DPHEP – summary of recent developments

M.Potekhin (NPPS)

Brookhaven National Laboratory

DPHEP: a bit of history

- DPHEP stands for Data Preservation in High Energy Physics and is an international collaboration, operating under the aegis of ICFA and actively supported by CERN. It started as a study group in 2008 and produced a LoI in 2009. The LHC experiments joined at about this time.
- Its goal is to develop, in a coordinated manner and based on practical experience, a set of recommendations for the global HEP community, regarding the long term Data and Analysis Preservation.
- The original Collaboration Agreement was created in 2013. DPHEP was then convened as a panel of ICFA and has worked in close coordination with that body.

DPHEP reports

DPHEP has produced three major reports, containing the updated recommendations developed by the Collaboration. The latest was published in 2023:

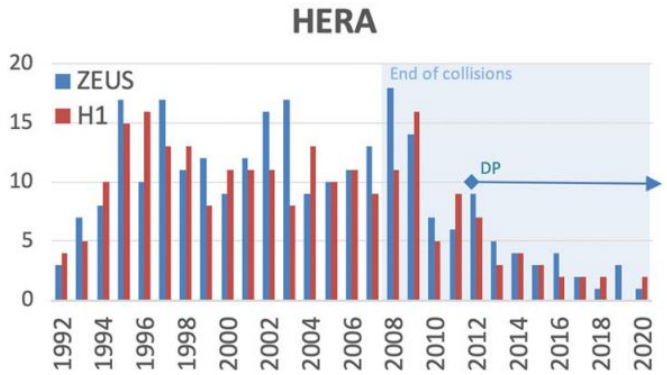
https://doi.org/10.1140/epjc/s10052-023-11885-1

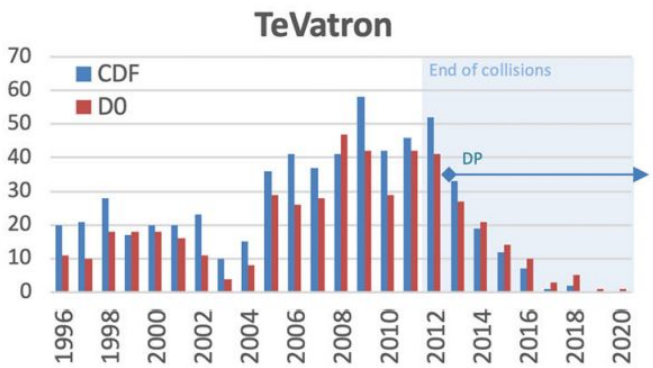
Preservation (according to C.Diaconu):
the process of transforming a "high intensity/rapidly changing" computing system into a "low intensity/slowly evolving" computing system with conserving the capacity of extracting new science from the "data"

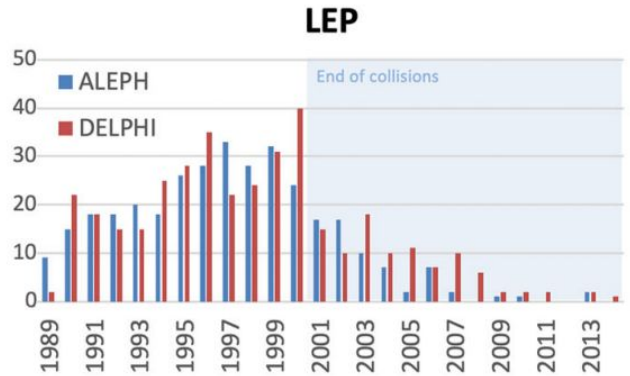
Also, see the recent presentation at CHEP
https://indico.cern.ch/event/1338689/contributions/6011102/attachments/2952294/5189998/DPHEP-CHEP-OCT2024.pdf

1 Overview . . . 2
2 What is data preservation? . . . 5
3 Methodologies of HEP data preservation . . . 6
3.1 Data preservation models and frameworks . . . 6
3.1.1 Preservation levels . . . 6
3.1.2 Frameworks . . . 7
3.2 Supervision models . . . 7
3.3 Preservation and openness . . . 9
3.4 Funding and valuing data preservation . . . 9
4 Experiment reports . . . 10
4.1 JADE . . . 10
4.2 ALEPH, DELPHI, L3, OPAL . . . 12
4.2.1 ALEPH . . . 13
4.2.2 DELPHI . . . 13
4.2.3 L3 . . . 14
4.2.4 OPAL . . . 14
4.2.5 LEP data and Key4hep . . . 15
4.3 H1 and ZEUS . . . 15
4.3.1 H1 . . . 15
4.3.2 ZEUS . . . 17
4.4 *BABAR* . . . 18
4.5 Belle I and II . . . 19
4.6 CDF and D0 . . . 20
4.7 PHENIX . . . 20
4.8 MINERvA . . . 21
4.9 BES III . . . 21
4.10 ATLAS, CMS, LHCb, ALICE . . . 22
4.10.1 ATLAS . . . 22
4.10.2 CMS . . . 24
4.10.3 LHCb . . . 25
4.10.4 ALICE . . . 27
5 DP technologies and projects . . . 27
5.1 HEPData . . . 28
5.2 CERN open data portal . . . 28
5.3 CERN analysis preservation . . . 28
5.4 REANA reproducible analyses . . . 30
5.5 ARCHIVER . . . 31
5.6 Bit preservation at CERN . . . 32
5.7 CERNVM . . . 32
5.8 CERNLIB . . . 33
6 DPHEP: the way forward . . . 34
A Glossary . . . 35
References . . . 38

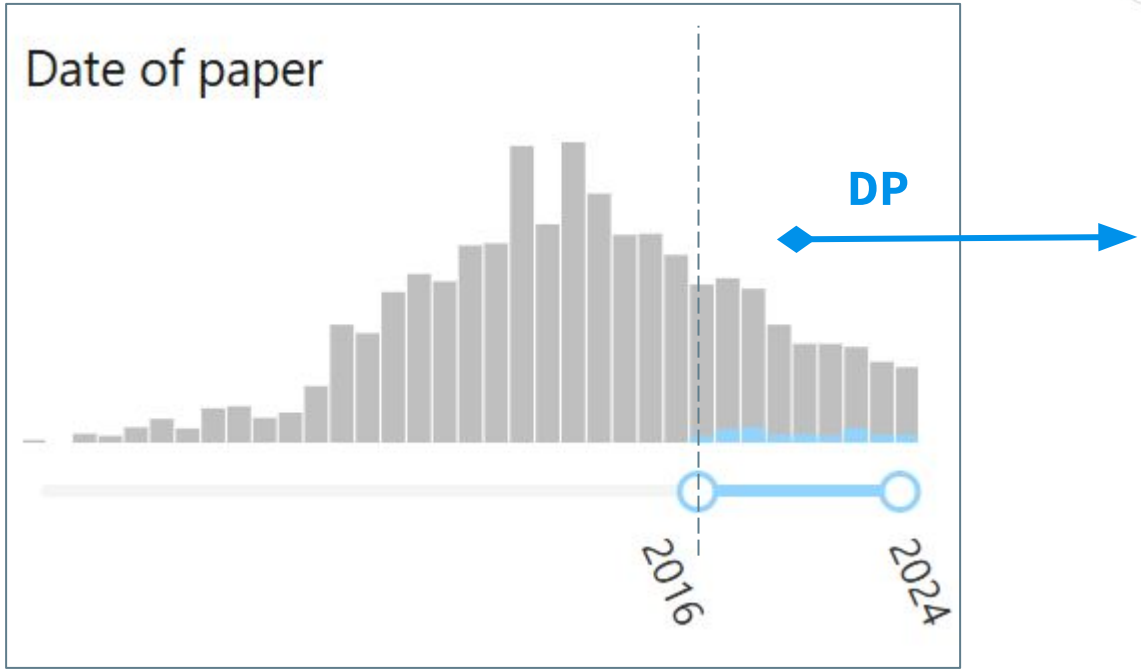
Science output after “end of collisions”







PHENIX output after “end of collisions”



Science output after “end of collisions” – ROI

- Substantial output is possible with only incremental extra funding of Data Preservation
- Typical quote: 10% return on 0.1% investment

DPHEP impressions

- New interest in preserved data due to the progression of the experimental programs e.g. HERA → EIC, LEP → FCC
- Successful “rescue” of a few legacy software components making it possible to continue maintenance of data such as from LEP
  - CERNLIB
  - PHIGS (dependency of Delphi and Opal software)
- Moving legacy data from proprietary formats to modern formats
- Emergence of access mode to the data, by request and in some cases in custom formats/“schemas”
- Continued prominence of the Open Data portal and more detailed plans for releasing data on that platform, by the LHC experiments

CERNLIB

- CERNLIB “has been rescued” (was an area of concern a few years ago, and fortunately has been addressed)
  - Prior to that, the last supported version came out in 2006
  - 32-bit – this had to be ported
  - Migration of the code base to GitLab
  - Modern build scheme based on CMake + CI
- Tight coordination, review of licenses etc were important in this work (e.g. Lund MC)
- Still technical caveats (e.g. related to 32-bit Zebra), pending retirement of X11
- Summary – a complex and labor intensive effort with a good chance of success

ALEPH

- CERNLIB dependency, resolved now
- Can recreate the old SL-based environments in containers running under Alma 9
- ALEPH data port – specifically mini-DSTs – to EDM4hep – via an intermediate text format
- In the process, implicit dependencies on numeric constants/corrections in the “ALPHA” code (analysis) are being eliminated
- Validation – work in progress

HERA

- Both Zeus and H1 continue to publish physics papers, years after data taking had ended – new collaborators still joining!
- H1
  - Projection: 4 new papers total in 2024
  - Continue to support Fortran and Geant3 dependencies
  - ROOT access to legacy data
  - Transition to SL6 in 2017, pending migration to SL9
- Zeus
  - One of the core and early participants of DPHEP
  - Starting in 2006, data is being migrated to a “common simplified ROOT format” – can be considered future-proof
  - MC component of the software “frozen” in containers
  - Lack of the available effort in the long term seen as the biggest problem

LHCb

- Continued work as described in the previous DPHEP meetings
- What makes it different: a “NTupling” service designed to solve the problem of potentially large data volumes to be delivered to the end user, making the long-term storage of certain types of derived data unfeasible.
- Solution – create intermediate data products on demand.
- The resulting NTuples can be customized. NB. in the real life LHCb workflow such data if effectively ephemeral, but here it is captured for the users’ benefit.

PHENIX DAP

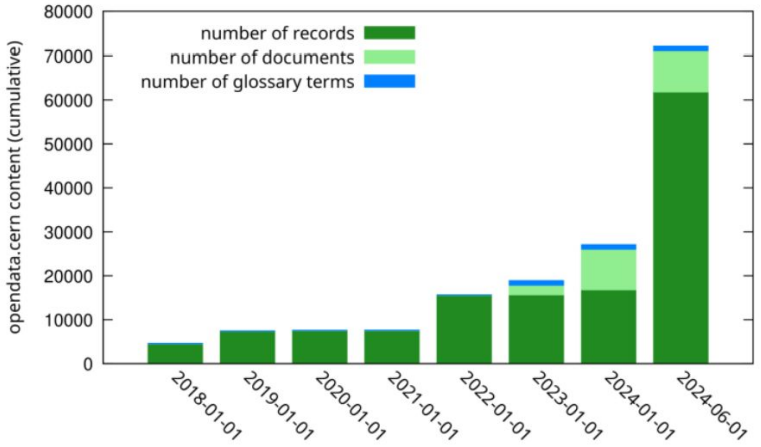
- PHENIX has started its Data and Analysis Preservation (DAP) effort in 2019, and periodic status updates have been presented in the previous DPHEP meetings, PV2023 and two ACAT conferences (2021 and 2024). Useful links:
  - https://indico.cern.ch/event/1043155/timetable/#9-bnl-rhic
  - https://iopscience.iop.org/article/10.1088/1742-6596/2438/1/012020
  - https://doi.org/10.5281/zenodo.7905555
- More than 25% of published papers (66/243) were produced after the end of data taking in 2016, and work is ongoing (although only a small fraction is formally DAP)
- A rare example of a full analysis preservation in REANA (next slide)
- A very successful HEPData effort

PHENIX: direct γ and π^0 analysis

- The main motivation for this study was an initially puzzling result regarding the nuclear modification factor in peripheral d+Au collisions. The background is explained in a PHENIX note: https://doi.org/10.5281/zenodo.8169171
- Ultimately it was determined that the key to understanding this result is to use the correct technique of estimating the centrality of collisions, making use of the Electromagnetic Calorimeter data. This has substantial scientific importance, and hence was chosen for preservation. REANA is used as the driver.
- This analysis featured a well documented flowchart and other crucial documentation which made this effort possible.
- SDCC provided containerization support

Open Data: http://opendata.cern/

- Repository of HEP data:
  - > 5 PT
  - > 70k entries
  - > 2.5 million files
  - 6 HEP experiments

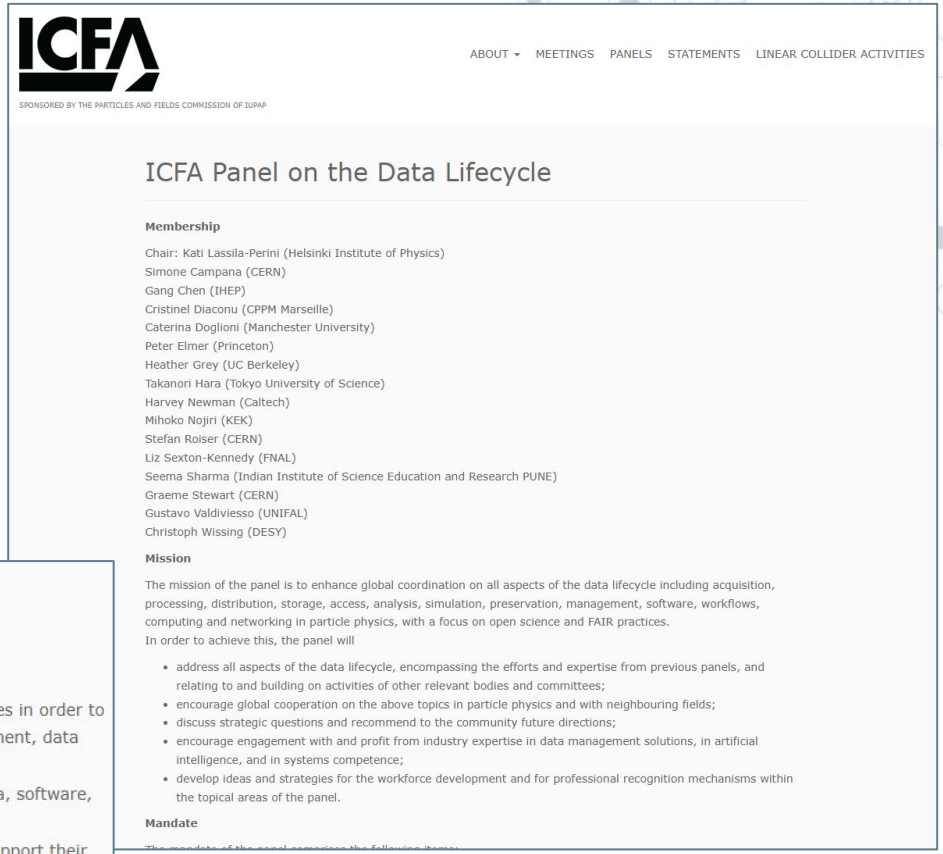


Experiment	Quota (TB)	Used (TB)	%	
ALICE	50	8	15%	
ATLAS	400	165	41%	
CMS	5150	4470	87%	
DELPHI	50	38	76%	new
JADE	1	0.65	64%	new
LHCb	910	800	89%	
OPERA	0.005	0.004	77%	
PHENIX	0.05	0.00008	0%	

New: data can be requested from cold storage – work in progress

The ICFA Panel

DPHEP will continue work under the renewed ICFA umbrella



ICFA

SPONSORED BY THE PARTICLES AND FIELDS COMMISSION OF IUPAP

ABOUT ▾ MEETINGS PANELS STATEMENTS LINEAR COLLIDER ACTIVITIES

ICFA Panel on the Data Lifecycle

Membership

Chair: Kati Lassila-Perini (Helsinki Institute of Physics)
Simone Campana (CERN)
Gang Chen (IHEP)
Cristinel Diaconu (CPPM Marseille)
Caterina Doglioni (Manchester University)
Peter Elmer (Princeton)
Heather Grey (UC Berkeley)
Takanori Hara (Tokyo University of Science)
Harvey Newman (Caltech)
Mihoko Nojiri (KEK)
Stefan Roiser (CERN)
Liz Sexton-Kennedy (FNAL)
Seema Sharma (Indian Institute of Science Education and Research PUNE)
Graeme Stewart (CERN)
Gustavo Valdiviesso (UNIFAL)
Christoph Wissing (DESY)

Mission

The mission of the panel is to enhance global coordination on all aspects of the data lifecycle including acquisition, processing, distribution, storage, access, analysis, simulation, preservation, management, software, workflows, computing and networking in particle physics, with a focus on open science and FAIR practices.
In order to achieve this, the panel will

- address all aspects of the data lifecycle, encompassing the efforts and expertise from previous panels, and relating to and building on activities of other relevant bodies and committees;
- encourage global cooperation on the above topics in particle physics and with neighbouring fields;
- discuss strategic questions and recommend to the community future directions;
- encourage engagement with and profit from industry expertise in data management solutions, in artificial intelligence, and in systems competence;
- develop ideas and strategies for the workforce development and for professional recognition mechanisms within the topical areas of the panel.

Mandate

Mandate

The mandate of the panel comprises the following items:

- Address the data lifecycle within a structured and integrated systems approach in HEP.
- Improve the awareness for the importance of the data lifecycle in HEP.
- Encourage and foster connections to other fields of science, to industry and to open science initiatives in order to profit from their expertise and competence in the following fields: big and distributed data management, data management systems, artificial intelligence, open science processes, data preservation systems.
- Help in organising practical support and act as point of contact for practical issues in the field of data, software, workflows and computing.
- Improve recognition of the nature and value of work on the data lifecycle in researchers' CVs and support their career development.

The full mandate of the panel can be found here.

Observations

- A successful Data and Analysis Preservation effort appears to always consists of two complementary — and distinct — parts:
  - Facility and Infrastructure
  - Expertise and engagement on the part of active participants of the experiment
- Both are only possible with appropriate allocation of resources
- The “Open Data” portal at CERN is one of the central DAP tools for HEP experiments e.g. LHC. We do not have an equivalent in the US. The potential to benefit NP experiments is substantial.

Summary

- The DPHEP Collaboration has been instrumental is creating guidance for Analysis Preservation effort for legacy experiments, and coordination of the actual software development necessary for that.
- This is a solid starting point for upcoming experiments in HEP/NP.
- The newly reformatted ICFA panel will continue effort in this direction.
- Experience shows that for this type of work to succeed, support and funding need to be made available at the institutional level.

Backup

Direct γ and π^0 analysis in PHENIX: the flowchart