Proceedings of the 2018 PV Conference,
15 – 17 May 2018, Harwell Space Cluster, UK

COLLABORATIVE LONG-TERM DATA PRESERVATION: FROM HUNDREDS OF PB TO TENS OF EB

Jamie Shiers

CERN, 1211 Geneva 23, Switzerland

Abstract

In 2012, the Study Group on Data Preservation in High Energy Physics (HEP) for Long-Term Analysis, more frequently know as DPHEP, published a Blueprint report detailing the motivation for, problems with and situation of data preservation across all of the main HEP laboratories worldwide. In September of that year, an open workshop was held in Krakow to prepare an update to the European Strategy for Particle Physics (ESPP) that was formally adopted by a special session of CERN's Council in May 2013 in Brussels and key elements from the Blueprint were input to that discussion. A new update round to the ESPP has recently been launched and it is timely to consider the progress made since 2012/2013, list the outstanding problems and possible future directions for this work.

INTRODUCTION

The current strategy for European Particle Physics, the main emphasis of which is clearly on physics, makes only a brief mention of data preservation:

*"… as well as infrastructures for data analysis, **data preservation** and distributed data-intensive computing **should be maintained and further developed**."*

We first describe how these somewhat enigmatic targets have been met in both a quantitative and qualitative fashion and then describe our goals for the up-coming revision of the strategy, including outstanding issues and concerns.

In addition, there have been a number of key developments both within and outside HEP that change the environment in which we work. These include:

1. Production services – at least at CERN – now exist for all of the main areas identified in the DPHEP Blueprint. These are described in an iPRES 2016 paper;
2. Significant additional emphasis is now placed on the need for data preservation, sharing and re-use through FAIR data management principles and so forth, both by funding agencies as well as research teams;
3. Cross-disciplinary collaboration continues to develop, following on from the Alliance for Permanent Access to Working and Interest Groups in the context of the Research Data Alliance;
4. The importance of Certification – a *sine qua non* – according to many viewpoints, continues to grow. At a January 2018 Science Europe workshop it was announced that a policy for listing only certified repositories – at least to the level of CoreTrustSeal – could be expected within 5 years (this is expected to be one of the recommendations of the FAIR Data Action Plan;
5. The volume of data to be preserved, as well as the "business cases" for so doing, have become much more quantifiable.

However, new challenges have arisen, emphasizing the fact that "data preservation is a journey – not a destination".

SITUATION AT THE TIME OF THE BLUEPRINT

At the time of the Blueprint, it was not uncommon for media migration to be under the full control and responsibility of the experiment / project in question. In other words, "bit preservation" as a service was not the norm and this is still the case at some laboratories. Indeed, one of the inputs to the 2012/13 ESPP update dismissed bit preservation as simply "copying the file a couple of times".

Although the use of Invenio-based services, such as INSPIREHEP, or more recently Zenodo or B2SHARE, for storing documentation was a fairly common practice, this covers only a small fraction of the experiment-specific documentation and "knowledge".

Validation of software was an area of concern with some promising work on (semi-)automatic validation frameworks (versus "museum systems").

FROM A STUDY GROUP TO A (HEP) COLLABORATION

In order to implement the recommendations of the Blueprint, the Study Group became a Collaboration with a Collaboration Board and an Agreement signed by most of the key laboratories and some of the funding agencies. A "2020 Vision" was presented in February 2013 to the International Committee for Future Accelerators (ICFA) – a high-level body formed by Directors for HEP institutes worldwide to whom DPHEP reports.

The most recent report – in March 2018 in Cambridge – highlighted a specific problem whereby the 2PB of data collected (and still actively analysed) by the BaBar collaboration at Stanford Linear Accelerator Center may have to find a new home in the relatively short term. Through the DPHEP Collaboration, the possibility of storing the 2PB of data in CERN's tape robots (and/or at an institute that is part of the BaBar collaboration) is being investigated. This has been given Director-level approval but needs further discussion, including implications such as making at least some of the data available through CERN's Open Data portal. (It is the author's strong opinion that the data should be copied to CERN during LHC LS2 – see below – and before the CERN Directorate changes in 2021!)

Other examples of "internal HEP" collaboration is the establishment of a CVMFS repository at CERN for non-CERN experiments where the necessary (modest) resources can no longer be provided by the former host laboratory.

PROGRESS SINCE THE DPHEP BLUEPRINT

As documented in the iPRES 2016 paper, CERN now offers production services in all of the above three areas. There are constant improvements to "bit preservation", even as the data volumes continue to grow and as additional concerns arise, such as the reduction of enterprise tape vendors from two (Oracle and IBM) to one (just IBM). Clearly, alternative but nevertheless cost effective solutions are being investigated.

The use of CernVM and CVMFS (to snap-shot all the different s/w versions and environment needed by the preserved data) has become almost a de-facto standard across HEP data preservation (and also production) efforts and has been a significant success story that was not predicted by the Blueprint. (A paper on its potential was, however, submitted to the ESPP update).

Open Data releases, most notably by the CMS experiment who have now released over 1PB of data, together with the associated documentation and software, has been another big win, with successful re-use of the data by a variety of communities (including leading to a publication not involving any members of the CMS collaboration).

ISO 16363 CERTIFICATION

As has been described elsewhere, CERN is pursuing ISO 16363 for its range of data preservation activities, including Scientific Data (e.g. that from the LHC and former LEP collider), its "digital memory" (videos, photographs, minutes etc.) as well as papers, reports, proceedings, circulars and so forth. Many of the metrics in ISO 16363 are matched by existing CERN (documented) practices, whereas in a small number of cases, e.g. business continuity and disaster preparedness, there is still work to do (the latter being done in coordination with other EIROforum institutes). Given the cost and value of CERN's data (together with the fact that we believe we are already quite closely aligned with this "most thorough" standard), we believe that this is the most suitable approach. We fully understand that it might not be appropriate for all projects / institutes / communities but a "lighter-weight" approach simply does not give the same level of assurance. (Think of the standards that you would like to see applied before blasting off in a unique space rocket). We are clearly indebted to others who have trod the certification path before us and the best advice that we could probably give is the same as for answering examination questions – *"read the question before you try to answer it"*.

THE HEP COMMUNITY WHITE PAPER

Published at the beginning of 2018, the Data Preservation chapter of the HEP Community White Paper that focuses on the challenges of the next decade or so, confirm the above achievements. For example, it notes *"**bit preservation with an acceptably low error rate can now be considered a solved problem**"*.

However, it lists other areas of future work that match closely with "the new world order", where increasing emphasis is placed on reproducibility of results (as well as re-use of the data for new analyses and purposes).

ANALYSIS CAPTURE AND PRESERVATION

One of the main new areas of work since the publication of the Blueprint has been on tools / services to capture enough information, including work-flows, to enable key analyses to be repeated and acceptably similar (not identical) results to be obtained. There is no space to describe this work in detail but further information can be found via https://analysispreservation.cern.ch/welcome and https://github.com/reanahub.

OPEN DATA AT THE MULTI-PB SCALE

Given that the LHC has just completed its 2nd "run[1]", with several more multi-year runs (and significantly more data to be acquired and analysed) before it, we can expect Open Data volumes well in excess of 10PB, possibly in excess of 100PB and conceivably even more.

Some issues around Open Data are largely independent of volume: for example, if an experiment releases 1% or 50% of a given dataset, the same amount of documentation and software will still be needed.

However, many people assume that Open Data implies "zero or low latency" that can have a significant resource impact on large data volumes. Is this required? Is it even useful? We know that recently released data is frequently accessed but this then falls away over a period of weeks. Interest may be re-kindled by future events and it may be more appropriate to have a more modestly sized cache for recent data, together with "featured data", that may include "data on request". This will require further discussion and clarification between data producers, consumers, service providers and of course funding bodies.

[1] LHC Run1 took place from 2009 to early 2013 and was followed by a "long shutdown" (LS1). Run2 started in mid-2015 and will continue until late 2018 when LS2 starts. Run3 should begin in mid-2021 and so on.

MIND THE GAP(S) – INCLUDING F & A

From a CERN viewpoint, the first data that has been successfully preserved is that from the 4 LEP experiments that took data from 1989 to 2000. During this period, the way in which data has been "findable" by the experiments, as well as the protocols through which it was "accessed", have changed considerably. These changes continue to take place during the LHC era every 3 – 5 years or so. This is at least partly due to the demanding performance requirements of HEP production and analysis which mean that specialised protocols and / or tools are often strongly favoured. Thus, it is important to consider also these needs – including the very large volumes of data and numbers of objects – before concluding that FAIR is fully understood and implemented. The debate will no doubt continue.

MULTI-DISCIPLINARY COLLABORATION

Multi-disciplinary collaboration has existed from many years through bodies such as the Alliance for Permanent Access, the Preservation and Archiving Special Interest Group, the Digital Preservation Coalition and of course this conference series, to name but a few.
Following on from an initial visit from experts at ESRIN to CERN in February 2018, what we are now proposing is more hands-on / technical work and information exchange, initially involving those EIROForum institutes involved in LTDP but potentially expanding to include other relevant parties. The information exchange that has already taken place in the EIROForum IT Working Group on Business Continuity is an example of the level of information exchange and we are proposing to hold possibly bi-annual technical working meetings (not conferences, not workshops – no suits, no ties) starting from later in 2018. CERN would be happy to host the first such meeting and subsequent events – if deemed successful – could rotate round other EIROForum members and / or interested parties.

ARCHIVING AND PRESERVATION IN THE CLOUD

Can we do it? Not in 4 pages, but both technically and financially this may be attractive. See the 4C project post-cards for some background hints.

INPUT TO THE NEXT UPDATE OF THE EUROPEAN STRATEGY FOR HEP

One of the key inputs is expected to be the successful certification of CERN as a Trustworthy Digital Repository. This, together with the associated Preservation Policy, Strategic Plan and surveillance audits will help ensure that preservation is "written into the fabric" of the organisation and that the necessary resources are provided long into the future.
Other inputs are expected to cover the new or re-enforced requirements regarding Open Data and reproducibility of results. Whilst the latter was already of concern at the time of the Blueprint it had not yet become a requirement from funding bodies. In addition, the "designated communities" were almost always the same as the producers: no Open Data policy, let alone release, had been made at that time. (Some collaborations / experiments were relatively flexible about who could become a collaboration member but this was still far from what today is understood by Open Data).

OUTLOOK AND CONCLUSIONS

We are well on the way to implementing our 2020 Vision, where all HEP archived data is easily findable and fully usable by the designated communities with clear (Open) access policies and possibilities to annotate it further. This will be built using best practices, tools and services that are well run-in, fully documented and sustainable, built in common with other disciplines and based on standards.

However, as has been previously noted, constant effort will still be required, in particular to handle the inevitable migrations, changes in technology and updated policies from funding agencies that will occur not only between now and then but also over the multi-decade period for which the data needs to be preserved.