



Data Preservation in High Energy Physics: a decade report



IRN Terascale @ Marseille, Luminy

25–27 oct. 2023
Luminy

Cristinel DIACONU
CPPM/CNRS-IN2P3/Aix-Marseille University

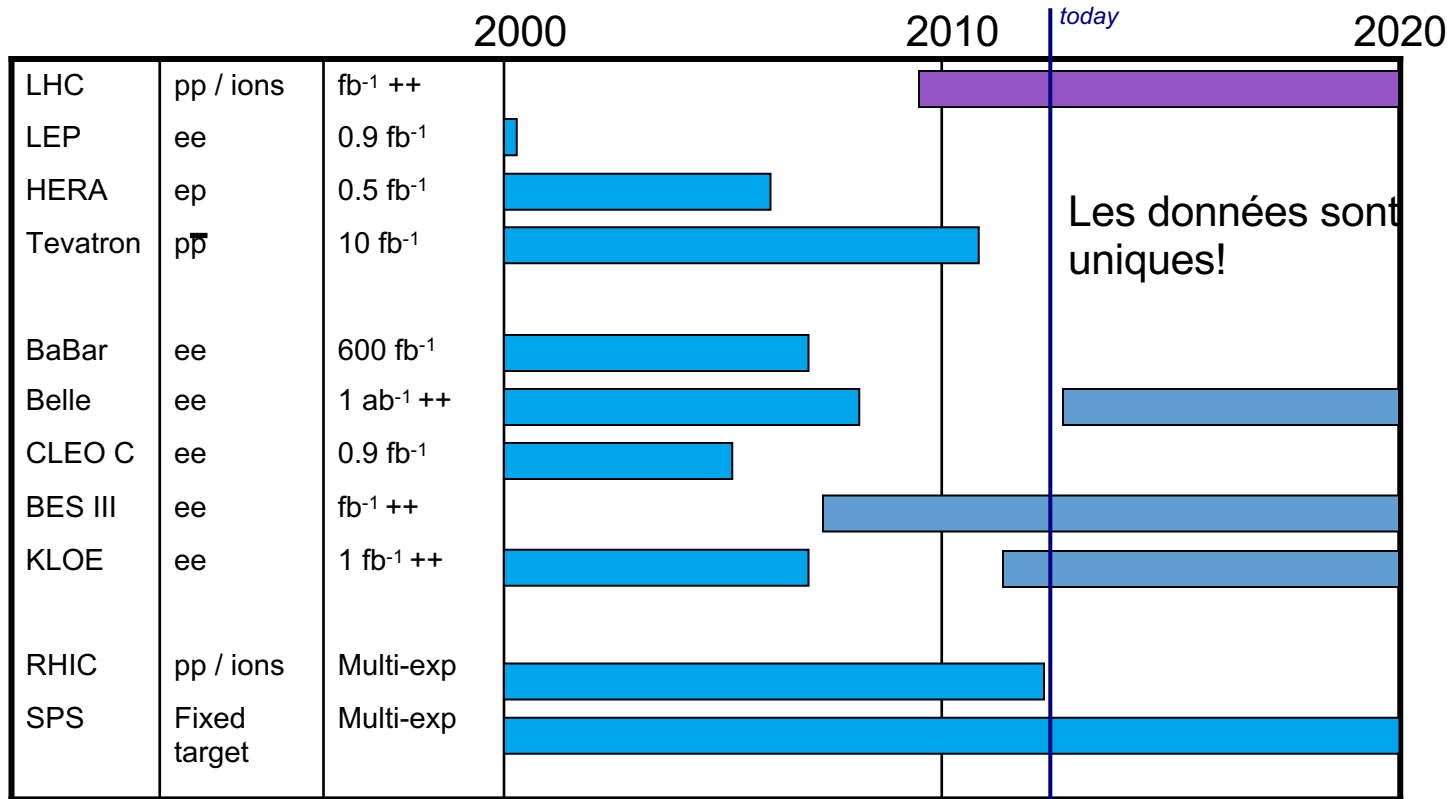


Study Group for Data Preservation and
Long Term Analysis in High Energy Physics



<http://dphep.org>

In 2013: HEP experiments in ± 10 ans : a change of phase was ongoing



[not all programmes, dates are approximate, just to give the picture]

The DPHEP Collaboration

- Collaboration Agreement was signed in 2014
 - Give a clear sign of the will of labs to collaborate in this common challenge
- Members:
 - 2014: CERN, DESY, HIP, IHEP, IN2P3, KEK, MPP
 - 2015 IPP/Canada , 2017 UK/STFC
 - Active labs from US, Italy
 - have not formally joined, but are represented in the Collaboration Board.
- The DPHEP collaboration continue to act as an ICFA panel, as indicated in the Collaboration Agreement
 - About 60 contact persons FA, Labs, experiments
 - Mandate prolonged to 2024



Collaboration Agreement for the DPHEP Project

BETWEEN:

The Partners of the DPHEP Project (the "Partners") set out in Annex 1 to the Collaboration Agreement,

CONSIDERING THAT:

(1) Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique;

(2) The Data Preservation and Long Term Analysis in High Energy Physics (DPHEP) project (the "Project"), an inter-experimental study group on HEP data preservation and long-term analysis, was initially formed by large collider-based experiments to investigate the technical and organizational aspects of HEP data preservation and convened by a Chair and a Project Manager as a panel of the International Committee for Future Accelerators (ICFA). Two reports were released, providing an analysis of the research case for data preservation and a detailed description of the various projects at experiment, laboratory and international levels;

(3) In its report of May 2012 (see Annex 2), the study group provided a concrete proposal for an international collaboration in charge of the Project and data management and policies in high-energy physics;

(4) The Partners have expressed their interest to take part in and contribute to the Project in order to implement the recommendations provided in the report referred to in Annex 2 and wish to formalize their collaboration through the present Collaboration Agreement;

(5) The mutual benefit of the Partners that shall result from collaboration between them;


HAVE AGREED AS FOLLOWS:

Organizational structure and decision mechanism

The organizational structure of the Project shall include the following entities:

- 1) International Advisory Committee (IAC)
- 2) Collaboration Board (CB)
- 3) Implementation Board (IB)
- 4) Project Manager
- 5) Chairperson

DPHEP Collaboration/ICFA Panel

DPHEP Data Preservation in High Energy Physics
@DPHEPColl

<http://dphep.org>

2009

Lol

[arXiv:0912.0255](https://arxiv.org/abs/0912.0255)

EPHEP 2009-001
November 16, 2009

Data Preservation in High-Energy Physics

<http://dphep.org>

Abstract

Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. An data management and group as HEP data preservation and long-term analysis was established as a part of the International Committee for Future Accelerators (ICFA). The group was formed by large collider-based experiments and investigated the technical and organizational aspects of HEP data preservation. An anniversary report was released in November 2009 addressing the general issue of data preservation in HEP. The report includes and reaches the intermediate report. It provides an analysis of the research issue for data preservation and a detailed description of the various aspects at experimental, laboratory and international levels. In addition, the paper provides a concrete proposal for an international organization in charge of the data management and policies in high-energy physics.

2012

Blueprint

[arXiv:1205.4667](https://arxiv.org/abs/1205.4667)

DPHEP 2012-001
May 2012

Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics

[www.dphep.org](http://dphep.org)

Abstract

Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. An data management and group as HEP data preservation and long-term analysis was established as a part of the International Committee for Future Accelerators (ICFA). The group was formed by large collider-based experiments and investigated the technical and organizational aspects of HEP data preservation. An anniversary report was released in November 2009 addressing the general issue of data preservation in HEP. The report includes and reaches the intermediate report. It provides an analysis of the research issue for data preservation and a detailed description of the various aspects at experimental, laboratory and international levels. In addition, the paper provides a concrete proposal for an international organization in charge of the data management and policies in high-energy physics.

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

2015

Collaboration MoU

[arXiv: 1512.02019](https://arxiv.org/abs/1512.02019)

https://doi.org/10.5781/trends.46158

DPHEP 2015-001
December 2015

Status Report of the DPHEP Collaboration: A Global Effort for Sustainable Data Preservation in High Energy Physics

www.dphep.org

Abstract

Data from High Energy Physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. An data management and group as HEP data preservation and long-term analysis was established as a part of the International Committee for Future Accelerators (ICFA). The group was formed by large collider-based experiments and investigated the technical and organizational aspects of HEP data preservation. An anniversary report was released in November 2009 addressing the general issue of data preservation in HEP and an extended blueprint paper was published in 2012. In July 2014 the DPHEP collaboration was formed as a result of the signature of the Collaboration Agreement by seven large funding agencies (which have since joined to ten as the process of acquisition) and in June 2015 the first DPHEP Collaboration Workshop and Collaboration Board meeting took place.

This status report of the DPHEP collaboration details the progress during the period 2013 – 2015 activities.

International Collaboration for Data Preservation and Long-Term Analysis in High Energy Physics

2023

Decade report

[arXiv: 2302.03583](https://arxiv.org/abs/2302.03583)

DPHEP 2023-001
February 2023

**Data Preservation in High Energy Physics
DPHEP Global Report 2022**

DPHEP Collaboration

Abstract

The document summarizes the status of data preservation in high energy physics. The paragraph and the paragraph below are derived from a perspective of more than ten years of experience with a sustained effort in international level. The status and the results are related to the preservation of data recorded at large-scale experiments and compared together with an ongoing changing effort in recent high energy physics experiments in smaller and smaller experiments. Various perspectives on data preservation in which are carefully treated in text version and PPTI pictures, are presented as well. It is prepared and an early phase of the document.

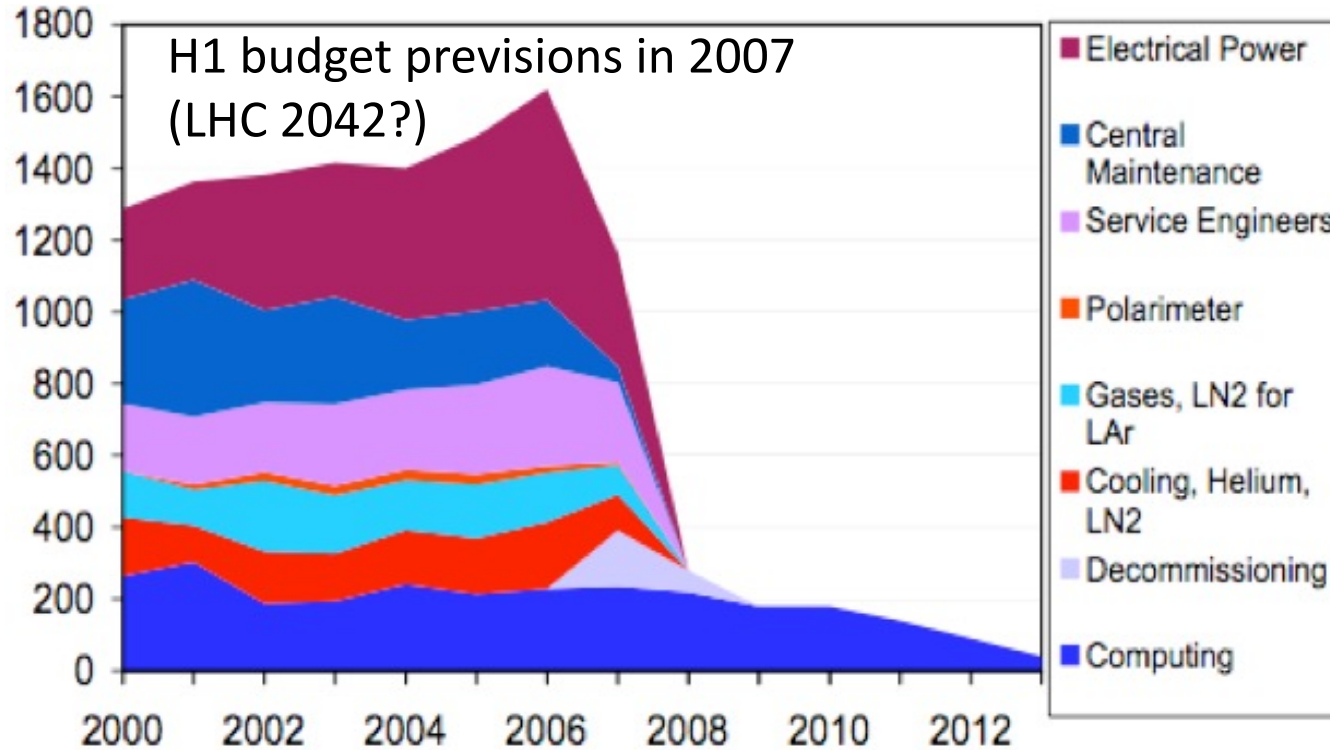
Submitted to Eur. Phys. J. C

[arXiv:submit/4724477 \[hep-ex\] 7 Feb 2023](https://arxiv.org/abs/2302.03583)

The copyright for the text of publications remains the property of the author or authors. If it is allowed as specified in the CC-BY-NC-ND license.

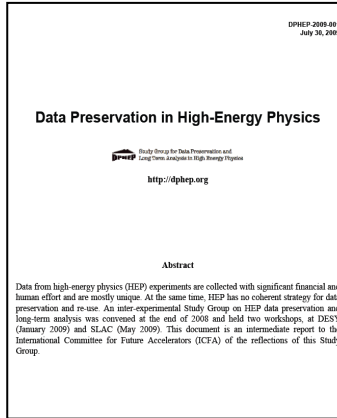
Eur. Phys. J. C 83, 795 (2023)

When it stops taking data



DPHEP Study Group (2009)

> [arXiv:0912.0255](https://arxiv.org/abs/0912.0255)



- An urgent and vigorous action is needed to ensure data preservation in HEP
 - Examples for the physics case explored
 - Data is rich and can be further exploited in most cases beyond the collaboration lifetime
- The preservation of the full analysis capability of experiments is recommended, including the preservation of reconstruction and simulation software
- An interface to the experiment know-how should be introduced: **data archivist** position in the computing centres
- The preservation of HEP data requires a **synergic action**: collaborations, laboratories and funding agencies
- An International Data Preservation Forum is proposed as a **reference organisation**. The Forum should represent experimental collaborations, laboratories and computing centres

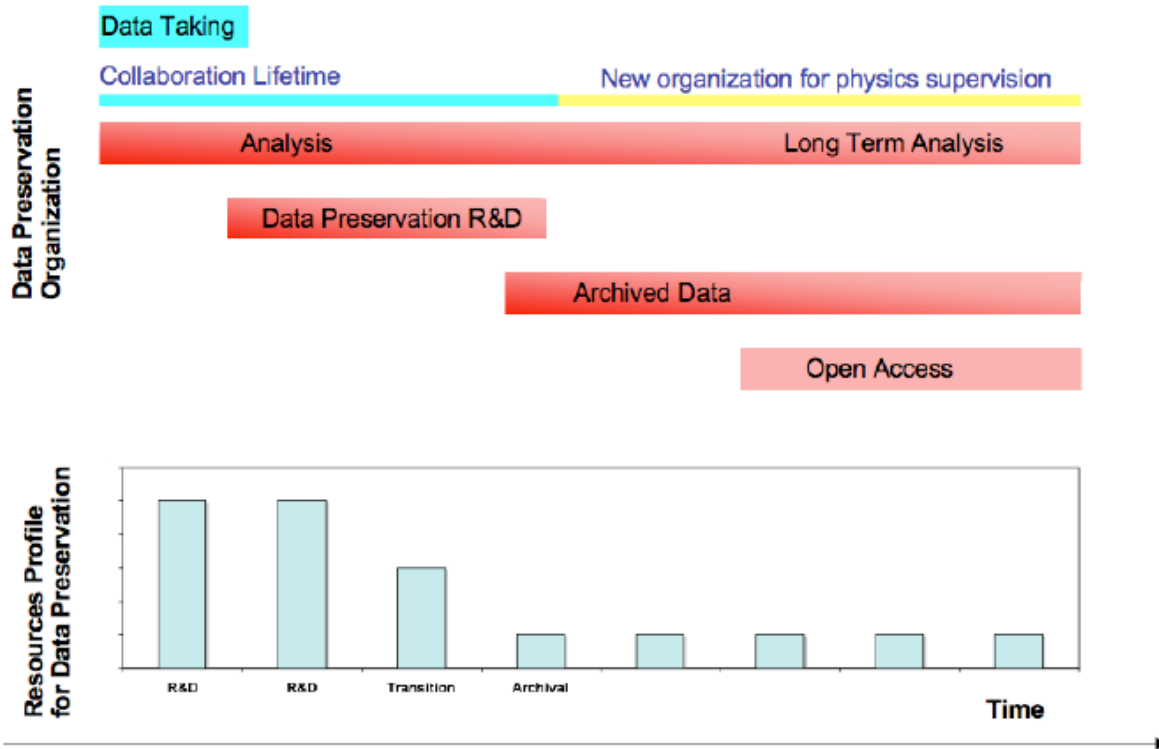


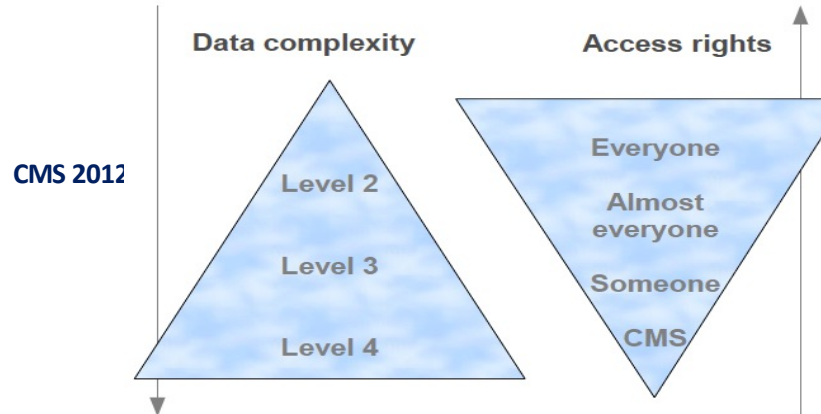
Figure 1: A possible model for data preservation organisation and resources presented as the milestones of the organisation and the resources evolution as a function of time.

Data Preservation in High Energy Physics

- What is “data”?
 - not (only) : “files”
 - but : “every digitally encoded information that was created as a result of planning, running and exploiting an experiment”
- What is “preservation”?
 - not: a freezer, a herbarium, a museum, an album, a cellar....
 - but: 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".
 - Requires clear plans and a long term organization
 - Within each collaboration and at international level (DPHEP)

Guidance into data complexity

Preservation Model		Use Case	
1	Provide additional documentation	Publication related info search	Documentation
2	Preserve the data in a simplified format	Outreach, simple analyses	Outreach, reanalysis
3	Preserve the analysis level software and data format	Full scientific analysis, based on the existing reconstruction	Technical Preservation Projects
4	Preserve the reconstruction and simulation software as well as the basic level data	Retain the full potential of the experimental data	



A matter of collaboration as well

- The supervision and knowledge transfer/capture is essential at long term
- Need to clarify the status and the rules
- Various stages of organisation can be defined:
 - 0: Organisation during experiment proposal.
 - 1: Organisation during data taking.
 - 2: Organisation after data taking
 - 3: Organisation after the collaboration funding scheme.
 - 4: Rescue organisational scheme. This organisation scheme is to be activated when:
 - the host laboratory stops support and announce no long-term commitment.
 - the official collaboration/data stewardship is stopped with no further plans (no step 3 is clearly defined).

Remarks:

- Taking no action necessarily implies decommissioning (deleting) the data.
 - “Securely” storing/freezing the files and the latest version of the software is certainly not a substitute for a preservation project.

Costs and Benefits

C1. Host laboratories allocate person power and computing resources.

in % to the construction/operation costs

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.

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.

- $\text{FoM} = \text{B1}/\text{C1}$

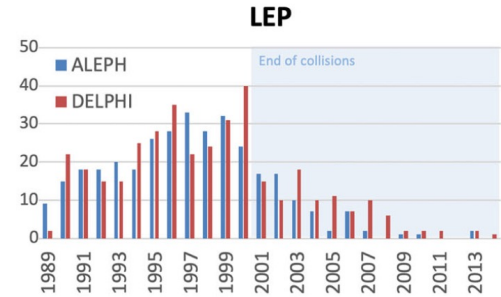
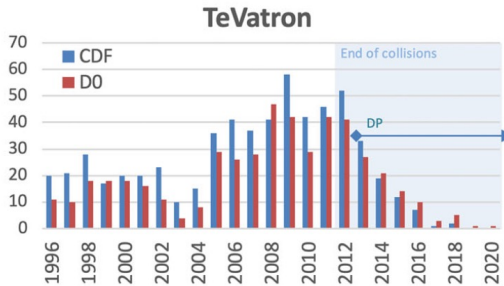
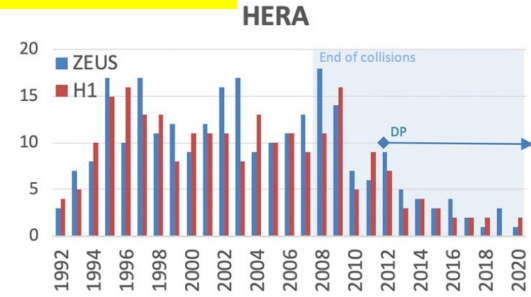
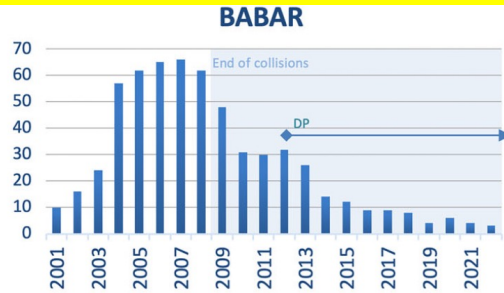
2023: Experiments Data Preservation Status

Laboratory/ Collider	Experiment	Data taking period	Preservation Level	Data Volume	Present status	Coll
DESY/PETRA	JADE	1979–1986	4	1 TB	Analysis running on preserved data; migrated from DESY to MPP	4
CERN/LEP	ALEPH, DELPHI, L3, OPAL	1989-2000	4	0.5 PB	Analysis running on preserved data	4
DESY/HERA	H1 ZEUS	1992 – 2007	4 3/ 4	0.5 PB 0.2 PB	Analysis running on preserved data	3
SLAC/PEP II	BABAR	1999–2008	4	2 PB	Analysis running on preserved data; migrated from home lab to different centers	4
KEK/KEKB	Belle I	1999-2010	4	4 PB	Analysis running on preserved data; Compatible with Belle II computing	2
FNAL/TeVatron	DØ CDF	1983–2011	4 4	8.5 PB 9 PB	Archived on tapes	4
BNL/RHIC	PHENIX	2000–2016	3	25 PB	Analysis running on preserved data	3
FNAL/v-beam	Minerva	2010–2019	3	10 TB	Analysis running	2
IHEP/BEPCII	BESIII	2009–2030	4	6 PB	Collecting and analyzing data	1
CERN/LHC	ALICE, ATLAS, CMS, LHCb	2010-2040	4	O(1EB)	Collecting and analyzing data	1

Conclusions after 10 years: the scientific output

DP is a **cost-effective way of doing fundamental research** by exploiting unique data sets in the light of the increasing theoretical understanding.

- DP leads to
 - a **significant increase in the scientific output** (10% typically)
 - for a minimal investment overhead (0.1%).
 - As predicted in 2013

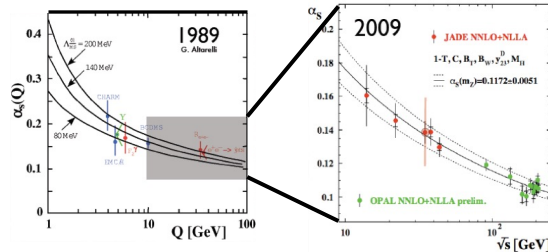


	Data taking stopped	Publications before 2012	Publications after 2012	Scientific return increase %
Babar	2008	471	154	33%
H1+ZEUS	2007	436	62	14%

JADE

- JADE DP stack is based on open standards, does not rely on specific SW and is extremely portable. One can run it completely on desktop.
- **“JADE – collider experiment on your desktop”.**

Data Preservation
model *circa* 1980-ies



The JADE Experiment at the PETRA e^+e^- collider -- history, achievements and revival

S. Bethke (Munich, Max Planck Inst.), A. Wagner (DESY)
Aug 23, 2022

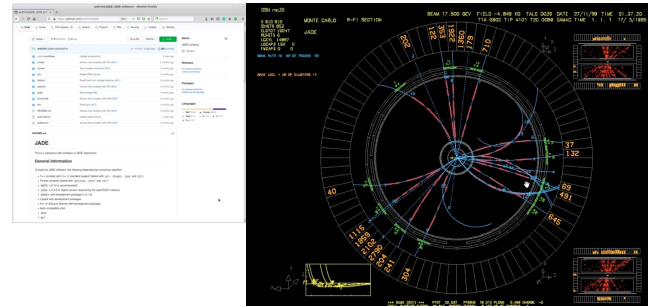
58 pages
Published in: *Eur.Phys.J.H* 47 (2022) 16
e-Print: [2208.11076](https://arxiv.org/abs/2208.11076) [hep-ex]

2021

JADE software: recent developments

More portability, testing and documentation.

- GNU and IBM toolchains support extended with preliminary Intel^{NEW} and NAG^{NEW}. GNU is still the most stable one.
- More CI tests^{NEW}.
- Updated the site and documentation^{NEW}.
- Support for CentOS^{NEW} and MacOSX10.15+ on x86_64^{NEW}



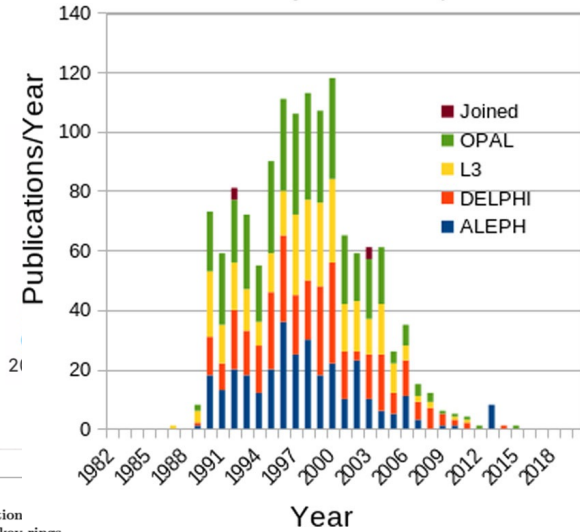
LEP

Papers using archived data

LTDP @LEP: Big Data Today - Peanuts Tomorrow

New physics with Archeodata

Publications by the LEP experiments

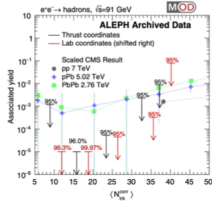


PHYSICAL REVIEW LETTERS 123, 212002 (2019)

2019

Measurements of Two-Particle Correlations in e^+e^- Collisions at 91 GeV with ALEPH Archived Data

Anthony Badea,¹ Austin Baty¹, Paoti Chang,² Gian Michele Innocenti,¹ Marcello Maggi,³ Christopher McGinn,¹ Michael Peters,¹ Tzu-An Sheng,² Jesse Thaler¹, and Yen-Jie Lee^{1,2}
¹Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
²National Taiwan University, Taipei 10617, Taiwan
³INFN Sezione di Bari, Bari, Italy



Physical interpretation of the anomalous Cherenkov rings observed with the DELPHI detector

V. F. Perpelitsa
 ITEP, Moscow
 T. Ekelof
 Department of Physics and Astronomy, Uppsala University
 A. Ferrer
 IFIC, Valencia University
 B. R. French
 brfrench@thaleswin.ch

On long-range pionic Bose-Einstein correlations – Including analyses of OPAL, L3 and CMS BECs –

Takuya Mizoguchi¹ and Minoru Biyajima²

¹National Institute of Technology, Toba College, Toba 517-8501, Japan
²Department of Physics, Shinshu University, Matsumoto 390-8621, Japan

February 23, 2021

2020

2021

→ FCCee

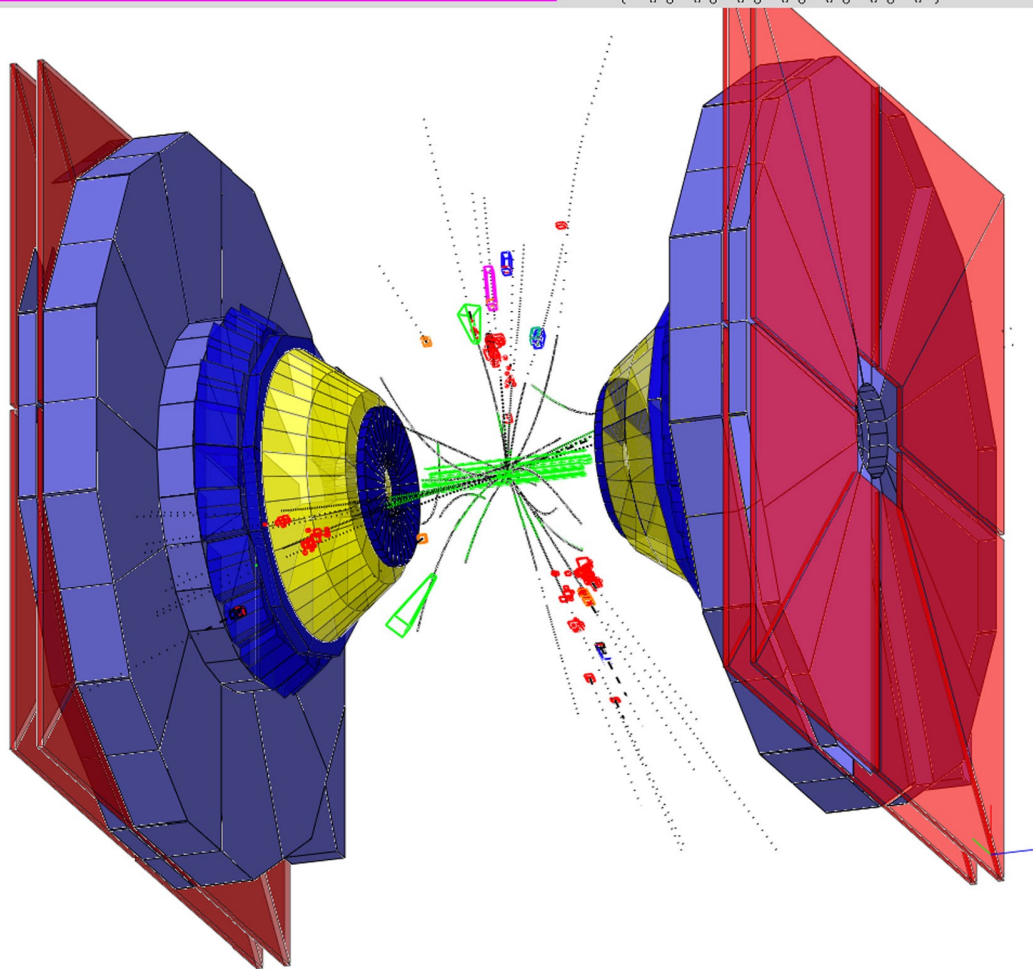
→ With real data from LEP



DELPHI Run: 109187 Evt: 3066
Beam: 103.0 GeV Proc: 20-Jul-2022
DAS: 23-Apr-2000 Scan: 26-Apr-2023
09:44:08 Tan+DST

	TD	TE	TS	TK	TV	ST	PA
Act	0	288	0	48	0	0	0
	(202	1288	0	0	48	0	0
Deact	0	0	0	0	0	0	0
	(0	0	0	0	0	0

DELPHI event display
with revised software



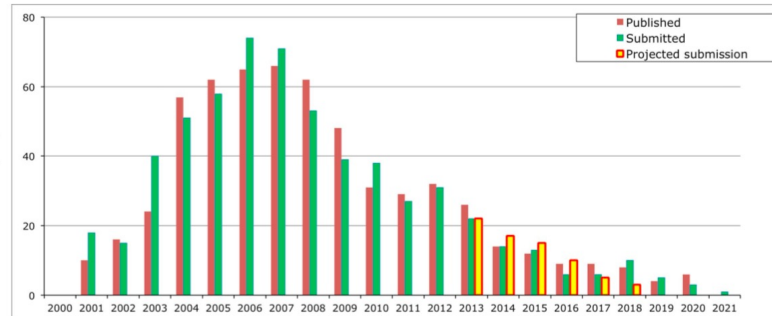
Babar (03/2021)

T. Cartaro

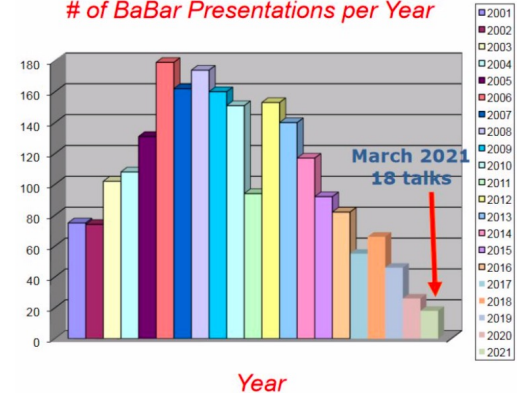


Publications

- 595 papers published or submitted
 - 9 papers published in 2017, 8 in 2018, 4 in 2019, 6 in 2020
 - 3 in the pipeline so far in 2021, few more expected later in 2021
- ~15 analyses active and on track for publication
 - Some are progressing slowly
 - 6 new analyses started last year and expect some more this year
- 25 talks in 2021
 - 7 talks at EPS-HEP, and more already assigned
 - 26 talks given in 2020 (17 cancelled due to COVID-19)
 - Often shared talks (and collaborative analyses) with Belle
- Quality of physics results still excellent



of BaBar Presentations per Year



But: SLAC LTDA decommissioned, moving to U. Victoria/CERN/CC-IN2P3/Grid-Ka
Open Data decided

Update of the BaBar publication skyline

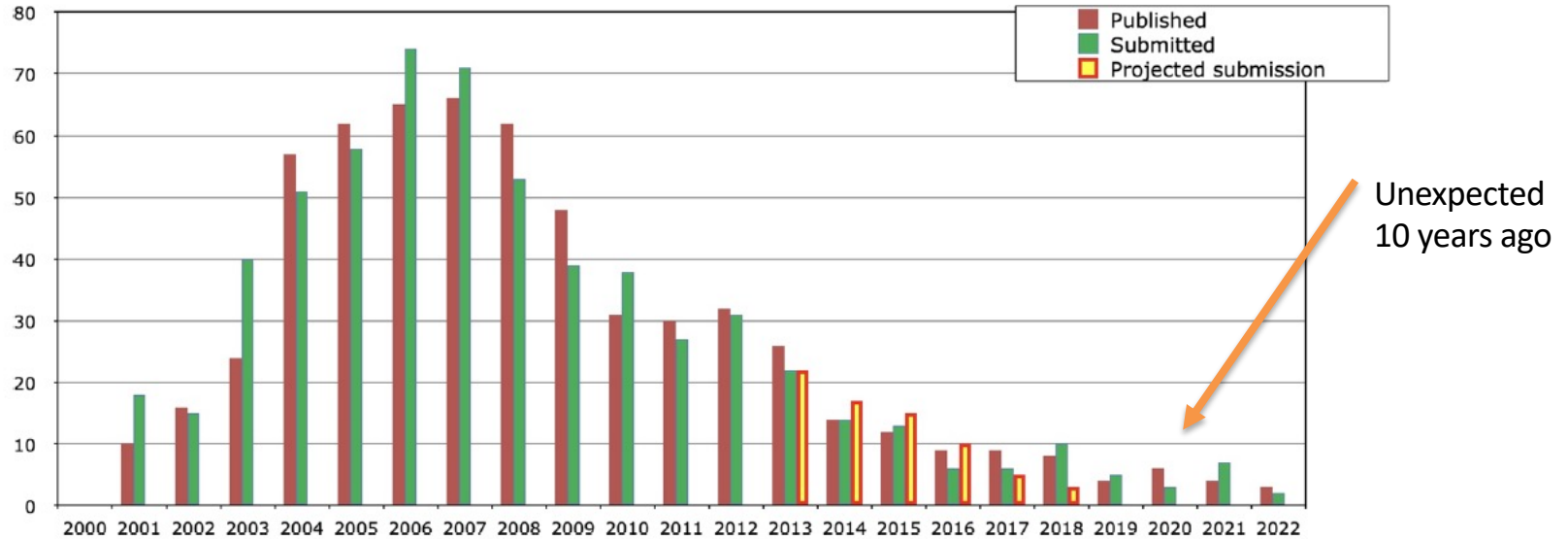


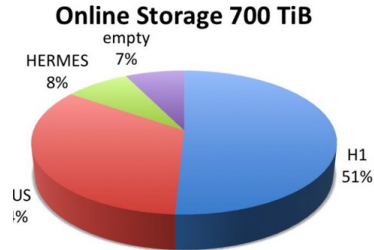
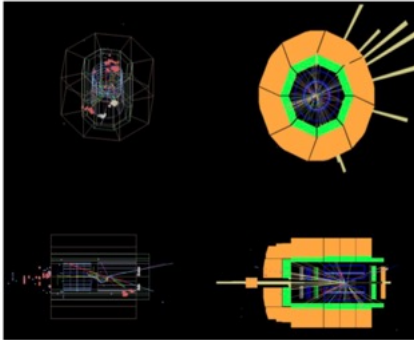
Fig. 6 BABAR submitted (green) and published (red) papers per year. In 2012 predictions for submissions (yellow) were made for the years 2013 to 2018. In 2012 it was predicted that no analysis would run after 2018

HERA: succesful DP, towards open data

- H1: “Level 4” DPHEP strategy
 - All data, full migration, including regular recompilation/validation
 - Recent “technology jump” succesfull : in line with modern tools
 - “LHC”-like tools, ready for opendata

- ZEUS : “Level 3/4” DPHEP strategy
 - Root ntuples produced in the preparatory phase
 - easy to maintain/use/test/open

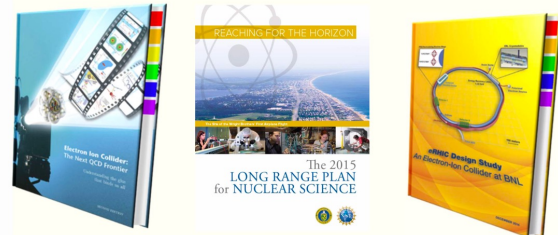
‘H1Red’ for simulated Pythia8.3 event



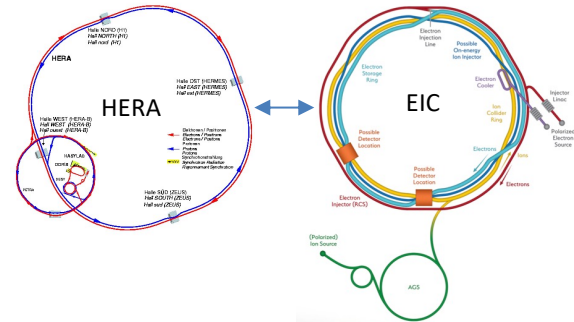
- New topics/collaborators (EIC)

Synergy with future experiment: EIC

- many EIC topics common with HERA



- some EIC members have recently joined ZEUS to work on common analysis topics with real ZEUS data



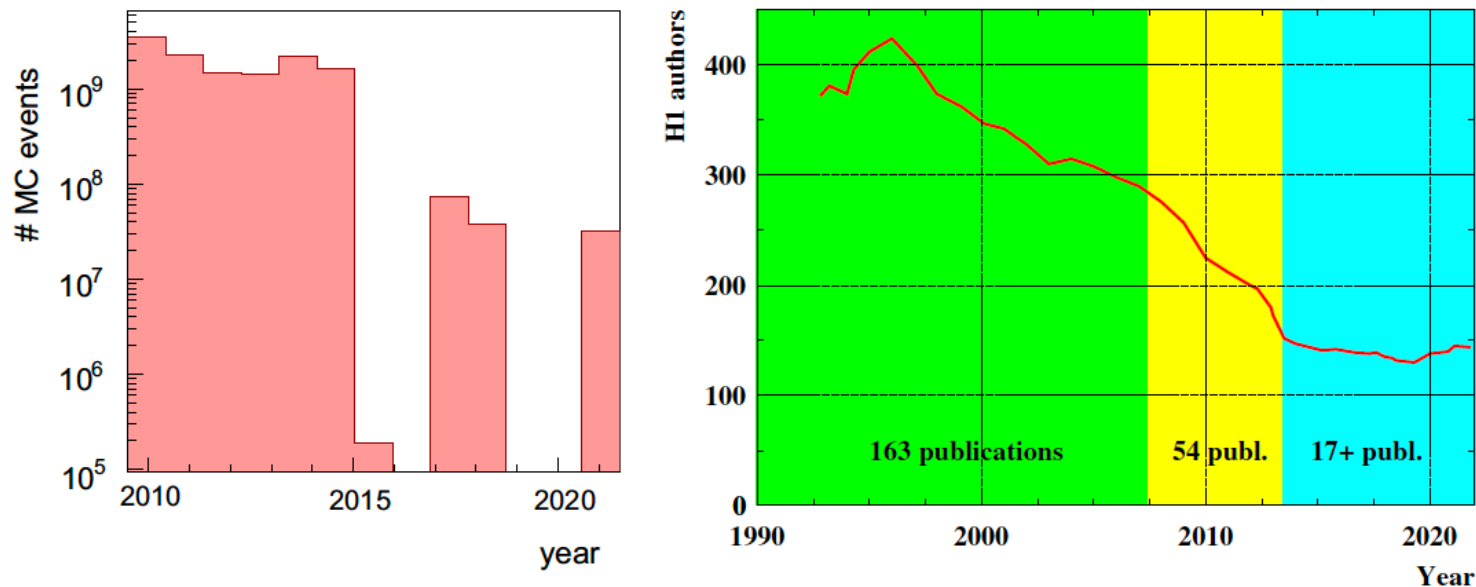


Fig. 4 Left: Number of Monte Carlo events produced centrally by the H1 Collaboration. The years without MC production are related to a change of the computing environment, or no MC requests. Right: Number of H1 authors is increasing since 2019 due to retained analysis capabilities and new interest in 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

Boosting the future experiments

Preserved data can be used to transfer knowledge, training/teaching, outreach or boosting new research programs

- **HERA → EIC**
 - “Scientists today have a **renewed interest in HERA’s particle experiments**, as they hope to use the data – and more precise computer simulations informed by tools like OmniFold – to aid in the analysis of results from future electron-proton experiments, such as at the Department of Energy’s next-generation **Electron-Ion Collider (EIC)**. “
- **Possibly**
 - **LHC → FCChh**
 - **LEP → FCCee**

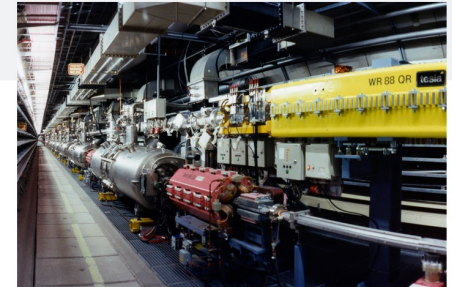
ARTICLE · MYSTERIES OF MATTER

How Do You Solve a Problem Like a Proton? You Smash It to Smithereens – Then Build It Back Together With Machine Learning

By Theresa Duque
October 25, 2022

New tool decodes proton snapshots captured by history-making particle detector in record time

CONTACT MEDIA@LBL.GOV →



Looking into the HERA tunnel: Berkeley Lab scientists have developed new machine learning algorithms to accelerate the analysis of data collected decades ago by HERA, the world’s most powerful electron-proton collider that ran at the DESY national research center in Germany from 1992 to 2007. (Credit: DESY)

<https://newscenter.lbl.gov/2022/10/25/solving-the-proton-puzzle/>

...and counting

News

News from the DESY research centre

2023/06/20

[Back](#)

Do quarks interact with the cosmos?

HERA data places limits on the interactions between quarks and cosmic background fields

DESY's HERA collider, decommissioned in 2007, is still providing valuable results to scientists. A newly released paper shows that quarks, which were the main particles under investigation at the electron-proton collider, do not visibly interact with potential cosmic background fields. This means that they don't violate a fundamental symmetry of nature, the rotation and Lorentz invariance. HERA was specifically well-suited for studying quarks, so these results set important limits for other experiments and searches.

ZEUS June 2023

BaBar April 2023
The 600th paper

PHYSICAL REVIEW D **107**, 072001 (2023)

Study of the reactions $e^+e^- \rightarrow K^+K^-\pi^0\pi^0\pi^0$, $e^+e^- \rightarrow K_S^0K^\pm\pi^\mp\pi^0\pi^0$, and $e^+e^- \rightarrow K_S^0K^\pm\pi^\mp\pi^+\pi^-$ at center-of-mass energies from threshold to 4.5 GeV using initial-state radiation

J. P. LEES *et al.*

PHYS. REV. D **107**, 072001 (2023)

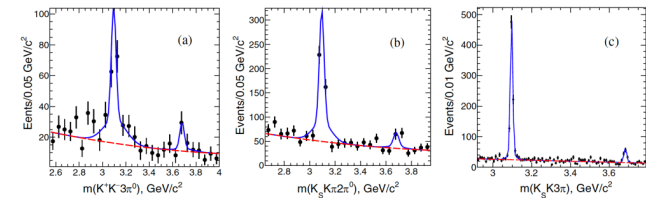
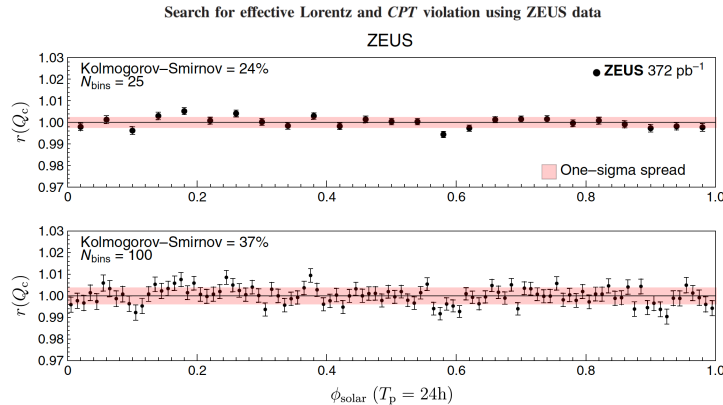


FIG. 16. The J/ψ invariant mass region for the (a) $K^+K^-\pi^0\pi^0\pi^0$, (b) $K_S^0K^\pm\pi^\mp\pi^0\pi^0$, and (c) $K_S^0K^\pm\pi^\mp\pi^+\pi^-$ events. The curves show the fit functions described in the text.

Unbinned Deep Learning Jet Substructure Measurement in High Q^2 ep collisions at HERA

H1 Collaboration · V. Andreev (Lebedev Inst.) [Show All\(148\)](#)

Mar 23, 2023

30 pages

e-Print: [2303.13620](#) [hep-ex]

Report number: DESY-23-034

H1 Mars 2023

LHC Data Preservation

- Data Preservation and Open Access policies (already since 2012-2014)
 - DP is a « specification » included in the computing models and plans for upgrades
 - HEP Software Foundation Roadmap
- Strong initiative on Open Data and Open Science policy
- Concrete implementation and technology-oriented survey
 - Very active multi-experiment projects
 - data re-use, réanalysis, réinterpretation, outreach etc.
 - **OpenData, Analysis Preservation, REANA...**

A Roadmap for HEP Software and Computing R&D for the 2020s

HEP Software Foundation¹

arXiv:1712.06982

CERN announces new open data policy in support of open science

A new open data policy for scientific experiments at the Large Hadron Collider (LHC) will make scientific research more reproducible, accessible, and collaborative

11 DECEMBER, 2020

nature physics

Explore Content Journal Information Publish With Us

nature > nature physics > perspectives > article

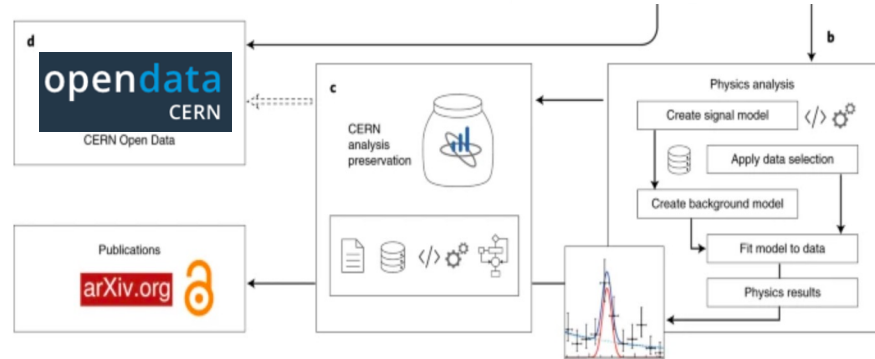
<https://www.nature.com/articles/s41567-018-0342-2>

Perspective | Open Access | Published: 15 November 2018

Open is not enough

me
tor
res.
ire,
ing
lers
er.
for

Filter by experiment	2017	2021
<input type="checkbox"/> ALICE	15	15
<input type="checkbox"/> ATLAS	101	109
<input type="checkbox"/> CMS	878	1167
<input type="checkbox"/> LHCb	3	4
<input type="checkbox"/> OPERA		904



os, Jose Benito
iguez
idis, Markus
t Iglesias, Kati

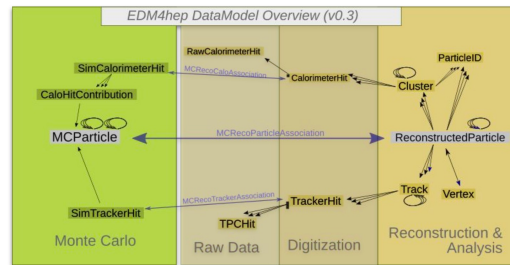
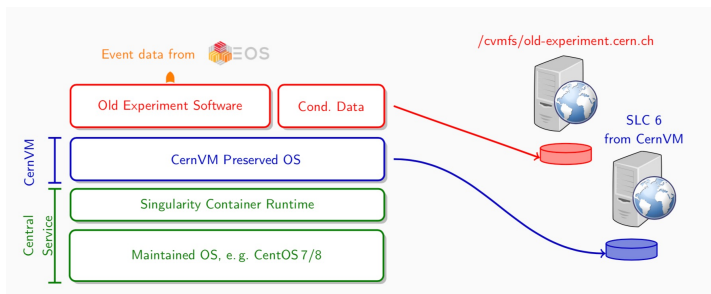


Other experiments expressed clear intention to join : LEP, JADE, H1/ZEUS, BaBar (HR is an issue)

Towards more standards

EDM4hep: the common language

CERNVM: the “freezer”



- The Event Data Model describes the structure of the data
 - Challenge: can we have the same for all HEP experiments? LCIO shared by ILC and CLIC
- Heavily inspired by LCIO and FCC-edm

key4hep / EDM4hep and DPHEP?

- Key4hep / EDM4hep: framework with longer perspective than a single experiment
 - Not just *another data format*, but one that might become a standard
- Requires “migration”, which may be a pain or not even possible
 - Workpower / Experts missing
 - Encapsulation may help here, both for migration and validation
- For LEP data, FCC-ee may provide a unique opportunity
 - Share to center-of-mass energies: 91.2 GeV, 160 GeV
 - Clear advantage in looking at what real data look like to understand bottle necks and limitations
 - Possible student projects
 - ALEPH: early investigations promising
 - ALPHA++ provides the relevant code for migration
 - Several ALEPH experts involved in FCC-ee studies

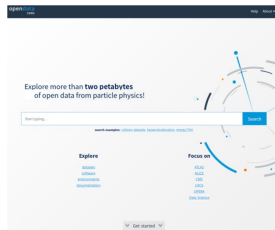
CERN Open Data portal: Status

Purpose: sharing event-level particle physics data and accompanying code for both education and research

Content: collision & simulated & derived datasets, software tools, analysis examples, VMs and containers, documentation, configuration, event display

Size: over 7600 records, over 900M files, over 2.4 PB

- ▶ CMS completed 2010–11 proton-proton data; released half of 2012 data; released 2010–11 heavy-ion data samples and corresponding pp reference datasets
- ▶ ATLAS released 13TeV educational samples
- ▶ Data science and Machine Learning (CMS, LHCb...)
- ▶ Non-LHC physics: OPERA neutrino physics data; interest from PHENIX (RHIC/BNL); JADE tests



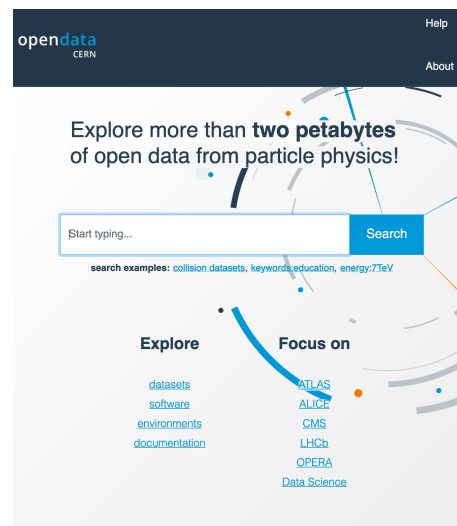
<https://opendata.cern.ch>

```
/tmp $ cernopendata-client download-files --recid 3388 --verify
=> Downloading file 3 of 11
-> File: /5360/FullFile.tar.gz
-> Progress: 60% size (5300)
=> Verifying file FullFile.tar.gz...
-> expected size 380, found 380
-> expected checksum adler32:FF63668a, found adler32:FF63668a
```

Command-line client to ease data download

@tiborsimko

2/3



CERN Open Data portal: Plans

- ▶ December 2020: A common statement on the open data policy by CERN management and ATLAS, ALICE, CMS, LHCb and TOTEM experiments.

<https://opendata.cern.ch/docs/cern-open-data-policy-for-lhc-experiments>

- ▶ Prepare for forthcoming increase in open data publishing.
- ▶ Introduce flexible hot/cold disk/tape storage solution. Part of dataset files on disk, part on tapes.
- ▶ Simplify ingestion and exposure of experiment datasets (Rucio, Dirac).
- ▶ Automate provenance testing and usage examples. (See the next presentation with REANA status overview.)

@tiborsimko

3/3

CERN announces new open data policy in support of open science

A new open data policy for scientific experiments at the Large Hadron Collider (LHC) will make scientific research more reproducible, accessible, and collaborative

11 DECEMBER 2020



Geneva, 11 December 2020. The four main LHC collaborations (ALICE, ATLAS, CMS and LHCb) have unanimously endorsed a new open data policy for scientific experiments at the Large Hadron Collider (LHC), which was presented to the CERN Council today. The policy commits to publicly releasing so-called level 3 scientific data, the type required to make scientific studies, collected by the LHC experiments. Data will have to be released approximately five years after collection, and the aim is for the full dataset to be publicly available by the close of the experiment concerned. The policy addresses the growing movement of open science, which aims to make scientific research more reproducible, accessible, and collaborative.

The level 3 data released can contribute to scientific research in particle physics, as well as research in the field of scientific computing, for example to improve reconstruction or analysis methods based on machine learning techniques, an approach that requires rich data sets for training and validation.

CERN Analysis Preservation and Reusable Analyses

nature physics

Explore Content Journal Information Publish With

nature > nature physics > perspectives > article

Perspective | Open Access | Published: 15 November 2018

Open is not enough

Xiaoli Chen, Sjinje Dalmeier-Tiessen, Robin Dasler, Sebastian Feger, Panfilos Fokianos, Jose Benito Gonzalez, Hani Hirvonen, Dinos Kousidis, Artemis Lavasa, Salvatore Mele, Diego Rodriguez Rodriguez, Tiber Simko, Tim Smith, Ana Trisovic, Anna Trzcinska, Ioannis Tsanaktidis, Markus Zimmermann, Kyle Cranmer, Lukas Heinrich, Gordon Watts, Michael Hildreth, Lara Lioret Iglesias, Kati Lassila-Perini & Sebastian Neubert

- **CAP** : preserve analysis
 - <http://analysispreservation.cern.ch/>
- **REANA** : improve workflow
 - Run research data analyses on containerised compute clouds
 - <http://reana.io/>

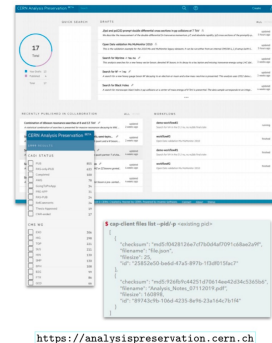
CERN Analysis Preservation framework

Purpose: capture and preserve all elements needed to understand and reuse an analysis even several years later; take a consistent snapshot linking all the knowledge

Usage: describe analysis + deposit n-tuples, code etc via CLI and web UI + share with colleagues = preserve knowledge

Community: pilot with ALICE, ATLAS, CMS, LHCb

- ▶ content restricted to collaborations
- ▶ metadata interconnected with collaboration databases
- ▶ associated knowledge, e.g. CMS statistics questionnaire
- ▶ helps addressing increasing number of funding agencies asking for comprehensive data management policies
- ▶ run by CERN Scientific Information Service (P. Fokianos, K. Naim)



@thuvetmbn

2 / 3

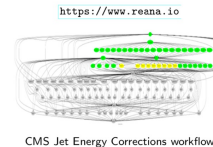
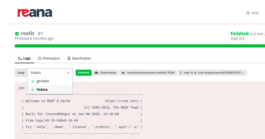
REANA reproducible analysis platform

Purpose: run declarative computational workflows on containerised compute clouds

Usage: data + code + environment + workflow = computational reproducibility

Community: pilot examples with ALICE, ATLAS, CMS, FCC, LHCb; ATLAS search groups (SUSY, EXOT, HDBS) now require workflow preservation as mandatory for analysis approval

- ▶ promotes pre-productibility during active analysis phase to facilitate future preservation
- ▶ integration with GitLab; CI/CD mode
- ▶ verification of analysis examples and data provenance chain (CMS AOD reprocessing)
- ▶ support for hybrid compute workflows with multiple backends (HTCondor, Kubernetes, Slurm)



@tiborsimko

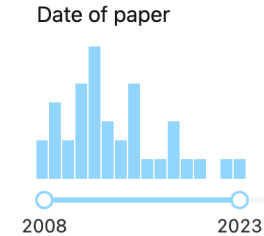
REANA running on supercomputers (e.g. NERSC)

3 / 3

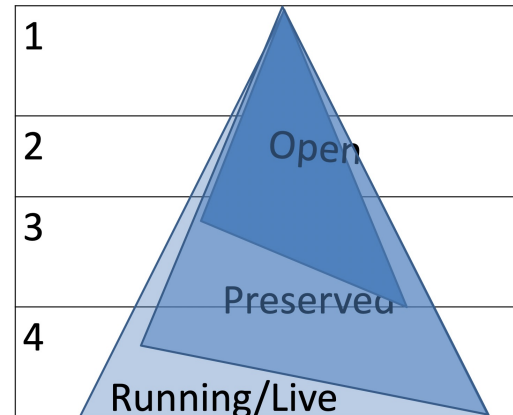
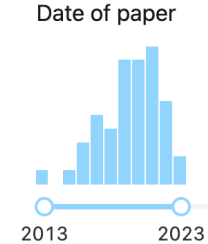
Preserved and Open Data

- Planning for preserved data improves the design of running and future experiments
- DP relies on and stimulates cutting-edge technology developments
- DP is strongly linked to **Open Science and FAIR** data paradigms
 - F findable A accessible R reproducible
 - Most difficult to obtain is “I” : interoperability
- Examples:
 - CERN Open Data Portal, Analysis Preservation (CAP), Reusable Analyses (ReAna), cernvm, key4hep etc.
 - Experiments with long DP practice intend to join open data projects
 - Lack of person power

find t **data preservation**
39 results



find t **open data**
50 results



Open Questions

- 1. Why the systems did not collapse after the data taking? The “common sense: “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 How is HEP data contributing to the human culture as a whole (like in arts, e.g. a painting, or a piece of music, which cannot be valued just in terms of investment, resources and financial transactions)
- 6. What global resources were used 5 and 10 years past the end of the experiment to keep systems alive and publish?
- 7. 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**?
- 8. 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?
- 9 Outreach and education done using real data sets?

Conclusions 1/2

- Significant/measurable impact of dedicated DP projects @expts./labs
 - Production of high quality and unique scientific results at very low (non-zero) cost
 - 10% output for less than 1% investment: ✓
 - **Long term organisation proves to be productive**
 - Signs of re-vigorating collaborations in the context of new projects
 - HERA-EIC; LEP-FCCee
 - Case for longer term preservation: data sets parking
 - CDF, D0, Babar, LEP, Jade : carefully follow the usability in time
- There is a full coherence (but not total overlap) between DP and **Open Data/Science**
 - LHC experiments consider both, looking forward to 2045
- The (DP)HEP future is also considered
 - FCC, EIC : transfer of knowledge in DP from LHC/oldies
- And more is possible on:
 - Education, training, outreach....(via open data)

Conclusions 2/2

- Although the general awareness of DP in HEP has largely increased in the past decade, and lots of activities and successes are reported, there is still a lack of coherence between the different experiments and projects.
- Transverse activities are still exceptions and cover only a minority of the existing data sets.
- Moreover explicit support of DP by some labs and funding agencies should be at least maintained and probably increase.
- Data preservation is one of the building blocks of the HEP scientific outcome and the DPHEP Collaboration intends to stimulate and support it.

Recent developments

- New ICFA panel, enlarging the scope:
 - “ICFA Panel on the Data Lifecycle”
 - Mission:
 -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.[...]
 - Mandate:
 - Address the data lifecycle within a structured and integrated systems approach in HEP[...]
 - Support the ongoing projects and collaborations started within the “Data Preservation in High Energy Physics” collaboration (DPHEP) and the “Standing Committee on Interregional Connectivity” (SCIC).
- IN2P3: Working group in Open Data (Chair A. Lemasson, Ganil)
 - Mandate: produce a status report and recommendations for IN2P3 data management policies
 - Help is needed for HEP!!!! Volunteers?

- B-up

The DPHEP 2020 Vision

- *The “vision” for DPHEP – first presented to ICFA **in February 2013** – a consists of the following key points:*
 - By 2020, all **archived data** – e.g. that described in DPHEP Blueprint, including LHC data – should be easily **findable** and fully **usable** by the **designated communities** with clear (Open) access policies and possibilities to annotate further
 - Best practices, tools and services should be well run-in, fully documented and sustainable; built in common with **other disciplines**, based on standards
 - There should be a **DPHEP portal**, through which data / tools accessed
 - **Clear targets & metrics** to measure the above should be agreed between **Funding Agencies, Service Providers** and the **Experiments (Collaborations)**.
 - Although there is clearly much work still to be done, this vision looks both achievable and the timescale for realizing it has been significantly reduced through interactions with other (non-HEP) projects and communities.

2012 (blueprint)

<p>Priority 1:</p> <p>Local Action in experiments, laboratories</p>	<p>Data preparation: 1-3 FTE/expt/2-3 years</p> <p>Data archivists: 0.5-1 FTE /lab</p>
<p>Priority 2:</p> <p>International organization</p>	<p>Project Manager: 1 FTE</p> <p>Technical support: 0.2 FTE</p> <p>Contributions from Labs: 0.2/lab (data archivists)</p>
<p>Priority 3:</p> <p>Transverse Projects (examples considered)</p>	<p>Project leaders: 1-2 FTE's/projects</p> <p>+ contributions from involved experiments 0.2 FTEs/expt.</p>

- According to the provisions 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 $O(10)$ FTE-years with a very marginal investment in material
- 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 $O(10^3)$ 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.
 - $C1 = O(0,1\%)$
 - $B1 = O(10\%)$
- $C1/B1 \rightarrow$ cost effective science
- Refinements possible
 - make the exercise for Open data as well

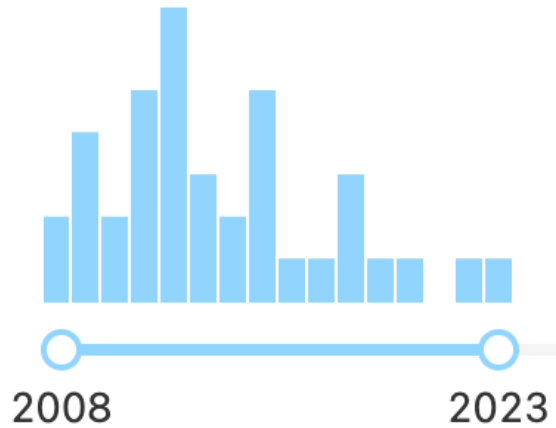
Preserved and Open Data

- Planning for preserved data improves the design of running and future experiments
- DP relies on and stimulates cutting-edge technology developments
- DP is strongly linked to **Open Science and FAIR** data paradigms
- Examples:
 - CERN Open Data Portal, Analysis Preservation (CAP), Reusable Analyses (ReAna), cernvm, key4hep etc.

Open and preserved

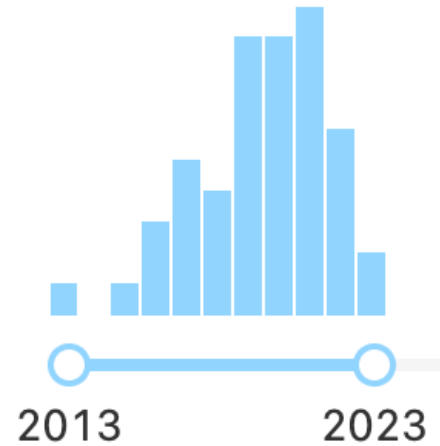
find t **data preservation**
39 results

Date of paper



find t **open data**
50 results

Date of paper



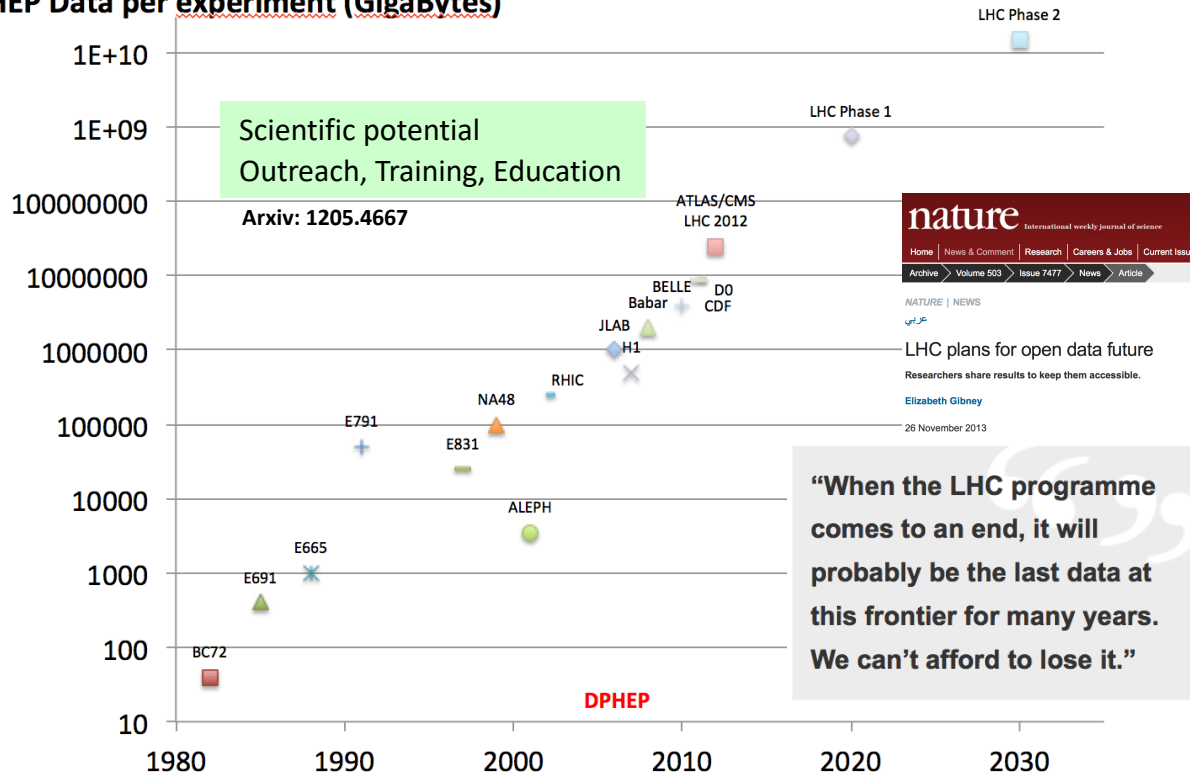
A word on FAIR

- The DPHEP objectives (2012) intrinsically comply with what has become to be known as FAIR principles (2016)
- Indeed, the data has to be
 - easy to find (F)
 - 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, 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.
 - DPHEP can certainly contribute to such a global approach

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.

HEP Data

HEP Data per experiment (GigaBytes)



Discussion incentives

- Preservation and sharing:
 - Let data escape into unknown/unusual world
 - “In time” (long term) → Preserved
 - “In space” (released to others) → Open
- Why would you do that?
 - Data contains more than planned for → more science
 - New audience, new ideas → more science
 - More technology, interdisciplinarity, skills, teaching, policy
- The motivation is shared by both P&O
 - How are those related?
 - DPHEP: P & O are complementary and rather strongly related aspects of a continuous output enhancement action around unique frontier science data
- DPHEP report 2022:
 - a strong interest to translate healthy and functional analysis systems into open data hosts , HERA, BaBar, RHIC
 - main pb: Person power
- There is room to think and act in common and global

DPHEP resources for DP

- 2012 Blueprint

	Project	Goals and deliverables	Resources and timelines	Location, possible funding source, DPHEP allocation
Experiment and laboratory Priority: 1	Experimental Data Preservation Task Force	Install an experiment data preservation task force to define and implement data preservation goals.	1 FTE installed as soon as possible, and included in upgrade projects	Located within each computing team. Experiment funding agencies or host laboratories. DPHEP contact ensured, not necessarily as a displayed FTE.
	Facility or Laboratory Data Preservation Projects	Data archivist for facility, part of the R&D team or in charge with the running preservation system and designed as contact person for DPHEP.	1-2 FTE per laboratory, installed as a common resource.	Experiment common person-power, support by the host labs or by the funding agencies as a part of the on going experimental programme. A fraction 0.2 FTE allocated to DPHEP for technical support and overall organisation.
Multi-experiment Priority: 3	General validation framework	Provide a common framework for HEP software validation, leading to a common repository for experiments software. Deployment on grid and contingency with LHC computing also part of the goals.	1 FTE	Installed in DESY, as present host of the corresponding initiative. Funding from common projects. Cooperation with upgrades at LHC can be envisaged. Part of DPHEP.
	Archival systems	Install secured data storage units able to maintain complex data in a functional form over long period of time without intensive usage.	0.5 FTE	Multi-lab project, cooperation with industry possible. Included in DPHEP person-power.
	Virtual dedicated analysis farms	Provide a design for exporting regular analysis on farms to closed virtual farm able to ingest frozen analysis systems for a 5-10 years lifetime.	1 FTE	The host of this working group should be SLAC. Funding could come from central projects and can be considered as part of DPHEP.
	RECAST contact	Ensure contact with projects aiming at defining interfaces between high-level data and theory.	0.5 FTE	Installed with proximity to the LHC, the main consumer of this initiative, with strong connections to the data preservation initiatives that may adopt the paradigms.
	High level objects and INSPIRE	Extend INSPIRE service to documentation and high-level data object.	0.5-1.5 FTE	Installed at one of the INSPIRE partner laboratories.
	Outreach	Install a multi-experiment project on outreach using preserved data, define common formats for outreach and connect to the existing events.	1 FTE central + 0.2 FTE per experiment	A coordinating role can be played by DPHEP in connection with a large outreach project existing at CERN, DESY or FNAL. The outreach contributions from experiments and laboratories can be partially allocated to the common HEP data outreach project and steered by DPHEP.
	Global Priority: 2	DPHEP Organisation	DPHEP Project Manager	1 FTE

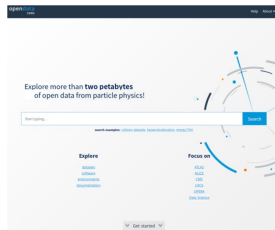
CERN Open Data portal: Status

Purpose: sharing event-level particle physics data and accompanying code for both education and research

Content: collision & simulated & derived datasets, software tools, analysis examples, VMs and containers, documentation, configuration, event display

Size: over 7600 records, over 900M files, over 2.4 PB

- ▶ CMS completed 2010–11 proton-proton data; released half of 2012 data; released 2010–11 heavy-ion data samples and corresponding pp reference datasets
- ▶ ATLAS released 13TeV educational samples
- ▶ Data science and Machine Learning (CMS, LHCb...)
- ▶ Non-LHC physics: OPERA neutrino physics data; interest from PHENIX (RHIC/BNL); JADE tests



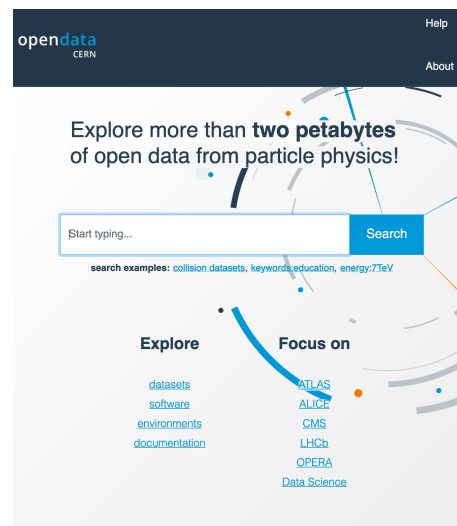
<https://opendata.cern.ch>

```
/tmp $ cernopendata-client download-files --recid 3388 --verify
=> Downloading file 1 of 11
-> File: /5360/FullFile.tar.gz
-> Progress: 610 KiB (500%)
=> Verifying file FullFile.tar.gz...
-> expected size 380, found 380
=> expected checksum adler32:ff6366ba, found adler32:ff6366ba
```

Command-line client to ease data download

@tiborsimko

2/3



CERN Open Data portal: Plans

- ▶ December 2020: A common statement on the open data policy by CERN management and ATLAS, ALICE, CMS, LHCb and TOTEM experiments.

<https://opendata.cern.ch/docs/cern-open-data-policy-for-lhc-experiments>

- ▶ Prepare for forthcoming increase in open data publishing.
- ▶ Introduce flexible hot/cold disk/tape storage solution. Part of dataset files on disk, part on tapes.
- ▶ Simplify ingestion and exposure of experiment datasets (Rucio, Dirac).
- ▶ Automate provenance testing and usage examples. (See the next presentation with REANA status overview.)

CERN announces new open data policy in support of open science

A new open data policy for scientific experiments at the Large Hadron Collider (LHC) will make scientific research more reproducible, accessible, and collaborative

11 DECEMBER 2020



Geneva, 11 December 2020. The four main LHC collaborations (ALICE, ATLAS, CMS and LHCb) have unanimously endorsed a new open data policy for scientific experiments at the Large Hadron Collider (LHC), which was presented to the CERN Council today. The policy commits to publicly releasing so-called level 3 scientific data, the type required to make scientific studies, collected by the LHC experiments. Data will be released approximately five years after collection, and the aim is for the full dataset to be publicly available by the close of the experiment concerned. The policy addresses the growing movement of open science, which aims to make scientific research more reproducible, accessible, and collaborative.

The level 3 data released can contribute to scientific research in particle physics, as well as research in the field of scientific computing, for example to improve reconstruction or analysis methods based on machine learning techniques, an approach that requires rich data sets for training and validation.

CERN Analysis Preservation and Reusable Analyses

nature physics

Explore Content Journal Information Publish With

nature > nature physics > perspectives > article

Perspective | Open Access | Published: 15 November 2018

Open is not enough

Xiaoli Chen, Sjinje Dalmeier-Tiessen, Robin Dasler, Sebastian Feger, Panfilos Fokianos, Jose Benito Gonzalez, Harri Hironaka, Dinos Kouzidis, Artemis Lavasa, Salvatore Mele, Diego Rodriguez Rodriguez, Tiber Simko, Tim Smith, Ana Trisovic, Anna Trzcinska, Ioannis Tsanaktsidis, Markus Zimmermann, Kyle Cranmer, Lukas Heinrich, Gordon Watts, Michael Hildreth, Lara Lloret Iglesias, Kati Lassila-Perini & Sebastian Neubert

- **CAP** : preserve analysis
 - <http://analysispreservation.cern.ch/>
- **REANA** : improve workflow
 - Run research data analyses on containerised compute clouds
 - <http://reana.io/>

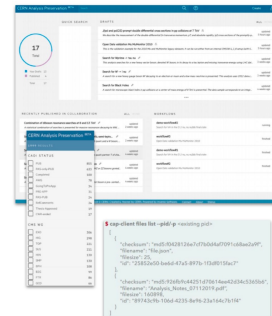
CERN Analysis Preservation framework

Purpose: capture and preserve all elements needed to understand and reuse an analysis even several years later; take a consistent snapshot linking all the knowledge

Usage: describe analysis + deposit n-tuples, code etc via CLI and web UI + share with colleagues = preserve knowledge

Community: pilot with ALICE, ATLAS, CMS, LHCb

- ▶ content restricted to collaborations
- ▶ metadata interconnected with collaboration databases
- ▶ associated knowledge, e.g. CMS statistics questionnaire
- ▶ helps addressing increasing number of funding agencies asking for comprehensive data management policies
- ▶ run by CERN Scientific Information Service (P. Fokianos, K. Naim)



<https://analysispreservation.cern.ch>

@thuvetmbn

2 / 3

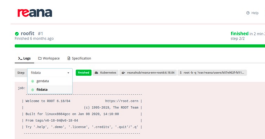
REANA reproducible analysis platform

Purpose: run declarative computational workflows on containerised compute clouds

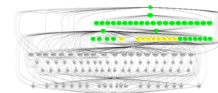
Usage: data + code + environment + workflow = computational reproducibility

Community: pilot examples with ALICE, ATLAS, CMS, FCC, LHCb; ATLAS search groups (SUSY, EXOT, HDBS) now require workflow preservation as mandatory for analysis approval

- ▶ promotes pre-productibility during active analysis phase to facilitate future preservation
- ▶ integration with GitLab; CI/CD mode
- ▶ verification of analysis examples and data provenance chain (CMS AOD reprocessing)
- ▶ support for hybrid compute workflows with multiple backends (HTCondor, Kubernetes, Slurm)



<https://www.reana.io>



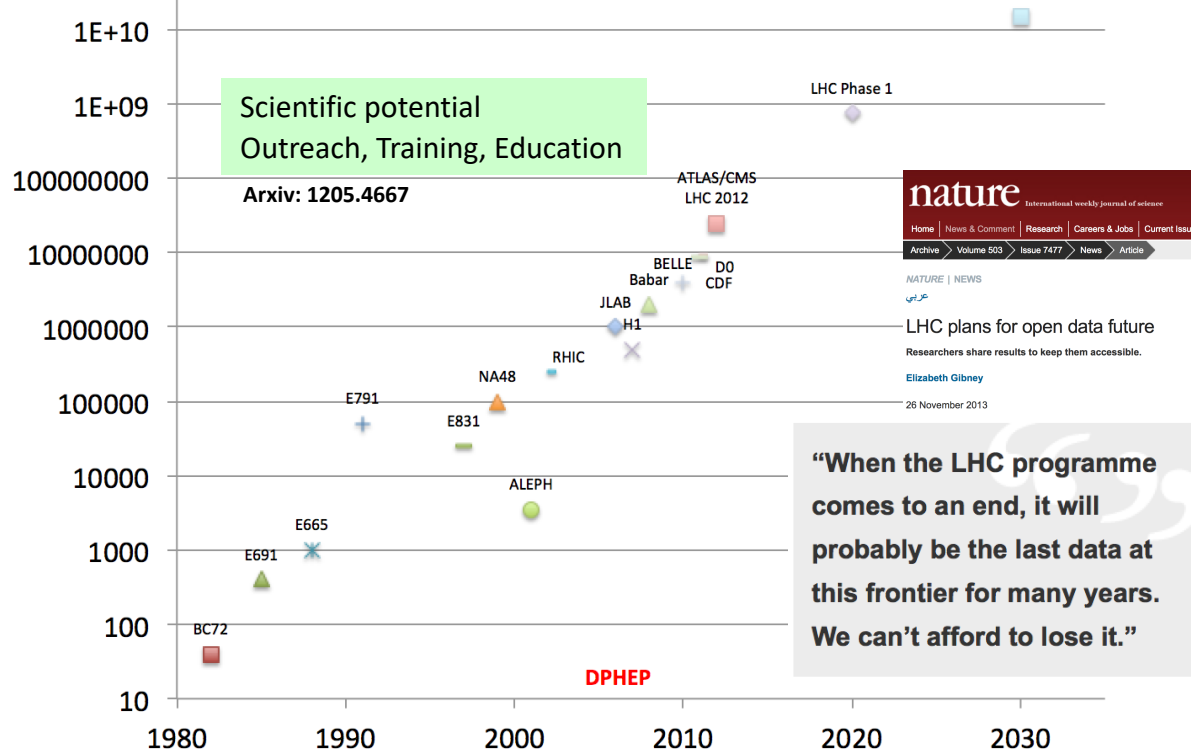
CMS Jet Energy Corrections workflow



REANA running on supercomputers (e.g. NERSC)

HEP Data

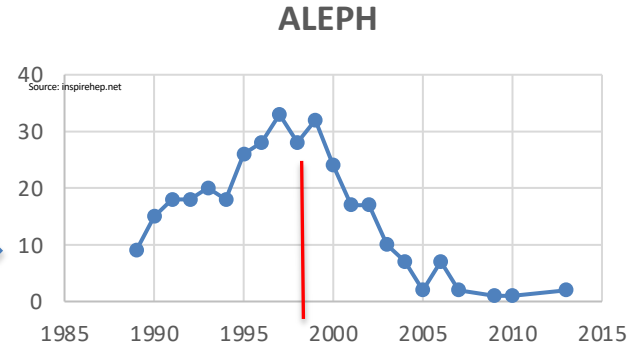
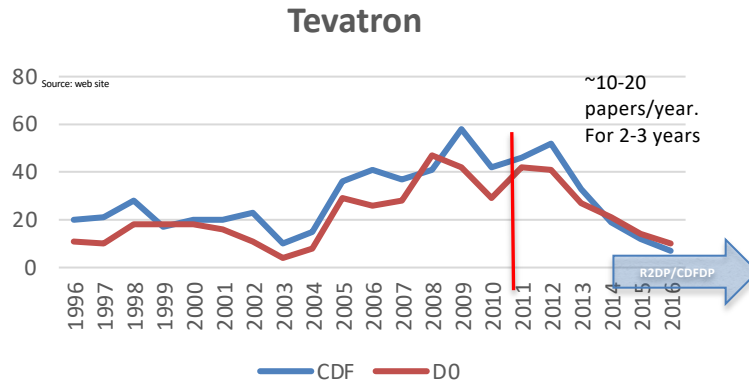
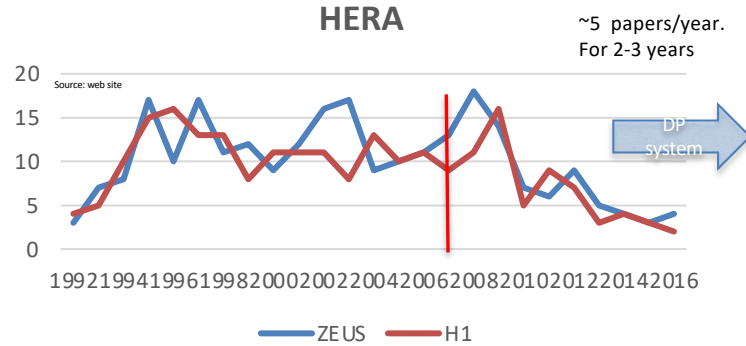
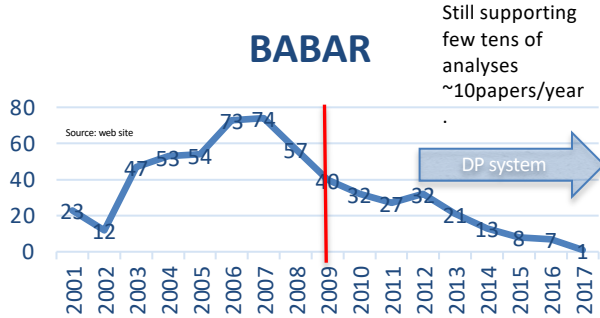
HEP Data per experiment (GigaBytes)



DPHEP timelines

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
	Start-up					Consolidation			DPHEP Collaboration		
HEP	HERA stops	Babar stops	LHC starts	Belle I stops	Tevatron stops				LHC Run 2		
DPHEP Group			ICFA Panel		LHC exp. joined	DPHEP Manger appointed at CERN		DPHEP Collaboration Agreements signed	1 st DPHEP Collaboration Meeting		2 nd DPHEP Collaboration Meeting
DPHEP Docs			DPHEP White Paper			Blueprint Report			DPHEP Status Report 2020 Vision		DPHEP 2017 Status Report
DP Projects within expts.		Babar DP starts		HERA DP starts	BELLE DP starts	CMS DP Policy CDF/DO DP starts Babar LTdap operational	ALICE, LHCb, DP Policies	ATLAS DP Policy H1/ZEUS DP systems operational	CERN/LHC Open Data	CERN/LHC Analysis Preservation Tevatron DP operational	

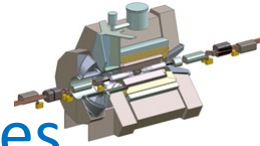
Scientific output: status 2017





2018 status

BABAR Highlights and Press Releases



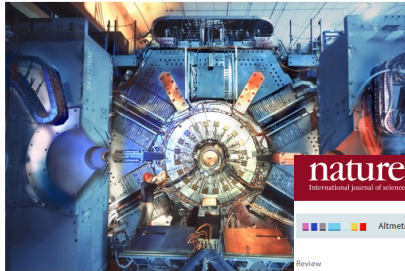
NEWS CENTER

New Study: Scientists Narrow Down the Search for Dark Photons Using Decade-Old Particle Collider Data

Analysis of data from the BaBar experiment rules out theorized particle's explanation for muon mystery

News Release Glenn Roberts Jr. (510) 5

November 2017



The BaBar detector at SLAC National Accelerator Laboratory. (Credit: SLAC)

Dataset:

Y(4S): 433/fb
Y(3S): 30/fb
Y(2S): 14/fb
Off resonance: 10%
Y(1S) accessed via
Y(2S,3S) → Y(1S) π⁺π⁻



A challenge to lepton universality in B-meson decays

Gregory Clezarek¹, Manuel Franco Sevilla², Brian Hamilton³, Robert Kowalewski⁴, Thomas Kuhr Vera Luth⁵ & Yutaro Sato⁷

Nature 546, 227–233 (08 June 2017) | Received: 15 December 2016
doi:10.1038/nature2 | 20 March 2017
Download Citation | June 2017 | online: 07 June 2017

Experimental particle physics | Phenomenology | Theoretical particle physics

Abstract



Home About Particle People

A communication resource from the world's particle physics laboratories

New study: Scientists narrow down the search for dark photons using decade-old particle collider data

8 November 2017 - Lawrence Berkeley National Laboratory

Analysis of data from the BaBar experiment rules out theorized particle's explanation for muon mystery

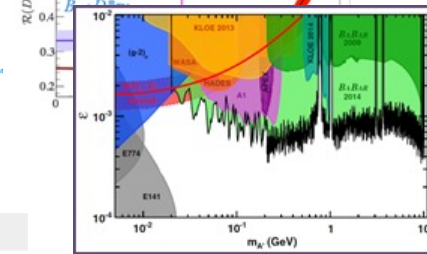
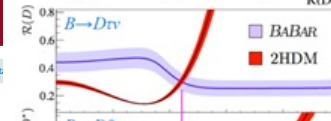
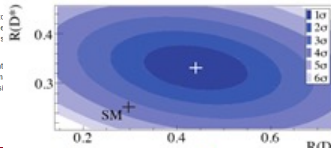
In its final years of operation, a particle collider in northern California was refused to search for signs of new particles that might help fill in some big blanks in our understanding of the universe.

A fresh analysis of this data, collected by physicists at the Department of Energy's Lawrence Berkeley National Laboratory (Berkeley Lab), limits some of the hiding places for one type of theorized particle – the dark photon, also known as the heavy photon – that was proposed to help explain the mystery of dark matter.

The latest result, published in the journal *Physical Review Letters* by the roughly 240-member BaBar Collaboration, adds to results from a collection of previous experiments seeking, but not yet finding, the theorized dark photons.

"Although it does not rule out the existence of dark photons, and definitively rule out their explanation for another property of the enigmatic particle known as the muon," University of Victoria professor

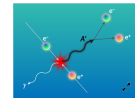
Dark matter, which accounts for an estimated 85 percent observed by its gravitational interactions with normal galaxies is much faster than expected based on their vis



Viewpoint: New Light Shed on Dark Photons

Douglas Bryman, University of British Columbia, Vancouver, British Columbia V2T2A6, Canada

A search for a hypothetical particle that could be related to dark matter has come up empty, putting new constraints on models that imagine a dark form of electromagnetism.

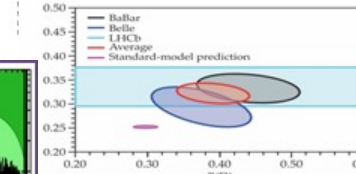


The Economist

The arrow of time

Backward ran sentences... To the relief of physicists, time really does have a preferred direction

By Lee 2013 | From the print edition
TIME seems to flow inexorably in one direction. Superficially, that is because things deteriorate with age – and this, in turn, is because there are immeasurably fewer ways to arrange particles in an orderly fashion than in a jumbled mess. Any change in an existing arrangement is therefore likely to increase its disorder.
Dig a little deeper, though, and things arrive become mysterious. A particle cannot, by itself, become disordered, so when you examine its behaviour in isolation the past and the future are hard to distinguish. If you film its movement and then give the film to someone else, we will not be able to work out just from the particle's behaviour which way to run the film through the projector.





BABAR needs Help!

- *BABAR* data actively being analyzed and high impact papers published (see slide 2). Expect this to continue to at least through 2021.
- SLAC management plans to stop hosting *BABAR* computing in February 2020 at which time the tapes with data will be ejected.
- DOE support ended in 2017, now running on international common funds (OCF).
- Looking for possibility of support and long term data preservation at
 - CERN,
 - GridKa (*BABAR* site for analysis and XRootD federated dataset main redirector),
 - University of Victoria (*BABAR* site for analysis, documentation, and tools support).
- *BABAR* lightweight VMs come with the latest software release and xrootd client included, running under the most common virtual machine players. Just add the data via the GridKa main XRootD redirector.

BABAR in Numbers

- 2PB of data on T10k-D tapes
 - raw, processed, Monte Carlo
 - **Unique dataset at the $\Upsilon(3S)$ resonance** (no plan at the moment to run at the $\Upsilon(3S)$ @ Belle II)
 - Full environment enclosed in VMs (SL5,SL6)
 - ~1TB of documentation, repositories, and dataset information (DBs, cvs, wiki, html)
 - Internal documents archived on INSPIRE
-
- 574 papers, ~10 papers/year past 3 years
 - 231 members (semi-frozen author list)
 - Including PhD students in Canada, Germany, Israel, Italy, Russia, US
 - Associated theorists mine data to test new ideas
 - ~20 analyses on track, ~10 more in the pipeline
 - Continue to have new analyses every year including joint *BABAR* -Belle analyses
 - **Students analyze *BABAR* data while working on Belle II and other experiments in construction/commissioning phase**