
- Clarifications regarding funding – of particular importance to past experiments where resources have already become sub-optimal
- The continuation of regular meetings and workshops, aligning as much as possible with related events (WLCG, CHEP, HEP Software Foundation etc.)
  - Proposal for “DPHEP track” @ CHEP 2016 made (successful @ CHEP 2012)
  
- ✓ **Input to the next round of ESPP – building on concrete experience, results and remaining challenges.**
- **Data Re-use, Sharing, Reproducibility of Results are key drivers**



## 3.5. Will there be need for an adjustment of the general CERN's data policy?

- CERN will establish a data policy that is in line with funding agency requirements, including in terms of Open Access (Science). This can be expected to be largely similar to that adopted by the 4 main LHC experiments, with a significant fraction of the data released after a reasonable embargo period.
  - The duration of the embargo period and the fraction of the data to be released would be determined based on experience, resource requirements and scientific, educational and cultural benefits.
  - Given that the total dataset of the (HL-)LHC will be in the Exabyte range, the volume of data to be released will eventually become significant and the appropriate resources must be factored into any planning.
- The development and use of the computing models and the infrastructure for HL-LHC does not depend on development of this policy.

Margaret Hamilton, lead software engineer of the Apollo Project, stands next to the code she wrote by hand and that was used to take humanity to the moon. [1969]

