Journal of Physics: Conference Series **513** (2014) 042041 doi:10.1088/1742-6596/513/4/042041

DPHEP: From Study Group to Collaboration

Jamie SHIERS

CERN, 1211 Geneva 23, Switzerland
Jamie.Shiers@cern.ch

Abstract. The international study group on data preservation in High Energy Physics, DPHEP, achieved a major milestone in 2012 with the publication of its eagerly anticipated large-scale report [1]. This document contains a description of data preservation activities from all major high energy physics collider-based experiments and laboratories. A central message of the report is that data preservation in HEP is not possible without long term investment in not only hardware but also human resources, and with this in mind DPHEP will evolve to a new collaboration structure in 2013. This paper describes the progress made since the publication of that report – shortly before CHEP 2012 – as well as the future working directions of the new collaboration.

1. Introduction

Shortly after the publication of the full DPHEP report – also known as the DPHEP Blueprint – a summary was submitted to the Open Symposium [2] on the update of the European Strategy for Particle Physics. It is encouraging to note that Data Preservation was retained as part of the updated strategy, as adopted by the CERN Council at the 16th European Strategy Session [3], held in Brussels in May 2013. The full strategy document [4] includes the following paragraph:

"The success of particle physics experiments, such as those required for the high-luminosity LHC, relies on innovative instrumentation, state-of-the-art infrastructures and large-scale data-intensive computing. Detector R&D programmes should be supported strongly at CERN, national institutes, laboratories and universities. *Infrastructure and engineering capabilities for the R&D programme and construction of large detectors, as well as infrastructures for data analysis,* ***data preservation*** *and distributed data-intensive computing* ***should be maintained and further developed****."*

In parallel, the International Committee for Future Accelerators (ICFA) [5] welcomed the Blueprint document and encouraged the move of DPHEP from a study group to an International Collaboration. The full text of this statement is reproduced below. At the time of writing, CERN has signed this agreement, with signatures expected from (at least) DESY, MPI, INFN and IN2P3 in the coming months.

2. ICFA Statement on Long-Term Data Preservation (March 2013)

"The International Committee for Future Accelerators (ICFA) supports the efforts of the Data Preservation in High Energy Physics (DPHEP) study group on long-term data preservation and welcomes its transition to an active international collaboration with a full-time project manager. It

encourages laboratories, institutes and experiments to review the draft DPHEP Collaboration Agreement with a view to joining by mid- to late-2013.

ICFA notes the lack of effort available to pursue these activities in the short-term and the possible consequences on data preservation in the medium to long-term. We further note the opportunities in this area for international collaboration with other disciplines and encourage the DPHEP Collaboration to vigorously pursue its activities. In particular, the effort required to prepare project proposals must be prioritized, in addition to supporting on-going data preservation activities.

ICFA notes the important benefits of long-term data preservation to exploit the full scientific potential of the, often unique, datasets. This potential includes not only future scientific publications but also educational outreach purposes, and the Open Access policies emerging from the funding agencies."

3. DPHEP Project Manager

As a result of these discussions, the CERN Director for Research and Computing, Dr. Sergio Bertolucci, wrote to the chair of the DPHEP Study Group, Dr. Cristinel Diaconu, offering to fund the DPHEP Project Manager position that was proposed in the DPHEP Blueprint for a period of 3 years, starting from 1 January 2013. It is foreseen that the initial deliverables that were defined in the Blueprint would be achieved during this period, following a startup phase during which the group migrated to the new structure. CERN also offered to participate in specific data preservation activities and in particular those mentioned in the submissions from IT and PH departments to the update of the European Strategy for Particle Physics.

4. The DPHEP Vision

The "DPHEP vision", which is used to frame our work, can be summarized as follows:

- By 2020, all archived data – e.g. that described in Blueprint, including LHC data – easily findable, fully usable by designated communities with clear (Open) access policies and possibilities to annotate further;
- Best practices, tools and services well run-in, fully documented and sustainable; built in common with other disciplines, based on standards;
- A key component of this vision is a **DPHEP portal**, through which data and tools are accessed.

5. DPHEP Deliverables

The initial deliverables of the DPHEP Collaboration are given in the table below. All of these objectives are being actively addressed and will be the subject of the remainder of this paper. For a variety of reasons – that hopefully will become clear – they will be addressed in reverse order.

Objective	Deliverable
1.Positioning as forum	Catalogue of technical knowledge and practical solutions Description of possible alternatives for governance.
2.Co-ordination of projects	Common R&D projects meet the expectations of the stakeholders.
3.Harmonisation and liaison	Synchronisation of preservation projects in the field. Identification of areas where external knowledge needs to be transferred to HEP.
4.Design sustainable future	Characterisation of discipline-wide toolkit for preservation Business plan for long-term preservation in HEP.
5.Outreach and advocacy	Understanding of needs/opportunities for medium- and small-sized collaborations. Concrete discussions with funding bodies/laboratories.

6. Outreach and Advocacy

In the recent past, various funding agencies and inter-governmental bodies have put increasing emphasis on Open Access: firstly to publications and increasingly to data. In parallel, a number of initiatives have been launched on data sharing and inter-operability. Whilst these discussions have gone beyond data preservation per se, they have nevertheless included it and provided concrete platforms for discussions with funding agencies. This work goes beyond what was initially intended in the list of objectives above, but has done much to build links with other disciplines and emphasize to funding agencies what we can bring – in terms of experience, as well as concrete technologies and even services – to the global problem. Two particular fora are worthy of note, namely the *Research Data Alliance (RDA)* and the *Alliance for Permanent Access (APA)*. It is not an exaggeration to say the contacts made through these two fora have had a significant impact on the design for sustainable solutions for the HEP community, as well as positioning us as important experts and contributors in the field. This work has been very visible to the funding agencies and in particular the EU (Europe) and the NSF (US).

7. Designing a Sustainable Future

Sustainability is a key issue for an activity such as long-term data preservation. This includes not only the costs and funding model(s), but also the commitment. These need to be "institutional level" – or even higher. It is for this reason that the support from the updated European Strategy for Particle Physics and ICFA is felt to be particularly important: we need commitments that go way beyond the term of any individual appointment (director, department head, group leader etc.) A specific example of long-term sustainability is that of "bit preservation". Under the auspices of HEPiX – the international coordination body for "Unix Information Exchange", a working group on bit preservation has been established. HEPiX itself is over 20 years old and it, or an evolution of it, can be expected to continue for decades hence. The working group has a mandate to coordinate efforts across laboratories, reducing costs and eliminating duplication of effort. It has also studied the costs of bit preservation with a concrete financial plan for about one decade (part of the "medium term plan" at CERN) and an outlook for much longer. Basically, there is a commitment to preserve the bits (at the current scale of 100PB and rising to a few EB) for at least several decades.

"Bit preservation" can be considered a concrete implementation of a "common project", as described in the DPHEP Blueprint.

Other sustainable elements include the use of the INSPIRE and HEPData services – although with other associated components. These too are so mainstream that they, or an equivalent – presumably superior – service will be guaranteed even in the long term.

8. Harmonisation and liaison

A number of disciplines other than HEP have been active in data preservation for long periods – even decades in cases such as astronomy. Others have come to the field more recently but have still significant experience and knowledge that complements that of HEP. Not unsurprisingly, HEP focuses very much on data related issues, including volume, as well as the daunting task of maintaining the offline software chain in a usable fashion – through a variety of techniques – for long periods. In contrast, other disciplines have concentrated more on areas that include provenance, governance, meta-data and so forth. Thus, there is a potential win-win situation for all parties in collaborating together. This is being carried out both through the RDA – where the focus includes understanding which other data-related technologies and "standards" maybe relevant – and the APA, which is more an inter-disciplinary forum.

Within the HEP community, and as witnessed by the data preservation-related posters at CHEP 2013, there are many potential areas for harmonization and collaboration. These include not only the use of the "common services", such as bit preservation, digital library and related services, but also in defining the requirements for "validation systems", in developing strategies for sustainable software, and so forth.

DPHEP will continue to be active in both these respects and will place increased focus on "common projects" as from early 2014.

9. Co-ordination of projects

A success of the DPHEP workshop that was held in conjunction with CHEP 2013 was the identification of areas were services already exist versus those where "common projects" were more appropriate. As mentioned above, this will be a key focus of the DPHEP Collaboration over the next period – up to CHEP 2015 and possibly beyond. These common projects could include the following:

- The DPHEP "portal" – a common entry point into archived data and the "knowledge base";
- A software validation system / framework – inspired by the work at DESY and elsewhere and updated to include requirements coming from running experiments;
- A virtual "museum system" – how to ensure the longevity of virtual machines in a coordinated, sustainable and durable way;
- A "CERNLIB consortium" to maintain the key components of the CERN Program Library – still heavily used by many pre-"third millennium" experiments.

10. Positioning as forum

DPHEP remains the dominant – but not unique – forum where data preservation issues are discussed and shared within the HEP community. The development of a DPHEP portal, the catalogue of sustainable solutions, together with the continuation of regular meetings and workshops will help secure this position. DPHEP is also being increasingly known in external communities and projects, an example being our detailed analysis of costs of bit preservation. These numbers are being shared with other projects – including those focusing explicitly on these aspects – and a "Full Costs of Curation" workshop is planned for January 2014. It is expected that the outcome of this workshop will not only help us build a concrete case (Use Cases; Business Cases; Costs and Cost Models) for discussion with funding agencies but also drive our work by identifying areas where optimization is possible and desirable.

11. Conclusions

Since CHEP 2012, DPHEP has made significant progress in the key areas outlined in the DPHEP Blueprint. A number of key, sustainable, services have been identified, together with areas more suitable for projects. Excellent contacts have been established and/or strengthened with other communities and projects. The "2020 vision" could even be achieved somewhat earlier – CHEP 2015 will be an important checkpoint.

Long-term data preservation in HEP is affordable and technically possible – provided that it is given the appropriate priority.

The "Full Costs of Curation" workshop, scheduled for January 2014, will enable us to write a roadmap for its implementation.

References

[1] Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics: http://arxiv.org/pdf/1205.4667.pdf.
[2] See http://espp2012.ifj.edu.pl/ for the programme and material of the workshop.
[3] See https://indico.cern.ch/internalPage.py?pageId=4&confId=244974 for details of the 16th European Strategy Session of the CERN Council.
[4] See https://indico.cern.ch/conferenceDisplay.py/getPic?picId=59&confId=244974.
[5] The Website of ICFA is maintained at http://www.fnal.gov/directorate/icfa/.
[6] The ICFA statement on Data Preservation can be found at: http://www.fnal.gov/directorate/icfa/ICFA_Statement_on_DPHEP_(wt).pdf
[7] The Research Data Alliance (RDA) – https://rd-alliance.org/node.

[8] The Alliance for Permanent Access (APA) – http://www.alliancepermanentaccess.org/.
[9] The DPHEP "Full Costs of Curation" workshop – https://indico.cern.ch/conferenceDisplay.py?confId=276820.