DPHEP-2023-01
February 2023

Data Preservation in High Energy Physics
DPHEP Global Report 2022

DPHEP Collaboration*

Abstract

This document summarizes the status of data preservation in high energy physics. The paradigms and the methodological advances are discussed from a perspective of more than ten years of experience with a structured effort at international level. The status and the scientific return related to the preservation of data accumulated at large collider experiments are presented, together with an account of ongoing efforts to ensure long-term analysis capabilities for ongoing and future experiments. Transverse projects aimed at generic solutions, most of which are specifically inspired by open science and FAIR principles, are presented as well. A prospective and an action plan are also indicated.

*See Appendix A for the list of collaboration members

Contents

1 Executive Summary

The issue of data preservation emerged with force as an important and community-wide issue in the field of high energy physics at the end of the first decade of this century, as a consequence of the end of several large collider programs, such as HERA, Tevatron, PEP-II etc. Previous preservation initiatives proceeded in an ad-hoc way due to the lack of experience with large and complex data sets (e.g. the JADE data set) and were constrained by a fast transition to new programs (LEP to LHC, for instance). By the end of the 2010's, the data preservation concept had been largely debated and formalised by an international working group, which rapidly was recognised as an ICFA [1] panel.

The working group produced a number of strong recommendations[1]: (i) urgent action had to be taken to organise the long term data preservation at the experiment and site levels, with identified resources, (ii) global action was needed to build an international collaboration and (iii) careful consideration of new, advanced technologies to address the issue of data preservation was needed. Most notably, a significant advance was made in defining the concept of "data preservation", that includes in fact all aspects related to a productive data analysis activity: digital data, metadata, publications, software, databases, documentation etc. A key issue was identified to be the organisation of data analysis activities in the long term, in particular by transforming and adapting the collaborations (new rules for governance, lighter procedures, flexible membership etc.), as well as considering to open the data for reuse by enlarged communities.

The ICFA panel initiated an international collaboration called Data Preservation in High Energy Physics (DPHEP) with the primary goal to foster the international collaboration and mutual support across HEP collaborations in enriching the scientific return of HEP data. DPHEP issues regular reports [2, 3]; the present document presents an overview of the DP activities worldwide.[2]

The status of the participating experiments and laboratories can be briefly summarized as follows:

DESY The two collider experiments adopted different preservation philosophies: H1 (migration and encapsulation) and ZEUS (encapsulation). Their systems and collaborations are both in a great shape and successful transitions to the DP systems are reported. The publications continue at an amazing rate and the publication plan for the next years includes a dozen potential papers. The objective is to keep systems "alive" through 2030. Remarkably, new institutes are joining the collaborations, in synergy with the EIC experiments.

CERN The LHC activity is in full swing and data preservation issues are treated in strong connection with open data approaches. A low rate but clearly identified LEP data and software activity is reported, with refreshed standards and technologies resulting from open data and open science initiatives created and developed on site. CERN is the host laboratory of DPHEP, maintains the DPHEP portal and ensures the operational management, which is essential for the collaboration. A rich panel of transverse projects have been developed at CERN towards an open usage of data and analysis, primarily from LHC experiments, but with a huge potential to incorporate data from other experiments.

MPI has long played a back-up role and and provides expertise on DP, with a multi-experiment framework explored (JADE, HERA, OPAL). Recent activities include "JADE on a desktop", a project dedicated to making JADE data as portable as possible.

[1] International Committee for Future Accelerators, https://icfa.hep.net/

[2] The document covers the contributions to the "Third DPHEP Collaboration Meeting" `https://indico.cern.ch/event/1043155/timetable/`.

KEK The transition and the overlap between Belle I/II experiments is the main feature of DP activities at KEK. BELLE I data is readable in the Belle II framework. The objective is to maintain Belle I data through 2023, at which point the precision will be exceeded by the new data.

BES3 The BES experiment at IHEP is expected to stop data taking by 2022. The objective is to preserve the data for 15 years. Strong support for national and international DP activities has been expressed.

BNL Reports contributions to DP activities in ATLAS. BNL and JLAB started a reflection on DP at the future EIC.

BABAR The long-term data analysis (LTDA) facility supported analyses since 2012. However, SLAC support ended in February 2021. The data is almost entirely copied to GridKa for user analyses. For recovery purpose, the 1.7 PB of data is also copied to CERN as well as hosted at CC-IN2P3. A user infrastructure for ongoing analyses and documentation is hosted by the HEP Research Computing group at the University of Victoria, Canada.

FNAL A transition to a DP system for both CDF (CDFDP) and D0 (R2DP) took place in 2012 but no recent activity was reported. Data is stored/saved at FNAL+Italy, but there is no intention to maintain the analysis facility. One can note the 500th D0 paper in 2021, as well as the recent W mass measurement from CDF, which are also illustrative of the interest for a long term preservation of unique data sets.

Figure 1 illustrates the scientific production at major experimental facilities as a function of year.

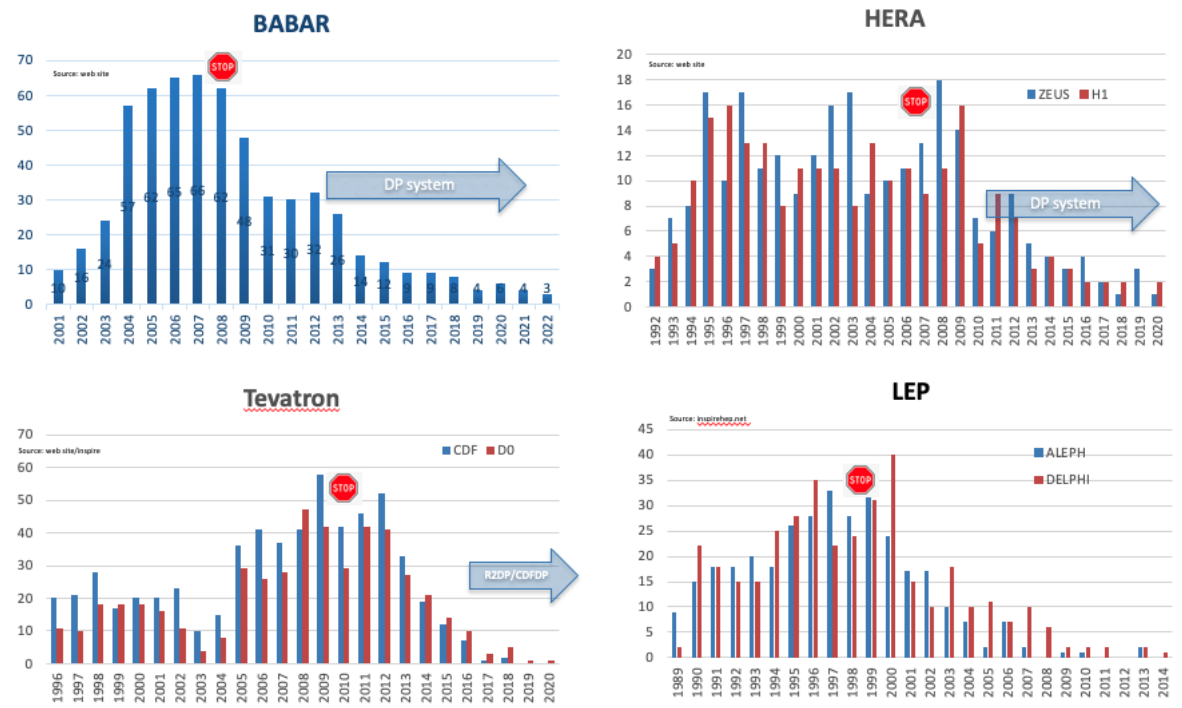

Figure 1: The publication record for four major experimental facilities. The end of the data taking is indicated by the red "stop" symbol. The coverage by a dedicated data preservation system is also shown as an arrow.

Based on the above input from the participating experiments and laboratories, the global status of data preservation in HEP can be summarized as follows:

- The implementation of DPHEP recommendations led to an enhanced scientific outcome in HEP. There are tangible effects of dedicated projects using the DPHEP recommendations (level classification and robust choices, formalisation of long term collaborations needs for data stewardship, new technological solutions such as virtualisation etc.). The concepts of DPHEP have been used in extending the scope of data preservation towards open science; for instance the CERN Open Data policy [4] has adopted the DPHEP classification. Moreover, the updates of the computing models at the LHC [5] incorporate data preservation as a specification, thereby increasing the chances for a smooth and productive transition into "preserved mode" at the end of the program. The community roadmap issued by the HEP Software Foundation covered the DP aspects as well [6].

- The investments/outcome balance is close to the estimates from 2012: about 10% of the overall publications have been obtained using a dedicated data preservation project (as opposed to a tolerated prolongation of a slowly freezing computing system), for which the global costs did not exceed a few per mille of the total experiments costs. This remarkable balance demonstrates the previous prediction [1] that a proactive data preservation action supports research at low cost [3].

- Some data sets are still in danger or have evolved towards an unusable state. For instance, the *BABAR* data set has been decommissioned from the host laboratory and is transferred to other computing centres. It can be noted that the DPHEP community has been active in searching for solutions. It is likely that the situation will repeat with other data sets in the future.

- The survey demonstrates that future/new experiments are likely to manifest an interest for the preserved data sets, for training, testing or even addressing new subjects. The emblematic JADE/LEP example is now being reproduced for HERA and EIC. Indeed, some EIC groups joined HERA collaborations (still active and organised around well-preserved data sets) in order to test new paradigms and train for the future experimental program. It is likely that the preparation for FCC-ee will benefit from the preserved LEP data sets, at least at the level of data model design and analysis frameworks preparation.

- The proof that complicated data analysis frameworks can be preserved with significant gains may stimulate more communities and experiments to join the reflection and implement suitable strategies. An example is offered in this document by MINERvA, a neutrino-nucleon scattering experiment at Fermilab.

- New types of analysis methods (e.g. machine learning) and data access (e.g. zenodo) are being tested on preserved data sets, thereby leading to new results or methods for long term preservation and open access.

- Open science policies implemented for recent or ongoing data sets are essential for the long term robustness of the data preservation. Moreover, the open science paradigms are being implemented for the already preserved (older) data sets. This approach requires nevertheless a significant effort, that could be supported in the context of a demonstrated scientific interest.

[3] It should be noted that those publications cover only the usage by the collaborations themselves. The subsequent usage of those publications also adds to the scientific impact of the preserved data, but it is not taken into account here.

– It should be noted, however, that the preservation systems remain fragile and the better understanding of the main weaknesses and an improving and proactive attitude towards DP did not remove the danger of a catastrophic loss. The key issue is the person power, that functions for most of the examples of preserved data sets in a voluntary mode. The "killer" technological steps are particularly dangerous (for instance disappearance of 32-bit platforms) and need permanent attention.

The overall conclusion is that the existence of an international structure such as DPHEP, oriented to a long term perspective for data preservation, minimally supported and hosted by a large laboratory (CERN), has served successfully as reference to well-defined projects with shorter time scale, that have obtained clear advances.

2 What is data preservation?

It may be useful to recall the scope and the goals of data preservation in HEP [2].

What is data?: the short answer is "everything that was created as a result of planning, running and exploiting an experiment". Indeed, there is a persistent confusion associating "data" to an operating system files, i.e. bits on a memory support such as disks, tapes, etc. This very superficial view is useless for any running experiment and cannot exclusively apply to any useful thinking about long term data preservation. Although the file system is of course essential, it is by far insufficient to perform a data analysis and therefore scientific work. The "data" in HEP – but also in any experimental work using digital/computing systems – is the multiverse of all inputs, outputs and tools used by a group of researchers in order to obtain a novel result:

- digital data files: raw and processed, control/configuration, meta-data, environmental parameters, operational data, databases, etc.,
- software in all its forms (front-end, trigger, middleware, reconstruction, classification include machine learning setups, high-level analysis, visualisation etc.),
- documentation files (internal/public notes, publications, manuals, contracts, photographs, technical drawings), and
- organisation and diffuse knowledge files: rules and procedures, contracts, minutes, meetings and slides, news, blogs, logbooks, address books, outreach material etc.

All these aspects have been considered and synthetically encapsulated in the so-called data preservation levels [2], briefly described in the next chapter. It is worth noting that some of those components may not exist in a digital format or may not be sufficiently structured during the data taking, and therefore need to be recreated in a robust format by the preservation project.

What is preservation?: the process of transforming a "high intensity / rapidly changing" computing system into a "low intensity / slowly evolving" computing system while conserving the capacity of extracting new science from the "data" (within its definition of above). "Preservation" in this context is *not* a freezer, nor a herbarium, a museum, an album, etc. It is, as the concrete examples of this document demonstrate, *a sustained and technologically demanding operation*. In other words, it is a complex project, that has to take into account the data typology, the research goals, the available resources and the collaboration decisions. It may involve choices in data aspects presented above and therefore deliberate data losses/dismissal, as well as significant restructuring and new processing. Furthermore, it cannot be a simple "playground with real data", only usable to reproduce already-published results. Therefore, the system has to preserve at least a part of (and ideally all) the unexplored knowledge of the accumulated experimental data (see the discussion on DP levels below). Under those conditions, the DP system has to tackle important tasks, such as:

- ensure physical existence of data from a digital point of view (see data definition above, all this has to be physically saved and secured at long term – and that includes software, of course) – note that this is the simplest and basically solved aspect of DP in HEP.
- as an obvious (and relatively easy to solve) aspect of the previous item: identify and provide computing and storage resources.

- ensure the functionality of the whole system, identify the potential risks and take appropriate measures as technology and community evolve. The level of complexity differs for the various aspects of the data. The simplest examples include the digital files, the documentation etc. that need only storage and access, i.e. rather standard operations independent of the experiment complexity in general. In contrast, specific experimental software and databases are much more difficult to keep functional across technological changes (hardware, operating systems etc.).
- ensure unambiguous and permanent data validation [7] and results reproducibility, naturally setting the ground for data reuse and open access[8].
- define and identify the human resources related to the research plan.[4]
- oversee and manage the collaborative work and manage the preserved data analysis activity according to the DP design.
- define and implement data access policies, i.e. for which purpose and under which formal regulations the data can be used, including opening the access to data to new collaborators and/or releasing the data to larger (not pre-identified) communities.
- observe and update the physics case of the preserved data. It should be noted that the technical solutions and the necessary choices on the information to be dismissed while designing a long term preservation system should decouple as much as possible from the epoch-related physics case. Indeed, the door should remain open for unexpected analyses.

The goals of a data preservation system as expressed in [2] intrinsically comply with what has come to be known as FAIR principles [9]. Indeed, the data has to be easy to find (F) and accessible (A), and therefore – in a HEP collaborative context – (re)usable (R). The interoperability (I), identified as one of the long term goals ten years ago, is becoming a built-in specification of the recent computing systems as well. Concrete steps have been achieved, with a few examples given in section 5, with a strong incentive originating from the open science policy or within structural projects such as WLCG. However, a clear strategy for a FAIR approach over the entire HEP field (including past, present and future experiments) is still to be defined.

The data preservation process implies a careful inventory of the existing "data archipelago", remapping its components in order to improve the navigation with the smallest possible effort in the long term. Important decisions have to be made, in particular on what should be preserved in line with the scientific objectives, the available resources and the collaboration perspectives[5]. A structured approach for addressing those crucial aspects is described in the next section.

3 Methodologies of HEP data preservation

This section summarizes the generic model of data preservation now used in many experiments to refer to the chosen long term data preservation models.

[4]This aspect is particularly interesting since the bulk of the long term activities are done on a voluntary basis, so they escape the usual needs-resources dialog with the funding agencies. This aspect is addressed also in the perspective of a cost-benefits analysis in section 3.4

[5]The DP process can be summarized as "Consider everything, prepare as much as you can, preserve what is possible".

3.1 Data Preservation models and frameworks

3.1.1 Preservation levels

Four generic DPHEP data preservation levels have been defined, supporting different use cases and expected efforts. The levels are organized in increasing complexity and are reported in Table 1:

Level	Model	Use Case
1	Provide additional information	Publication-related information search
2	Preserve the data in simplified form	Outreach, simple training analysis
3	Preserve the analysis-level software and data format	Full scientific analysis based on existing reconstruction
4	Preserve the reconstruction and simulation software and raw data	Full potential of the experimental data

Table 1: Definition of the preservation models, in order of complexity.

Guided by those generic levels, the experiments can choose the use case they intend to support and adapt the preservation model to the corresponding level. Higher levels include lower levels, so that if level 4 is chosen all the uses cases covered by levels 1-3 are also covered.

The first level involves providing information in addition to the published result with the purpose of improving the understanding of the result and/or the ability to use the result for a further high-level analysis. The additional information may include the exact values used of a published plot, extra tables, data hidden in assumptions use to derive the published, or result meta-data related to the running conditions, for example. A challenge associated with this level is the technology choice for storing the additional information. Global information infrastructures, such as the ones used by the running experiment or others used by the community, such as INSPIRE, may be beneficial for the robust preservation process. Care has to be taken to make sure that the technology chosen is kept alive or can be migrated to keep access to the additional data persistent. Usually solutions allowing export to text files in ASCII format have the highest grade of preservation.

The second level consists of preserving the data in a different format, such that it can be read by non-experiment specific software. Typically only some information is kept, for example the run information (integrated luminosity, collision energy, etc.) and the 4-vectors of the reconstructed event particles, the total energy and so on. The format can be as simple as a CSV table in ASCII, for a maximum portability, or some common binary format, such as ROOT TTrees, which have proven to be readable over at least two decades. The format for level 2 preservation is typically not enough for a full analysis; it is typically used for outreach and educational purposes. Moreover, the basic formats considered to be standard and stable at a given time may still evolve in the longer term. This and the previous level are useful for reinterpretation analyses based on published synthetic data (such as data points, simplified ntuples, likelyhood functions etc.).

The third level consists of preserving the experiment data in the original data format used for analysis as well as the software required to read and process those data. This is typically sufficient to perform a complete analysis if the related reconstruction and detector calibration are adequate for the purpose. The requirements on the experiment software preservation are much stronger than in lower levels. They can be mitigated by using the existing software to create intermediate objects which can be read and, for example, visualized with recent software.

The fourth level consists of preserving the data and the software in a format that the full chain

is available, for example a new reconstruction to include new calibration constants, or improved algorithms. This requires also the possibility to generate new sets of simulated events with new and/or improved versions of the generator codes. This level is the most demanding in terms of software preservation, requiring the preservation of the whole environment, including many dependencies. The benefits are evident, retaining full flexibility for future use.

3.1.2 Frameworks

Preserving access to data means preserving the data themselves (the bits) and the ability to use them. Bit preservation consists of making sure that the data are stored in a safe mode. Certification processes to ensure that the bits are in healthy state exist, such as the ones developed by the space community [10, 11] which have also been considered for HEP [12]. These protocols require that the data can be read by test programs run regularly and that they are kept on healthy hardware. Data centers such as the one at CERN copy the data to new storage (tapes) every 18 months, which allows for sanity checks, for example making sure that all the files are still available, and to keep up-to-date with storage technology.

For a HEP experiment, ensuring the ability to use the data is more than just being able to read them back and write them out to a new medium. For the data to remain useful all the activities typical of an HEP experiment (generation of signals and backgrounds, simulation, reconstruction, analysis, etc.) must remain possible. This means that the experiment software ecosystem must continue to run. Because the experiments all use customized software, keeping these ecosystems alive is an important challenge that HEP collaborations face.

There are essentially two ways to achieve this: freeze the last validated experiment software and create the conditions to continue to run the frozen software; or continuously port the software to the latest stable operating system.

Freezing the software requires keeping the possibility to have access to the original operating system. This is usually achieved via virtual machines or keeping alive nodes still running the original operating system (these can be laptops with a bunch of disks attached – the so-called suitcase model). This solution has some drawbacks; for example, the evolution of security requirements might break the connection of these machines to the rest of the world.[6]

Porting the software means adapting the existing software to a new operating system environment. A crucial part of porting is validation, which is making sure that the software behaves as expected. Porting the experiment software to a new OS is part of the experiment lifetime activities, so one could expect that the process is already established and should not present many difficulties. However, even for the most automated of the experiments, the validation of a new OS environment relies on the feedback of the physics groups, which is something fading out as soon as the experiment goes into preservation mode. A generic validation framework was discussed in [2] and implemented in several examples of DP systems presented in this document.

3.2 Supervision models

One of the key ingredients of the data management plans for long term preservation includes the status of the collaboration that steers the experiment. It is interesting to note that the organisation of experimental collaborations depends not only on the size and complexity of the experimental device, but also on their status in the data taking process.

Various stages of organisation can be defined:

[6]Problems of this kind are not only academic: SLC5-based CMS open-data analysis stop working when the data servers raised the TLS requirements to levels that the SLC5 clients could not cope with [13].

0: Organisation during experiment proposal. The collaboration structures itself such that the plans for a new experiment are pursued. The role of the physics case working group is critical in the first steps, with the experimental aspects being progressively enforced. From R&D, production and global construction, the technical groups evolve during the data taking into running, maintenance and performance groups. The collaborations already produce significant amount of data (simulations, data bases, plans etc.), while the actual data and documentation design process is still ongoing.

1: Organisation during data taking. The organisation during the data taking is focused on the running efficiency (experimental performance, data quality, data preparation and availability) and on extracting the full scientific information according to the publication plans. The data is usually processed several times and the corresponding simulation follows in an even more frequent way. The community is fully focused on the scientific output. As for the previous stage, the data preservation was historically not fully treated, but more recent experiments do take into account those aspects and have developed dedicated task forces and adopted public policies for data preservation and access. This organisational form is perfectly illustrated at present by the LHC experiments.

2: Organisation after data taking. The end of data taking induces the need to adapt both immediately and longer after the data taking, during analysis and collaboration funding times. A strong decrease in the global common and institute-level funds is observed (no more current bills, technical personnel are moved to other projects). A pressure to "move on" is generally manifested from the laboratories and communities. In that context, not less, but more organisation is needed. The publication plans have to be fully consolidated with person power commitments (the reliability of which has to be carefully evaluated, to avoid overoptimistic schedules). The usual competition across the collaboration to obtain highlight results is not a management asset anymore (no multiple groups on the same subject, reduced ability to trigger a high-intensity initiative such as a conference rush or an urgent data processing, etc.). Each subject is basically covered by a group (or a single person) and a list of open subjects is normally emerging as well – the latter naturally provides a clear and concrete argument for data preservation[7] . In this context, the collaborations usually re-organise for a more flexible, though rigorous, frame. This stage can be done in two sub-steps, depending on the available work-force and support from the participating institutes and the host laboratory.

3: Organisation after the collaboration funding scheme. Usually, the collaborations exist officially during the funding periods as agreed by the Memoranda of Understanding. Experience shows, however, that the scientific collaborations may extend well after the end of the official funding (HERA and *BABAR*, for instance). In that case, the collaboration model of governance usually changes, giving a more prominent role to the active collaborators as the institutional bodies (funding agencies, institutional boards, etc.) diminish or stop their involvement. This form of organization strongly relies on the support from the host laboratory, since the data stewardship has to be ensured by the host laboratory computing centre, usually as a continuation and with limited but sufficient technical support. For the distributed computing case (LHC, for instance), the funding agencies and the corresponding computing centers responsible for data hosting need to continue some support for the long term data stewardship in a coordinated way.

4: Rescue organisational scheme. This organisation scheme is to be activated when:

– the host laboratory stops support and announce no long-term commitment.

[7]The list of open subjects (i.e. not covered by person power commitments) tends to decrease slower than expected or even to grow after the end of the data taking for at least two reasons: lost of person power and new subjects emerging.

– the official collaboration/data stewardship is stopped with no further plans (no step 3 is clearly defined).

Examples include LEP (a clear step 3 has not been defined, although elements of technical support and continued publications process are present) and *BABAR* (data support stopped at SLAC). The collaborations have found solutions by the initiative of individuals and proactive groups that continue to access and steward data sets (LEP at CERN, JADE at MPI Munich) or by moving data sets to other facilities and preserving the physics case (*BABAR* at GridKa/Germany, CC-IN2P3/France and HEP-RC Victoria/Canada). No universal recipe can be applied here, but some actions still have to be taken in the form of a preservation project. Indeed, if/when the above listed conditions are encountered, taking no action necessarily implies decommissioning (deleting) the data.[8]

These variants show that during the life of a collaboration, the organisation plays a fundamental role in the data longevity. While a few collaborations have succeeded in defining strong collaboration models that persisted for a long time (3), it is likely that, in the longer term, the "data children" will have to live their lives (4). In that case, the ambition of a full preservation model may need to be revised to a more simplified configuration that preserves an essential part of the information, mostly for historical purposes. In this context, the matter of standardisation and common format is very relevant (see for instance section 4.4.5).

The data preservation issue used not to be discussed at all in the initial steps of the experiment (0 and 1), since it was most commonly considered as an end-of-the-run operation, which is far from optimal for the long term. This has changed in the past decade, with data preservation becoming one of the specifications of the experimental design. An illustrative example, although not referring to brand new experiments, is the update of the LHC experiments computing models for run 2 [5]. The preparation of future experiments at FCC and EIC, as well as the community roadmap [6] confirm this trend as well. Moreover, in parallel and very often as a common and synergistic effort with data preservation, initiatives for opening data for larger communities and outreach have strongly emerged in the past years. Those initiatives also impact the internal organisation and in turn offer a higher reliability for the long term data preservation.

3.3 Preservation and openness

Open data, open software, and open science principles help to facilitate long-term data and knowledge preservation from several different points of view.

Firstly, opening data may lead to simple "lots of copies keep stuff safe" usage scenarios. This increases confidence against data loss and helps to ensure data survival especially for smaller data set size scenarios. However, opening data and keeping them accessible for truly long term has inherent maintenance costs. The open data platforms may have to ensure the data integrity against "bit rot" but also need to be involved with future data format conversions to prevent obsolescence. The truly long-term data preservation and availability may therefore require guarantees by the open data providers to ensure long-term access to preserved data. The open data policies of experimental collaborations may have to take carefully into account the long-term data availability for periods extending beyond the experiment funding lifetime. It is interesting to explore connections with data management policies of hosting laboratories.

Secondly, opening data provides an opportunity to make it more robust. Opening data for larger communities beyond the context of the original experiment involves deepening data curation and

[8]The host laboratories and collaborations should be warned that any rescue operation vaguely imagined for a later stage is nearly impossible: storing and freezing the files and the latest version of the software is certainly not a substitute for a preservation project.

data stewardship processes with the aim of making data more understandable, not only to a new generation of researchers, but also to non-specialists. A good example may be the rising importance of collaborations between particle physicists and the machine learning community. The data must be opened and described in a way that is understandable to other communities. The interoperability can be facilitated by richer data semantics. The robustness of the data-opening process is further assisted by carefully documenting the data provenance and the data validation principles.

Thirdly, opening data facilitates making preserved data more "actionable". The ultimate goal of data preservation is to facilitate future data reuse. The FAIR data stewardship principles advocate making data Findable (i.e. discoverable by both humans and machines); Accessible (i.e. available and obtainable by common protocols); Interoperable (i.e. syntactically and semantically understandable to a wide community of users) and Reusable (i.e. sufficiently described and shared for future reuse). The opening of preserved data make them "actionable" by simply lowering the barrier to providing data validation scripts or data analysis operational examples, the periodical execution of which helps to ensure the FAIR-ness of data as a function of time. Data preservation and openness thus often go hand in hand in facilitating future data reuse beyond the original data acquisition and analysis contexts.

3.4 Funding and valuing data preservation

The specific support for data preservation has different sources:

C1. Host laboratories allocate person power and computing resources.

C2. Collaborating laboratories participate in the effort: replicate or take over data and computing systems and provide technical assistance.

C3. Researchers and engineers participate outside their main research area.

C4. Innovative computing projects, including pluri-disciplinary open science initiatives, may offer attractive opportunities for data preservation and are therefore an indirect source of support.

C5. The proximity of a follow-up experiment clearly helps in structuring and supporting a data preservation project.

Given the success of most of the data preservation projects registered with DPHEP and presented below, the issue of allocated resources is a very interesting one. The most successful experiments have benefited from explicit host laboratory support in the initial phase (C1). This extra support allowed a definition of a specific project, for which the investments can be accounted for as "data preservation costs". According to the previsions from DPHEP initial documents and in agreement with the few projects observed in the past years, the direct investments in dedicated DP projects correspond to $\mathcal{O}(10)$ FTE-years with a very marginal investment in material [9] The C1 item can be compared with the total experimental costs that are, for the kind of collaborations considered here (HERA, *BABAR* etc.) of a few $\mathcal{O}(10^3)$ FTE-years (plus the constructions costs, usually corresponding to multi-hundred millions). With this perspective, one can very approximately estimate that the investment in a DP project corresponds to at most a few per mille from the total cost of the experiment.

Those costs are to be compared to potential benefits:

[9] The costs for maintaining older data sets within hots laboratories computing centers beyond the costs foreseen in the MoU can be considered as negligible, since those data sets became relatively small and could be considered as a common "social security" service.

B1. New publications – counting here those executed with a strong involvement of the dedicated DP systems.

B2. Publications made by other groups/people using the new publications produced at B1.

B3. Preserving the scientific expertise and the leadership in the field of the experiment, possibly boosting the transition to a new experiment

B4. Technology expertise in robust data preservation. Improved ability to plan for new experiments and preserve their scientific potential at long term.

The listed items cannot all be straightforwardly quantified.

Therefore, the cost/benefits balance can only be counted in a very simplistic way, for instance with the ratio C1/B1. Normalizing this ratio to the entire experiment, it comes out from the few exemplary cases cited below that the cost-benefit ratio of the preserved data are very favourable, since investments of the order of a few per mille lead to about 10% gains in the publication record.

Going beyond this quick estimate, which offers nevertheless a very encouraging indication of the positive balance in favour of the preservation of data, a refined analysis of the benefits and costs would be very interesting in order to understand the overall landscape. This further analysis, which is not within the scope of the present document, would have to answer a number of interesting questions:

1. Why the systems did not collapse after the data taking? The “common sense”, expressed by some prominent scientists at the end of the data taking, was to “publish your last paper and leave”. Still, a small but motivated community voluntarily kept data alive for many years and extracted unique science from it, beyond the “local ntuples” philosophy that eventually perpetuates only very specialised analyses.
2. How are the human resources accounted for by the funding agencies or labs? Is doing analysis on preserved data subversive, tolerated or highly valued?
3. How are the publications valued in the “long-term” analysis mode of a collaboration? What is the impact of those publications? Are the authors able to claim visibility and recognition?
4. How is the value of this (new) science displayed? What is the full cost (and who is supporting it) to promote this 10% of additional science?
5. What global resources were used 5 and 10 years past the end of the experiment to keep systems alive and publish?
6. Are the DP requirements compatible with the running experiments conditions? How much extra investments are needed to make ”fresh” data suitable for a long term preservation and how those investments can be optimised further when considering open data and open science aspects?
7. How are future projects supporting, stimulating and shaping data preservation projects and how are the cost and benefits of this transfer of knowledge accounted for?

All these questions and many more, when considered in the perspective of the scientific outcome that continues to come from some experiments more than 15 years after data taking stops, indicate a “data preservation miracle”, where science continues to be extracted at low cost from data sets for which the access and the complexity are not obstacles for a small but highly motivated community.

4 Experiment reports

This section contains reports from a number of high-energy collider experiments connected to DPHEP.

4.1 LHC

The LHC experiments have adopted since several years data preservation policies both for the ongoing runs and also as specifications for the future upgrades. A clear synergy, pointed out from the very first DPHEP reports, between preservation and open science is successfully practiced by all LHC experiments, as described in this section. Furthermore, transverse projects focused on generic solutions for data re-use and re-analysis are focused on the LHC, as will be described in the next chapter.

4.1.1 ATLAS

The ATLAS approach to data preservation is informed by a larger policy framework designed to ensure the long-term impact of the collaboration's physics program during and beyond its lifetime. Thus, while the raw data generated by the experiment are preserved, ATLAS also invests in analysis, metadata, and software preservation, resulting in a spectrum of data and software products tailored towards a number of audiences. All public data products released by ATLAS aim to adhere to FAIR principles by leveraging community web infrastructure to aid in locating relevant data (findable) and using standard access protocols (accessible). When possible, community-developed common data formats are chosen (interoperable). Finally, data and software products are explicitly designed to be input for new research rather than an archival preservation of prior investigations (reusable).

Data generated by the ATLAS experiment are categorized according to the levels defined in a previous DPHEP report [2], to which varying preservation and access policies are attached. Level 1 data represents data released publicly alongside a published result by the collaboration. First and foremost this includes the publications themselves, which are made available as Open Access under the SCOAP3 initiative. In addition to the paper itself, additional data are released publicly on community cyberinfrastructure such as HEPData [14] to facilitate reuse by physicists. This includes digitized tables and figures from the publication, such as yield tables and theory parameter bounds derived from the data. Additionally, data are released pertaining to the analysis design that allow the approximate reimplementation of the data analysis procedure, such as tabulated selection efficiencies, key multivariate observable definitions such as neural networks and decision trees, or cutflow data. More than half of ATLAS analyses currently include HEPData records. ATLAS also routinely releases collaboration-validated software implementations approximating the actual event selection as Rivet routines [15] and SimpleAnalysis classes [16]. The current Rivet analysis coverage of ATLAS is 35% [17] (the highest of any tracked experiment). For SimpleAnalysis, ATLAS provides the SimpleAnalysis framework which is released as Open Source software [18], and more than 35 search analyses have been included already. As a key data product for re-use, ATLAS also releases the statistical models underlying the search or measurement (see e.g. Refs. [19, 20]). This captures all systematic effects and allows external researchers to reproduce the statistical procedure at full fidelity or re-use ATLAS results in follow-up studies such as global statistical inference. The focus of public Level 1 products is the use of long-term, stable human- and machine-readable formats. The Level 1 records are assigned stable identifiers (DOIs) and are individually citable by third-party researchers.

Augmenting the public Level 1 data, ATLAS preserves detailed internal publication-specific information through two complementary approaches. A comprehensive metadata record in its internal GLANCE database captures the full analysis life cycle from inception to final jour-

nal publication and collects diverse metadata such as analysis team membership, presentation, approval reviews, code repositories, physics keywords, etc. In addition, a growing number of analyses are preserved as a software product for future reuse and reproducibility. Here, analyses are described using a declarative computational workflow language, using parameterized task templates and Linux Container Images. The latter are able to capture not only the analysis software but its full set of dependencies (OS, language runtimes, compilers, etc). A particular use-case is the reinterpretation of analyses in terms of alternative signal models using RECAST [21, 22]. Here a tight integration with the REANA platform developed by CERN IT [23] was developed.

ATLAS does not host L1 data directly but relies on the continued development and maintenance of the archival cyberinfrastructure, most crucially HEPData, to provide this service to the community.

At Level 2, ATLAS compiles special-purpose data sets targeted for outreach and education beyond the HEP community and releases them publicly on the CERN Open Data Portal (see e.g. Refs. [24, 25]). The data sets are deliberately simplified in both content and format to facilitate broad usage in non-academic settings, minimizing technical boundaries. As a consequence these data are not suitable for research usage. Alongside the data sets, a range of software examples such as interactive notebooks, data analysis macros, or visualization code are prepared and maintained (see e.g. Ref. [26]).

Reconstructed event-wise data sets beyond the scope of a single analysis are categorized as DPHEP Level 3. These represent the main data used by ATLAS members wishing to prepare a physics result and are available internally in a range of data formats. The data are produced from raw data sets using the ATLAS reconstruction software Athena, which is made available as Open Source and whose development is publicly accessible [27]. In addition, Athena is citable through a corresponding Zenodo record [28]. A given release of the software is packaged and preserved independently from the source repository using industry standard formats for global distribution. In some cases, container images of the software are available [29]. Both observed collision events and simulated events are processed through tightly controlled versions and configurations of the software, and for any given data set, detailed metadata and provenance information is stored in the ATLAS Metadata Interface (AMI). Based on these metadata, reproduction of derived data sets is straightforward if necessary. To ensure a coherent data set of all data produced during a data-taking period, the reconstruction software only undergoes performance improvements, while major changes and algorithmic improvements are incorporated during large-scale reprocessing campaigns. Such reprocessing of old data with new software is a key step towards internal data preservation and allows ATLAS to maintain consistent multi-run data sets that may be processed by a single software release. This was recently undertaken to unify the format of the Run 2 data set with the upcoming Run 3 data set.

For public access to Level-3 data, ATLAS has implemented the guidelines of the broader CERN Open Data Policy [30]. In this scheme, ATLAS will follow a staged collaboration-defined data release schedule that begins to publicly release Level-3 collision and simulation data within five years after the close of a data-taking period and ensures a full release of the complete data set after the close of the experiment. The time delay allows for an initial exploitation of the data by the collaboration and ensures that the released data have benefited from all algorithmic improvements and calibration studies of the collaboration. The data are planned to be released in a calibrated format designed also to be used internally for the bulk of ATLAS analyses. Alongside the data, ATLAS releases software to perform analysis at the same level of fidelity, including the assessment of systematic uncertainties. The full ATLAS data analysis releases are Open Source [27] and additionally are preserved regularly using Linux Containers [29].

In addition to the bulk release of collision and simulation data in a fully calibrated format, ATLAS has also produced several special purpose data sets for R&D purposes. Two examples are the data sets prepared for the TrackML project [31], a public data challenge for the development of machine-learning based tracking algorithms, and the Higgs ML challenge [32]. Similarly, a data set was recently released for the training of machine-learning-based calorimeter simulation [33, 34]. Such data sets may consist of fewer or only a subset of events but include more low-level information than the bulk release.

Level 3 data are released through the common CERN Open Data Portal with CERN as the host laboratory assuming custodial responsibility over the released data. Similarly, the corresponding software and containers are hosted on CERN infrastructure. It is anticipated that any public metadata associated with each analysis and data set would likewise be hosted on common infrastructure, for example in the CERN Analysis Preservation Portal (CAP) [35]. Like HEPData, these infrastructure components are crucial to the success of the Open Science program of ATLAS but not directly maintained by the collaboration.

The DPHEP Level 4 category describes the raw data from the detector before reconstruction. These data are not immediately usable for physics research and are rarely accessed by collaboration researchers. They are therefore not released publicly during the lifetime of the experiment. These data are, however, preserved and serve as input for future reprocessing by the experiment, and they will become public after the close of the experiment for archival and historical purposes.

In preparation for the LHC Run 3, the ATLAS software underwent a major evolution, from multi-process to multi-threaded code. This allowed significantly improved memory usage in the ATLAS software [36]. To ensure efficient use of the ATLAS Run 2 data, it was decided that prior to the Run 3 data taking, the Run 2 data set and corresponding Monte Carlo samples should be reprocessed. This reprocessing campaign was completed in 2022; however, it required a major effort, in particular for the data and the changing conditions throughout the data taking period. Periodically reprocessing the data in this manner has proven key to avoiding a loss of interoperability in internal data sets or additional unnecessary analysis complexity.

For the ATLAS collaboration, the preservation of its data and the search for perennial means for its efficient scientific use is of great importance. The intrinsic complexity of the data, metadata and data structure and the software to analyze it pose a major challenge to finding practical data and analysis preservation solutions. The collaboration believes that the current level of effort invested in data and analysis preservation and the planned effort to support the release of open Level 3 data are appropriate to match the current scientific needs. Extending the scope of the ATLAS data preservation efforts would require additional human and computing resources from external sources.

4.1.2 CMS

The CMS experiment is addressing the data preservation and reusability through regular open data releases, putting into action the experiment's Data preservation, reuse and open access policy [37]. The policy defines the collaboration's approach for maintaining data collected by the experiment usable in long term. It was first approved in 2012, as the first of such policies in the HEP domain, and it has been updated in 2018 and 2020. CMS has mandated a Data Preservation and Open Access Group responsible for managing the implementation of the policy.

The CMS data policy is in agreement with the CERN open data policy, and like all LHC experiments, CMS publishes results in Open Access journals. Additional data, at so-called Level

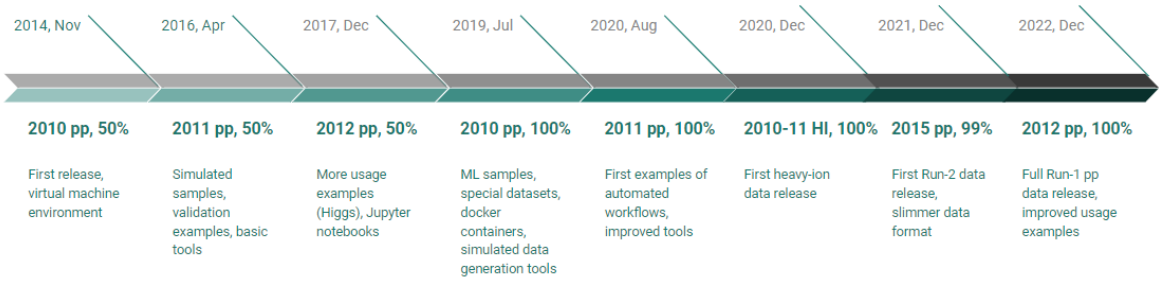

Figure 2: CMS data release timeline.

1, to facilitate immediate re-use and the combination of these results are provided through, and archived in the long-term, by trusted third parties such as HEPData. At Level 2, simplified data formats for several levels of immediate re-use such as limited analyses, education, and outreach are available, accessible, and preserved through the CERN Open Data portal [38].

The Level 3 consists of data that can be used to reproduce published analyses and perform new analyses, and CMS was the first HEP experiment to release data of research quality and has done pioneering work in the domain since 2014. The CMS data releases take place regularly through the CERN Open Data portal. As defined in the CMS policy, CMS will normally make 50% of its data available 6 years after they have been taken. The proportion will rise to 100% within 10 years, or when the main analysis work on these data in CMS has ended. However, the amount of open data will be limited to 20% of data with a similar centre-of-mass energy and collision type while such data are still planned to be taken.

The Level 4 consists of raw data, as stored directly after data-taking without further processing. Custodial copies of these data are stored at CERN and at the corresponding custodial computing Tier 1 for each data set. CMS can release small samples of raw data potentially useful for studies in the machine learning domain and beyond together with level 3 formats. If storage space will be available, raw data can be made public after the end of all data taking and analysis

The CMS open data releases contain full reprocessing of collision data from each data-taking period and the simulated data corresponding to these data. They are made available in the format and with the same data quality requirements that analyses of the CMS collaboration start from. The public data are accompanied by a compatible version of the CMSSW software and additional information necessary to perform a research-level physics analysis. Example code and some specific guide pages are provided to explain and instruct the use of this associated information.

The first CMS data release took place in 2014, followed by regular releases since then, as shown in the timeline in Fig 2. At the time of writing of this document, all Run-1 proton-proton data from 2010-2012 and first heavy-ion data are in the public domain, and the first batch of Run-2 data from 2015 has been released. Run-1 data (2010-2012) are in Analysis Object Data (AOD) format and some early special data in RECO format from which AOD is a subset. Starting from Run-2, a slimmer MiniAOD format is in use, and a reduced NanoAOD format averaging to about 1-2 kB per event will also be made available.

The released data are accompanied with rich metadata, describing their quantity and quality, and importantly, the full provenance information. The provenance information records the exact parameters and software conditions used at the time of data-taking, for collision data, and in the event generation and simulation step for Monte Carlo data. The input parameters are recorded also for the subsequent processing steps. This information together with the open-source CMSSW software makes it possible to trace back all data handling, although doing so is

certainly a tedious task.

A virtualised computing environment, compatible with the software version with which the original data can be analysed, is provided and maintained. The CMS provides docker software containers [39, 40], regularly tested and updated, and Virtual Machine (VM) images, based on the CernVM software appliance [41]. They are part of the data release, as well as any additional information necessary to perform a research-level physics analysis. The additional data products are needed in different steps of the analysis, for example for data selection, as correction factors to be applied to physics objects, or for evaluating the final measurable results in terms of cross sections.

CMS data released through the CERN open data portal satisfy FAIR principles for Findable, Accessible, Interoperable, and Re-usable data and metadata to a large extent. But due to the complexity of experimental particle physics data, the FAIR principles alone do not guarantee the re-usability of these data, and additional effort is needed to pass on the knowledge needed to use and interpret them correctly.

This knowledge includes, first of all, learning the computing environment and software for the first step of data selection. Due to the experiment-specific data format, the first step will almost inevitably be done using the CMS software in a computing environment compatible with the open data. Open data users can download a software container image and run it on their computer, independently of its operating system. Recent developments for Windows Subsystem Linux (WSL2) have also made this feasible in Windows, in addition to Linux and macOS. The CMS open data group has invested a good amount of work in setting up these containers so that the first user experience with the CMS open data remains smooth.

After having set up the computing environment, open data users will need to learn the intricacies of experimental particle physics data. How to select the data of interest, how to identify the particles properly, how to understand the efficiencies and uncertainties in the analysis process, how to estimate the backgrounds, and how to address many other challenges with experimental data. To address these questions, the CMS open data group is putting together a comprehensive CMS open data guide [42]. In addition, regular workshops are conducted for physicists interested in research use of these unique public data. These workshops aim to cover the skills needed to get started with the experimental particle physics data from the CMS experiment and get participants initiated with hands-on exercises. Sessions practising running CMS data analysis jobs in a cloud computing environment are also included.

To ease the task of getting properly started with open data analysis, CMS is setting up all open data examples as workflows that can be run an automated way. This also makes it possible to regularly test and verify that all data and computing assets are working as expected. Similarly, work is ongoing to set up test workflows to monitor that reprocessing raw data and regenerating new Monte Carlo data remains possible in the legacy environment.

To encourage wider utilisation of the knowledge embedded in physics analysis work within the collaboration, analysis procedures, workflows, and code will be preserved internally in version-controlled code repositories and software image registries (e.g. CERN GitLab) which are currently being set up. They will be made available and searchable to collaborators through CERN analysis preservation services [35]. The open data releases can include selected analysis workflows, at an example level. The reproducibility of the preserved workflows can be tested in systems like REANA [23].

CMS open data are widely in use. The number of scientific publications by open data users external to the CMS collaboration is comparable to that of an internal working group in the

collaboration, thus broadening the scientific value of these data. The usage can be monitored through citations to the data records that all have a digital object identifier. However, the counting mechanism for repositories such as Inspirehep is still imprecise. CMS open data also offer the possibility of benchmarking HEP tools in a realistic context, an opportunity that many development teams have benefited from. In addition, they have been used at university-level education, for example, particle physics courses or as a new type of physics laboratory exercise. Simplified data sets derived from CMS open data have been used to enrich the curriculum at the high-school level and in many particle physics outreach initiatives.

4.1.3 LHCb

The LHCb collaboration is committed to the community efforts to make LHC data available to the public. Its collaboration board has ratified the CERN Open Data policy [4] in 2020. Data releases are made available through the CERN Open Data portal [38] and trusted repositories, such as HEPData [14] or RIVET [43], and comprise data on three levels of complexity. Publications are open access and routinely accompanied with supporting materials and releases of machine readable results to HEPData, where appropriate. All plots and tables of LHCb papers are accessible in machine readable form through the LHCb published results pages[10].

For educational purposes and outreach selected data sets with various degrees of information depth are published on the open data portal. These datasets are usually intended for a specific application, such as the International Masterclasses, and contain heavily filtered information necessary for the respective purpose.

In December 2021 LHCb has approved the first release of reconstructed data from Run I of the LHC. For LHCb the corresponding level or complexity has been defined as the output of the stripping or, where applicable, the turbo stream, which is the same level of abstraction available to the researchers inside the LHCb collaboration. The stripped data contains offline selections of several thousand selection algorithms called stripping lines. The stripping lines are grouped into streams, corresponding to similar physics signatures. In the current scheme, releases are done on a stream by stream basis. The Electroweak, Radiative and Leptonic streams have been approved for release so far, corresponding to about 200 TB of data. The release to the open data portal and the accompanying documentation is currently being finalized.

The software to analyse these data is available as open source [44] and documented through the LHCb starterkit. However, only limited support can be provided by LHCb to potential external users.

The development of the open data curation tools, documentation and metadata schemes is organised within the LHCb Data Processing and Analysis (DPA) project. All related efforts invested by collaboration members count as service work.

To improve the sustainability of these efforts two issues with the currently planned open data release have to be addressed. First, the data volumes scheduled for release by the LHCb collaboration are quite significant and would amount to higher storage requests on the open data portal than those of ATLAS and CMS combined. Second, despite the best documentation efforts, the access to the data provided by current tools, while appropriately flexible for experts, lacks an intuitive interface, that would facilitate the exploration of the data by external users.

In order to address both issues a new tool, called the NTuple-Wizard, is expected to enter beta-testing within the LHCb collaboration in May 2022. The basic idea of this tool is to allow external users to submit queries to the LHCb data, which will be processed by the LHCb

[10] `http://cern.ch/lhcbproject/Publications/LHCbProjectPublic/Summary_all.html`

analysis productions system accessing the original replicas of the data on the grid. For security reasons the users cannot directly submit analysis jobs to the production system but instead are guided through configuring an analysis in a fully interactive web application. This application also provides search features to locate specific processes of interest in the data, which can be visualised graphically as so called decay trees. The application allows to configure a large fraction of the analysis tools available in LHCb to extract quantities of interest (e.g. kinematic variables or particle identification classifier outputs) from the data. The wizard finally generates a job configuration, which is passed to the LHCb analysis production team without further user interference, where it can be validated and finally run on LHCb compute resources. The result is a plain ROOT ntuple, which is made available to the user.

When the beta phase is successful the NTuple-Wizard is planned to be incorporated into the CERN open data portal. Aside from the technical issues involved in this integration, there are a few open questions related to the management of the generated data, which need to be worked out. The Wizard and the analysis production system provide provenance tracking, but schemes to store this information as meta data accompanying the ntuples still need to be decided.

The concept aims at allowing to make the existing replicas accessible, thus reducing the need for dedicated open data copies of the data. Moreover, the web application provides guidance and documentation on the meaning of the extracted information. In the future standardised (semantic) meta data, describing the content of the generated ntuples could be generated as well. Finally, the wizard allows fine grained control over which data is made available, beyond the current by-stream release scheme.

LHCb requires a minimal analysis preservation policy for all published analyses. The analysis software is archived on CERN gitlab in the physics working group workspaces. Input data, which is typically filtered to the ntuple-level is archived on EOS or grid storage. Upstream samples and processing steps, such as the offline stripping selections are managed centrally with a dedicated bookkeeping system based on DIRAC.

With the start of Run 3 LHCb has adopted the real time analysis (RTA) strategy based on a complete online event reconstruction. The default data processing path will not include an offline reconstruction step anymore and the output of the high level trigger (HLLT) will be roughly equivalent to the level of abstraction provided by the stripping offline selection in Runs 1 and 2. The concept has already been successfully piloted by the turbo stream data produced since 2015. From the analysis preservation point of view the move to RTA means that large parts of the data preparation are by construction done centrally managed with all necessary configurations becoming part of the LHCb software stack.

In order to capture the user analysis parts, which typically deal with interpretation of the data, rather than data preparation, current efforts focus on the Snakemake workflow engine, which has become quite popular within LHCb and is now also supported by REANA. A future vision would be to enable the majority of LHCb analyses to preserve at least the workflow producing the central value of the analysis for deployment to REANA.

For the capture of meta-information generated during the internal discussions and review a new database system, called Analysis Lifecycle Management (ALCM) is under construction in LHCb, which will eventually replace the current public results pages and offer the possibility to export public extracts of the information to systems like CAP.

4.1.4 ALICE

Recently, ALICE together with the other CERN's Large Hadron Collider (LHC) collaborations has adopted a new policy [4] supporting a consistent approach towards the openness and preser-

vation of experimental data. Essentially, all LHC experiments share a common strategy for the release of Published Results (Level 1), Outreach and Education Data (Level 2), and Reconstructed Data (Level 3), while Raw Data (Level 4) are not considered usable in a meaningful way outside the collaborations.

In compliance with the CERN Open Access Policy, all ALICE publications are available with Open Access and data points together with additional information including the analysis code are made public at the time of publication using the HEPData portal [14] and Rivet toolkit [43]. In addition, dedicated subsets of ALICE data for the purposes of education and outreach have been released in the TTree ROOT data format and are accessible through the CERN Open Data Portal (CODP) [38] together with the needed software to analyze them via virtual machine, container technology, and web-applications. The early implementation of the ALICE open access policy for Reconstructed Data led to public availability of 5% of 2010 Pb–Pb and 7% of 2010 proton-proton (pp) collisions data sets in Event Summary Data (ESD) format on the CODP. The software availability has been achieved by using virtualization technology based on microCernVM bootable machine image with CernVM-FS client and the related documentation has been published on CODP.

To best align the implementation of open data with the new software developments, the ALICE collaboration plans to adopt a new data format to make the ALICE Reconstructed Data public. A large conversion campaign of data from Runs 1 and 2 has been performed and the data sets with the new format is available from April 2022. Such a new format, based on the ALICE's O2 Project, has been developed for Run 3 and it is more compact and more performant with respect to the previous formats. In addition, having all old data sets converted into the new format will ensure the long-term usability of the old data sets with the new Run 3 analysis framework.

Documentation and software availability constitute the other key elements of the ALICE data preservation strategy allowing the future collaborators, the wider scientific community, and the public to analyze data for educational purposes and for eventual reassessment of the published results. ALICE has already put in place procedures including the use of Docker containers and the software deployment on CVMFS for the software continuous integration and the release validations. Such procedures are also the bases of the Data and Analysis Preservation for the members of the ALICE collaboration, while the wider scientific community and the public cannot benefit from these procedures having no access to CVMFS. For a wider range of users, lightweight images containing specific versions of the new Run 3 ALICE analysis framework and other basic software will be also created and placed on CODP to simplify access.

Further developments to achieve the preservation of the different stages of the analyses have been based on CERN Analysis Preservation (CAP) [35]. Tests of simple serial workflows have been performed creating proper JSON files from analysis trains and uploading them to CAP. The option to rerun ALICE analyses in REANA [23] using inputs from CAP and running code within Docker containers has been tested but it requires more detailed studies with the use of Common Workflow Languages (CWL) to describe more complex workflows.

4.2 HERA

Computing needs and data sizes of the HERA experiments are relatively small, compared to other experiments on the DESY IT analysis and storage system host. On best effort, DESY IT will continue hosting the data needed for analysis and offer computing resources for analysis. It has shown that the HERA experiments integrate well with the other HEP experiments in the common NAF cluster. Thus, they only put a small additional support and resource load, and at the same time benefit from the continuously evolving computing infrastructure. The DESY NAF supports container technologies, so that also freezing of experiment code is possible.

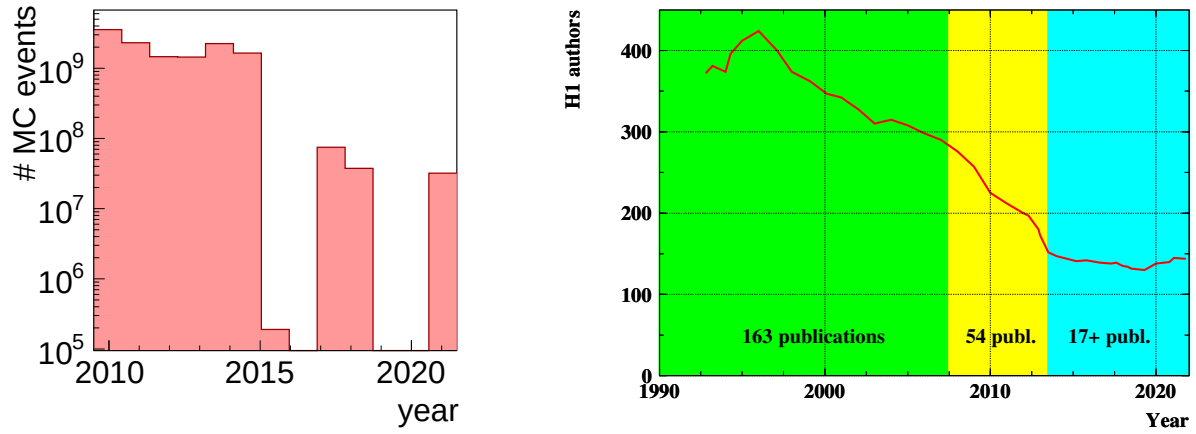

Figure 3: Left: Number of Monte Carlo events produced centrally by the H1 Collaboration. The years without MC production are related to a change of the computing environment, or no MC requests. Right: Number of H1 authors is increasing since 2019 due to retained analysis capabilities and new interest in *ep* physics. The colored areas indicate the data taking period (green), the period with active funding (yellow) and the period under the new collaboration agreement in *data preservation mode* (cyan). The number of corresponding publications is also indicated.

DESY IT benefits from the HERA DPHEP experience in other projects leveraging the use of FAIR data and sustainable analysis of data.

4.2.1 H1

The H1 experiment [45, 46] recorded a unique data set of lepton–proton collisions at HERA in the years 1992 to 2007. The complete RAW collision data comprises around 75 TB, the set of compressed DST data amounts to about 20 TB and the analysis level files are about 4 TB. A sizeable and universally employed software stack for the processing and analysis of the H1 data also exists, which was initially developed in the years 1988-2012. This is comprised of a series of core software packages in `Fortran` and the object–oriented analysis core framework, `H1oo`, which is written in `C++` and based on `ROOT`. Following a DHPEP level 4 preservation policy, the H1 Collaboration continues to maintain these data, all related software, including simulation and reconstruction code, as well as all relevant documentation on the data, MCs, software, detector, operation, meetings and collaboration life. A continuously updated webserver[11] provides access to all resources for the collaboration members and for external visitors.

The H1 Collaboration works under a renewed collaboration agreement with individuals as the only entity, and it is coordinated by a spokesperson, two deputies and a scientific secretary, and DESY acts as host laboratory. Since 2012, an additional twenty H1 papers (9% of the total) have been published, and presently several data analyses are ongoing or started recently. In 2021, a new Monte Carlo production campaign with about $4 \cdot 10^7$ events was performed to support one of the ongoing analyses and the full capability of the present software environment was proven. Access to the data, software and internal documentation is granted through a DESY computing account, which also provides access to computing resources for data analysis activities. Data, software or internal documentation are not planned to be made public because of missing or unresolved copyrights of large parts of the software and documentation. However, the H1 Collaboration is open to new members and saw even an increase in signing authors for the recent publications compared to those from a few years ago.

[11] https://www-h1.desy.de

Any documentation is provided through a dedicated webserver at DESY, where more than 12 000 digital documents and notes, as well as about 4000 presentations of internal meetings, and the original internal webpages are maintained. Relevant documentation in other locations, like institutional web spaces, private repositories, or workgroup servers, were carefully migrated to the common storage. The offline documentation is stored in 150m of shelve space in the DESY library archive.

Since 2012, several software migrations were performed including the migration of the entire software and environment to 64–bit operating systems, up to Scientific Linux 6 and CentOS7. Previous stable software releases are kept in a central repository and can be used for bit-level validations or for production within containerized workflows. A comprehensive update of the software and analysis frameworks was performed recently and is summarized in Ref. [47], and is briefly summarized in the following.

Whilst the full data analysis capability and flexibility was retained, after one decade, the overall status of the H1 software was shown to have a few shortcomings from the present point of view. The programming languages and standards such as `C++98` were found to be unattractive for young people to learn and liable to slow down new developments and data analysis efforts. In addition, outdated dependencies such as `ROOT5` complicated the usage of modern data analysis techniques and tools. Furthermore, the risk of incompatibility due to different compiler standards or different interfaces such as MC event generators only increases with time. As no externally maintained package repository was used, new packages had to be provided manually and the compilation and maintenance still required a profound and detailed knowledge about the specific structure of the H1 software, as well as some insight into the historic development. The renewed structure of the H1 software stack and environment [47] allows for an easy transition to even newer platforms, like AlmaLinux 8 or CentOS8 or 9, and are kept up-to-date with DESY's central Linux platforms.

Considering the above and further arguments, the status of the H1 software and the data analysis capabilities was revisited in 2020. All core software packages have since been successfully migrated to a modern computing platform, based on `amd64` (x86-64) `CentOS7`, using the `GNU` 9.2 compiler. Remaining external dependencies were updated to the latest releases and are now obtained from the `LCG/AA` package repository [48, 49], greatly reducing the number of H1 specific solutions. The common object–oriented data analysis framework `H1oo` is now based on `ROOT6` and supports the `C++17` standard. An online code documentation for `H1oo` is available, whilst interactive analysis in `C++` through `CLING` is now also possible, as well as for the first time in `Python` (v3). Several benchmark analyses codes were migrated, like those of jet production [50, 51] or inclusive DIS cross section measurements [52], and serve as a valuable and validated starting point for new data analyses. The programs and libraries are now provided to the members of the H1 Collaboration on shared global file systems for convenience. All H1 software packages are now maintained in `git` and new build instructions for a complete rebuild of the entire software stack have been prepared. Container solutions have also been implemented for backward compatibility and software tests. New Monte Carlo event production campaigns with the full detector simulations and run-dependent conditions can be, and are, performed at the DESY computing cluster(s) and using the modernized H1 software stack. See Fig. 3 for a historical summary of the H1 MC production campaigns.

Many H1 data analysis activities are still ongoing to this day, and new analysis projects are beginning due to the uniqueness of this scientific data set. There is an increasing interest of the HEP community in *ep* scattering, in particular by physicists from EIC, and this interest is also reflected in the recent addition of new collaboration members. This is directly seen from the number of H1 authors, which, after a minimum number of 130 in 2019, is now steadily increasing

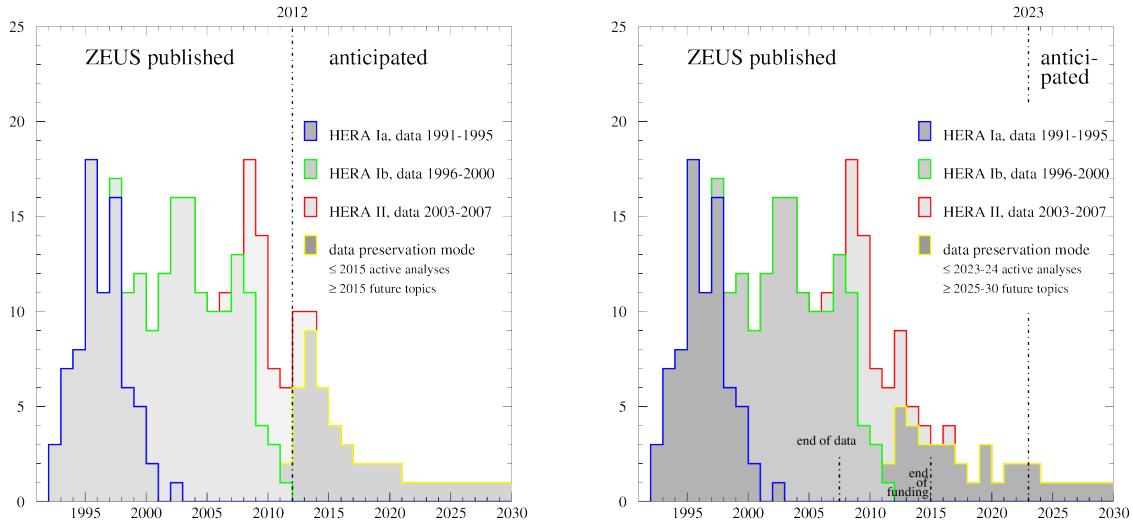

Figure 4: Number of ZEUS papers published and anticipated to be published per year. Original 2012 version (left), compared to current 2023 version (right).

again, see Fig. 3.

By carrying out modifications to the software architecture, H1 is confident in providing high quality data analysis capability of the unique HERA data in the future, using modern analysis tools, like deep-neural networks for reconstruction or unfolding [53, 54], and recent programming languages on state of the art platforms.

4.2.2 ZEUS

The ZEUS data and knowledge preservation project [55] was internally started in 2006 and was generalized through founding contributions to the eventual 2012 DPHEP study group document [2] and the subsequent DPHEP collaboration agreement [3].

Some of the physics goals to be pursued further were formulated at the workshop on Future Physics with HERA data in 2014 [56, 57] and many of these have steadily been implemented since. They were and are being complemented by results on new ideas which were not yet in the focus of attention at that time, often implemented by new groups, e.g. from the EIC, Heavy Ion and theory communities, freshly joining the collaboration.

An extensive documentation on the ZEUS detector, analysis techniques, available data and MC generated samples is available from www-zeus.desy.de and www-zeus.mpp.mpg.de.

Like almost all HEP papers, the ZEUS physics papers are stored on arXiv [58] and accessible on INSPIRE [59]. Fig. 4 shows the evolution of the expectation on the number of publications with time as projected at the time of the study group document [2] compared to the latest update of this projection including the publications actually achieved. The two agree quite reasonably, enhancing confidence into the further projections for the next decade. Integrating up to that time, about 10% of the total ZEUS physics output (already at >5% now) is expected to have resulted from the dedicated data preservation effort, after the end of funding, without which 80–90% of these results would never have existed. This is a great return for the <1% additional investment originally made (or rather diverted from the base budget) for data preservation. The ZEUS data preservation strategy lies half way between the ‘level 3’ and ‘level 4’ preservation strategies as defined by DPHEP [2]. All the data are being preserved at ‘level 3’, i.e. in the

form of reconstructed and calibrated low level basic objects (calorimeters deposits, tracks, lepton candidates, ...), as well as higher level composite objects (e.g. jets, missing ET, particular meson decays, ...) which can also be rebuilt from the available basic information. These data comprise so-called 'common ntuples' created from the 360 M real events recorded by the ZEUS detector between 1996 and 2007.

They are stored in a unified flat ROOT ntuple format which can be read and processed with almost any past or future version of the ROOT software package, or any other package providing a ROOT data format interface such as e.g. root-numpy, pyroot, RDataFrame or uproot. Consequently so far essentially no maintenance updates were needed since the original setup in 2006, there is no need for virtual machines or containerization, and in this respect the ZEUS analysis software is always as modern as ROOT or any future backwards compatible successor (note that ROOT itself is backwards compatible to earlier HBOOK/PAW formats dating all the way back to the late 1970s through a simple 'h2root' converter).

All simulated data sets available up to the time of the end of funding in 2014 are also stored in the same format. The relations between the MC generated events and the description of the simulated processes is documented in dedicated internal ZEUS web pages and the lists of files belonging to each generated MC sample are also stored in a standalone sqlite3 database. The generation of small additional simulated data sets including detector simulation (level 4, through an encapsulated and/or containerized approach) is possible [60] and has been used successfully. Dedicated Code needed for the 'zevis' event display and for automated handling of the 'cninfo' data base (both very useful but not crucial) are available in github and are kept updated by MPP.

All basic real and simulated data, with a total size about 250 Tb, are stored and made available in two different geographical locations (DESY/Hamburg and MPP/Garching) using dcache technology [61], and can be used through respective generic computing facilities, as detailed above for the DESY case. Optionally, they can also be accessed through the ZEUS Grid instance.

Currently, the data are accessible only to ZEUS members or individual ZEUS associates, but plans exist to release at least part of the data as Open Data through the recently funded German national PUNCH4NFDI project [62].

In summary, the ZEUS data preservation project, in line with the HERA data preservation project as whole and the general DPHEP strategy, was and is very successful, and large parts of it have been implemented as forseen with very limited dedicated resources; (more resources would have allowed and would still allow the success to be even bigger). More than 30% of the total HERA physics results were produced in the 14 years since the end of data taking, and more results are coming. This was made possible by the essential support of the host lab during the final phase of funding and the continued IT support, and is currently sustained to a large extent by person power originating from external sources.

4.3 *BABAR*

The *BABAR* detector operated at the PEP-II asymmetric-energy electron-anti-electron storage ring at the SLAC National Accelerator Laboratory, and collected physics data from October 1999 until April 2008. The data were collected mostly at the Y(4S) resonance at 10.56 GeV/c^2, the B-anti-B meson production threshold. CP violation and B physics were the main research program but, at that energy, the cross sections for tau-anti-tau lepton and c-anti-c quark production are of comparable magnitude, making of *BABAR* a flavor factory, able to access also lepton physics and charm physics. Continuum (u,s, and d quark prodution) and ISR physics are also accessible. The *BABAR* data set also includes a scan of the Y(nS) resonances, in particular *BABAR* has the

largest Y(3S) data set ever collected and, to date, there are no plans for taking data at the Y(3S) peak at Belle-II (or any other current experiment).

BABAR data are in ROOT format, and the software is written in C++ Object Oriented. While the software is based on 32-bit, it can run in 32-bit mode also on 64-bit architectures when the 32bit compatibility layer is available. The 32-bit binaries in the latest release work on SL6 64-bit and could potentially be ported to a CentOS7 or CentOS8 derivative, but there is no plan currently to do so due to the limited availability of expert man power. Currently, we operate the software and perform physics analyses using SL6 virtual machines. With a fully working software release and the raw data, *BABAR* could aim to a "Level 4" preservation model, the main problem at this stage being man power and limited resources for the needed infrastructure.

In February 2021, SLAC support for *BABAR* stopped and the Collaboration worked on porting the infrastructure on other, more sustainable platforms. In particular:

1. Data
   - Processed data, real and simulated (1.2PB), hosted at GridKa (Germany), remotely accessed via XRootD for physics analysis
   - CC-IN2P3 (France) hosts a full copy of the data (2PB including the raw data, for recovery purposes only, MoU in place until 2025 with renewal option)
   - CERN (Swiss) hosts a full copy of the data as well (transfers still on-going but nearing completion)
2. Collaboration tools are hosted on a variety of platforms now
   - InspireHEP private collections for archiving internal documentation of published results
   - CERN indico for meetings
   - CERN e-groups for information exchange and to control access to internal meetings on indico
   - Caltech hosts general mailing lists to reach all *BABAR* members and associates
   - Google tools for membership and active analysis tracking
   - The HEP Research Computing (HEP-RC) group at the University of Victoria, Canada, hosts web based documentation (historical discussion forums, HTML pages, and wiki)
3. Analysis resources: The University of Victoria HEP-RC group hosts servers for providing user accounts, storage areas, access to the data, and resources for analyses. Servers have been provided by the HEP-RC group and the *BABAR* group at the University of Texas at Dallas. In addition, a few machines previously used by *BABAR* at SLAC have been moved from SLAC to Victoria to be integrated into the new analysis system.

While the Collaboration retains the ownership of the data, we also support a *BABAR* Associates Open-Access Program which allows any interested individual or group to easily join *BABAR* and receive the full support of the Collaboration and access data, software, and all internal documents. This approach has already been successful in a number of cases.

The *BABAR* Collaboration continues to exploit the rich data set and continues to produce a wealth of world-class results. To date, the *BABAR* Collaboration has published 598 papers and 2 more are currently accepted for publication. About 20 new analyses have started in the last 4 years. Figure 5 presents the predicted and current status of *BABAR* publications.

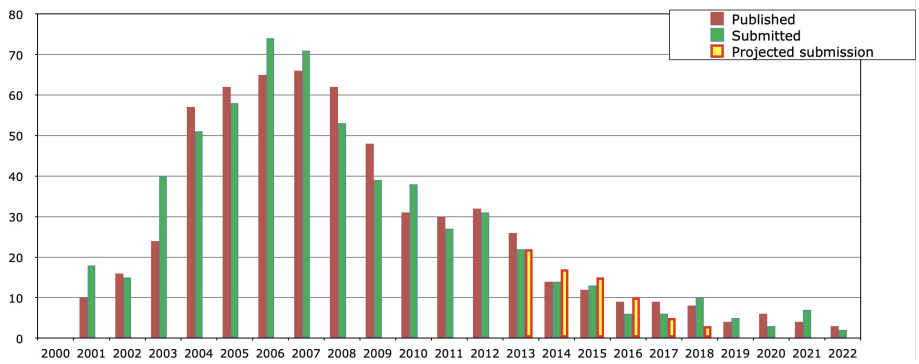

Figure 5: *BABAR* submitted (green) and published (red) papers per year. In 2012 predictions for submissions (yellow) were made for the years 2013 to 2018. In 2012 it was predicted that no analysis would run after 2018.

4.4 LEP

All four LEP experiments initially stored data in CASTOR at CERN with two copies on tape. As part of archive service modernisation, the tape-resident data is about to be migrated to CTA, the successor of CASTOR. Already in 2015, the frequently used data types have been copied to the EOS systems, making it available to users without the need to stage any tapes. Access to data on tape and disk is regulated by the respective experiment policies and is generally restricted.

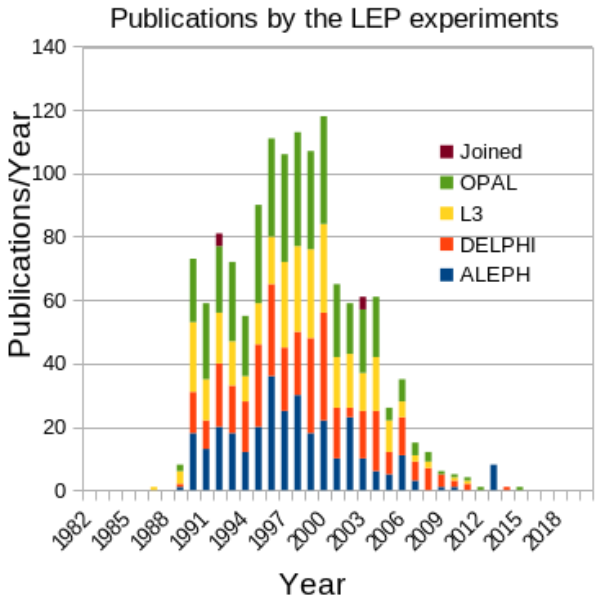

Figure 6: LEP publications per experiment. Note that those publications only reflect the work done by the collaborations themselves. Further usage of those publications, also enhancing the impact of the preserved data, is not accounted for by this figure.

4.4.1 ALEPH

The ALEPH libraries to access the collected data and to produce simulated data were compiled with g77 for the 32-bit x86 processors and are still functional on platforms supporting/emulating this architecture. The most recent validated operating system is SLC6, which is available under CVMFS and can be used, for instance, with Singularity. The ALEPH software is also available

in CVMFS and GITLAB at CERN, and rely heavily on the 32-bit/g77 version of CERNLIB. To move to a more recent architecture will require the re-compilation of the entire software followed by an in-depth validation campain. Part of the measured data and simulated data produced by the ALEPH experiment were made available in a simplified format, which is accessible in EOS with much reduced dependency on the original software. This approach promises to be a longer-term solution to access the existing datasets translated in a more recent HEP data format. In case of need of new simulated data with upgraded or new models the full software stack implementing the entire ALEPH workflow is still necessary.

4.4.2 DELPHI

The DELPHI experiment has moved its software stack to modern technologies, e.g. GITLAB at CERN for the sources and CVMFS for the binaries, while archiving older binaries on EOS. During 2022 the full stack was ported to 64-bit as support for 32-bit libraries is vanishing. This was possible thanks to the efforts to revive CERNLIB, see chapter 5.7. The experiment software heavily relies on CERNLIB thus a validated and complete 64-bit version of CERNLIB is a pre-requisit for the future of the DELPHI software stack. The revised DELPHI software stack supports data analysis frameworks, simulation, reconstruction and event visualisation. Builds are available for CentOS7, CentOS-Stream 8 and 9 and Alma 8 and 9 in both 32 bit and 64 bit, and Ubuntu 18, 20 and 22 (64 bit only). A special challenge was the removal of a dependency on a commercial software package in the event display. When logging into CERN interactive services with a DELPHI registered account, the login scripts will automatically select the correct version of the stack and set all environment variables as needed. It is also possible to initialise the stack manually by simply sourcing a script on CVMFS.

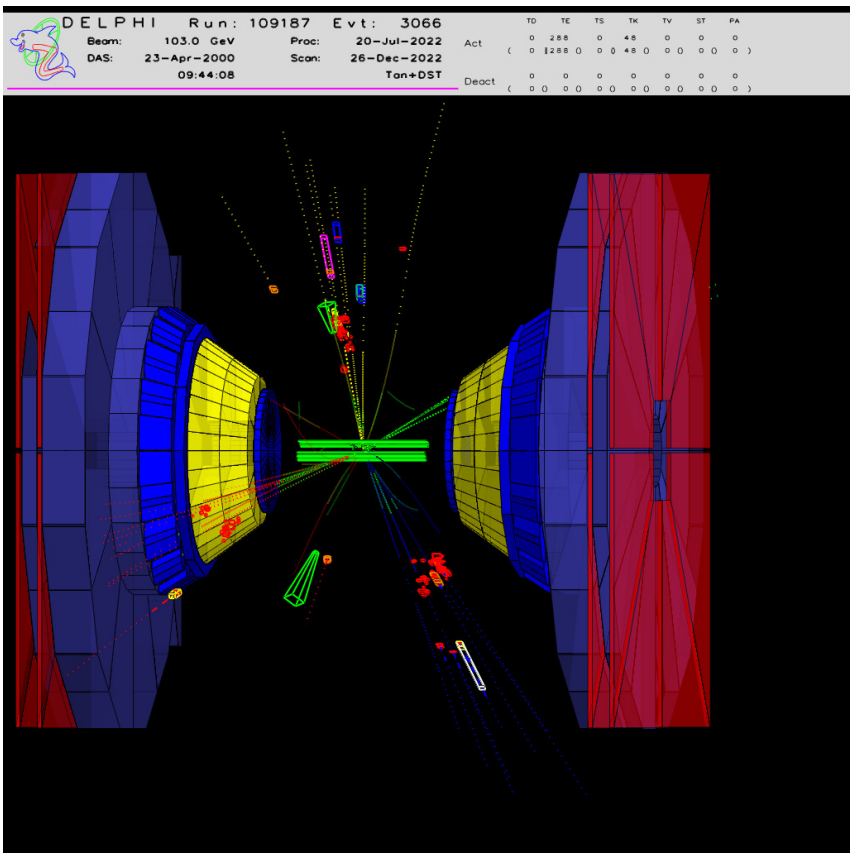

Figure 7: Example of a DELPHI event, reconstructed from raw data using the revised software stack.

Figure 4.4.2 shows an example of a DELPHI event which has been extracted and reconstructed from raw data using the 64bit stack.

Validation of the new stack is still ongoing and initial results look promising, however, more work is required.

In parallel, DELPHI provides containers and instructions to run old binaries inside these containers. There are also older SLC6 binaries archived on CVMFS which can be used from a CERNVM image which is shared with OPAL.

On the longer term, the OpenData initiative is a possible option, albeit only for educational purposes. Work on extracting data samples and converting them into an appropriate format has started in 2022 and is currently ongoing.

DELPHI documentation, including scanned and OCR processed technical and internal notes dating back to 1982, are available in CDS.

On analysis preservation, some analysis codes were preserved along with their output ntuples. These can be useful as additional checks to validate the revised software stack.

4.4.3 L3

L3 legacy data are preserved and stored in EOS at CERN (/eos/experiment/l3/) in several formats. Data in the original L3 compressed format (DVNs) were also converted into ROOT files and are stored in EOS as well. These ROOT files can be easily processed using existing ROOT libraries and utilities. Detailed documentation on the content of these files exists, although is not publicly available yet. The latest L3 policies allowed public use of these data under supervision by L3 members before publication, in order to ensure proper interpretation of the information.

4.4.4 OPAL

OPAL has preserved its data and complete software stack, with the exception of the event display, and kept it working up to CentOS8. CERNLIB and in particular ZEBRA are needed for access to the data. A 32-bit SLC6 based VM has been defined, but creating new instances does require the option to boot and install SLC6 based nodes, showing the weakness of relying on VMs for data archiving. Keeping the whole software alive and porting it to newer OS versions appears to be a safer approach, but also requires porting the required external libraries (CERNLIB, ZEBRA, etc.). For OS versions newer than SLC6 the availability of the revived CERNLIB, see chapter 5.7, is essential. The data has been migrated to EOS, and the software environment is being migrated from AFS to CVMFS. HBOOK/PAW based Ntuples exist for some specific types of analysis. A proper validation suite is not available.

The documentation has been stored in CDS.

4.4.5 LEP data and Key4hep

As seen in the previous sections, the approach for accessing LEP data is very different from one experiment to the other, and is typically limited to the remaining specialists. In some cases, reduced tuples have been extracted and provided to the external requesters in simple formats; but the process is not automated and it is difficult to imagine a generic solution for FCC-ee based on such an approach.

To address these difficulties, it has been suggested that a possible solution could be connected to the recently started key4hep project [63]. Key4hep aims to create a common low maintenance, customizable ecosystem of software components, interoperating through a common event data model, EDM4hep, providing the language for transient and persistent storage.

Key4hep is the framework used for FCC-ee data processing at all steps, from generation to analysis. Converting the LEP data, at least the ones used for analysis, in EDM4hep would have automatically enable the FCC community to access the LEP data with the tools they already know. The migration would have a significant impact also for long term data preservation of LEP data, helping to achieve the FAIR data principles by improving on: Accessibility, detaching from library and OS obsolency; Interoperability, promoting a single standard framework; Re-usability, requiring less specific expertise.

A preliminary investigation has shown that the migration could be feasible, at least at level 3, i.e. to "perform complete analyses when the existing detector reconstruction and simulated data sets are adequate for the pursued goal". The approach, which could be fully automated, would be to go through XML-like intermediate files, follow by conversion to EDM4hep.

The migration process would require resources and investment, but the return could be huge, not forgetting that the current high interest from EW/Higgs factories studies may provide a unique possibility.

4.5 JADE

The JADE experiment was one of the experiments located at the PETRA e^+e^- storage ring at DESY in Hamburg, Germany. The JADE detector comprised accurate tracking, fast multi-hit electronics, measurement and identification of photons, electrons and muons. A schematic view of the JADE detector is given in Fig. 8 and a more detailed description can be found in Ref. [64] The experiment took data between 1979 and 1986 in the center-of-mass range

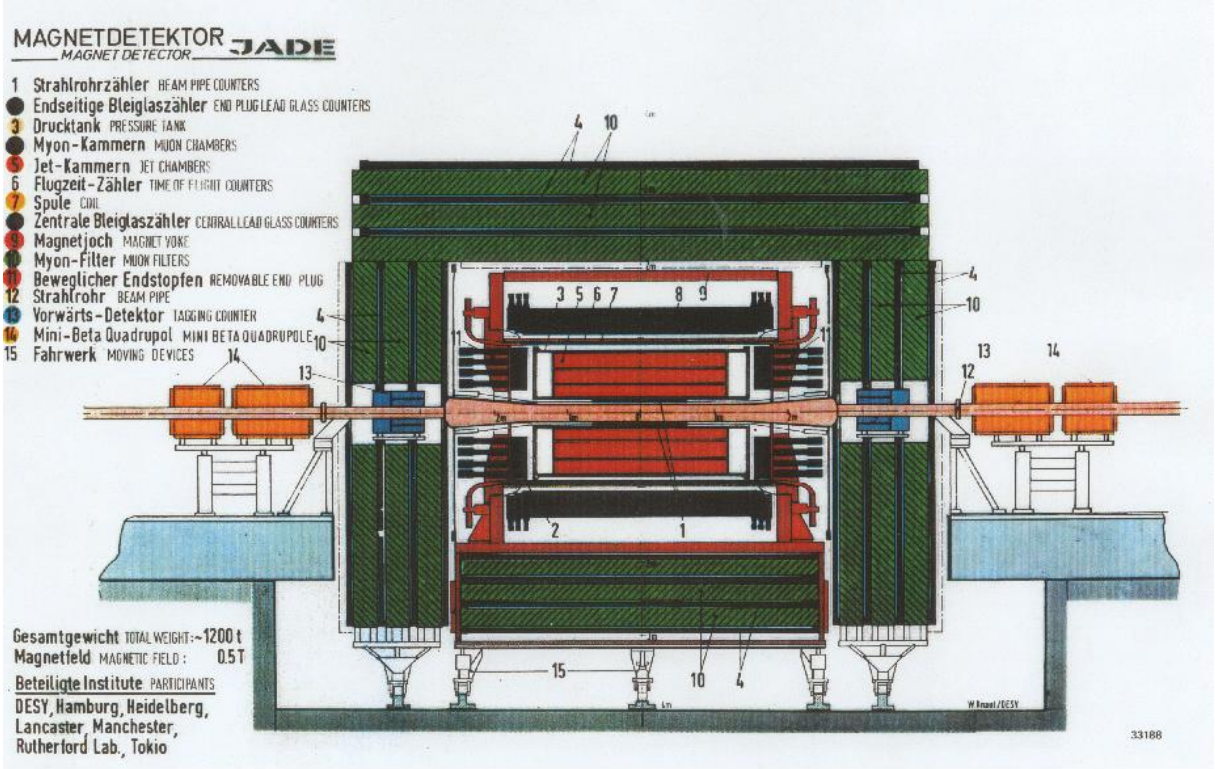

Figure 8: Longtitudual cross-section of JADE detector. The diameter of the Jet Chamber is about 1 m.

between 12 and 46.6 GeV. The results from the JADE experiment were published within a regular collaboration structure between 1979 and 1991. Important scientific results of the JADE collaboration are the (co-)discovery of the gluon, the establishment of jet-physics and tests of Quantum-Chromodynamics. Other highlights are studies of the hadronisation process via the

string effect, electro-weak precision tests, two-photon physics and searches for then not confirmed particles of the Standard Model, i.e. the top quark and Higgs boson, and searches for New Physics like Super-Symmetry, free quarks etc.

Preserved JADE data

The data includes the collision events recorded by the JADE detector at energies between 12 and 46.6 GeV and the Monte-Carlo (MC) simulated events. Most of the preserved MC simulated event samples were generated in the 2000s, during the initial resurrection of the JADE software. These samples were produced with then contemporary MC event generators Herwig6 [65], Pythia6 [66], Jetset [67], with parameter settings as used by the OPAL experiment. Now in the 2020s these samples can be superseded with modern and more precise MC simulations. Both the data and MC simulated events are also preserved in a processed form of ROOT [68] or HBOOK/PAW [69] n-tuples, which are suitable for many QCD-related analyses and are compatible with similar n-tuples of the OPAL experiment, see Sec. 4.4.4.

The preserved data includes the calibration information and the luminosity tables. The integrated luminosities at the main energy points range from 1.46/pb at 14 GeV to 150/pb at $33.8-36$ GeV. This corresponds to O(1000) hadronic events at the low energy points and about 35.000 hadronic events at $33.8-36$ GeV. As of 2020s, the JADE data with the total size just below 1 Tb can be considered as 'tiny'. With such a small data size, it makes little sense even to discuss the costs/resources that should be dedicated to the physical storage of data. However, for the convenience of access to the available data, the data are stored (and password/certificate protected) in local discs in MPP, MPCDF[12] archive filesystem, in the MPCDF dCache storage system and in the MPCDF ownCloud cloud storage. The diversity of storage instances provides an excellent opportunity for the access and analysis of the data from the Grid (dCache [70] instance), or from an local desktops (ownCloud). Technically the data consists of about 35.000 files with a total size of 806 Gb.

Preserved JADE software

The preserved JADE software consists of the original codes designed to process the JADE data, some MC event generators from the 1980-2000, the detector simulation routines for the JADE experiment, some calibration codes, the event display, and the analysis routines which were used to create the ROOT/PAW n-tuples mentioned above. In addition to that some interface codes to modern MC event records were added.

The original JADE software has evolved on many different platforms and after the first resurrection [71] in the middle of the 2000s consisted of approximately 50.000 lines of Fortran code running on IBM AIX4.3 systems with the IBM Fortran runtime. The environment the JADE software required for compilation and execution required also GNU or AIX binutils, C runtime and CERNLIB, see Sec.5.7. The JADE event display required a specific graphics package HIGZ [72].

The main goal of the next update of JADE software was not only to port it from the AIX system to a modern Linux environment, but to assure its portability and eliminate the need of any complex and/or exotic environment requirements. Therefore, JADE build system based on make was replaced with the CMake build system. The codes were updated to be compatible with modern GNU Fortran and several other compilers. Not all of the used Fortran compilers are free, and not all Fortran runtimes have the support of the essential features for the JADE software (e.g. mixed endianess I/O), therefore only the GNU Fortran and Intel Fortran toolchains and runtimes are practically useable.

[12] Max-Planck Computing and Data Facility, Gießenbachstraße 2, 85748 Garching, Germany

Thanks to the portability of the CMake build system it became possible to compile the JADE codebase not only on Linux but also for the first time on MacOSX. The dependence of the original codes on the CERNLIB and HIGZ was undesirable, so to avoid these dependencies the required CERNLIB and HIGZ functions were emulated with a help of the ROOT analysis framework, with an optional dependence on CERNLIB. With the performed code updates it became possible to create new MC generated event samples for JADE using modern MC event generators. And to certain degree re-reconstruct the original JADE data. Some functionality of the JADE event display was also restored. The computing model of the JADE data re-analysis would include the processing of the data (and/or MC simulated events) into plain ROOT n-tuples and subsequent steps performed using the ROOT framework.

The JADE codes were publicly accessible from the MPP JADE site for a long time,[13] therefore after the updates, the codes were put in a public git repository account in GitHub.[14] The usage of GitHub has allowed for regular automated builds of the JADE software on modern Linux and MacOSX platforms. All the dependencies needed for the software are available from open source projects.

JADE documentation and cultural heritage

The physics papers of JADE are available on InSpire. Many of those are scanned copies of the hard copies from CERN or KEK libraries. The JADE web site and the GitHub site mentioned above include as well: the full list of JADE physics papers, technical notes, scanned logbooks, data and software preservation documentation. In addition, the technical notes are available as hard copies at MPI.

In addition to the equipment, also the data and the software of a physics experiment can be considered as part of the history of physics and of physicists. A part of the archive of photographs from the times of the active collaboration was made public on the JADE web site at MPP.

Most recent JADE analyses

Although many e^+e^- experiments were conducted after the end of data taking period of JADE, the energy range covered by JADE remains unique and could only be accessed in future experiments, e.g. in just a few days of data taking at FCC-ee. Therefore, the JADE data was a source of important QCD studies in the last decades. As an example, in the most recent JADE analyses [73, 74] a relatively competitive extraction of the QCD strong coupling constant $\alpha_S(M_Z)$ was performed using these data.

JADE preservation policy

As of 2022 MPP offers support for any possible re-analysis efforts of the JADE data under supervision of the JADE members. We recommend that the results are published according to the collaboration guidelines discussed in 2009. The JADE group at MPP is eager to join the most modern developments of the Data Preservation in the context of the JADE Data preservation and therefore contributing to the CERN OpenData initiative is under active discussion.

In March 2022 the JADE collaboration approved making the JADE data public and the decision was backed by the DESY Directorate. Therefore anyone can re-analyze the data. Any results obtained with JADE data and/or software should be supplemented with a note " We thank the JADE collaboration and DESY for making the data and corresponding software publicly available. The data analysis presented here has not been reviewed by these entities and is the sole responsibility of the authors. "'

[13] `https://www.mpp.mpg.de/en/research/data-preservation/jade`
[14] `https://github.com/andriish/JADE`

4.6 CDF/D0

The CDF and D0 experiments were multipurpose detectors at the Fermilab Tevatron proton–antiproton collider. They completed data collection in 2011 and jointly participated in the Run II Data Preservation Project (R2DP), described in detail in Ref. [75]. Each experiment has a total preserved data set of approximately 10 PB, and Fermilab has migrated both of them to LTO8 media. Both experiments produced over 500 publications, including two for CDF and three for D0 in the past two years.

The goal of R2DP was DPHEP Level 4 preservation through at least the end of life of Scientific Linux 6 (2020). Containerized environments in conjunction with CVMFS allow for a longer timeline, as long as appropriate SL6 containers are available. CDF code built for SL6 has also been validated to run on SL7. Fermilab officially declared CDF and D0 "completed" experiments in 2021, though the CVMFS repositories and data sets remain available for additional analysis. Fermilab's Core Computing Division has also begun producing public webpages for "historical" experiments, of which CDF will be an early example.

4.7 The PHENIX Experiment at RHIC

The main challenge in the area of Data and Analysis Preservation (DAP) in PHENIX is that the preservation effort started in earnest fairly recently (2019) and and in the final stages of the lifecycle of the experiment. PHENIX took its last data in 2016 and has been conducting active analyses since, producing 13 journal articles in 2019-2020 and a substantial number of conference contributions. This work is being done against the background of a gradually diminishing number of active contributors and a large and complex software environment maintained by the facility, as well as a sophisticated analysis apparatus. Over the past two decades the legacy web services became obsolete in terms of both technology and content, and hard to maintain. Knowledge management has become a substantial challenge.

In order to address these issues, the PHENIX Collaboration has undertaken an effort to put in place Data and Analysis Preservation procedures and practices including

- Use of Docker containers to preserve specialized and/or legacy computing environments and enhance software portability. Because of the large volume of the software stack, parts of it were refactored into packages deployed on CVMFS. In addition to fully fledged images of the complete stack (currently kept in a private registry at BNL), lightweight images containing specific versions of ROOT and other basic software were also created and placed on Docker Hub to simplify access.
- Leveraging REANA as a mechanism to capture final stages of select analyses for preservation, validation and user training. Testing of simple serial workflows has been done, and work is underway to implement more complete analyses using more complex workflows which require the use of CWL (Common Workflow Language).
- Active supervision and management of the materials created and submitted by the Collaboration to the CERN HEPData portal (Level 1 in standard DAP nomenclature), with a broad team Involvement.
- Joining the CERN OpenData portal and using that platform to host self-contained packages which include PHENIX special purpose limited datasets and basic examples of analysis software (Level 2 in standard DAP nomenclature).
- A vigorous team effort to migrate PHENIX research materials from legacy information systems approaching end-of-life to a robust and well maintained digital repository, opting

to use the Zenodo instance at CERN. There are currently approx. 500 items migrated in this manner.

– Development and deployment of a new Collaboration website for easy access to curated materials including those obtained from legacy resources, optimized for long-term stability and ease of maintenance and based on the static website generator technology.

Common across all these work areas is the strategy of using community-developed and supported tools, frameworks and services while keeping in-house development to the absolute minimum. In that regard, collaboration with DPHEP provides the most value and is central to the achievements of DAP in PHENIX. In doing this work, PHENIX enjoys the support and contributions from the SDCC facility at BNL.

4.8 BES III/ IHEP

BESIII [76] at the BEPCII accelerator is a major upgrade of BESII at the BEPC for the studies of hadron physics and τ-charm physics with the highest accuracy achieved until now. The peak luminosity of the double-ring e^+e^- collider, BEPCII, is 10^{33} $cm^{-2}s^{-1}$ at center-of-mass energy 3.78 GeV.

BESIII is an unique experiment currently operating in tau-charm energy zone in the world. It started to collect data in May 2009, and its end of data acquisition will be extended to 2030.

The data preservation of BESIII will follow LEVEL 4 of DPHEP. These data is expected to be preserved for another 5-10 years after the end of data acquisition. IHEP computer center have preserved about 4 PB raw data, 1 PB data of other types on tapes which are managed by IBM 3584 library. Migration from CASTOR to EOSCTA has been initiated at September, 2021. Since there is a large gap between CASTOR 1.x and EOSCTA, real data copy is unavoidable in the process of migration. After the migration, the bit preservation technology stack will be aligned with CERN-IT and good practices of CERN can be reused in the future.Tape upgrade from LTO4 to LTO7 will be finished in this migration at the same time. BESIII Offline Software System (BOSS) is the offline data processing software system of BESIII. It is developed on top of Gaudi. It releases stable versions every year while the latest version is 7.0.8. Currently, OS version of physical computing node has been upgraded to CentOS 7. Earlier BOSS versions and their depenencies are kept in CVMFS and earlier OS version (Scientific Linux 5 and 6) are provided by Singularity containers. Environments of releases $\leq 6.5.5$ can only be recovered in virtual machines.

There are several HEP experiments at IHEP which have overlap with BESIII in terms of physicist, software developers and IT supports. Their DPHEP policies will generally follow the experience of BESIII. In the near future, their will be 2-3 neutron/photon sources running simultaneously at IHEP and its south branch. DPHEP policies of these new scientific facilities require more attentions. In 2019, the project of "National high energy physics data center" was proved by Chinese Ministry of Science and Technology. This project will provide support of manpower and funding for the DPHEP at IHEP. With support of this project, IHEP-CC can make further explorations on software technologies of open data, outreach and reusable data analysis.

4.9 Belle I / II

The Belle II experiment [77] at the SuperKEKB accelerator in Tsukuba, Japan has started the data taking with the full Belle II detector in 2019. Since then, we have accumulated roughly 270 fb^{-1} datasets by the end of 2021 [78] and started providing the new physics results [79, 80].

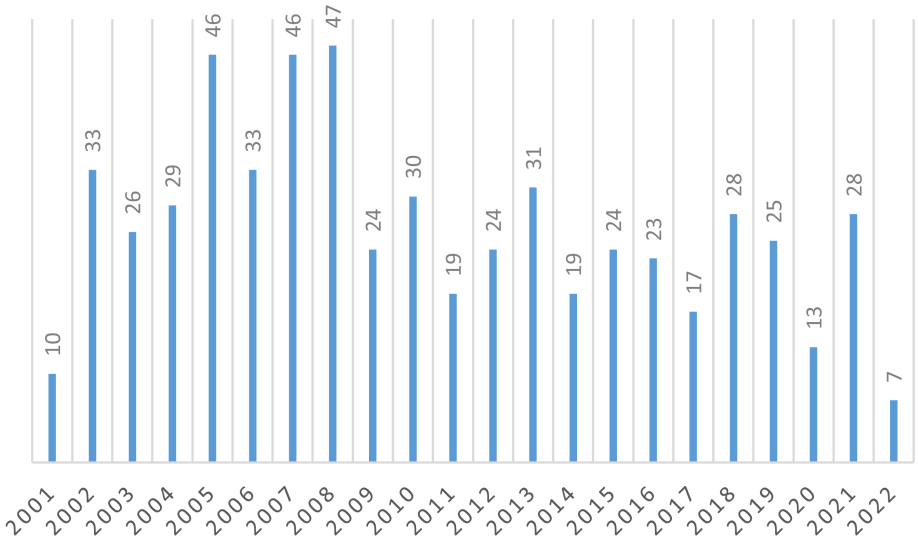

Figure 9: Number of the published paper per year from Belle.

Literally, the Belle II is the successor of the Belle experiment [81], which accumulated more than 1 ab^{-1} [82] in total. In addition, some of the Belle data such as Y(6S) on-resonannce data are very unique data sets in the world. Therefore, Belle decided to preserve all RAW and mDST (mini. Data Summary Table for physics analysis) data sets for the Belle II experiment and make these Belle data and software accessible to the Belle II collaborators at least until the Belle II data exceeds the Belle data. Currently, these Belle data are stored at the KEK central computing (KEKCC) system, which is also the main computing system for the Belle II experiment. Although Belle II adopted the distributed computing system, Belle data is accessible only from KEKCC, now. Belle software has been fixed in 2009 except for some pathches. The Belle software was originally developed with SL5, but it was migrated to SL6 and CentOS7 because it is necessary for the signal MC production even now. In parallel, the Belle data I/O tool was developed and integrated in the Belle II software framework, so that the Belle II collaborators can read and analyze them only with the Belle II software in which the recent analysis tools are available. Thanks to these data and analysis preservation efforts, Belle physics analysis activities are still quite vital even 10 years after the data-taking was finished.

Figure 9 shows the number of published physics paper per year from Belle since 2001. We keep sending out the new physics results and constantly publishing 20 papers per year.

For the future Belle data preservation, we maybe need to migrate all of the data from the current KEKCC to the new system because the replacement of the entire KEKCC system is scheduled in 2024 Summer. We also have to consider the data and analysis preservation beyond 2024 but the detail is not decided yet.

Concerning Belle II, we have launched the task force to evaluate the strategy of the data and analysis preservation. Under this activity, we are discussing the possible computing model both in the post-SuperKEKB-running period and post-Belle II experiment lifetime, the period of time for accessibility of the preserved data, the necessary analysis infrastructure, the estimated cost and effort, and so on.

4.10 MINERvA

MINERvA [83] is a neutrino scattering experiment at Fermilab that recorded data between 2009-2019. The MINERvA collaboration has published more than thirty neutrino interaction cross

section measurements [84] that will inform and tune interaction models for future oscillation experiments such as DUNE and HyperKamiokande. To ensure that its data is usable into the 2030's, the MINERvA collaboration began a data preservation project in 2019 [85]. The project has three components:

- Preservation of the high- and low-level reconstructed objects in a ROOT-based analysis ntuples corresponding to the entire MINERvA dataset.
- A software library known as the MINERvA Analysis Toolkit (MAT) with utilities for transforming the ROOT-based tuples into cross section measurements.
- A second software library (MAT-MINERvA) based on MAT that can be used to reproduce existing MINERvA analyses and form the basis for new analyses.

As of 2022, the two software libraries have been developed and are available on Github [86, 87], and the datasets are available on Fermilab-hosted DCache servers.

5 DP Technologies and projects

Although HEP has a long tradition of common software development, in particular for data access, simulation and analysis (e.g HIGZ [72], RooT [68], PAW [69], GEANT3 [88]), the computing systems and the data structures are largely experiment specific. This variety is of course related to an efficient experiment running and originates mainly from the objective of prompt physics results. This objective is to be obtained using the resources allocated to each experiment.

However, this approach induces a fragility towards a long term data preservation, since some systems are "custom made" and difficult to maintain in the longer term. It may also raise further difficulties when the data is open to a larger audience, beyond the experimental collaboration, because adjustments are needed towards lighter/standard interfaces - otherwise the learning curves for users may be overwhelming. Moreover, the data preservation ("cold") systems may require new technologies and new methodologies, beyond ("hot") systems used during the experiment lifetime.

Therefore, as identified also in the early DPHEP documents, there is a clear need for transverse projects, to be exploited by several or even all experiments, potentially proposing new standards and aligning the goals for long term preservation with those of the open and FAIR data. A few examples of such transverse projects are presented in this section.

5.1 HEPData

HEPData is the primary open-access repository for publication-related (Level 1) data from particle physics experiments, with a long history going back to the 1970s. The HEPData project underwent a complete transformation in 2017 to a new platform (hepdata.net) hosted on CERN computing infrastructure [89]. Another transition was made in 2020 to deploy the web application via a Docker image on a Kubernetes cluster shared with the INSPIRE-HEP project. Funding is provided by the UK Science and Technology Facilities Council (STFC) to Durham University (UK) for staff to maintain the operation of the hepdata.net site, provide user support, and develop the open-source software (@HEPData on GitHub) underlying the web application. In the past, data preparation in a standard format and upload to the repository were also handled by HEPData staff at Durham University, but now these tasks are delegated to the experimental collaborations.

Data submitted to HEPData (as YAML) is primarily in a tabular form that can be interactively plotted in the web application and automatically converted to other common formats (CSV, JSON, ROOT, YODA). The interactive nature of HEPData means that data tables must be kept sufficiently small ($\sim$MB or less) that they can be quickly loaded in a web browser. In practice, tables with more than $\sim$10,000 rows (for example, a covariance matrix for a measurement with $\sim$100 bins) cannot be practically rendered in a web browser. However, moderately large tables or non-tabular data files can be attached to a HEPData record as additional resources (in any format), where the original files can be downloaded but the interactive nature is lost. To avoid the multiple problems caused by the attempted upload of large files, an overall size limit of 50 MB is currently imposed on the uploaded archive file. Publication-related (Level 1) high-energy physics data that is not suitable for HEPData, due to either being too large or predominantly in a non-tabular format, should be submitted to another data repository like Zenodo, which currently plugs a gap to host HEP data (and software) that does not fit into other repositories.

5.2 CERN Open Data Portal

The CERN Open Data portal [38] was launched in 2014. The portal manages and disseminates more than two petabytes of open data from particle physics. The majority of the data content

comes from the LHC experiments. The data are being released in periodic batches in close collaboration with LHC experiments and their data preservation teams following the published open data policies. This usually implies a certain embargo period that allows for the exploitation of data within the collaboration as well as for the data curation and verification procedures.

The released open data are described as bibliographic records following a custom metadata schema. The portal focuses on describing the usual metadata (title, authors, keywords, year), the technical metadata (file formats, file sizes, file locations and checksums) as well as high-level metadata necessary to understand the data context (how was the data selected, how it can be used, the data semantics). The bibliographic records are minted with a Digital Object Identifier (DOI) to ease data citation and referencing.

The CERN Open Data portal content currently contains over 15 thousands bibliographic records describing over 1.2 million files of over 2.8 petabytes. The content represents raw data samples (Level 4),the collision and simulated data sets (Level 3), as well as simplified derived data sets and event display files (Level 2). Whenever possible, the data is accompanied with associated documentation, the data provenance information [90] as well as the corresponding data production configuration files (see Fig. 10). The portal also describes software tools and provides data analysis examples together with the associated Virtual Machines and Docker container environments where the data analysis examples can be run. This aims to simplify data reuse for theoretical physicists or machine learning specialists that may undertake data analyses without being necessarily well acquainted with the detailed internal knowledge of the LHC experiment that produced the data.

The curated open data content is being used for both educational and research purposes. The CERN Open Data portal offers embedded event display interfaces allowing particle physics students to visualise detector events and perform basic histogramming on site. The research-grade data is available for download by HTTP and XRootD protocols. A command-line client was developed to allow researchers to automatise download procedures and to verify the integrity of downloaded data for large collision and simulated data sets.

Besides the four LHC experimental collaborations (ALICE, ATLAS, CMS, LHCb), the data sets of relevance to the Machine Learning communities are classified under a separate experiment-independent Data Science collection the interest in which is constantly growing. Beyond LHC particle physics, the CERN Open Data portal nowadays disseminates the OPERA neutrino physics data sets [91] as well as the first batch of open data from the PHENIX collaboration, demonstrating the widening of the open data coverage from pure CERN LHC particle physics content towards the wider HEP open data content at large.

5.3 CERN Analysis Preservation

CAP is a service built at CERN to meet the specific analysis preservation needs of research teams at the Large Hadron Collider (LHC) experiments. Generated at the world's largest research instrument, through multinational research collaborations consisting of thousands of scientists and engineers, the research data of the LHC experiments is unique, precious and of unparalleled complexity. While the dynamic organizational setting of CERN makes it particularly challenging for the data, software and associated knowledge around a physics analysis to be preserved in a comprehensive reusable manner, the LHC experiments aim to adopt a consistent approach towards the openness and preservation of experimental data.

CAP was developed to preserve information related to a scientific analysis as it is produced, making it easier to describe, find, exchange and hand-over information in fast paced research environments with highly fluctuating personnel. The service responds to two parallel demands.

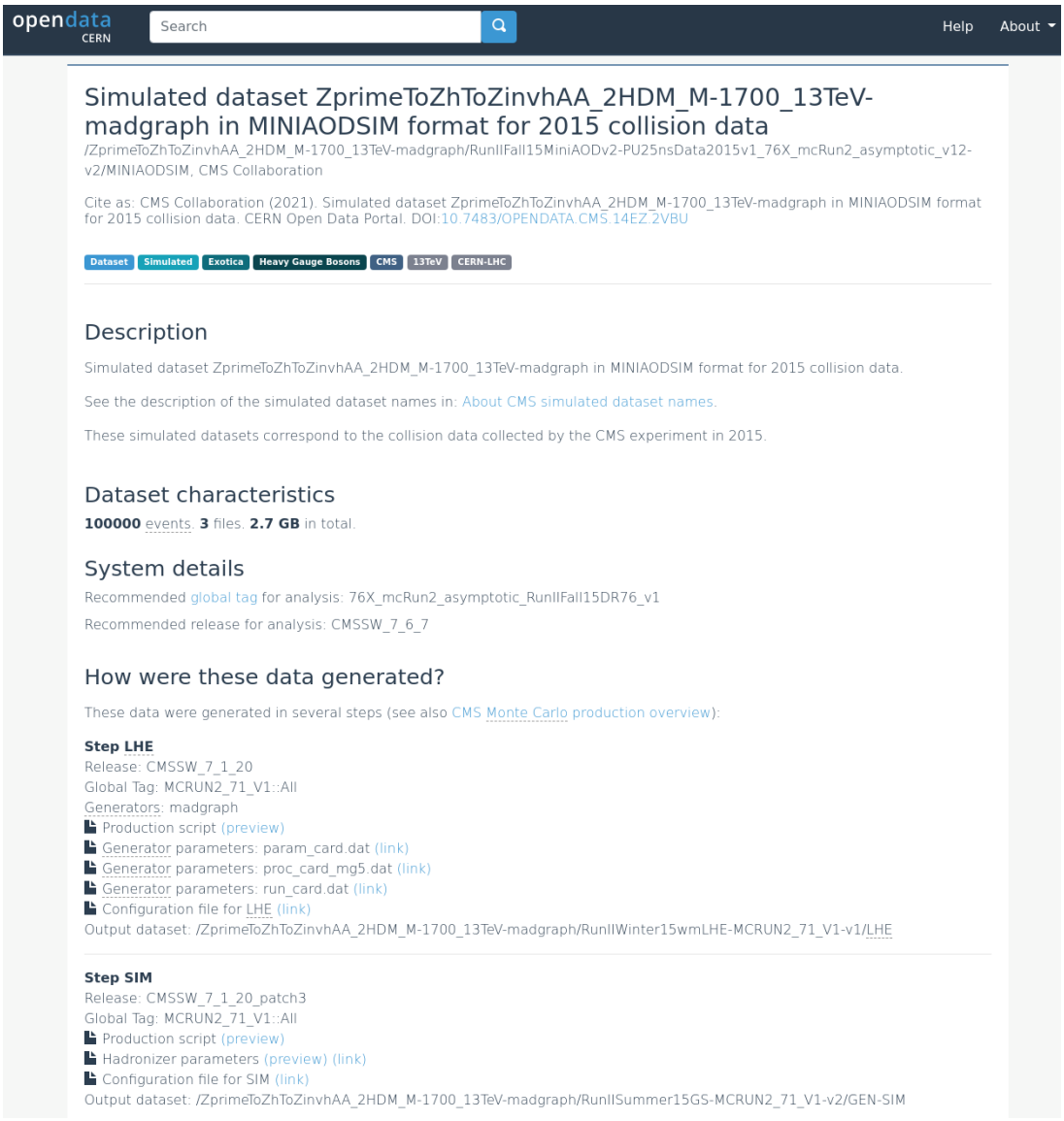

Figure 10: Example of a CMS simulated data set released on the CERN Open Data portal. The full data set provenance information is captured and made available to users.

First, the internal needs of the community, i.e. the LHC experiments, where the high throughput of analyses results in significant challenges in terms of capturing and preserving data analyses and enabling reproducibility of research results, which might be needed at a future point. This is evidenced by the large number of cases where knowledge around analyses was lost once researchers moved on, as some analysis materials had not been properly nor timely made accessible. Second, the external demands: the service allows experimental teams to address the increasingly pressing requirements of funding agencies worldwide who have put in place data management policies regarding research data and knowledge preservation for future reuse and reproducibility.

Developed in close partnership with a diversity of research teams at the LHC, CAP has been designed to flexibly adapt to varying experimental workflows, e. g. building on centralized tools that are already in place or taking over the role of information aggregator. The service facilitates the capture of information about scientific analyses to facilitate reuse and reproduction even many years after its initial publication, permitting to extend the impact of preserved analyses through future revalidation and recasting services. To facilitate the adoption of analysis preservation as a standard part of the scientific process, CAP has been designed to seamlessly integrate into collaboration workflows, so that information can be captured at the earliest stage, and throughout an analysis life-cycle. Designed as a secure environment, experimental teams retain full control of their information and data. CAP enables users to apply the appropriate access restrictions so that users are always in control of when and if their work is shared or published.

Concrete use cases are:

- Community analysis information hub: Bring together information from various tools and databases. Search, compare, retrieve and share information.
- Validation and reuse: Instantiate preserved analyses and computational workflows on compute clouds to allow their validation or execution with new sets of parameters to test new hypotheses.
- Streamlined handover and onboarding: A person having done an analysis is leaving the collaboration and has to hand over the know-how to other collaboration members. A newcomer would like to join a group and requires information on past analyses.
- Optimised preservation-publication workflow: Prepare more complex outputs for public releases. Easy information exchange between CAP and public-facing publishing platforms.
- Support for Open Science policies: Aggregate and preserve information to comply with any internal or external policy requirements.
- Multipurpose reviewing tool: Enable an internal review committee to check and repeat an analysis.

Built over several years of focused development and tested across research teams at the LHC, CAP is a mature and robust solution for the preservation of analyses to support data preservation and management to meet funder expectations and research integrity concerns in a world moving increasingly towards open science. CAP has been endorsed by the four main LHC experiments and through its demonstrated success in the highly complex domain of high energy physics, at one of the world's largest and complex research organizations, the CAP service shows potential for broader application across research disciplines. Furthermore, the technology and tools used to develop the service are not only open source, but are discipline agnostic. More specifically, CAP is built on top of the open source framework Invenio, which already has a wide variety of

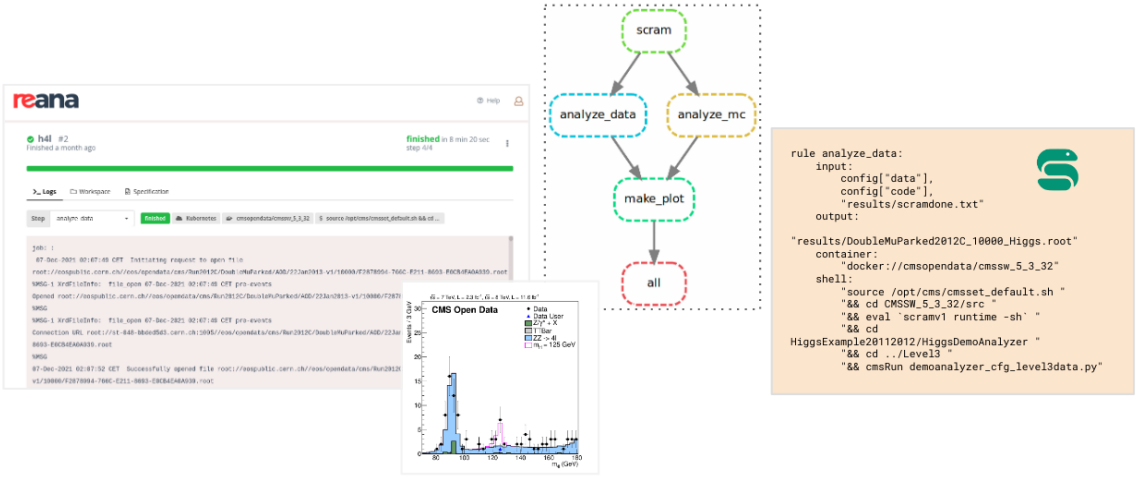

Figure 11: Example of Higgs-to-four-leptons analysis of CMS open data running on REANA reproducible analysis platform. The analysis consists of four steps and is expressed in the Snakemake workflow specification language. The specification for one of the steps is illustrated on the right.

applications in many disciplines, and all the underlying source code for CAP is openly available on GitHub.

5.4 REANA Reproducible Analyses

REANA is a reproducible analysis platform [23] that aims to facilitate reusable science by allowing researchers to structure and run parametrised computational data analysis workflows. REANA was launched in 2017 as a sister project to ensure the reuse of content preserved in the CERN Open Data portal and the CERN Analysis Preservation framework services [8].

The data analysis process can be generally described by means of specifying input data, the analysis code, and the computational recipe used to arrive at the final results trough a sequence of computational steps. This process can be expressed by means of Directed Acyclic Graph (DAG) where graph vertices represent units of computation with their inputs and outputs and graph edges describe the interconnection of various computational steps [92]. REANA supports several such DAG standards (CWL, Snakemake, Yadage), parses the workflow specification described by the user and dispatches its computational steps to various supported compute backends (Kubernetes, HTCondor, Slurm). The reproducibility of computations is assisted by means of using software containers (Docker, Singularity) that fully encapsulate the original computational environments of each analysis step.

The REANA approach was successfully tested in several data production and data analysis scenarios. For example, REANA was used for ATLAS reinterpretation searches for new physics or CMS reconstruction and jet energy corrections [93]. In the CERN Open Data portal, REANA is used both on the "data production" side to ensure the correctness of preserved data sets' provenance information [90] as well as on the "data analysis" side by running several data analysis examples such as CMS open data Higgs-to-four-lepton example analysis (see Fig. 11).

REANA promotes early adoption of computational reproducibility principles by researchers. The integration with source code management platforms such as GitLab allows researchers to develop analyses on GitLab and run either full data analysis tasks on REANA or at least test the correctness of analysis workflow after each code change. If an analysis is developed in this "continuous integration" manner [93], the preservation of knowledge associated with the data analysis as well as the future deposit of analysis assets into digital repositories are largely facil-

itated. REANA therefore complements the data preservation repositories by promoting active "preproducibility" of data analyses during the active analysis phase rather than only relying on passive data deposition and subsequent "reproducibility" once the analysis is completed [94].

REANA was developed with particle physics use cases in mind, but the platform profits from synergies with general reproducible data analysis patterns in other scientific disciplines, such as astronomy, bioinformatics and life sciences [95].

5.5 Bit Preservation at CERN

During 2021 CERN migrated all archival storage from the earlier CASTOR system to the new CERN Tape Archive (CTA) which consequently now assumes CERN's bit preservation responsibilities. All preservation use cases, including of course the LEP data, were moved into CTA.

In terms of bit preservation CTA provides essentially the same guarantees as CASTOR. Indeed, the migration was purely a matter of moving metadata and the data on tapes remained unaffected. File access permissions are managed slightly differently in CTA and this triggered discussion on the appropriate model for authorising access to preservation data. A model where a privileged curator is able to maintain a list of reader accounts (using CERN's e-groups system) is being elaborated.

During 2021 CERN deployed a new tape library in a building separate from the computer centre. This library will be commissioned for use with CTA, which will allow dual-replica data to benefit from geographical separation between the replicas.

CTA now holds the data that CERN has received from *BABAR* for the purposes of data preservation. The integrity of this archive has yet to be validated by *BABAR* and thus the transfer is not yet considered complete.

5.6 CERNVM

CERNVM provides a portable platform to develop and execute HEP experiment applications [96]. The CERNVM platform consists of a virtual appliance with a minimal operating system (OS) and a distributed file system (CVMFS) that provides a global shared area for scientific applications and conditions data [97]. Both the appliance and the distributed file system distribute data processing environments to the heterogeneous, distributed computing infrastructure used by running experiments. The CERNVM technology is designed to take software preservation aspects into account and to allow for an easy transition from an actively maintained experiment software stack to the preservation phase [98].

The CERNVM appliance provides both an image for a full virtual machine (VM) for use with hardware virtualization technology such as KVM and Amazon EC2 as well as a container image for use with container tools such as Docker and Kubernetes. Full VMs give access to hardware emulation technology, so that software stacks compiled for deprecated CPU architectures can run on contemporary hardware, albeit at a significant performance cost (often $\times 10$ and more slower). Container virtualization has a negligible performance overhead; it relies, however, on a stable interface from the Linux kernel to user-land applications.

Retrofitting the CERNVM technology to software stacks from the LEP experiments, the approach has shown to bridge more that 20 years. A CERNVM environment with a CMS software stack from 2011 to process LHC run 1 data was created with minimal effort, since the CVMFS and CERNVM infrastructure has been in use from the start by LHC experiments.

Key in the CERNVM approach is the split between the minimal appliance (the OS platform) and the experiment software directory tree provided by the CVMFS network file system. The

naïve use of virtualization technology, in contrast, which bundles OS and application software in a single VM or container image, is easily susceptible to container management problems. Due to the complexity and variety of experiment application software, the naïve approach results either in huge images (e.g., 100GB+ for all ATLAS releases) that are difficult to distribute or in a proliferation of special purpose containers that are difficult to book-keep. Software stacks provided through CVMFS, on the other hand, are well-maintained by the experiments' software librarians and the necessary binaries for any given task are loaded on-demand at runtime. As CVMFS is a versioning file system, software stacks installed on CVMFS are automatically preserved remain available for future use.

Several lessons were learned from exercising CERNVM and CVMFS for the preservation of LEP and LHC experiment applications:

1. It is important to distinguish between *scientific* applications (e.g., event generators, detector simulation, histogramming tools) and *system* applications (e.g., web clients, data access software). In contrast to scientific applications, system applications generally do not freeze well. This is due to the fact that system applications often interact with the outside world or at least the (virtualized) hardware. For instance, web clients may unsuccessfully try to connect with an outdated SSL protocol to HTTPS servers. Hard disk utilities may be subject to integer overflows on today's volume sizes.

2. As a result, it is important to identify network dependencies of data processing workflows before freezing. There may be dependencies, for instance, to online databases (such as conditions databases), grid configuration, or data access servers. A stable frozen software environment should only use the POSIX file system as a means to interact with the outside world. The experiment software as well as all input and output data should be stored on network file systems, which can be mapped into virtualized environments. In CERNVM, CVMFS is used to store software and detector conditions data and EOS is used for experiment data.

3. Lifting authorization barriers (such as X.509 certificate authorization) has been useful in the context of data preservation efforts. Authorization protocols are particularly ill-suited for freezing due to sudden upgrades of protocols or software following the revelation of vulnerabilities. Also, there are often expiry mechanisms in security protocols such as certificate lifetimes. Software and data on CVMFS is publicly readable by design; providing the data on public EOS servers instead of protected shares significantly simplifies the provisioning of historic data processing environments.

Looking ahead, new challenges arise for the preservation of the LHC experiment software of current and future runs. Compared to Run 1, the software complexity rose substantially, especially with respect to external, non-HEP software such as mathematical and ML libraries and scientific Python tools (Fig. 12). Growing software stacks can be mitigated by strengthening the role of software librarians and generally a disciplined approach to software dependency management. Secondly, the last few years saw a rise of greater architectural variety compared to the previous decades which were dominated by the Intel X86_64 architecture. Today, there is a mix of Intel, ARM, and POWER CPUs, accompanied by GPUs and (rarely) FPGAs. When transitioning a software stack into a frozen state, it will likely be necessary to streamline the available binaries to one or few reference architectures.

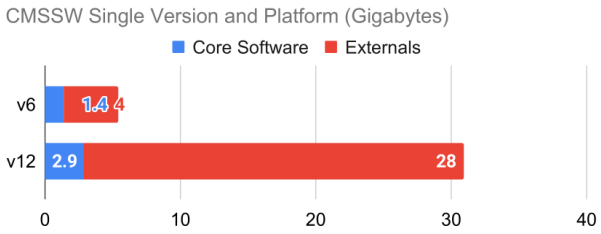

Figure 12: Growth of the CMS software stack in approximately 5 years.

5.7 Managed service migration/retirement: the CERNLIB example

The management of the retired external software packages poses a problem for the rapidly developing but not optimally managed physics-related software.

One of the approaches that could be used to maintain the external software packages is to create a digital archive/library implemented on top of the popular software versioning platforms as github or GitLab.

CERNLIB is one of the key software components that have been extensively re-used by many collaborations and still plays an key role for the preservation activities e.g. for LEP- but also HERA-era experiments. The CERNLIB libraries and tools have since their inception been migrated to a large set of platforms and also several code management systems – still preserving most of the original functionality and development(commit) history.

The CERNLIB contains multiple software sub-libraries, many of which are actually independent software packages. Logically CERNLIB is split into a set of sub-packages

- MATHLIB, a library with mathematical routines
- PACKLIB, general purpose library
- KERNLIB, general purpose library
- GEANT, a version of the Geant3 [88] simulation toolkit.
- MCLIB, which contains a set of most used MC event generators.
- GRAFLIB, a library with interfaces to graphical routines
- PAWLIB, Physics Analysis Workstation

Each sub-package is split into smaller sub-sub-subpackages, very often with their own requirements and compilation options. Depending on the host system, the CERNLIB might requre the following software

- C preprocessor, C99 and Fortran77 compilers
- SED and ZIP utilities
- CMAKE and any build system – for sc CMake builds
- IMAKE and MAKE – for sc Imake builds
- BLAS with Lapack compiled with the same Fortran compiler (optionally)

- Freetype libraries and headers
- X11 libraries and headers of version R6.6+
- Motif libraries and headers
- OpenSSL libraries (and headers for CMake builds)

The community of users may soon be facing the challenges of a forced migration (caused by hardware and OS obsolescence) and consequently validation on recent 64-bit platforms. Since such a validation poses significant technical and personnel effort, the idea to consolidate the validation activities on a single, community maintained software base with a small set of actively validated platforms emerged at a recent DPHEP community workshop. The effort in a consolidated, up-to-date issue tracking and, as much as possible, automated build and validation pipeline would then allow to exploit the small remaining expert effort for the benefit all remaining CERNLIB deployments and prepare for a period in which expertise in CERNLIB internals is rare or not anymore readily available.

After the end of active development of the CERNLIB in 2007, it was widely used by many collaborations and individuals. The intensive usage of CERNLIB on the constantly updated operating systems resulted in a development of large number of code patches. In the end of 200x most of these patches were consolidated by Kevin McCarthy and used to create Debian packages for CERNLIB. Those patches, as well as patches from the DESY theory group, Fedora and Debian projects became the starting point for the updates of CERNLIB in 2022. All the updates were implemented in the newly created CERNLIB GIT repository hosted by CERN[15] and later extended by various improvements by the CERNLIB/DPHEP initiative. The repository is a merger of previously separate repositories with CERNLIB sub-libraries and contains all the development history of CERNLIB since 1990s. The code is also supplemented with a continuous integration system (CI) that performs builds of CERNLIB on selected platforms, namely older Scientific Linux 4, 5, 6; CentOS 7, 8, 9; Alma 8, 9; Fedora 35; in 32- and 64-bit variations, as well as Ubuntu 18, 20, 22; in 64-bit.[16] While some of the images used in CI are stock images from the vendors, the older Scientific Linux 4, 5, and 6 container images are maintained under the DPHEP cover.

While CERNLIB contains codes written in Fortran and C, there are little worries about future ability to compile the updated CERNLIB code. However, the native CERNLIB build system, IMAKE [99], is relatively outdated and can potentially disappear in the nearest future. Therefore, while the IMAKE scripts are still maintained in the CERNLIB, significant efforts were put recently into a re-implementation of the build system of CERNLIB in CMAKE [100]. CMAKE is an open-source, free cross-platform build tool with a history for more than 20 years. As of 2022 CMAKE is de-facto the standard for software written in C/C++/Fortran. With the CMAKE-based build system it became possible to perform checks of CERNLIB code with many more compilers on multiple historical and modern platforms. This approach is a carbon copy of the techniques used to port the JADE software, see Sec. 4.5. As a result, the updated CERNLIB version can be compiled on RHEL4+ i686, x86_64, ppc64 systems and modern 64bit MacOSX using GNU 3+, CLang, Intel and NVidia compiler collections. However, only certain combinations of the compilers and systems are fully backed by the DPHEP and included in the CI.

[15] `https://gitlab.cern.ch/DPHEP/cernlib/cernlib`

[16] Upstream Canonical dropped certain 32bit development libraries, therefore for Ubuntu only 64-bit variations are available.

Distribution The CERNLIB code is available in `https://gitlab.cern.ch/DPHEP/cernlib/cernlib` and should serve as a drop-in replacement for the CERNLIB 2006. The updated CERNLIB can be also packaged into RPM package. While no further development of CERNLIB is foreseen, we encourage the current and future users of CERNLIB to give a feedback on the updated CERNLIB and submit their suggestions, bugreports and bugfixes.

5.8 ARCHIVER

The project ARCHIVER [101, 102] develops innovative services for Long Term Digital Preservation of scientific datasets using the Pre-Commercial Procurement instrument funded by the European Commission. R&D was performed competitively by commercial suppliers, across different implementation phases. The services were developed by Arkivum and Libova co-designed with input from research clusters members: CERN, PIC/IFAE and DESY are members of the ESCAPE cluster, DESY is also a partner in the ExPaNDs cluster, while EMBL-EBI is a member of EOSC-Life. These research performing organizations deployed use cases from Astrophysics, High-Energy Physics, Life Sciences and Photon-Neutron Sciences. The use-cases driving the ARCHIVER consortium's need for research and development of innovative data preservation services extended the preservation ecosystems of research organizations to create more dynamic solutions to be deployed primarily by using a hybrid model, on-premise or exclusively in cloud environments, operated by European SMEs that were enhanced and assessed against best practices such of CoreTrustSeal and DPC RAM, so that datasets remain FAIR (Findable, Accessible, Interoperable, Reusable) for decade timescales or more.

The services resulting from the ARCHIVER R&D are addressing the gaps that put data at long-term risk as they prevent the construction and operation of sustainable Trusted Digital Repository services, affecting organizations both large and small who are tasked with being custodians of valuable research artifacts. Thanks to its outstanding results, the ARCHIVER project has been recognized by the international community and been awarded by the International Council on Archives, category for Collaboration and Cooperation, at the DPC awards ceremony in Glasgow, on the 12th of September 2022. The model developed in ARCHIVER is considered very beneficial for the research community and contributes to address concerns about sharing and re-use of FAIR data to reproduce research as a pillar for Open Science and long-term preservation of the EOSC federated data, by delivering sustainable, production quality long-term data preservation services for user communities that fills a gap in the existing European Open Science Cloud portfolio.

6 DPHEP: the way forward

DPHEP is an unique forum ensuring continuity, expertise exchange and knowledge preservation within a field where the usual HEP collaborative models are modified to deal with a rather uncommon situation: obtain new scientific results at constant level of complexity with reduced resources and therefore refactored analysis environments. With an overview of over ten years, one can observe that the concepts, the methodologies, the technologies and the collaborative configurations required to preserve data at long term require a significant amount of innovation. While DPHEP focused so far on collider experiments, it is obvious that other collaborations – in particle physics, astroparticles, cosmology, nuclear physics etc. – face similar problems in preserving large and complex data sets. An increased synergy within and beyond the HEP community remains possible and desirable. The links with the industrial and economical worlds are an interesting perspective as well, still to be explored.

Moreover, data preservation is one of the multiple possible actions towards improving the current and future computing systems in HEP and beyond. As it becomes more and more obvious from the case studies presented above, the data preservation activity is not only performing research on existing unique data sets, it is also shaping the future, in coherence with the open science policies and methodologies. Furthermore, there is a clear potential for training the new generations and improving their technological skills.

The DPHEP Collaboration commits to drive a culture of open sharing of knowledge, data, software, algorithms, infrastructures and other research resources. The objectives for the next period are to:

1. Improve the awareness and stimulate improvements of data preservation:
   - Explore and document the aspects related to data preservation: scientific motivation, organisation options, technologies, standards, outreach and education.
   - Attract new collaborators, enlarge the community, organise workshops, issue regular Global Reports.
   - Ensure links to other communities.
2. Reinforce and support the ongoing laboratory-based projects, encourage inter-laboratory cooperation and strengthen the links to other projects in different host institutions (external computing centers, Universities etc.).
   - Keep alive data sets that (can) still produce science and keep track of parked data sets.
3. Support/develop the DP aspects for future experiments and encourage the transfer of knowledge.
4. Encourage open data and open science as a way to preserve data and knowledge.

Data preservation is one of the building blocks of the HEP scientific outcome and the DPHEP Collaboration intends to stimulate and support it.

Acknowledgements

The authors would like to thank the respective funding agencies for support in pursuing the topic. In particular, the endorsing of DPHEP as an ICFA panel has been and continues to be an essential base of sustainable international collaboration in the area of data preservation across collaborations and laboratories.

References

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

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

[3] **DPHEP** Collaboration, S. Amerio *et al.*, "Status Report of the DPHEP Collaboration: A Global Effort for Sustainable Data Preservation in High Energy Physics", arXiv:1512.02019 [hep-ex].

[4] CERN, "CERN Open Data Policy for the LHC Experiments." CERN-OPEN-2020-013, Geneva, Nov, 2020. https://cds.cern.ch/record/2745133.

[5] **LHC** Collaboration, "Update of the Computing Models of the WLCG and the LHC Experiments",. http://cds.cern.ch/record/1695401.

[6] **HEP Software Foundation** Collaboration, J. Albrecht *et al.*, "A Roadmap for HEP Software and Computing R&D for the 2020s", *Comput. Softw. Big Sci.* **3** no. 1, (2019) , arXiv:1712.06982 [physics.comp-ph].

[7] D. Ozerov and D. M. South, "A Validation Framework for the Long Term Preservation of High Energy Physics Data", *J. Phys. Conf. Ser.* **513** (2014) , arXiv:1310.7814 [hep-ex].

[8] X. Chen *et al.*, "Open is not enough", *Nature Phys.* **15** no. 2, (2019) .

[9] M. Wilkinson *et al.*, "The fair guiding principles for scientific data management and stewardship", *Scientific Data* **Article No.160018** no. 3, (2016) . 10.1038/sdata.2016.18.

[10] "ISO 14721: OAIS – a reference model for what is required for an archive to provide long-term preservation of digital information." https://www.iso.org/standard/57284.html.

[11] "ISO 16363: Audit and certification of trustworthy digital repositories." http://www.iso16363.org/.

[12] J. Shiers and M. Hildreth, "DPHEP - Certification: Motivation, Benefits and Status." https://indico.cern.ch/event/658060/contributions/2889539/.

[13] G. Ganis and J. Blomer, "Achilles' heels in LTDP or how do we address risks." https://indico.cern.ch/event/658060/contributions/2915946/.

[14] "HEPData: Repository for publication-related High-Energy Physics data." https://www.hepdata.net.

[15] "Rivet: Analyses reference." https://rivet.hepforge.org/analyses.html.

[16] **ATLAS** Collaboration, "Simplified ATLAS SUSY Analysis Framework." http://simpleanalysis.docs.cern.ch.

[17] "Rivet: Analysis coverage." https://rivet.hepforge.org/rivet-coverage.

[18] **ATLAS** Collaboration, "SimpleAnalysis software release." https://zenodo.org/communities/atlas_experiment/.

[19] **ATLAS** Collaboration, "Implementation of simplified likelihoods in HistFactory for searches for supersymmetry",. http://cdsweb.cern.ch/record/2782654.

[20] **ATLAS** Collaboration, "Reproducing searches for new physics with the ATLAS experiment through publication of full statistical likelihoods",. http://cdsweb.cern.ch/record/2684863.

[21] K. Cranmer and I. Yavin, "RECAST: Extending the Impact of Existing Analyses", *JHEP* **04** (2011) , arXiv:1010.2506 [hep-ex].

[22] **ATLAS** Collaboration, "Reinterpretation of the ATLAS Search for Displaced Hadronic Jets with the RECAST Framework",. http://cdsweb.cern.ch/record/2714064.

[23] "REANA: Reproducible research data analysis platform." https://www.reana.io/.

[24] **ATLAS** Collaboration, "ATLAS 13 TeV samples collection Gamma-Gamma, for 2020 Open Data release. CERN Open Data Portal." http://opendata.cern.ch/record/15006.

[25] **ATLAS** Collaboration, "ATLAS 13 TeV samples collection exactly three leptons (electron or muon), for 2020 Open Data release. CERN Open Data Portal." http://opendata.cern.ch/record/15004.

[26] **ATLAS** Collaboration, "ATLAS Virtual Machine, for 2020 Open Data release. CERN Open Data Portal." http://opendata.cern.ch/record/15008.

[27] **ATLAS** Collaboration, "Athena." https://gitlab.cern.ch/atlas/athena.

[28] **ATLAS** Collaboration, "Athena", May, 2021. https://doi.org/10.5281/zenodo.4772550.

[29] **ATLAS** Collaboration, "Athena Container Repository." https://gitlab.cern.ch/atlas/athena/container_registry.

[30] **ATLAS** Collaboration, "ATLAS Data Access Policy. CERN Open Data Portal." https://opendata.cern.ch/record/413.

[31] M. Elsing *et al.*, "TrackML Particle Tracking Challenge." https://sites.google.com/site/trackmlparticle/.

[32] **ATLAS** Collaboration, "Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014. CERN Open Data Portal." http://opendata.cern.ch/record/328.

[33] **ATLAS** Collaboration, "FastCaloGAN Training Project", October, 2021. https://doi.org/10.5281/zenodo.5589623.

[34] **ATLAS** Collaboration, "Datasets used to train the Generative Adversarial Networks used in ATLFast3. CERN Open Data Portal." https://opendata.cern.ch/record/15012.

[35] "CERN analysis preservation." https://analysispreservation.cern.ch/.

[36] **ATLAS** Collaboration, "Performance of Multi-threaded Reconstruction in ATLAS",. http://cdsweb.cern.ch/record/2771777.

[37] **CMS** Collaboration, A. M. Sirunyan *et al.*, "2020 CMS data preservation, re-use and open access policy. CERN Open Data Portal." https://opendata.cern.ch/record/415.

[38] CERN, "CERN open data portal." http://opendata.cern.ch.

[39] **CMS** Collaboration, A. M. Sirunyan *et al.*, "Container image." https://gitlab.cern.ch/cms-cloud/cmssw-docker-opendata/container_registry.

[40] **CMS** Collaboration, A. M. Sirunyan *et al.*, "Container image." https://hub.docker.com/u/cmsopendata.

[41] **CMS** Collaboration, A. M. Sirunyan *et al.*, "VM images." http://opendata.cern.ch/search?page=1&size=20&q=VM&subtype=VM&type=Environment&experiment=CMS&file_type=ova.

[42] **CMS** Collaboration, A. M. Sirunyan *et al.*, "CMS Open data guide." https://cms-opendata-guide.web.cern.ch/.

[43] "Rivet – the particle-physics MC analysis toolkit." https://rivet.hepforge.org.

[44] LHCb, "Davinci physics analysis application." https://twiki.cern.ch/twiki/bin/view/LHCb/DaVinci.

[45] **H1** Collaboration, I. Abt *et al.*, "The Tracking, calorimeter and muon detectors of the H1 experiment at HERA", *Nucl. Instrum. Meth. A* **386** (1997) .

[46] **H1** Collaboration, I. Abt *et al.*, "The H1 detector at HERA", *Nucl. Instrum. Meth. A* **386** (1997) .

[47] **H1** Collaboration, D. Britzger, S. Levonian, S. Schmitt, and D. South, "Preservation through modernisation: The software of the H1 experiment at HERA", *EPJ Web Conf.* **251** (2021) , arXiv:2106.11058 [hep-ex].

[48] A. Pfeiffer, "Overview of the lcg applications area software projects", in *IEEE Symposium Conference Record Nuclear Science 2004.*, vol. 4, pp. 2020–2023 Vol. 4. 2004.

[49] S. Roiser, A. Gaspar, Y. Perrin, and K. Kruzelecki, "Servicing HEP experiments with a complete set of ready integreated and configured common software components", *Journal of Physics: Conference Series* **219** no. 4, (Apr, 2010) .

[50] **H1** Collaboration, V. Andreev *et al.*, "Measurement of multijet production in ep collisions at high Q^2 and determination of the strong coupling α_s", *Eur. Phys. J. C* **75** no. 2, (2015) , arXiv:1406.4709 [hep-ex].

[51] **H1** Collaboration, V. Andreev *et al.*, "Measurement of Jet Production Cross Sections in Deep-inelastic ep Scattering at HERA", *Eur. Phys. J. C* **77** no. 4, (2017) , arXiv:1611.03421 [hep-ex]. [Erratum: Eur.Phys.J.C 81, 739 (2021)].

[52] **H1** Collaboration, F. D. Aaron *et al.*, "Inclusive Deep Inelastic Scattering at High Q^2 with Longitudinally Polarised Lepton Beams at HERA", *JHEP* **09** (2012) , arXiv:1206.7007 [hep-ex].

[53] **H1** Collaboration, V. Andreev *et al.*, "Measurement of lepton-jet correlation in deep-inelastic scattering with the H1 detector using machine learning for unfolding", arXiv:2108.12376 [hep-ex].

[54] M. Arratia, D. Britzger, O. Long, and B. Nachman, "Reconstructing the kinematics of deep inelastic scattering with deep learning", *Nucl. Instrum. Meth. A* **1025** (2022) , arXiv:2110.05505 [hep-ex].

[55] **ZEUS, DESY DPHEP Group** Collaboration, J. Malka, "The ZEUS data preservation project", in *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference and 19th Workshop on Room-Temperature Semiconductor X-ray and Gamma-ray Detectors*, pp. 2022–2023. 2012.

[56] A. Bacchetta *et al.*, "Summary of Workshop on Future Physics with HERA Data", in *Future Physics with HERA Data for Current and Planned Experiments*. 1, 2016. arXiv:1601.01499 [hep-ex].

[57] A. Geiser, "Possible future HERA analyses", in *Future Physics with HERA Data for Current and Planned Experiments*. 12, 2015. arXiv:1512.03624 [hep-ex].

[58] "arXiv.org." https://arxiv.org.

[59] "iNSPIRE: Discover High-Energy Physics Content." https://inspirehep.net.

[60] **ZEUS** Collaboration, A. Verbytskyi, "The ZEUS long term data preservation project", *PoS* **DIS2016** (2016) , arXiv:1607.01898 [hep-ex].

[61] D. Krücker, K. Schwank, P. Fuhrmann, B. Lewendel, and D. M. South, "Data preservation for the HERA experiments at DESY using dCache technology", *J. Phys. Conf. Ser.* **664** no. 4, (2015) .

[62] "PUNCH4NFDI: Particles, Universe, Nuclei and Hadrons for the NFDI." https://www.punch4nfdi.de.

[63] "Key4hep: A Common Software Ecosystem." https://github.com/key4hep.

[64] B. Naroska, "e^+e^- Physics with the JADE Detector at PETRA", *Phys. Rept.* **148** (1987) .

[65] G. Corcella et al., "HERWIG 6: An Event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)", *JHEP* **01** (2001) , arXiv:hep-ph/0011363.

[66] T. Sjostrand, S. Mrenna, and P.Z. Skands, "PYTHIA 6.4 Physics and Manual", *JHEP* **05** (2006) , arXiv:hep-ph/0603175.

[67] T. Sjostrand, "High-energy physics event generation with PYTHIA 5.7 and JETSET 7.4", *Comput. Phys. Commun.* **82** (1994) .

[68] I. Antcheva et al., "ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization", *Comput. Phys. Commun.* **182** (2011) .

[69] R. Brun et al., "PAW: PHYSICS ANALYSIS WORKSTATION: INCLUDING VERSION 1.03/05." CERN-Q-121, 2, 1988.

[70] G. Behrmann et al., "A distributed storage system with dCache", *J. Phys. Conf. Ser.* **119** (2008) .

[71] P.A. Movilla-Fernandez, "Studien zur Quantenchromodynamik und Messung der starken Kopplungskonstanten $\alpha(s)$ bei $S^{1/2} = 14 - GeV - 44 - GeV$ mit dem JADE-Detektor", other thesis, Aachen, Tech. Hochsch., 11, 2002.

[72] R. Bock et al., "HIGZ: HIGH LEVEL INTERFACE TO GRAPHICS AND ZEBRA." CERN-Q-120, 3, 1988.

[73] **JADE** Collaboration, J. Schieck et al, "Measurement of the strong coupling $alpha_S$ from the three-jet rate in e^+e^- - annihilation using JADE data", *Eur. Phys. J. C* **73** no. 3, (2013) , arXiv:1205.3714 [hep-ex].

[74] **JADE** Collaboration, C. Pahl et al., "Study of moments of event shapes and a determination of alpha(S) using e+ e- annihilation data from Jade", *Eur. Phys. J. C* **60** (2009) , arXiv:0810.2933 [hep-ex]. [Erratum: Eur.Phys.J.C 62, 451–452 (2009)].

[75] S. Amerio *et al.*, "Data preservation at the Fermilab Tevatron", *Nucl. Instrum. Meth. A* **851** (2017) , arXiv:1701.07773 [hep-ex].

[76] **BESIII** Collaboration, M. Ablikim *et al.*, "Future Physics Programme of BESIII", *Chin. Phys. C* **44** no. 4, (2020) , arXiv:1912.05983 [hep-ex].

[77] **Belle-II** Collaboration, T. Abe *et al.*, "Belle II Technical Design Report", arXiv:1011.0352 [physics.ins-det].

[78] **Belle-II** Collaboration, "Belle II luminosity." https://confluence.desy.de/display/BI/Belle+II+Luminosity#BelleIILuminosity-Totalrecordedintegratedluminosity.

[79] **Belle-II** Collaboration, I. Adachi *et al.*, "Search for an Invisibly Decaying Z' Boson at Belle II in $e^+e^- \to \mu^+\mu^-(e^\pm\mu^\mp)$ Plus Missing Energy Final States", *Phys. Rev. Lett.* **124** no. 14, (2020) , arXiv:1912.11276 [hep-ex].

[80] **Belle-II** Collaboration, F. Abudinén *et al.*, "Search for Axion-Like Particles produced in e^+e^- collisions at Belle II", *Phys. Rev. Lett.* **125** no. 16, (2020) , arXiv:2007.13071 [hep-ex].

[81] **Belle** Collaboration, A. Abashian *et al.*, "The Belle Detector", *Nucl. Instrum. Meth. A* **479** (2002) .

[82] **Belle** Collaboration, "Belle luminosity." https://belle.kek.jp/bdocs/lumi_belle.png.

[83] **MINERvA** Collaboration, L. Aliaga *et al.*, "MINERvA neutrino detector response measured with test beam data", *Nucl. Instrum. Meth. A* **789** (2015) , arXiv:1501.06431 [physics.ins-det].

[84] MINERvA Collaboration, "Publications web page", 2023. https://minerva.fnal.gov/recent-minerva-results/.

[85] R. Fine *et al.*, "Data Preservation at MINERvA", arXiv:2009.04548 [hep-ex].

[86] MAT, "Git repository", 2023. https://github.com/MinervaExpt/MAT.

[87] MAT-MINERvA, "Git repository", 2023. https://github.com/MinervaExpt/MAT-MINERvA.

[88] R. Brun et al., "GEANT3." CERN-DD-EE-84-1, 9, 1987.

[89] E. Maguire, L. Heinrich, and G. Watt, "HEPData: a repository for high energy physics data", *J. Phys. Conf. Ser.* **898** no. 10, (2017) , arXiv:1704.05473 [hep-ex].

[90] T. Šimko *et al.*, "Open data provenance and reproducibility: a case study from publishing CMS open data", *EPJ Web Conf.* **245** (2020) .

[91] G. De Lellis, S. Dmitrievsky, G. Galati, A. Lavasa, T. Šimko, I. Tsanaktsidis, and A. Ustyuzhanin, "Dataset of tau neutrino interactions recorded by the OPERA experiment", *EPJ Web Conf.* **245** (2020) .

[92] T. Šimko, L. Heinrich, H. Hirvonsalo, D. Kousidis, and D. Rodríguez, "REANA: A System for Reusable Research Data Analyses", *EPJ Web Conf.* **214** (2019) .

[93] T. Šimko, L. A. Heinrich, C. Lange, A. E. Lintuluoto, D. M. MacDonell, A. Mečionis, D. Rodríguez Rodríguez, P. Shandilya, and M. Vidal García, "Scalable Declarative HEP Analysis Workflows for Containerised Compute Clouds", *Front. Big Data* **4** (2021) .

[94] P. B. Stark, "Before reproducibility must come preproducibility", *Nature* **557** (2018) .

[95] T. Šimko, K. Cranmer, M. R. Crusoe, L. Heinrich, A. Khodak, D. Kousidis, and D. Rodriguez, "Search for computational workflow synergies in reproducible research data analyses in particle physics and life sciences", in *14th International Conference on e-Science*, pp. 403–404. 2018.

[96] P. Buncic, C. A. Sanchez, J. Blomer, A. Harutyunyan, and M. Mudrinic, "A practical approach to virtualiztion in HEP", *The European Physical Journal Plus* **126** no. 1, (2011) .

[97] J. Blomer, C. Aguado-Sanchez, P. Buncic, and A. Harutyunyan, "Distributing LHC application software and conditions databases using the CernVM file system", *Journal of Physics: Conference Series* **331** no. 042003, (2011) .

[98] F. Berghaus, J. Blomer, S. D. Tiessen, G. C. Melia, G. Ganis, J. Shiers, and T. Simko, "CERN services for long term data preservation", in *Proc. 13th int. conf. Digital Preservation (iPres'16)*. 2016.

[99] P. DuBois, *Software portability with imake - practical software engineering (2. ed.)*. O'Reilly, 1996.

[100] Kitware, "Cmake", 2022. `https://cmake.org/`.

[101] "ARCHIVER Project." `https://archiver-project.eu/`.

[102] "ARCHIVER white paper draft:." `https://docs.google.com/document/d/1jc1J3ouLI1LtRKGQhVsRl3XoIrkFFF5bwhcX4UEZ-Q4/edit#`.

A The DPHEP Collaboration

DPHEP Collaboration

T. Basaglia[1], M. Bellis[I,2], J. Blomer[1], J. Boyd[1], C. Bozzi[3], D. Britzger[4], S. Campana[1], C. Cartaro[5], G. Chen[6], B. Couturier[1], G. David[II,7], C. Diaconu[8], A. Dobrin[9], D. Duellmann[1], M. Ebert[10], P. Elmer[11], J. Fernandes[1], L. Fields[21], P. Fokianos[1], G. Ganis[1], A. Geiser[12], M. Gheata[9], J. B. Gonzalez Lopez[1], T. Hara[13], L. Heinrich[1], M. Hildreth[21], K. Herner[14], B. Jayatilaka[14], M. Kado[1], O. Keeble[1], A. Kohls[1], K. Naim[1], C. Lange[20], K. Lassila-Perini[15], S. Levonian[12], M. Maggi[16], Z. Marshall[18], P. Mato Vila[1], A. Mečionis[1], A. Morris[17], S. Piano[16], M. Potekhin[7] M. Schröder[1], U. Schwickerath[1], E. Sexton-Kennedy[14], T. Šimko[1], T. Smith[1], D. South[12], A. Verbytskyi[4], M. Vidal[1], A. Vivace[1], L. Wang[6], G. Watt[19], T. Wenaus[7]

Affiliation Notes

[I] Also at: Siena College
[II] Also at: Stony Brook University

Collaboration Institutes

[1] CERN, Geneva, Switzerland
[2] Cornell University, USA
[3] INFN Ferrara, Italy
[4] Max-Planck-Institut für Physik, München, Germany
[5] SLAC National Accelerator Laboratory, USA
[6] Institute of High Energy Physics, IHEP, CAS, Bejing, China
[7] Brookhaven National Laboratory, BNL, USA
[8] Aix Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France
[9] Institute of Space Science, ISS, Bucharest, Magurele, Romania
[10] HEP Research Computing, University of Victoria, BC, Canada
[11] Princeton University, USA
[12] Deutsches Elektronen Synchrotron, DESY, Hamburg, Germany
[13] High Energy Accelerator Research Organization, KEK, Tsukuba, Japan
[14] Fermi National Accelerator Laboratory, Batavia, USA
[15] Helsinki Institute of Physics, Finland
[16] INFN Trieste, Italy
[17] University of Bonn, Germany
[18] Lawrence Berkeley National Laboratory, Berkeley, USA
[19] IPPP, Durham University, Durham, UK
[20] Paul Scherrer Institut, Villigen, Switzerland
[21] University of Notre Dame, Notre Dame, USA