PROCEEDINGS OF SCIENCE

Data preservation in High Energy Physics: a collaborative perspective

Cristinel Diaconu[a,c] **and Ulrich Schwickerath**[b,c,*]

[a] *CPPM, Aix-Marseille Université and CNRS/IN2P3*
[b*] *CERN,*
Esplanade des particules, Geneva, Switzerland
[c] *on behalf of the DPHEP Collaboration*

E-mail: diaconu@cppm.in2p3.fr, Ulrich.Schwickerath@cern.ch

Data preservation is essential for present and future experimental facilities, enabling cost-effective fundamental research by leveraging unique data sets as theoretical and experimental understanding advances. This contribution summarizes the status of data preservation in high energy physics from a perspective of 15 years of experience with a structured collaborative effort at international level. There is clear evidence that the data preservation enhances the scientific output of experiments during and after data acquisition, improves computational efficiency and strengthens the scientific activity. Lessons from past experiments and recent developments will be presented, as well as prospectives and recommendations.

PoS(EPS-HEP2025)570

*Speaker

ISSN 1824-8039. Published by SISSA Medialab.

https://pos.sissa.it/

1. Introduction

Looking back over the past decades, there are several examples where data of past experiments became interesting again. Reasons can include improved understanding of the theory, new theoretical models which may be tested against ancient data or interest in old data and concepts while planning new experiments. Data from older experiments can also be used for educational purpose. A often cited example is the story of the recovery of the JADE experimental data, which was then combined with LEP results to proof the running of the strong coupling constant. Similarly, there is new interest coming up in LEP data, for example in view of improved theoretical calculations, and also from the FCC community, when it comes to planning of the detectors and using real data to plan the future format of the scientific data to be recorded by the future experiments at FCC-ee [1]. Furthermore, the well-preserved HERA data will contribute to future EIC preparations, for example by enabling tests of new kinematic reconstruction techniques [2].

2. The DPHEP collaboration

Also triggered by the JADE data recovery, which took about 8 years, in 2009 a Letter Of Intent was published [3], establishing the DPHEP study group, with the aim to improve the situation on preserving data in high energy physics. The group published a status report in 2012 [4], and three years later the newly created DPHEP collaboration was born [5]. The DPHEP collaboration organises workshops every two to three years, and reports to the International Committee for Future Accelerators (ICFA). A decade report was published in 2023 [6], and DPHEP recently contributed as well to the ESPP [8]. The workshops are a place for raising awareness and exchanging experiences, where people doing the actual work meet and discuss.

Level	Description	Use case
1	Provide additional documentation	Publication related info search
2	Preserve simplified format	Outreach, simple analyses
3	Preserve analysis level software and data format	Full scientific analysis, based on existing reconstructed data
4	Preserve reconstruction and simulation software and basic level data (raw data)	Retain full potential of the experimental data

Table 1: Preservation levels and their meaning, as defined by the CERN Opendata policy.

3. Preservation levels

In [4], the DPHEP study group defined four different preservation levels, see table 1. The lowest level refers to documentation related things, including the published physics results. This level aims at the public, thus everybody should have access to this. The second level goes beyond this, as it allows to run simple analyses on simplified, preserved data. Typically, this level requires extra knowledge about the format of the preserved data and how to use it, thus complexity is increased, and less people are able to use it. Level 3 and level 4 go deeper. Level 3 allows to run full analyses, and requires access to data, while in level 4 also reconstruction and simulation work flows are preserved, which allows full exploitation of the historical data.

4. Risks and preventing them

Preserving the data has different aspects. These can be categorised as:

- Bit-preservation
- Software preservation
- Documentation
- Analysis preservation

Each of these categories come with its challenges and risks. Bit-preservation simply means keeping the data accessible and in a consistent state. While this is mostly under control, it is important to protect oneself from mishaps which may happen due to physical or system failures, or human errors. Having multiple copies at different locations is a good way to address this.

For software preservation, different solutions exist, starting from keeping a container with archived binaries to keeping the full software stack in source code and porting to and validating against new operating system versions as they appear. The latter approach is most flexible, as it allows to adapt to the always changing infrastructure, however, it is also the most expensive solution. Keeping the software in binary form in containers or virtual machines works as long as external constraints allow this. Risks stem from changing storage access or CPU architecture changes.

Documentation needs to be kept up to date with the current state-of-the-art way of doing things, and as complete as possible, to allow non-collaboration members to use and analyse the data, even after no expert is left who could help. This can be quite a challenge.

Finally, analysis preservation is a challenge on its own, since analyses are frequently done in a non-centralised way, for example by students or smaller working groups. The code may not be written with a later publication in mind, and frequently gets lost when it is kept on private areas, when the people in charge leave.

All four areas require action from both the host laboratory and the collaborations while data collecting is ongoing. This is especially true for analysis preservation. As an example, there are very few original analysis codes preserved from the LEP experiments, which would be very useful to have now where interest in this data is increasing again, triggered by the publication of some of this data.

A list of recommendations of best practices, heavily based on experiences from data preservation over the past 10 years, has recently been compiled and published by ICFA [7]. This document can be used by host labs and experiments as a guidance for long term preservation.

5. Cost and Benefit

The cost for the preserving data which has already been taken has to be interpreted relative to the total cost of the experiment. From experience, this fraction is marginal, and dominated by a small amount of person power to design and maintain the data preservation system [9]. The benefit can be estimated by counting the number of papers published after the data taking has ended, thus using preserved experimental data, and set that in relation to the total scientific output of the experiment. Historical data shows that about 20% of the scientific output is created after the data taking has stopped. Striking examples are the BaBar and HERA (H1 and ZEUS) experiments which stopped in 2008 and 2007, respectively, and are still producing significant scientific output [9].

Dedicated preservation systems are needed, and indeed require a careful planning and non-negligible resources. However, those resources are much less than the full costs of the experiments, and are typically of the order of 0.1%. The low cost of preserving data with respect to the additional scientific output makes data preservation a very cost effective way of doing fundamental research.

6. Open data

Open data and data preservation are related and complementary, however, they are not the same beast, and should be viewed as different things. Making the data publicly available is an excellent way to ensure continued very long term access to preserved data. On the other hand, preserved data which nobody is allowed to access any longer, because the collaboration has stopped to exist and nobody eligible is left, can be considered as dead data and should be deleted.

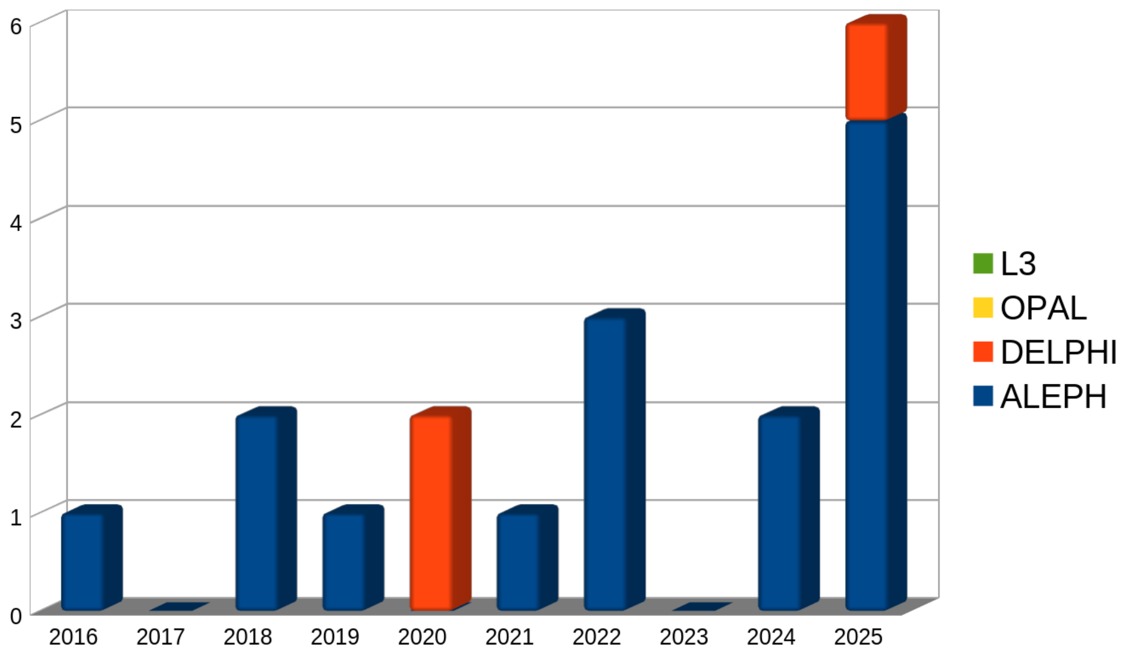

Figure 1: Third-party publications based on LEP experimental data and produced outside the collaboration. Source: `https://arxiv.org`

It is an interesting exercise to look at publications using LEP experimental data in the last years. Figure 2 in [9] shows that the LEP collaborations stopped publishing around 2016. Most LEP experiments used to have fairly restrictive data access policies, with the exception of the ALEPH experiment. Keeping this in mind, fig. 1 shows publications of third party people quoting to have used LEP data. Third party publications are clearly dominated by ALEPH data, with a tendency of an increase during the past two years. The DELPHI experiment has changed their data access policy in 2024, and published all their data on the Open Data Portal at CERN in autumn 2024. There is quite some interest in that data, more than originally expected. A first third party contribution using this data has been published in 2025, and more can be expected to appear in the next years [10]. In addition, the OPAL experiment has changed its access policy in 2025, following the example of ALEPH and DELPHI. OPAL data is expected to become available publicly during 2026.

7. Conclusions

Preserving data from high energy physics is a cost effective way to boost the scientific output of the experiments. From historical data it can be derived that at least 20% of the scientific output of typical experiments is produced after the end of data taking, and at least 10% is unforeseen by then. Keeping archived data alive and accessible also helps to seed future experiments, such as FCC or EIC.

References

[1] F. Gaede, T. Madlener, P. Declara Fernandez, G. Ganis, B. Hegner, C. Helsens, A. Sailer, A. Stewart, 'EDM4hep - a common event data model for HEP experiments', doi:10.22323/1.414.1237, 2022

[2] V. Andreev *et al.* [H1], "Machine Learning-Assisted Measurement of Lepton-Jet Azimuthal Angular Asymmetries in Deep-Inelastic Scattering at HERA," [arXiv:2412.14092 [hep-ex]].

[3] R. Mount *et al.* [DPHEP Study Group], 'Data Preservation in High Energy Physics,' [arXiv:0912.0255 [hep-ex]].

[4] Z. Akopov *et al.* [DPHEP Study Group], 'Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics,' [arXiv:1205.4667 [hep-ex]].

[5] S. Amerio *et al.* [DPHEP Collaboration], 'Status Report of the DPHEP Collaboration: A Global Effort for Sustainable Data Preservation in High Energy Physics', doi:10.5281/zenodo.46158 [arXiv:1512.02019 [hep-ex]].

[6] T. Basaglia *et al.* [DPHEP Collaboration], 'Data preservation in high energy physics,' Eur. Phys. J. C **83** (2023) no.9, 795 [arXiv:2302.03583 [hep-ex]].

[7] C. Campana *et al.*, 'Recommendations for Best Practices for Data Preservation and Open Science in HEP', url=https://arxiv.org/abs/2508.18892

[8] 'ESPP Submission: Data Preservation in High Energy Physics, DPHEP Collaboration' Alexandre Arbey (IP2I, Lyon) et al. e-Print: 2503.23619 [hep-ex]

[9] C. Diaconu, G. Ganis, U. Schwickerath, 'Data Preservation in High Energy Physics: A more than ten years perspective', doi:10.1051/epjconf/202533701033, "https://doi.org/10.1051/epjconf/202533701033"

[10] 4th DPHEP workshop, https://indico.cern.ch/event/1432766/