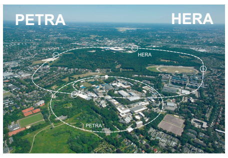
DPHEP Data Preservation in High Energy Physics

Roman Kogler, David South, Michael Steder

The Case for Data Preservation



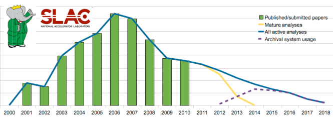
HEP data are unique

Particle physics experiments are designed to probe the structure of matter, the nature of fundamental interactions and ultimately to extend our understanding of nature. Since the advent of collider experiments the available range in energy and intensity has been enlarged by many orders of magnitude. However, the development, building and commissioning of colliders and the corresponding detectors takes considerable human, technological and financial effort. So far every collider and its associated scientific program have been unique in energy range, process dynamics or experimental techniques. Data collected from these experiments continue to be crucial to our understanding of particle physics, ranging from precision measurements to searches for new signatures beyond the Standard Model. The data preservation effort aims to ensure long-term availability of these data after the end of the experimental collaborations.

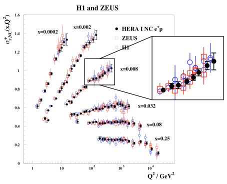
Data preservation can increase the physics potential of experiments

Long-term possibility for analysis

Precision analyses continue long after the data taking is finished, making use of the full statistical power and best knowledge of systematic uncertainties.

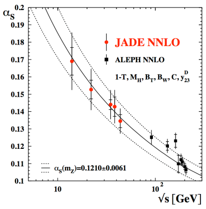


Combination of results among different experiments



Combining similar measurements from different experiments increases statistical significance and reduces systematic uncertainties via cross-calibration techniques to arrive at a more precise result.

Data re-use



Theoretical developments, new analysis techniques or the latest experimental observations may motivate the re-use of experimental data from previous HEP installations. A successful re-analysis of JADE data from the PETRA collider has lead to a precise determination of the strong coupling in an energy range that is still unique today.

Education, training and outreach

Accessibility of experimental data in a simplified format will allow students not originally part of an experimental collaboration to use these data for educational purposes. HEP data can help to introduce the general public to the field of particle physics and improve the public understanding of science.

What is HEP Data?

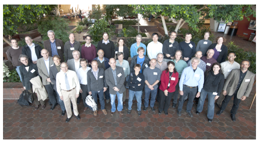
HEPDATA: REACTION DATA Database
Durham University
Journal of High Energy Physics
SPIRES
EPJ C

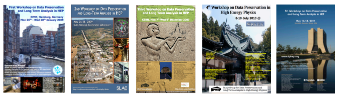
Raw collision data, DSTs, tape and disc storage devices, detector schematics and blueprints, detector simulation, MC generators, event reconstruction and analysis software, expert knowledge of collaboration members, experimental results, scientific publications, online databases and archives, internal digital and non-digital documentation, manuals, slides and notes,hypernews messages, …

DPHEP

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

DPHEP is a collaboration of experiments, laboratories and computing centres with about 100 contact persons. The study group, which is endorsed by ICFA, aims to review and document the physics objectives and technological aspects of data preservation in HEP in close cooperation with similar international initiatives in other fields.





Following a series of workshops the recommendations of the group for past, present and future facilities will be published soon. More information at **www.dphep.org**

Models of Preservation

Preservation Model	Use Case
1. Additional information	Publication related information
2. Provide data in simplified format	Outreach, training
3. Preserve the analysis level software and data format	Full scientific analysis possible, based on existing reconstruction
4. Preserve the full simulation and reconstruction software as well as the basic level data	Retain the full potential of the experimental data

Increasing complexity, benefit and cost

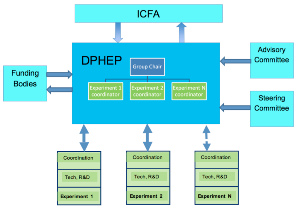
Governance

Tasks for a collaboration during its lifetime

- Supervision of the data preservation process
- Definition of the future collaboration structure
- Transition to the new operational model
- Establishment of authorship rules and supervision of physics output

Within the global HEP community

- The issue of open access to preserved HEP data
- Endorsement of preservation strategies from collaborations, laboratories and funding agencies
- Development of the International Data Preservation Forum, DPHEP



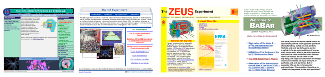
Future Documentation

Non-digital documentation

- Collection, cataloguing and central storage of relevant documents, e.g. minutes of meetings, presentations, …
- Digitisation of important sources, e.g. logbooks, blueprints, …

Digital documentation

- Former online monitoring and shift tools
- Web-based documentation, electronic logbooks, presentations in meetings, minutes, …



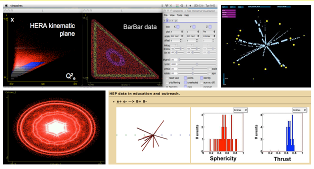
External services

- Documentation and publication related documents hosted by HEP services
- Additional possibilities to store internal information using e.g. INSPIRE

Outreach and Training

Complementary to existing HEP outreach projects, development of tutorials and exercises using real experimental data for educational purposes.

- Definition of a common, simplified data format
- Implementation of tools and user friendly interfaces
- Projects already begun within the BaBar and Belle collaborations, as well as joint projects within the DPHEP community



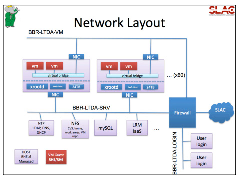
Technologies

The evolution of hardware, software and analysis models is a challenge for long-term data preservation. Any archival system should be able to cope with future technological changes. In the case of complex data preservation models (levels 3 and 4), the access to the data, MC generators as well as analysis level, reconstruction and simulation software needs to be maintained.

This task is made more complicated by the rapid development and limited hardware lifetime in the computing world. The appearance of new technologies such as CPU parallelism, the switch from grid to cloud computing, the advent of 64 bit processing as well as changing protocols for data access increase the complexity of the task in hand.

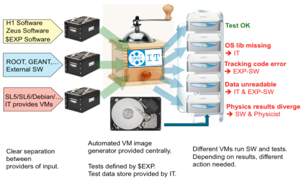
Data preservation strategies for data and experimental software

The supervision and custodianship of the data should be defined in order to preserve the data integrity and a working interface. Two strategies for software preservation are possible:



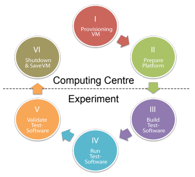
Freezing of software using virtualisation techniques, such as the approach taken by BaBar, needs only an initial investment where running software is provided but recompilation is not foreseen. However, this approach relies on the longevity of current protocols, the software environment and the virtualisation system employed.

A strategy of continuous migration, which is envisaged by the H1 collaboration, requires the long-term intervention of a data archivist. This is more involved than freezing the software, but by adjusting to future technological changes the lifetime of the software may be considerably extended.



Validation

Regular tests and validation are essential for successful software migrations. Generic solutions for validating experimental software are being developed at DESY, where a joint project involving the computing centre and the HERA experiments is in progress.



Virtual environments have proven to be useful for encapsulating validation systems with a clear separation of the experimental software and the working environment provided by the host lab. More information can be found in the talk by Yves Kemp at ACAT 2011, “A Validation System For Data Preservation in HEP”.

ZEUS hermes BELLE ALICE ATLAS CMS LHCb DESY KEK Jefferson Lab Institute of High Energy Physics Chinese Academy of Sciences Fermilab SLAC NATIONAL ACCELERATOR LABORATORY CERN