DPHEP-2014-001-V1
January 2014

DPHEP Status Update

Executive Summary

Over the past 12 months, significant progress has been made in the following areas:

- Understanding in detail the costs of curation and preservation, in terms of materials and personnel, for both past and current experiments. A "business plan" is in preparation (March 2014);
- Moving from proposed "common projects", as identified in the DPHEP Blueprint document, to concrete services and solutions, with a sustainable outlook. A workplan is in preparation (March 2014);
- Building and strengthening bi-directional links with other disciplines and projects. As an example, the DPHEP Project Manager is co-chair of the Research Data Alliance Interest Group on Preservation and has been elected to the Executive Board of the Alliance for Permanent Access – both multi-disciplinary data and knowledge preservation bodies.

A better understanding of the role and benefits of the DPHEP Collaboration has been achieved; the DPHEP Collaboration agreement has been finalised and has been signed by CERN, with signatures expected from (at least) DESY, INFN and IN2P3 in the foreseeable future.

Costs of Curation

It is not our intent to give a detailed business plan for long-term data preservation here but this will be summarised in a forthcoming document, expected by March 2014. However, it is understood than the "materials" budget for long-term data preservation, which may start as quite significant, tends to zero (after one to two decades), whilst the personnel budget is consistent with that outlined in the DPHEP Blueprint. As a concrete example, the annual costs of data preservation of the resurrected JADE data and software are around $100K. Similarly, a few per cent of the WLCG budget would be sufficient to fund common data preservation activities for those past experiments where the data and "knowledge" is still available (including JADE, LEP, Hera, Tevatron, BaBar) as well as on-going activities for the LHC experiments. This can easily be compared with the "scientific value" of such preservation, to build a clear business case.

Common Projects, Services and Solutions

At the above-mentioned workshop, it was agreed to structure on-going DPHEP activities around the following lines:

- A DPHEP portal – built using inspiration and guidance from existing portals from other scientific disciplines – as the main entry point to archived data, access policies, documentation and other needed material;

- Consistent use of tools such as INSPIRE, HepDATA and Rivet for documentation, publication-level data and links to saved analyses;
- The use of CernVM/CernVMFS to allow reproduce simulation / reconstruction and analysis on archived data, together with validation frameworks and sustainable software techniques;
- "Bit-level" preservation, coordinated across the laboratories via a HEPiX working group with transparent and published reliability data;
- Work on "open data" formats, validation and analysis tools, to broaden the usability of preserved data and greatly increase its potential longevity.

It is understood the close collaboration between experiments and laboratories on the above items can not only cover the full spectrum of needed data preservation activities, but also result in significant manpower savings (through the elimination of duplication) and also in more robust and flexible solutions.

Links with Other Disciplines

In addition to the above-mentioned formal roles, DPHEP presentations have been made at a number of International Conferences and workshops during 2013, and DPHEP has established close links with the EU 4C project – A Collaboration to Clarify the Costs of Curation. These activities have helped to make others aware of our work – and in particular our areas of expertise, such as in large-scale bit preservation, where we currently lead the (scientific) field. It has also allowed us to benefit from the work and ideas of others, including the Astronomy and Space Agency communities – both well advanced in many fields of knowledge capture and data preservation – as well as in formulated and presenting cost models and business cases (APARSEN and 4C projects).