DPHEP-2013-001-V1
February 2013

Long-Term Data Preservation in High Energy Physics: A 2020 Vision

www.dphep.org

Abstract

This report presents a 2020 vision for *Long-Term* Preservation of Data in High Energy Physics. The goal is that by 2020 all existing efforts on Long-Term Preservation in HEP will be fully harmonized, both internally and also with other disciplines. The target is that all *archived* data – such as that described in the DPHEP Blueprint, including LHC data – remains fully usable by well-defined *designated* communities. The best practices, tools and services should all be well run-in, fully documented and sustainable. Any exceptions to the above should be clearly described. A key element of the strategy is to adopt the widely used *Open Archival Information System reference model*, first produced in 2002 and for which an ISO standard exists. The reference model was last revised in June 2012.
This "vision" is believed to be achievable, subject to some basic conditions regarding resources in the immediate future.
It is useful to introduce a project/discipline independent definition of "long-term". Simply put, this means preparing for – and adapting to – disruptive changes.

Study Group for Data Preservation and Long Term Analysis in High Energy Physics

Jamie Shiers, CERN, DPHEP Project Manager

Introduction

2012 saw the publication of the Blueprint Document from the DPHEP study group (DPHEP-2012-001). A summary of this document was used as input to the Krakow workshop on the future European Strategy for Particle Physics and it is to be expected that Data Preservation will be retained in the updated strategy to be formally adopted by the CERN Council in May 2013.

The same year also saw a number of other important events directly related to data preservation and data management in the wider sense.

In July, the European Commission published a Recommendation on "access to and preservation of scientific information" and a workshop was held in October to investigate the level of European and International coordination that was desired in this area (with an obvious conclusion). In conjunction, various funding agencies are adding requirements on data preservation and open access plans to future funding calls.

In parallel, an activity now named "The Research Data Alliance" (RDA) was established with support from the US, the EU and Australia (other countries in Asia-Pacific are expected to join) "to accelerate and facilitate research data sharing and exchange" and a working group on data preservation is in the process of being established. There are very clear signs that the output of the RDA will have a direct influence on funding as part of the EU's Horizon 2020 programme and presumably also elsewhere in the world.

Activities related to these events also allowed us to strengthen our collaboration with numerous other disciplines: not only those with whom we have long had ties, such as Astronomy and Astrophysics, but also other scientific communities as well as arts and humanities, all of whom are deeply involved in data preservation activities.

Basking in the scientific results of the LHC in 2012, there is a clear halo effect to be exploited.

Following the proposal by the CERN Director for Research and Scientific Computing to support the requested position of a DPHEP project manager (for an initial period of 3 years), DPHEP is moving from being a study group to be an active international and interdisciplinary collaboration. An initial set of goals is listed in the DPHEP Blueprint and is currently being actively worked on.

This report outlines the rapid progress that was made particularly in the second half of 2012 as well as the exciting opportunities for the future. It is framed in terms of a simple "vision" based on the Open Archival Information System model.

The DPHEP Collaboration Agreement

As part of the transition from a study group to an active collaboration, a multi-lateral Collaboration Agreement, based on text from the DPHEP Blueprint and other existing such agreements, has been drafted by the CERN Legal Service. After a period for comments and corrections, it is foreseen that most parties will be able to sign this agreement on the timescale of the June ICFA meeting.

The current draft is appended to this document for information and comment.

The OAIS Reference Model

This model introduces a number of simple concepts that are useful, such as:

- Producers and consumers of archived data (no explanation needed);
- The *knowledge base* of the expected consumers (also referred to as the *designated community*). One should always assume a lower knowledge base

than that of the producer, even if producer and consumer are the same (but separated in time);
- *Information packages* that describe the contents of the archive and how it can be used. The main such package is the *Archive Information Package* which describes a set of data and contains pointers to all information, software, metadata and so forth that is required to use the data for its intended purpose.

The model has already been adopted by numerous and highly diverse communities and there appear to be no obstacles to its adoption within HEP. Furthermore, any joint projects with other disciplines are likely to require at least loose adherence to this originally *de facto* but now *de jure* standard.

The "2020 Model"

The proposed model is deceptively simple. Significant work is required to implement it but many of the elements already exist, at least in some form.

- The Archive Information Packages are described by simple XML documents, stored in one (or preferably more) *Invenio* (or similar) instances;
- Reliable storage systems with rule-based engines manage the back-end, including running regular integrity checks, invoking simple applications defined by the "producer" and issuing the necessary warnings in case of exceptions;
- Semi-automated systems assist with "simple" changes, such as new operating system or compiler versions, new release of database or other software etc, e.g. the DESY Generic Validation Framework (that could also be of value to "live" experiments in dealing with constant migrations);
- Regular and meaningful scientific / outreach / challenge / training exercises use the archived information;
- The full scientific and cultural potential of the data is thus maintained.

This sounds deceptively simple: in fact there are many issues that need to be addressed in even the most trivial of cases. For example, "repack" of data onto new media is a risky and error prone operation: in October 2012 nearly 1000 LHCb files were "lost" during a repack operation (but fortunately all but 2 were recoverable from other sites in WLCG). However, this is far from the only case that has occurred and constant attention is currently required.

A much larger issue concerns changes in the offline environment that we know from experience are bound to happen. Here we need to learn from experience: all completed experiments successfully addressed major changes in the offline environment, moving from mainframes to proprietary Unix systems to x86 architecture and Linux, rewriting major parts of the simulation, reconstruction and analysis software (e.g. moving from Patchy + Zebra + Fortran to C++) and major changes in the data format have also been successfully accomplished. What we have not systematically done is prepare for change: doing so now for LHC would have a small incremental cost for the first (LS1?) migration but would have a (possibly significantly) lower integrated cost over the lifetime of the experiments, as well as positioning the data much better for long-term preservation. It simply needs to be part of the overall plan and resourced accordingly.

A Sustainable Cloud Data Infrastructure

An essential part of any long-term data preservation strategy is a sustainable storage infrastructure. Numerous national libraries have digital archives of around 1PB each, with national archives around the 10PB mark. The LHC "archive" is now close to 100PB.

Not only is it outside the core business of many institutes to build up multi-PB archives, but also there are potentially economies of scale in offering a collective service. Attempts to build a financial model for commercial cloud-based storage highlight not only the relatively high cost of the storage itself but also the prohibitive copy-in / copy-out costs.

A solution that has been discussed briefly within the data preservation community is to establish an academic and research cloud storage mutual. Start-up funds could come, e.g. from the EU, but the on-going costs would come from the consumers of the service. High copy-in / copy-out costs would be avoided through the use of National and International Research and Education Networks and the service would be offered "at cost" – with fully transparent pricing based on hardware acquisition and operations costs. The service could be offered by subscription or based on a mutual amongst a group of organisations sharing the costs *pro rata*.

Initial discussions, including with the EU as a potential funding agency, have been positive.

This is not an integral part of the overall 2020 vision. However, it offers to other communities a tangible example of the advantages of working with leaders in the large-scale storage and data management arena. Subject to power and cooling constraints, one could imagine that one or more large laboratory could offer to host an instance of such a service.

Collaboration with other Disciplines / Activities

One of the most surprising discoveries of recent months is the amount of interest and level of on-going activities in the data preservation area. The Science and Technologies Facilities Council (STFC) in the UK is active in a number of these, as is the Digital Curation Centre, also in the UK. The Space Agencies – who developed the OAIS model – are also particularly active and have been successful in attracting EU funding in both FP6 and FP7. The Astronomy and Astrophysics communities have well-established policies for making data publically available, as well as standard formats and tools. Earth Science – a much more heterogeneous community than HEP – is also attracting significant attention from funding agencies, including for EU-US-AU collaboration. A number of fora (e.g. the Alliance for Permanent Access), conferences (e.g. the annual International Data Curation Conference, iPRES (International Conference on Preservation of Digital Objects), PV (a bi-annual event on "ensuring long-term **P**reservation and adding **V**alue to scientific and technical data") and workshops (e.g. topical workshops on data preservation) exist: we have started to actively participate in these and our contributions have been welcomed. Other key events of recent months include the e-Infrastructure Reflection Group meeting on "data", which again was a chance not only to discuss with other communities but also to provide input to the Horizon 2020 programme.

The Crux of the Matter

Of all the different elements of a successful long-term data preservation strategy, by far the most complex – and least studied to date – is that of maintaining the software

and associated environment usable for long periods and in adapting it to changes, particularly in the period when the original authors and experts are no longer available.
On the timescale of 2020, it is possible that an SL(C)6 based environment will offer the needed stability (the final end of life of RHEL6, with which SL(C)6 claims binary compatibility, is November 20 2020) and it is not unreasonable to believe that existing offline software can be made to work correctly in such an environment – in some cases this is already done.
However, this – and / or the use of virtual machines or "museum systems" – does not fully address the "long-term" aspect.
Addressing the associated issues for longer periods should logically be a key element of any future project in this area, e.g. one funded under the Horizon 2020 programme, with strong collaboration with leading experts in this domain.

S.W.O.T. Summary

The table below loosely characterizes the project's Strengths, Weakness, Opportunities and Threats.

Strengths	DPHEP is well established within the community and recent contacts to other disciplines are very encouraging.
Weaknesses	Effort is very scarce within the project at a time when manpower is already stretched to the limit elsewhere.
Opportunities	Through a convergence of events there are clear possibilities for significant funding and collaboration in the EU's Horizon 2020 programme and most likely corresponding programmes in other areas of the world.
Threats	Failure to invest now would jeopardise attempts to "rescue" LEP data as well as to take other preservation events (BaBar, Tevatron, Hera etc.) to a stable and sustainable state. It could also limit our ability to prepare for – and hence participate in – future projects.

Summary of Recommendations

1. Adopt the OAIS model across HEP Long-Term Data Preservation Projects;
2. Actively participate in the Research Data Alliance and its Working Groups with the intent to influence EU, US and other funding agencies;
3. Build on existing contacts with other disciplines to share best practises and where possible also tools and services (e.g. data and "library" storage services);
4. Ensure that there are sufficient resources (2013 / 2014) to allow the vision of Long-Term, Sustainable Archives to be realized.
5. Address true long-term preservation by R&D into handling change in all areas of the archive environment and in particular that of the software and offline infrastructure.