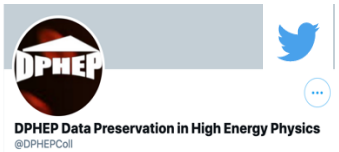
Data Preservation in High Energy Physics: report and perspectives



http://dphep.org

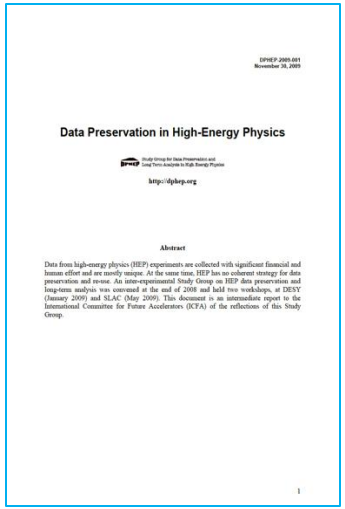
C. Diaconu (CPPM, Aix-Marseille University and CNRS/IN2P3)
G. Ganis (CERN)
U. Schwickerath (CERN)
for the DPHEP Collaboration

DPHEP Collaboration

Inter-experimental study group on HEP data preservation and long-term analysis
initiated as a panel of the International Committee for Future Accelerators (ICFA) in 2008

2009
LoI

arxiv:0912.0255


DPHEP-2009-001
November 30, 2009

Data Preservation in High-Energy Physics

http://dphep.org

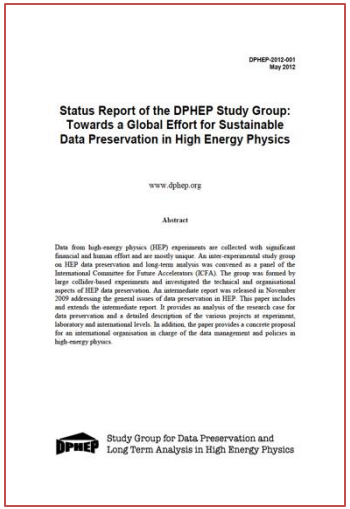
Abstract

Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. At the same time, HEP has no coherent strategy for data preservation and re-use. An inter-experimental Study Group on HEP data preservation and long-term analysis was convened at the end of 2008 and held two workshops, at DESY (January 2009) and SLAC (May 2009). This document is an intermediate report to the International Committee for Future Accelerators (ICFA) of the reflections of this Study Group.

1

2012
Blueprint

arXiv: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

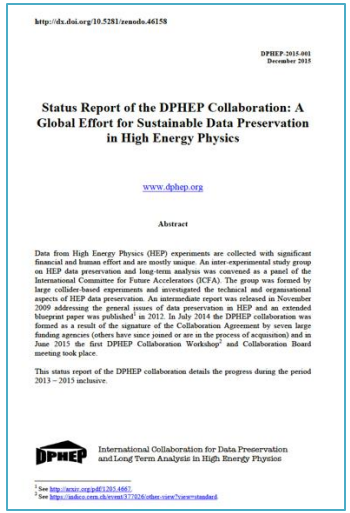
Abstract

Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. An inter-experimental study group on HEP data preservation and long-term analysis was convened as a panel of the International Committee for Future Accelerators (ICFA). The group was formed by large collider-based experiments and investigated the technical and organisational aspects of HEP data preservation. An intermediate report was released in November 2009 addressing the general issues of data preservation in HEP. This paper includes and extends the intermediate report. It provides an analysis of the research case for data preservation and a detailed description of the various projects at experiment, laboratory and international levels. In addition, the paper provides a concrete proposal for an international organisation 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


http://dx.doi.org/10.5281/zenodo.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 inter-experimental study group on HEP data preservation and long-term analysis was convened as a panel of the International Committee for Future Accelerators (ICFA). The group was formed by large collider-based experiments and investigated the technical and organisational aspects of HEP data preservation. An intermediate report was released in November 2009 addressing the general issues of data preservation in HEP and an extended blueprint paper was published[1] 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 (others have since joined or are in the process of acquisition) and in June 2015 the first DPHEP Collaboration Workshop[2] and Collaboration Board meeting took place.

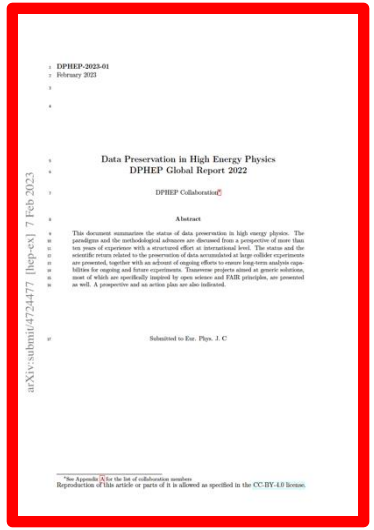
This status report of the DPHEP collaboration details the progress during the period 2013 – 2015 inclusive.

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

[1] See http://arxiv.org/pdf/1205.4667.
[2] See https://indico.cern.ch/event/377026/other-view?view=standard

2023
Decade report

arXiv: 2302.03583


DPHEP-2023-01
February 2023

Data Preservation in High Energy Physics
DPHEP Global Report 2022

DPHEP Collaboration[†]

Abstract

This document summarises 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 advocacy 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.

Submitted to Eur. Phys. J. C

arXiv:submit/4724477 [hep-ex] 7 Feb 2023

[†]See Appendix A for the list of collaboration members
Reproduction of this article or parts of it is allowed as specified in the CC-BY-4.0 license.

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

ICFA Panel 2024: DPHEP joined Data Lifecycle Panel

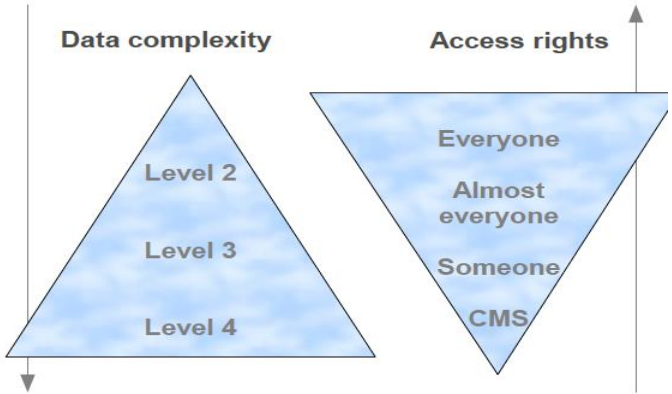
Data Preservation in High Energy Physics

- 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	

CMS 2012



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: DP Organisation during experiment proposal.
- 1: DP Organisation during data taking.
- 2: DP Organisation after data taking and during analysis-only mode.
- 3: DP Organisation after the collaboration funding scheme.
- 4: DP 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 == 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 - specifically to DP.
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.

- FoM = B1/C1

Colaboration in rescue mode

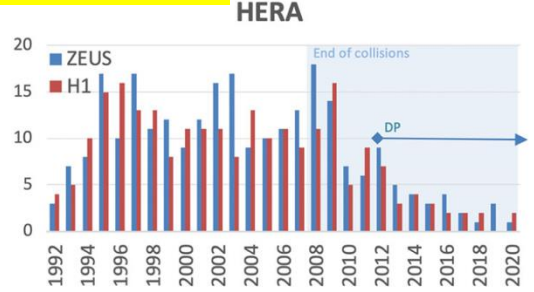
2023: Experiments Data Preservation Status

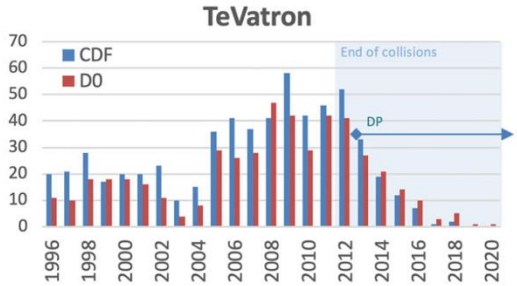
Laboratory/ Collider	Experiment	Data taking period	Preservation Level	Data Volume	Present status	Coll
DESY/PETRA	JADE	1979–1986	4	1 TB	Analysis running on preserved data; migrated from DESY to MPP	4
CERN/LEP	ALEPH,DELPHI,OPAL L3	1989-2000	4 3/4	0.5 PB	Analysis running on preserved data	3/4
DESY/HERA	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	3 or 4	O(1EB)	Collecting and analyzing data	1

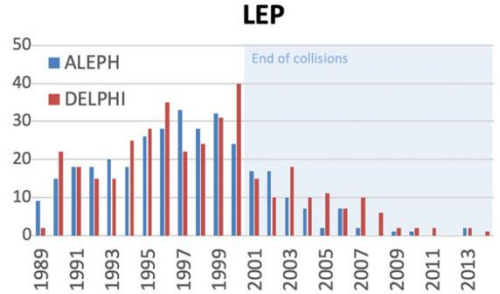
Conclusions after 10 years: the scientific output

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 2012-2022	Scientific return increase %
Babar	2008	471	154	33%
H1+ZEUS	2007	436	62	14%

Ongoing. Still

News

News from the DESY research centre

2023/06/20

Back

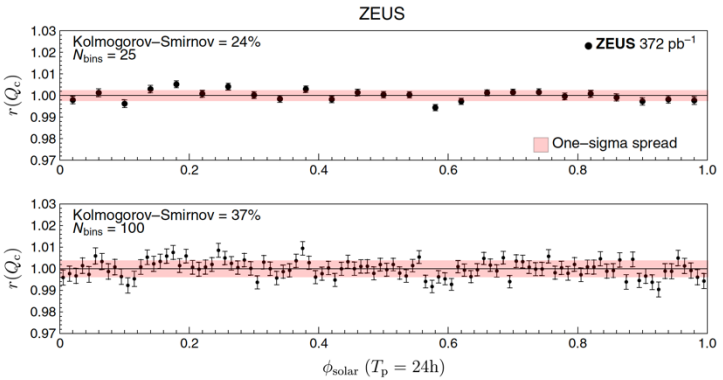
Do quarks interact with the cosmos?

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

Search for effective Lorentz and *CPT* violation using ZEUS data



BaBar April 2023
The 600th paper

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

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

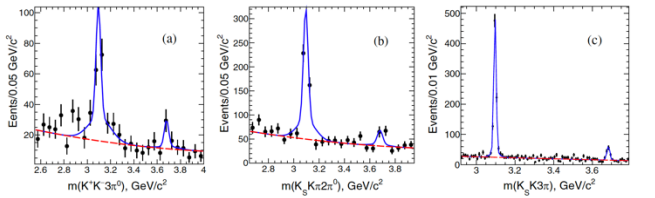


FIG. 16. The J/ψ invariant mass region for the (a) $K^+K^-\pi^0\pi^0\pi^0$, (b) $K^0_S K^\pm\pi^\mp\pi^0\pi^0$, and (c) $K^0_S K^\pm\pi^\mp\pi^+\pi^-$ events. The curves show the fit functions described in the text.

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

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

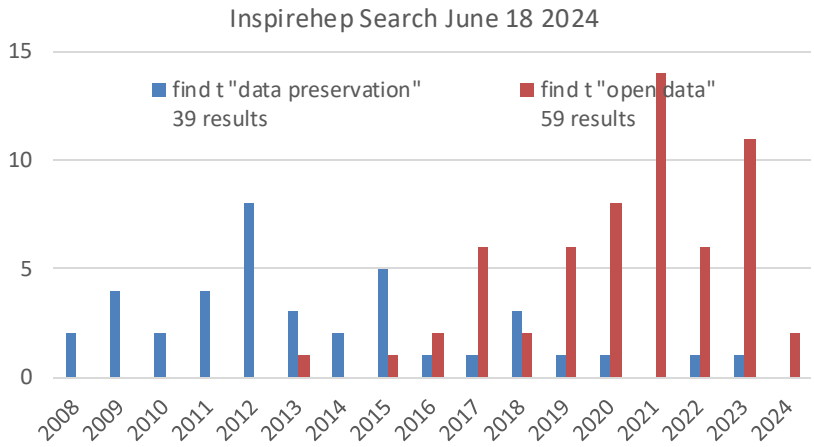
CONTACT MEDIA@LBL.GOV

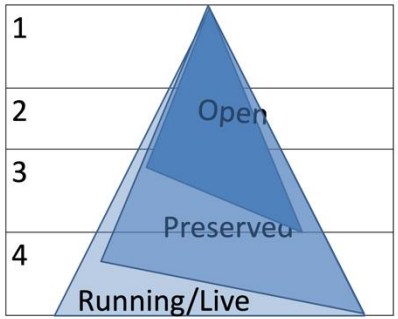
Looking into the HERA tunnel: Berkeley Lab scientists have developed new machine learning algorithms to accelerate the analysis of data collected decades ago by HERA, the world's most powerful electron-proton collider that ran at the DESY national research center in Germany from 1992 to 2007. (Credit: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

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

Preserved and Open Data

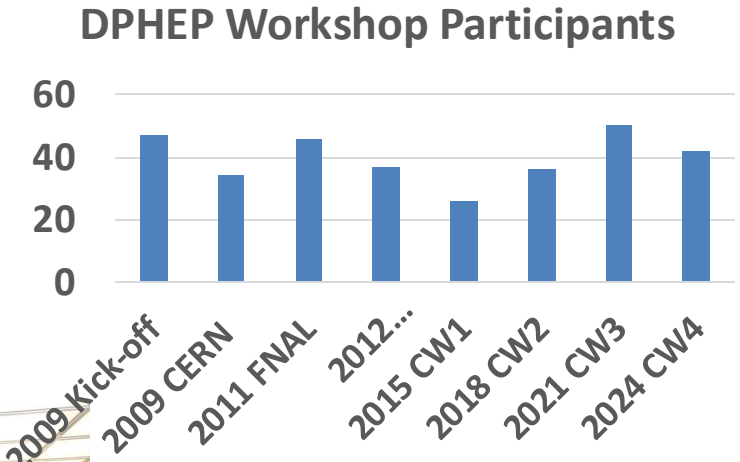
- Planning for preserved data improves the design of running and future experiments
- DP relies on and stimulates cutting-edge technology developments
- DP is strongly linked to **Open Science and FAIR** data paradigms
  - F findable A accessible R reproducible
  - Most difficult to obtain is “I” : interoperability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

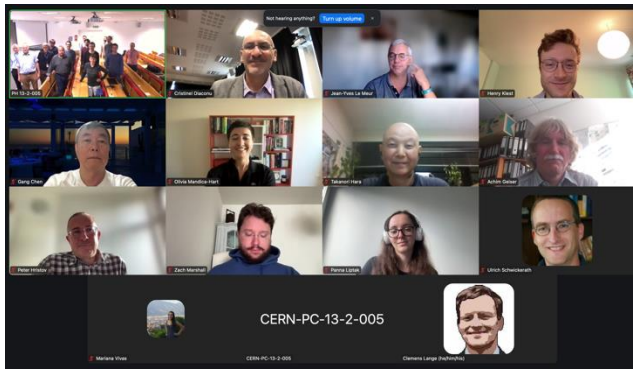




4th DPHEP Collaboration Workshop
Oct 2-3, 2024, CERN

https://indico.cern.ch/event/1432766/ report for DL panel





Highlights

- Significant progress has been made since the last workshop, both on data preservation and opening of data
- While CMS has pioneered the publication of open data, the other LHC experiments are rapidly catching up now: ATLAS, LHCb, ALICE
- Open data policies are increasingly applied also beyond LHC; useful synergy and *data opening calendar*
- **LEP Data is back!**

- In many contributions continued funding of DP was mentioned as an issue. An example is the BaBar experiment, whose software is running on outdated hardware but there is no funding to replace them
- Transfer across generations is visible
  - HERA ➔ EIC; LEP ➔ FCC
- Ongoing review of CERN OC3, which may offer some opportunities for DPHEP.
- The survey is encouraging, following a bottom up approach starting from the people involved in the practical work.
  - Document the output, echo to 10y report's open questions

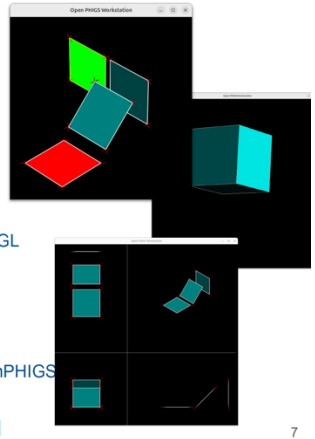
Highlights:LEP &co.

- LEP Data is back!
- CERNLIB rescued
- Visual OPAL and DELPHI via rescued PHIGS alpha version
- ALEPH Data ➔ EDM4hep (➔ FCC-ee)

OpenPHIGS for data preservation

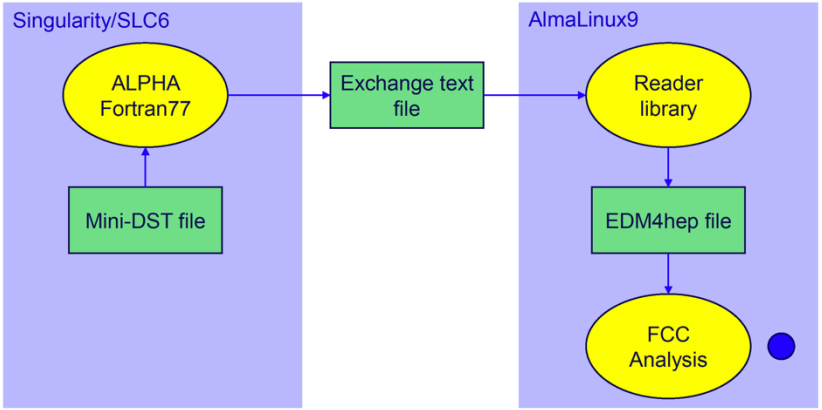
Caveats:

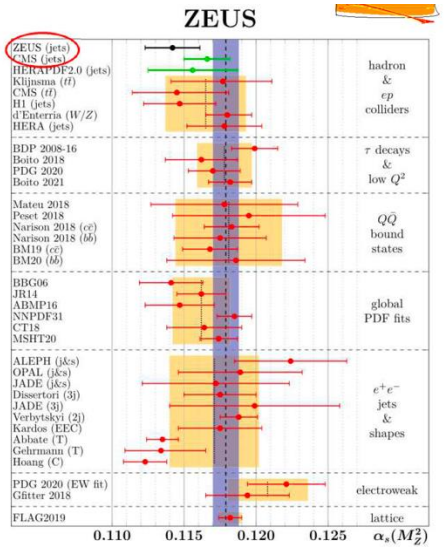
- Had to drop some functionality
  - Printing as PostScript is not possible
  - Replaced by TGA or PNG which is not the same resolution wise
- Still incomplete implementation
- Still bugs to be solved
  - E.g. filling only works for concave surfaces, as is the case in OpenGL
- Not fully compliant to the standard
- Still at pre-alpha level in terms of code maturity
- Still no documentation
  - Could import a subset of the man pages from the PEX based OpenPHIGS implementation
- Eventually not the same look and feel as with the original



7

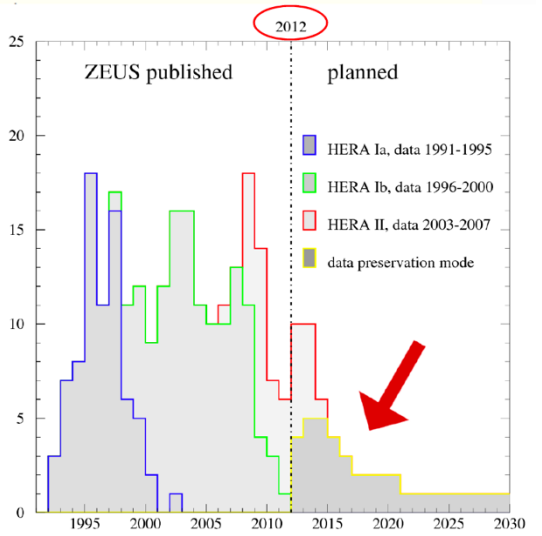
Workflow





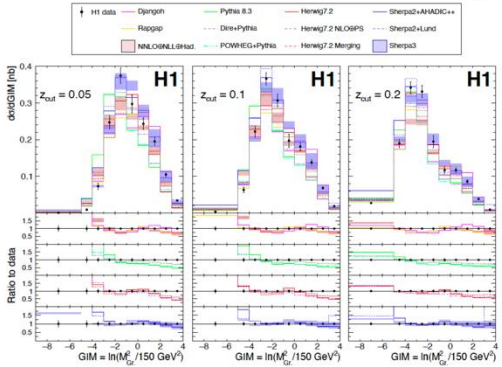
HERA: olympic shape

Two DP model, very similar succes s
Attract new collaborators (17 years after the end of collisions)

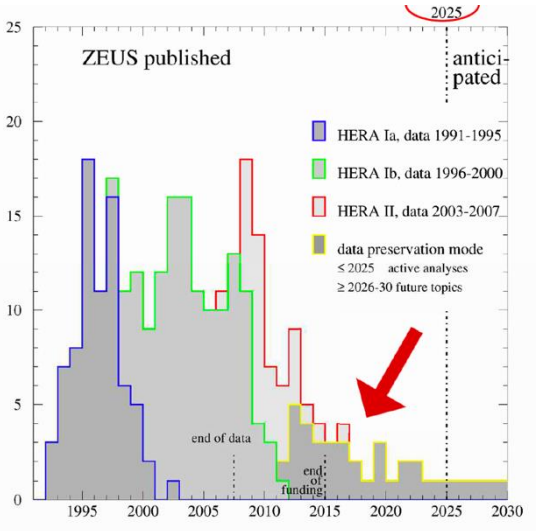


Recent physics results (2024)

- Recent H1 results: investigations of the hadronic final state
  - Jet substructure (2023)
  - Angular moments (2023)
  - Empty hemisphere events
  - 1-jettiness eventshape
  - Groomed eventshapes: [2403.10134], EPJ C84 (2024) 718



Groomed invariant mass and 1-jettiness are measured. First time grooming is tested in ep. Models have some difficulties to describe our data.



BaBar data: active, travelling, community support/help/rescue, fragile

Data Access

- Data available to analyses: ~1.5PB
  - no storage available at UVic for it
- GridKa agreed to host the BaBar data to be used by analyses
  - BaBar site since a long time
    - had already some data on site; anything missing was copied to GridKa
  - also the metadata db to find the data files needed in an analysis was already there
    - only update of content needed
  - BaBar environment configuration specifies XRootD and db access point for the data and db
- Framework at UVic needs to access data at GridKa via streaming...
  - works surprisingly well for normal event data
    - workflow: read event, process, read event, process,...
  - but conditions data is also read via streaming
    - large amount of data each job needs to read

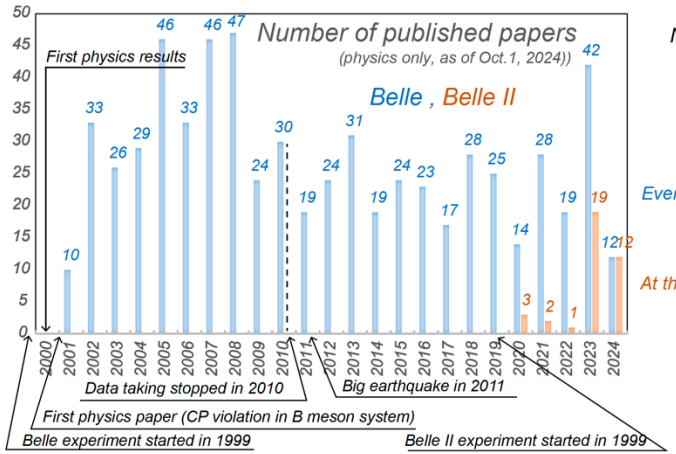
Hardware overview:

- XRootD proxy server: old machines
- XRootD redirector: VM on an old machine
- login machine: VM on an old machine
- BaBar interactive VM: VM on an old machine
- NIS server: VM on an old machine
- web server: on VM on an old machine
- babar wiki: VM on an old machine
- babar Hypernews: VM on an old machine
- NFS server: one new server, multiple old machines

Redundancy/Reliability:

- protect against disk failure
- protect against server failure

Active physics analysis with Belle data



Not in the left plot (Belle)
Accepted : 2
Submitted : 9
To be submitted : 7
(and many on-going analyses)

*Even after 14 years, **Belle's data** is still being analyzed **vigorously**.*

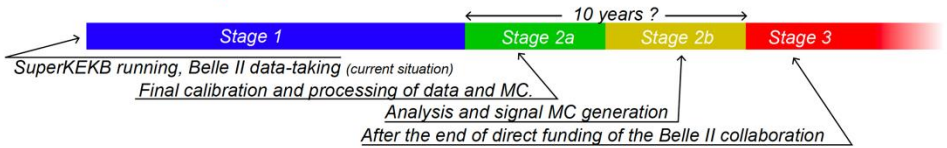
*At the same time, **data analysis of Belle II** has become **more active***

DP / AP activities in Belle II

in June 2021, Belle II formed the Data Preservation Task Force

Charge

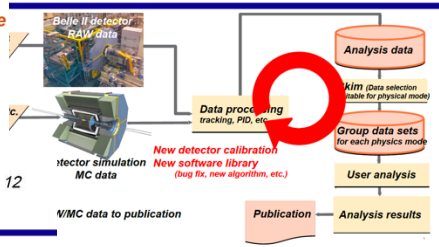
1. *The expected impact of the data preservation plan on Belle II Physics publications,*
2. *The computing model required to enable the preservation plan, including raw data reprocessing and MC production, both in the post-SuperKEKB-running period and the post-Belle II lifetime,*
3. *The data that should be preserved,*
4. *The period of time for accessibility of the preserved data,*
5. *The analysis infrastructure that should be preserved,*
6. *The estimated cost and effort of Belle II data and analysis preservation, and*
7. *The outreach potential of open Belle II data*



The Task Force presented Belle II with four priority recommendations for consideration

for detail, please check https://indico.belle2.org/event/7653/contributions/44071/

Difficulties on DP / AP



Extra Cost (budget, human)
Extra Computing resource

Extra Efforts

- *Data archive with keeping accessibility*
- *Analysis preservation*
- *Common data format (HEP data is too complicated)*

Fair evaluation in the collaboration ?
Career path ?

d of experiment dedicated direct funding,
will continue to manage them? (no-longer collaboration)

A sustainable "business" model that incorporates long-term data / analysis preservation is necessary

v to make "data preservation" attractive for young researchers and/or researchers in other field e.g. informatics?)

Data preservation

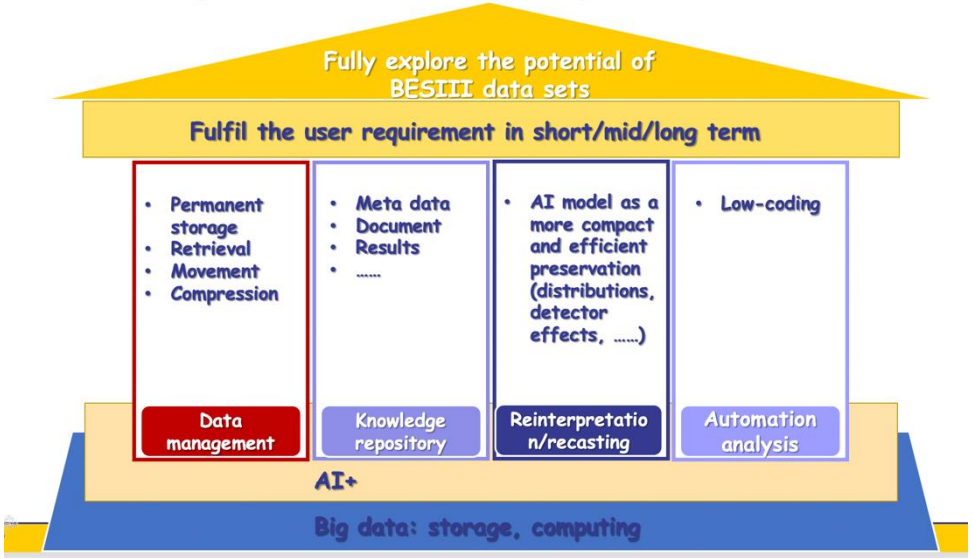
- BESIII adopts DPHEP Level-4 model
  - The full potential of data
    - RAW, DST (data, incl.MC),
    - metadata (calibration databases, ……),
    - software, documents
  - Adhere to the FAIR principles

DPHEP Collaboration: T. Basaglia, M. Bellis J. Blomer et al.: Data Preservation in High Energy Physics
Eur.Phys.J.C 83 (2023) 9, 795

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

BES III

AI-empowered data ecosystem for BESIII



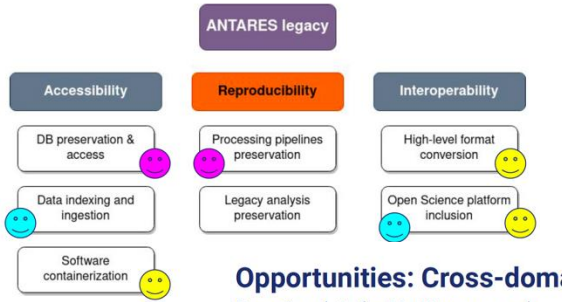
Antares : astro-neutrinos

Who is doing it

Funding and opportunities - first considerations

Funding options

- Partially integratable in KM3NeT infrastructure development (INFRADEV2)
- Included in ACME call (HORIZON-INFRA-2023-SERV-01), currently starting
  - *4.2.10. Access to neutrino data of ANTARES telescope.*
  - *4.3.2. ANTARES and KM3NET neutrino telescope data analysis services*
- Still looking for funding

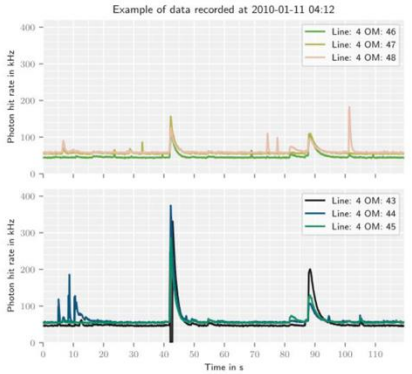


Opportunities: Cross-domain research

Deep-Sea data for Maritime research

Database preservation & access

- Database contains valuable information on environmental conditions in the Deep Sea and bioluminescence rate
- ORACLE database hard to preserve (supported versions, licensing …)
- Considering containerization
- Aiming to provide interface for Maritime Science: DEEPSEA project @OSCARS, not funded



Studying bioluminescence flashes with the ANTARES deep-sea neutrino telescope. Limnol Oceanogr Methods, 21: 734-760. https://doi.org/10.1002/lom3.10578

Conclusions

- 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 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)

bckp

DPHEP (Hi)story

Agenda

Wednesday Oct 2nd

- 14:25 **ALEPH**; Jacopo Fanini
- 14:45 **CERNLIB;** Andrii Verbytskyi, Ulrich Schwickerath
- 15:00 **DELPHI** ; Dietrich Liko, Dr Ulrich Schwickerath
- 15:15 **OPAL** ; Matthias Schroeder
- 15:30 **DELPHI and OPAL event displays**; M.Schroeder

- 15:45 → 16:00 Preserved Coffee

- 16:00 **ZEUS**; Achim Geiser
- 16:20 **H1** ; Speaker: Henry Klest
- 16:40 **JADE** Andrii Verbytskyi /Richard Hildebrandt
- 17:00 **PHENIX** ; Maxim Potekhin
- 17:20 **BaBar** ; Marcus Ebert

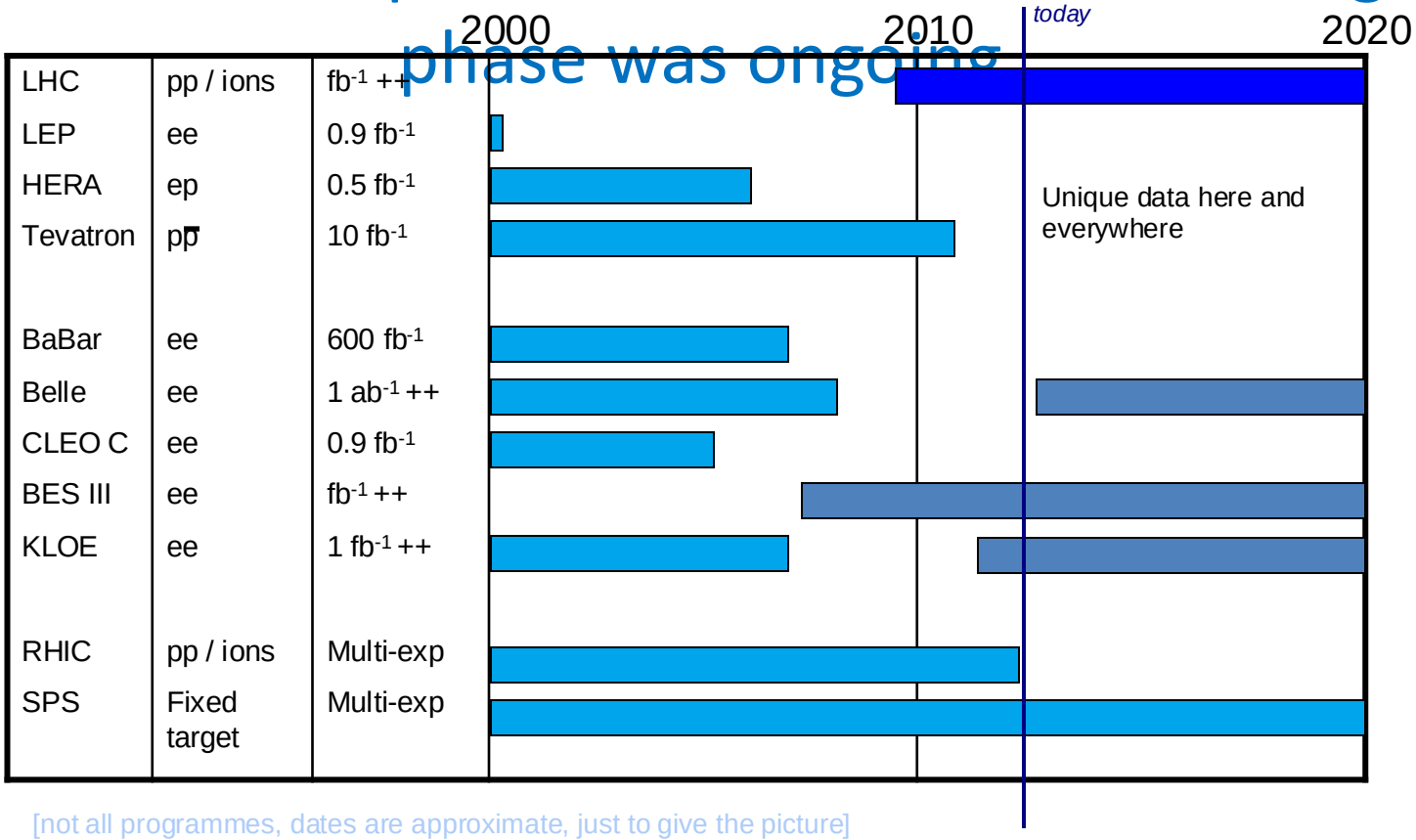
Thursday Oct 3rd

- 09:00 **KEK / Belle I & II** ; Takanori Hara
- 09:20 **BESIII** Gang Chen
- 09:40 **CERN Open Data portal** Pablo Saiz
- 10:00 **REANA** Marco Donadoni (CERN)
- 10:20 **CERN Analysis Preservation porta** P. Fokianos
- 10:45 → 11:00 Preserved Cofee
- 11:00 **CERN Open Data: Policy/implementation**; J. Boyd
- 11:20 **ALICE** ; David Dobrigkeit Chinellato
- **11:40 ATLAS**; Zach Marshall
- 12:00 **LHCb**; Dillon Fitzgerald
- 12:30 → 14:00 Preserved Lunch
- 14:00 **Preserving ANTARES legacy data** ; Jutta Schnabel
  14:20 PUNCH4NFDI ; Achim Geiser
- 14: **40 CMS** ; Julie Hogan
- 15:00 **ICFA Data Lifecycle Panel** ; Kati Lassila-Perini
- 15:25 **DPHEP Collaboration**

Open Questions (homework)

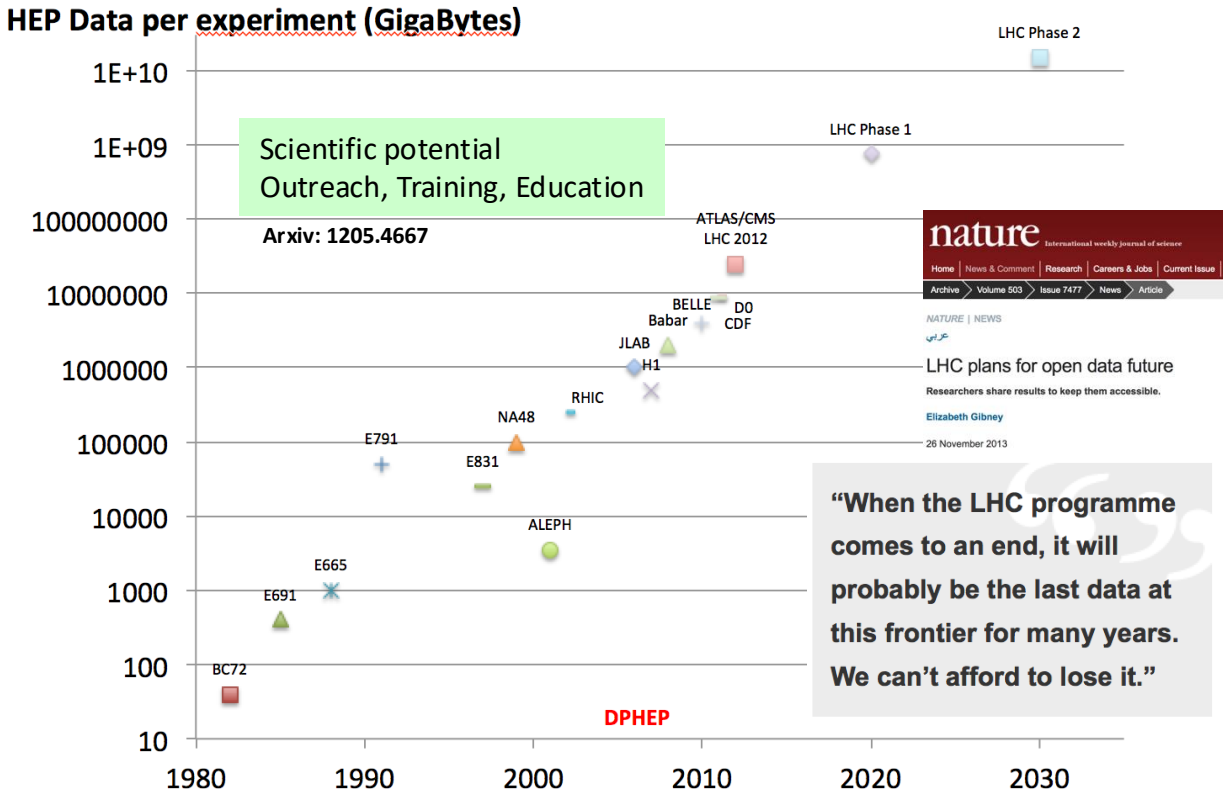
- 1. Why the systems did not collapse after the data taking? The "common sense: "publish your last paper and leave".
  - Still, a small but motivated community voluntarily kept data alive for many years and extracted unique science from it, beyond the "local ntuples" philosophy that eventually perpetuates 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 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?

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



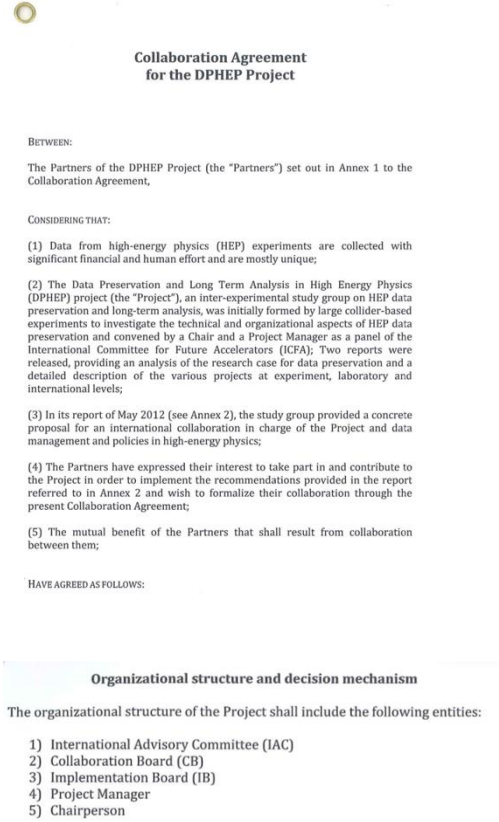
[not all programmes, dates are approximate, just to give the picture]

HEP Data


HEP Data per experiment (GigaBytes)
1E+10
1E+09
100000000
10000000
1000000
100000
10000
1000
100
10
1980
1990
2000
2010
2020
2030
LHC Phase 2
LHC Phase 1
Scientific potential
Outreach, Training, Education
Arxiv: 1205.4667
ATLAS/CMS
LHC 2012
BELLE
D0
Babar
CDF
JLAB
H1
RHIC
NA48
E791
E831
ALEPH
E665
E691
BC72
DPHEP
nature
International weekly journal of science
Home
News & Comment
Research
Careers & Jobs
Current Issue
Archive
Volume 503
Issue 7477
News
Article
NATURE | NEWS
LHC plans for open data future
Researchers share results to keep them accessible.
Elizabeth Gibney
26 November 2013
"When the LHC programme comes to an end, it will probably be the last data at this frontier for many years. We can't afford to lose it."

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;

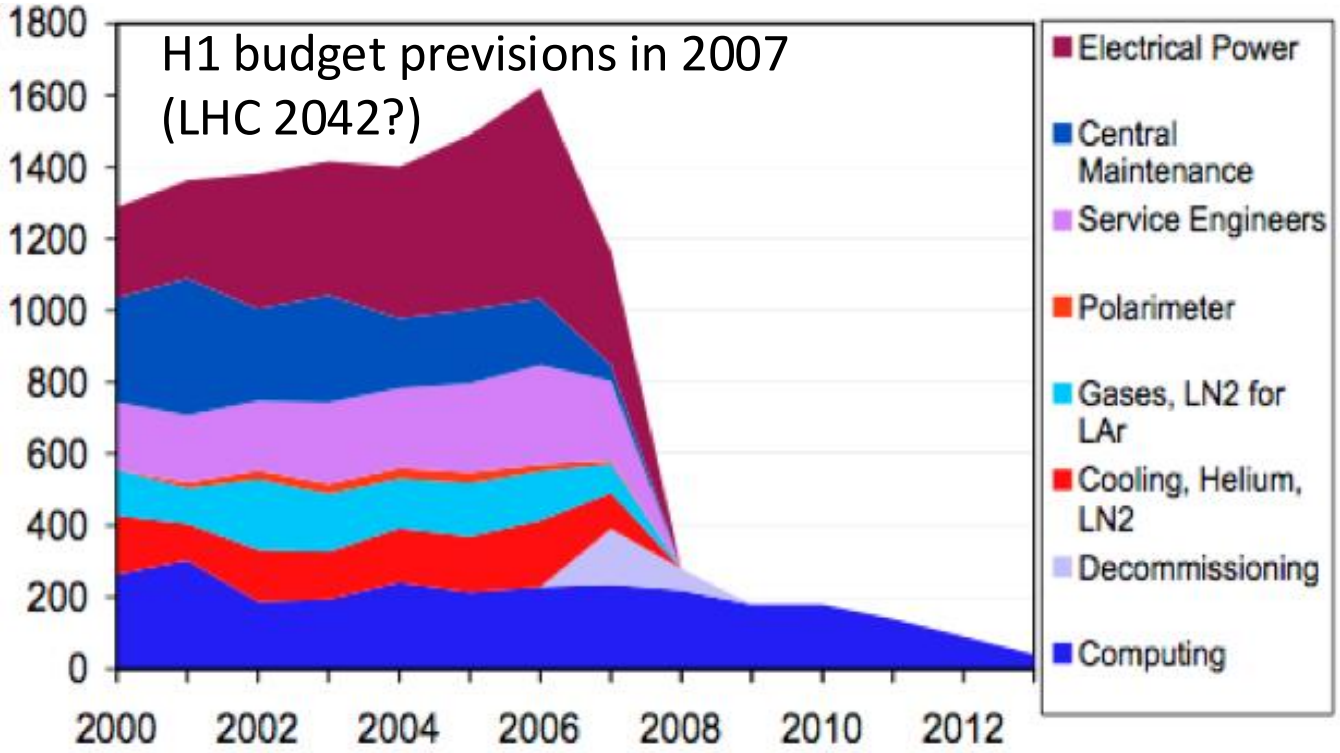
HAVE AGREED AS FOLLOWS:

Organizational structure and decision mechanism

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

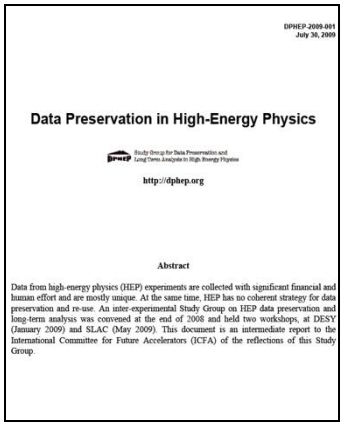
1) International Advisory Committee (IAC)
2) Collaboration Board (CB)
3) Implementation Board (IB)
4) Project Manager
5) Chairperson

When it stops taking data



DPHEP Study Group (2009)

> *arXiv:0912.0255*



DPHEP-2009-001
July 30, 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. At the same time, HEP has no coherent strategy for data preservation and re-use. An inter-experimental Study Group on HEP data preservation and long-term analysis was convened at the end of 2008 and held two workshops, at DESY (January 2009) and SLAC (May 2009). This document is an intermediate report to the International Committee for Future Accelerators (ICFA) of the reflections of this Study Group.

- 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

2009 Report

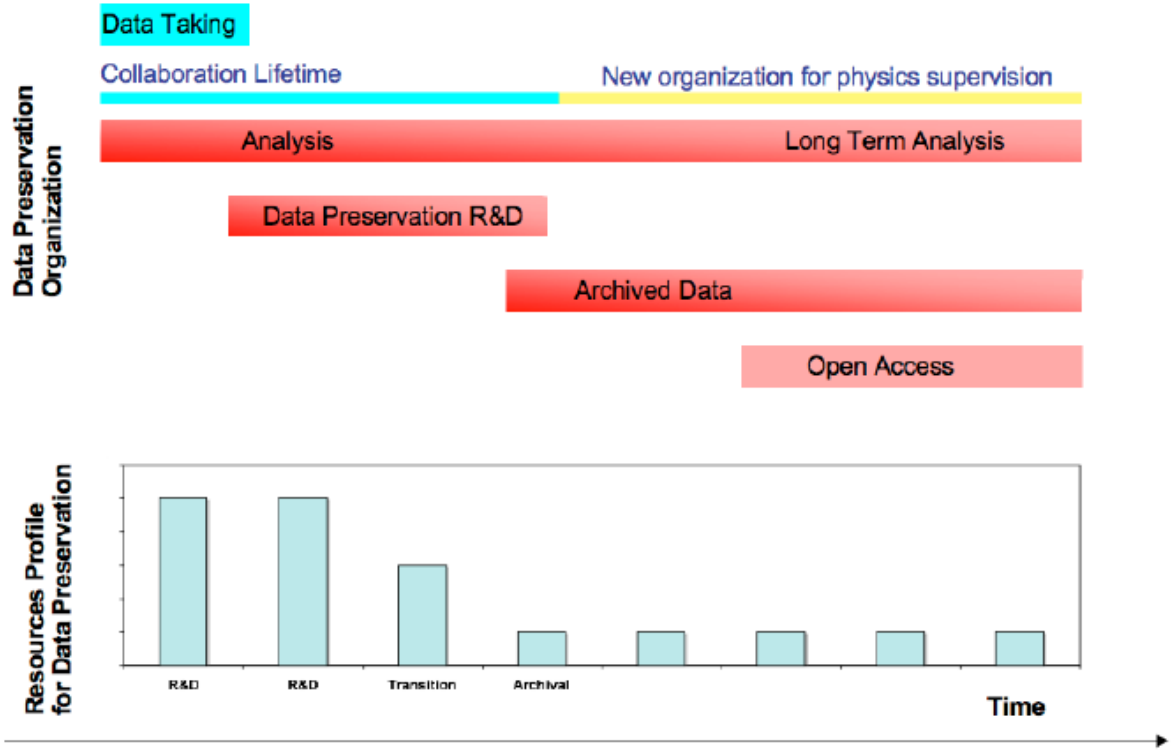


Figure 1: A possible model for data preservation organisation and resources presented as the milestones of the organisation and the resources evolution as a function of time.

Recent developments

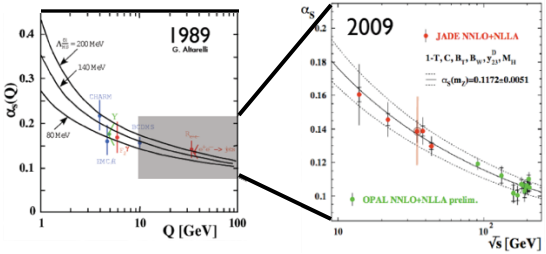
- New ICFA panel, enlarging the scope:
  - "ICFA Panel on the Data Lifecycle"
  - Mission:
    - ….enhance global coordination on all aspects of the data lifecycle including acquisition, processing, distribution, storage, access, analysis, simulation, preservation, management, software, workflows, computing 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).

Experiments Status

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



2021

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

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

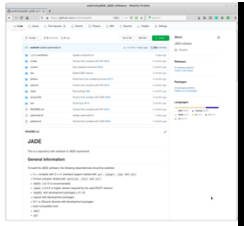
Aug 23, 2022

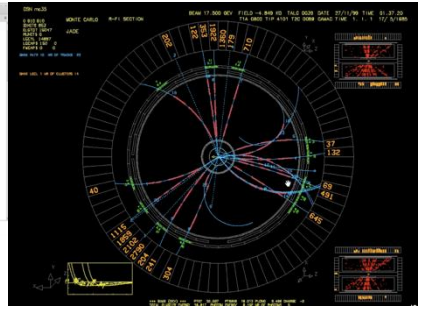
58 pages
Published in: *Eur.Phys.J.H* 47 (2022) 16
e-Print: 2208.11076 [hep-ex]

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 CentOS8[NEW] and MacOSX10.15+ on x86_64[NEW]



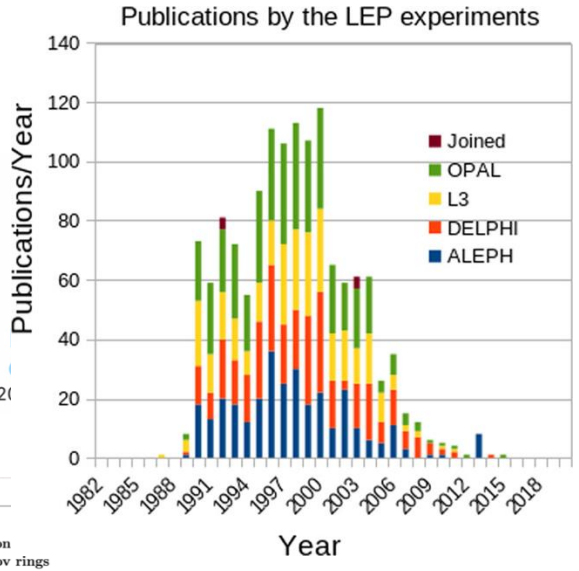


LEP

Papers using archived data

LTDP @LEP: Big Data Today – Peanuts Tomorrow

New physics with Archeodata

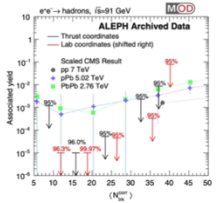


PHYSICAL REVIEW LETTERS **123,** 212002 (2019)

2019

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

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



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

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

2020

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

Takuya Mizoguchi[1] and Minoru Biyajima[2]
[1]National Institute of Technology, Toba College, Toba 517-8501, Japan
[2]Department of Physics, Shinshu University, Matsumoto 390-8621, Japan

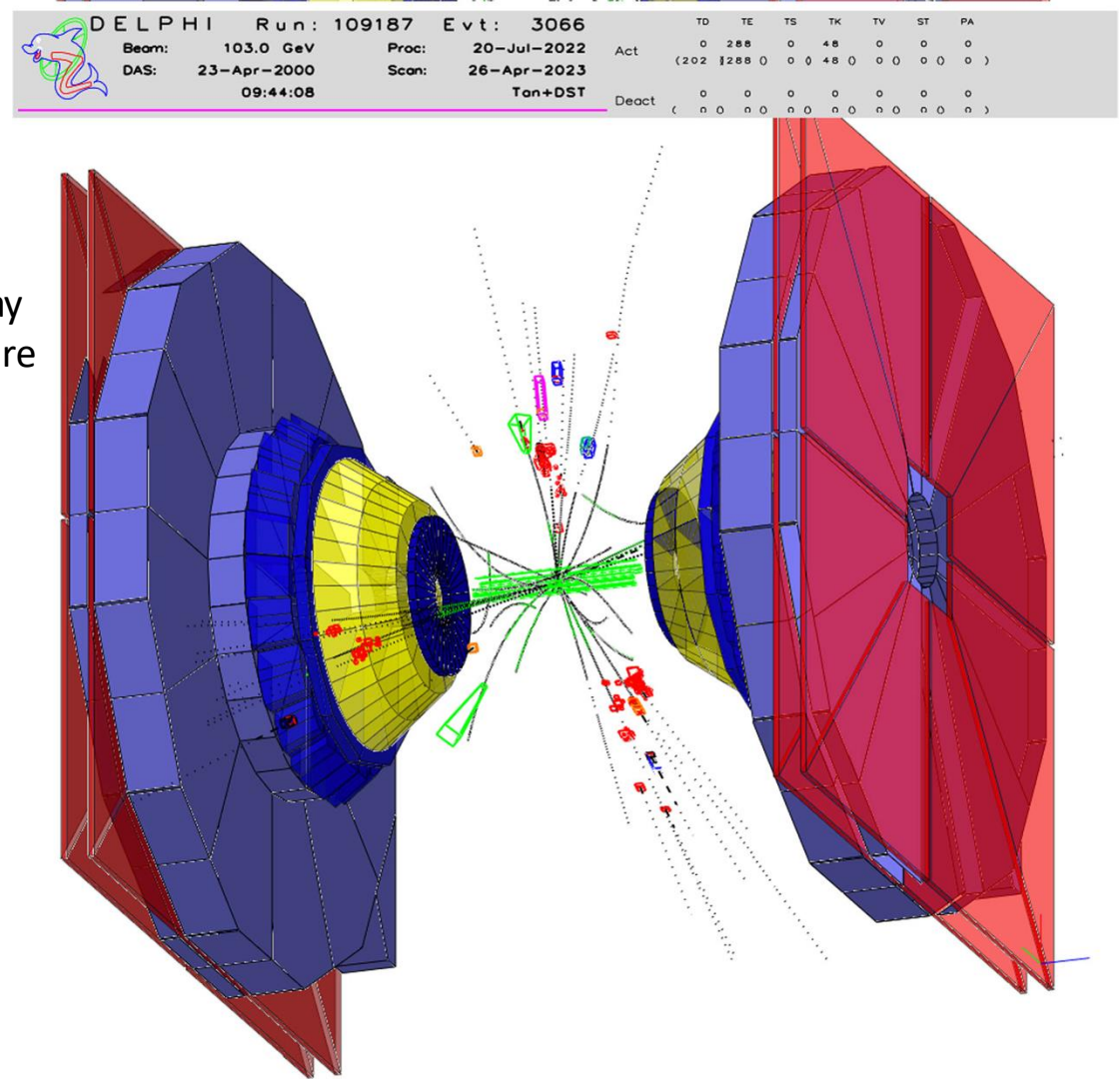
February 23, 2021

2021

➔ FCCee

➔ With real data from LEP

DELPHI event display with revised software

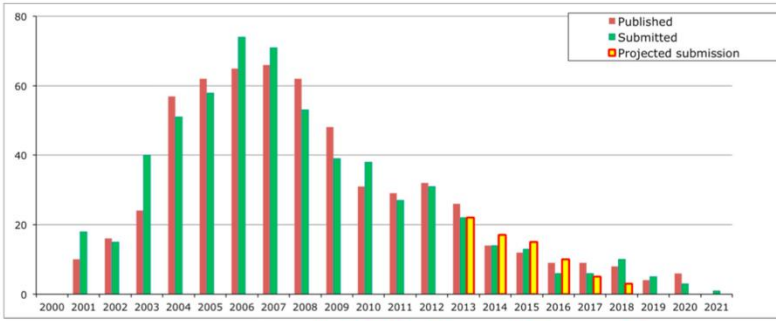


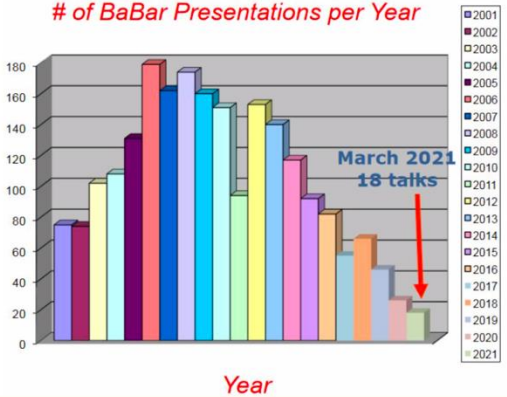
Babar (03/2021)

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

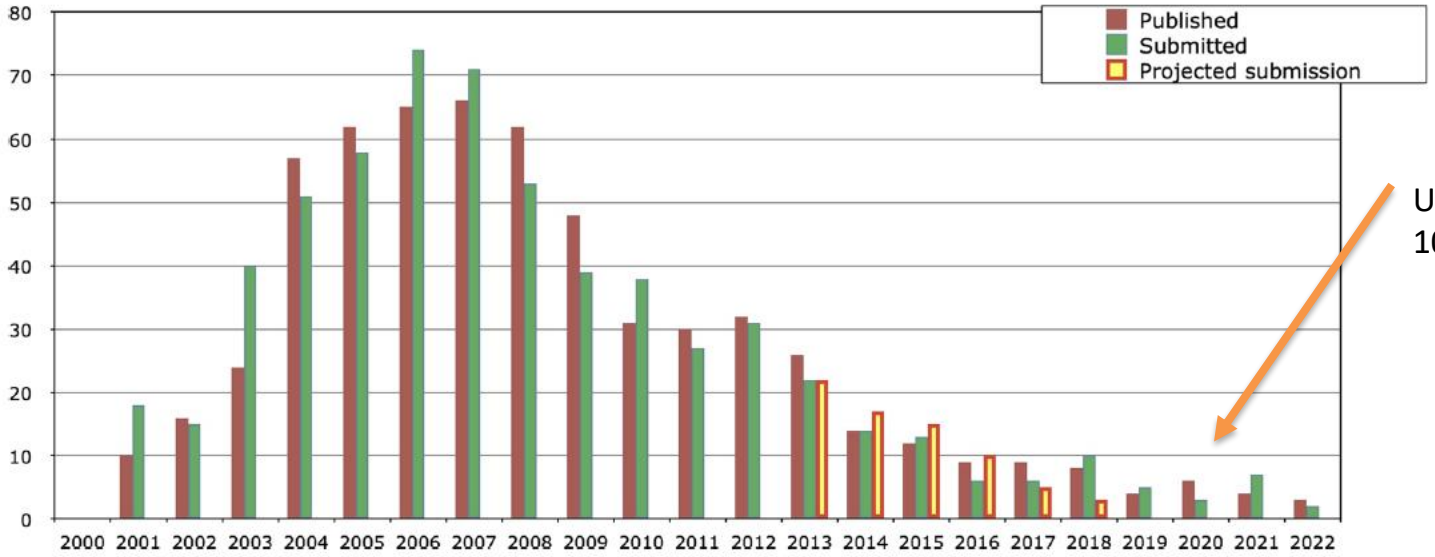




But: SLAC LTDA decommissioned, moving to U. Victoria/CERN/CC-IN2P3/Grid-Ka

Open Data decided

Update of the BaBar publication skyline



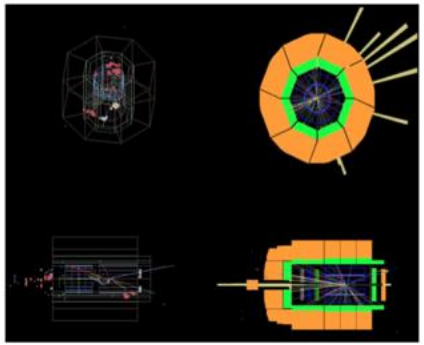
Unexpected
10 years ago

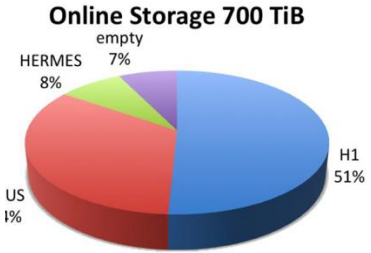
Fig. 6 *BABAR* submitted (green) and published (red) papers per year. In 2012 predictions for submissions (yellow) were made for the years 2013 to 2018. In 2012 it was predicted that no analysis would run after 2018

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

'H1Red' for simulated Pythia8.3 event



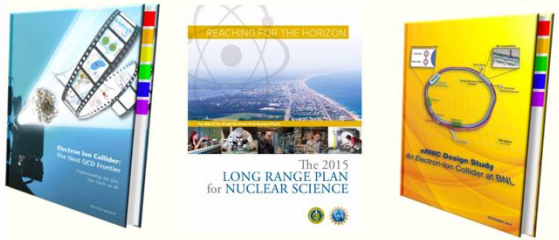


  - New topics/collaborators (EIC)

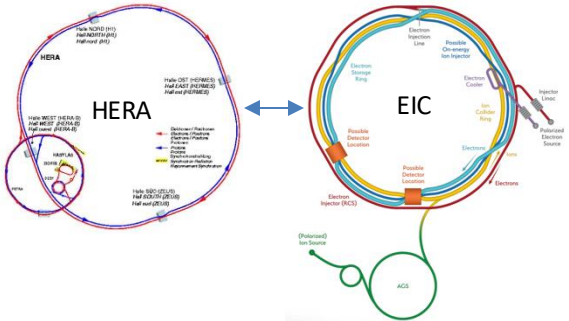
- ZEUS : “Level 3/4” DPHEP strategy
  - Root ntuples produced in the preparatory phase
  - easy to maintain/use/test/open

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



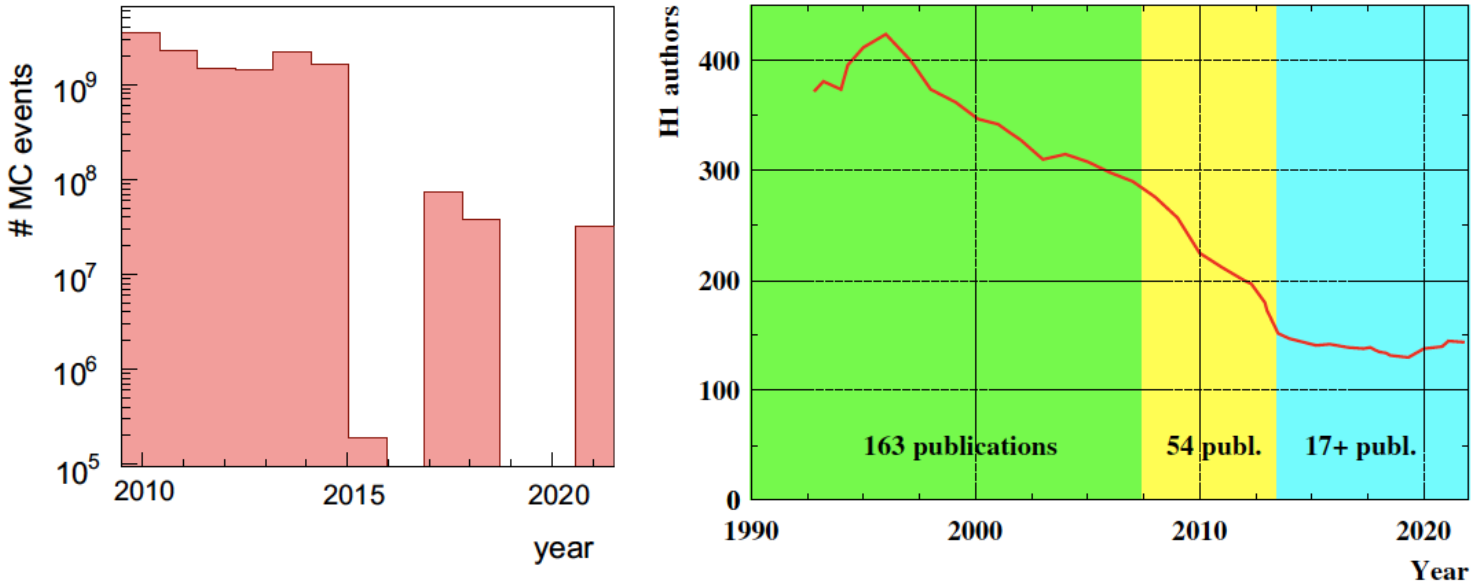


Fig. 4 Left: Number of Monte Carlo events produced centrally by the H1 Collaboration. The years without MC production are related to a change of the computing environment, or no MC requests. Right: Number of H1 authors is increasing since 2019 due to retained analysis capabilities and new interest in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

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 Foundation[1]

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

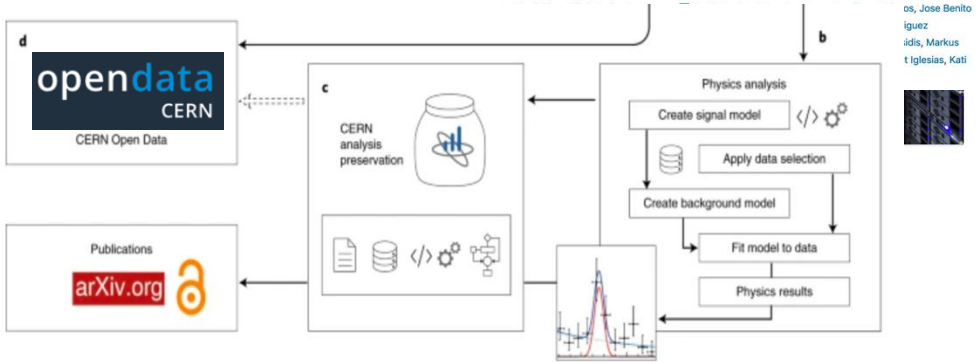
Explore Content · Journal Information · Publish With Us

nature › nature physics › perspectives › article

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

Perspective | Open Access | Published: 15 November 2018

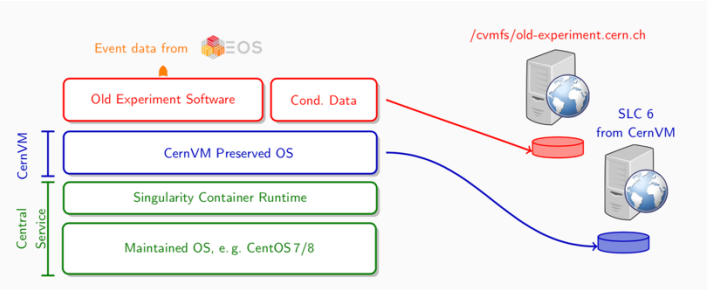
Open is not enough



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

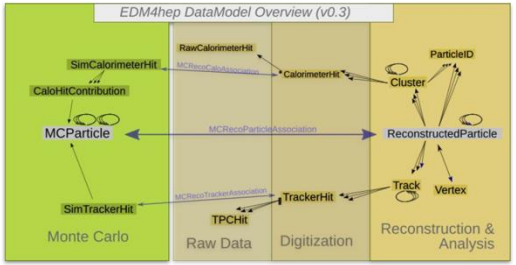
Towards more standards

CERNVM: the “freezer”



EDM4hep: the common language

- 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



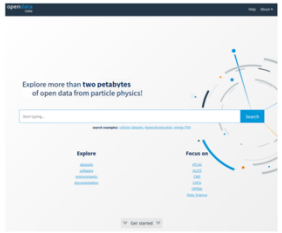
key4hep / EDM4hep and DPHEP?

- 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

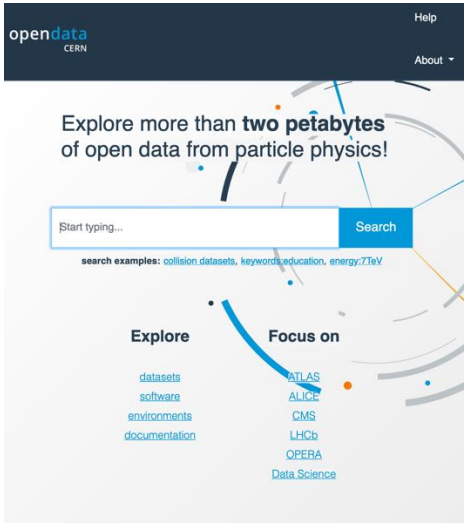
- CMS completed 2010–11 proton-proton data; released half of 2012 data; released 2010–11 heavy-ion data samples and corresponding pp reference datasets
- ATLAS released 13TeV educational samples
- Data science and Machine Learning (CMS, LHCb...)
- Non-LHC physics: OPERA neutrino physics data; interest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



`https://opendata.cern.ch`

```
/tmp $ cernopendata-client download-files --recid 5500 --verify
==> Downloading file 1 of 11
 -> File: ./5500/BuildFile.xml
 -> Progress: 0/0 kiB (100%)
==> Verifying file BuildFile.xml...
 -> expected size 305, found 305
 -> expected checksum adler32:ff03668a, found adler32:ff03668a
```

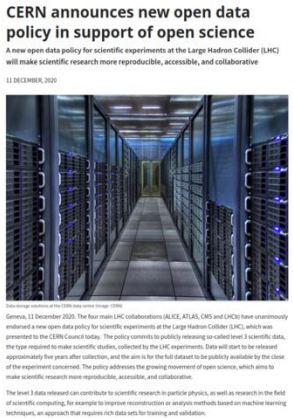
Command-line client to ease data download



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).
- Automatise 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

Data storage solutions at the CERN data centre (image: CERN)

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 start 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 Us

nature > nature physics > perspectives > article

Perspective | Open Access | Published: 15 November 2018

Open is not enough

Xiaoli Chen, Sünje Dallmeier-Tiessen, Robin Dasler, Sebastian Feger, Pamfilos Fokianos, Jose Benito Gonzalez, Harri Hirvonsalo, Dinos Kousidis, Artemis Lavasa, Salvatore Mele, Diego Rodriguez Rodriguez, Tibor Šimko, 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

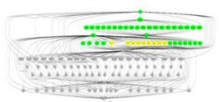
@tiborsimko 2 / 3

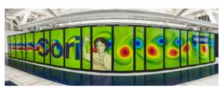
REANA reproducible analysis platform

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-producibility 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)

@tiborsimko 3 / 3

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)

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

- 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 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 ➔ 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.

A word on *FAIR*

- The DPHEP objectives (2012) intrinsically comply with what has became 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.

Discussion incentives

- Preservation and sharing/open:
  - Let data escape into unknown/unsual 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 continous output enhancement action around unique frontier science data
- DPHEP report 2022:
  - a strong interest to translate healthy and functional analysis sytems into open data hosts , HERA, BaBar, RHIC
    - main pb: Person power

- There is room to think and act in common and global

DPHEP ressources for DP

- 2012 Blueprint

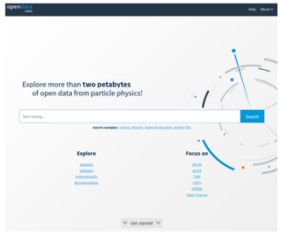
	Project	Goals and deliverables	Resources and timelines	Location, possible funding source, DPHEP allocation
Experiment and laboratory *Priority: 1*	Experimental Data Preservation Task Force	Install an experiment data preservation task force to define and implement data preservation goals.	1 FTE installed 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	A position jointly funded by a combination of laboratories and agencies.

Table 8: Resources required by projects of the DPHEP study group.

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

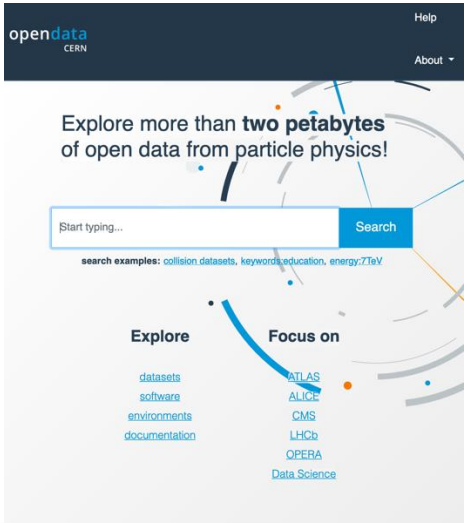
- CMS completed 2010–11 proton-proton data; released half of 2012 data; released 2010–11 heavy-ion data samples and corresponding pp reference datasets
- ATLAS released 13TeV educational samples
- Data science and Machine Learning (CMS, LHCb...)
- Non-LHC physics: OPERA neutrino physics data; interest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



`https://opendata.cern.ch`

```
/tmp $ cernopendata-client download-files --recid 5500 --verify
==> Downloading file 1 of 11
 -> File: ./5500/BuildFile.xml
 -> Progress: 0/0 kiB (100%)
==> Verifying file BuildFile.xml...
 -> expected size 305, found 305
 -> expected checksum adler32:ff03668a, found adler32:ff03668a
```

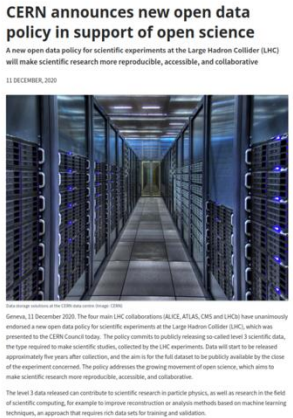
Command-line client to ease data download



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).
- Automatise 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

Data storage racks at the CERN data centre (image: CERN)

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 start 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 Us

nature > nature physics > perspectives > article

Perspective | Open Access | Published: 15 November 2018

Open is not enough

Xiaoli Chen, Sünje Dallmeier-Tiessen, Robin Dasler, Sebastian Feger, Pamfilos Fokianos, Jose Benito Gonzalez, Harri Hirvonsalo, Dinos Kousidis, Artemis Lavasa, Salvatore Mele, Diego Rodriguez Rodriguez, Tibor Šimko, 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

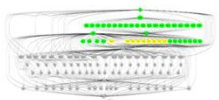
@tiborsimko 2 / 3

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-producibility 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

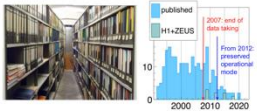

REANA running on supercomputers (e.g. NERSC)

@tiborsimko 3 / 3

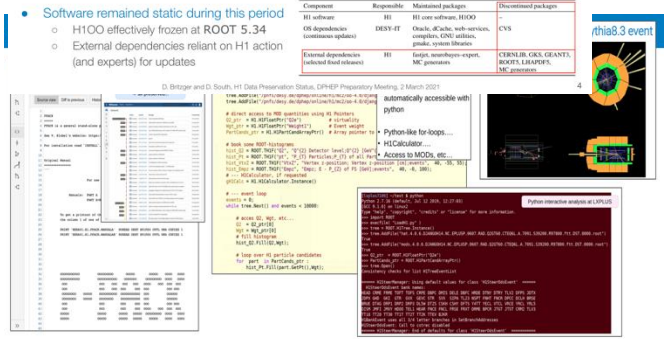
HERA: succesful DP, towards open data

Preserved Operational Mode: 2012-2020

- Significant "level 1" effort made on documentation
- New collaboration model: H1 Physics Board
- H1 physics publications continued
  - 28% of 233 papers after data taking ended
  - 66 after 2007, including 18 after 2012
- Level 4 preservation model includes recompilation of software and migration to newer OS
  - Main OS used in 2012: 32-bit Scientific Linux (DESY) 5, based on RHEL5 (EoL: March 2017)
- Migration to 64-bit SLD5 from 2012, careful work requiring detailed validation (*sp-system*, Vitaliy)
  - Successful move to 64-bit SLD6 (EoL: Nov 2020) and later CentOS7 made possible due to this effort
- Software remained static during this period
  - H1OO effectively frozen at ROOT 5.34
  - External dependencies reliant on H1 action (and experts) for updates



Component	Responsible	Maintained packages	Discontinued packages
H1 software	H1	H1 core software, H1OO	-
OS dependencies (continuous updates)	DESY-IT	Oracle, dCache, web-services, compilers, GNU utilities, gmake, system libraries	CVS
External dependencies (selected fixed releases)	H1	fastjet, neurobayes-expert, MC generators	CERNLIB, GKS, GEANT3, BOOST, LHAPDF5, MC generators

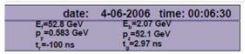


- H1 has unique data and continues to produce physics publications, long after data taking ended
- A recent software modernisation program has been performed to allow this to continue using modern analysis tools, recent programming languages and on state-of-the-art platforms

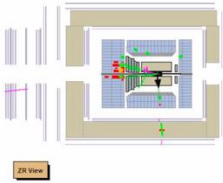
D. Britzger and D. South, H1 Data Preservation Status, DPHEP Preparatory Meeting, 2 March 2021 7

Common Ntuple analysis model

- ZEUS Common Ntuple: **Motto: keep it simple!**
  flat (simple) ROOT-based ntuple (same format as PAW ntuple converted with h2root)
  containing high level objects (electrons, muons, jets, energy flow objects, ...)
  as well as low level objects (tracks, CAL cells, ...)



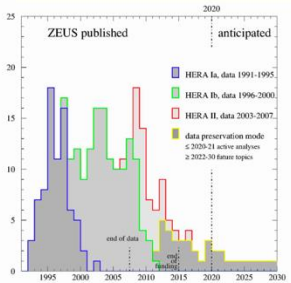
- Well tested !
  almost all recent ZEUS papers (24 out of 25) based on Common Ntuples
- Easy to maintain
  transition sl5 -> sl6 -> sl7 completely transparent
  (just use newer ROOT version)
- "Easy" to use
  most recent ZEUS papers based on results produced by master students, PhD students or postdocs from remote institutes, e.g. related to EIC or Heavy Ion communities, using resources at DESY or MPP:
  analysis on DESY NAF/BIRD computing farm
  or at MPI/Garching



- Low threshold for access to data by external groups

02. 03. 21 A. Geiser, DPHEP meeting 1

ZEUS physics papers



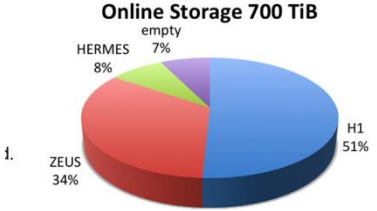
majority of papers produced
in "data preservation mode"
already since 2012 (25 papers)

since end of DESY funding 2014:
2015-20: 14 papers,
1 with > 500 citations
2021: expect 2-4 papers
long term: ~1-2 papers/year -> ~2030
expect ~10% of total ZEUS output
~80-90% of these would never exist without dedicated data preservation

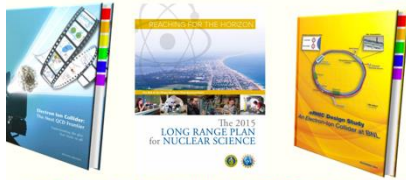
ZEUS data preservation program is a success!
some small official resources could double the output and/or allow Open Data

02. 03. 21 A. Geiser, DPHEP meeting 2

...periment: EIC

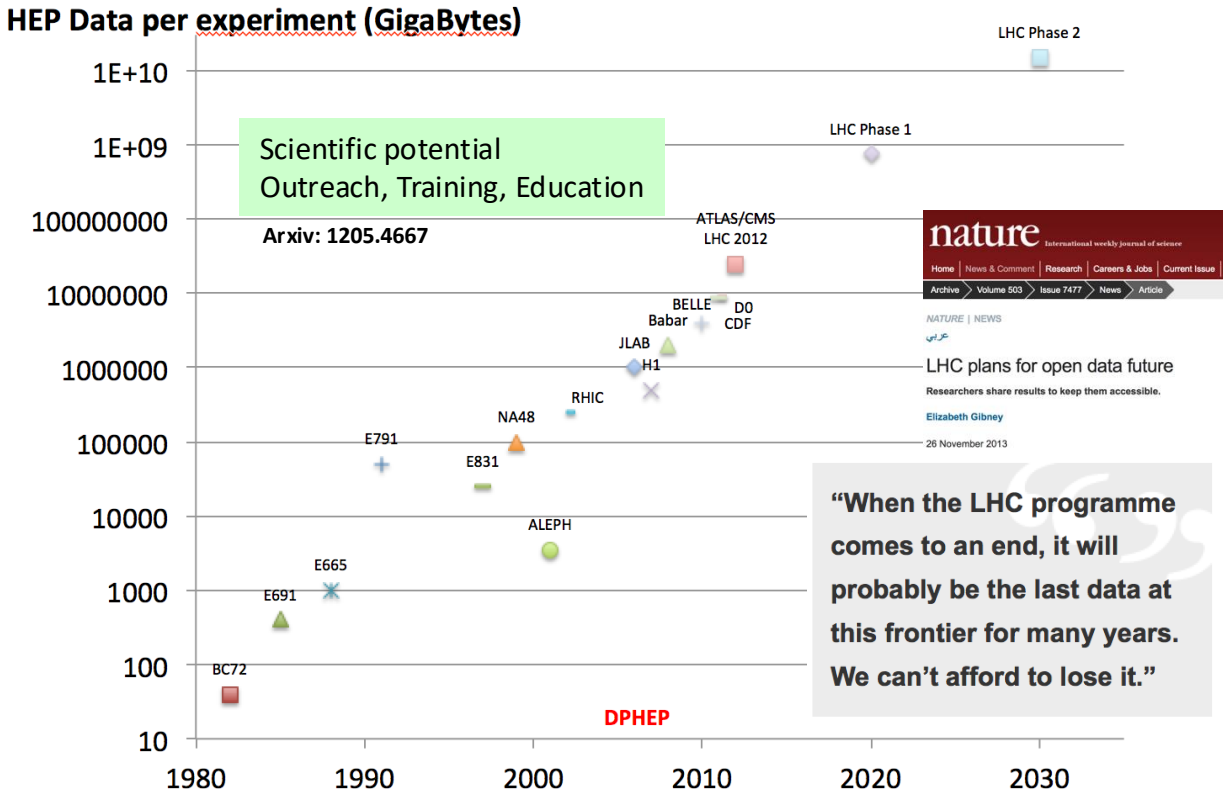


- many EIC topics common with HERA



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

HEP Data

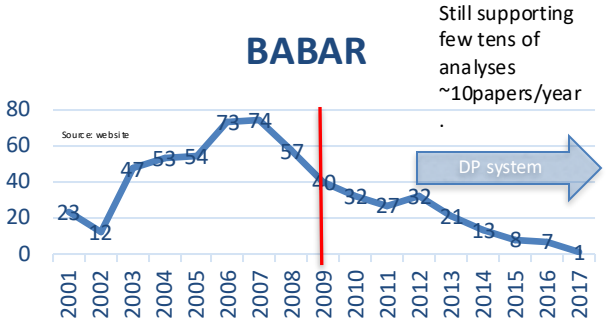

HEP Data per experiment (GigaBytes)
1E+10
1E+09
100000000
10000000
1000000
100000
10000
1000
100
10
1980
1990
2000
2010
2020
2030
LHC Phase 2
LHC Phase 1
Scientific potential
Outreach, Training, Education
Arxiv: 1205.4667
ATLAS/CMS
LHC 2012
BELLE
D0
Babar
CDF
JLAB
H1
RHIC
NA48
E791
E831
ALEPH
E665
E691
BC72
DPHEP
nature
International weekly journal of science
Home
News & Comment
Research
Careers & Jobs
Current Issue
Archive
Volume 503
Issue 7477
News
Article
NATURE | NEWS
عربي
LHC plans for open data future
Researchers share results to keep them accessible.
Elizabeth Gibney
26 November 2013
"When the LHC programme comes to an end, it will probably be the last data at this frontier for many years. We can't afford to lose it."

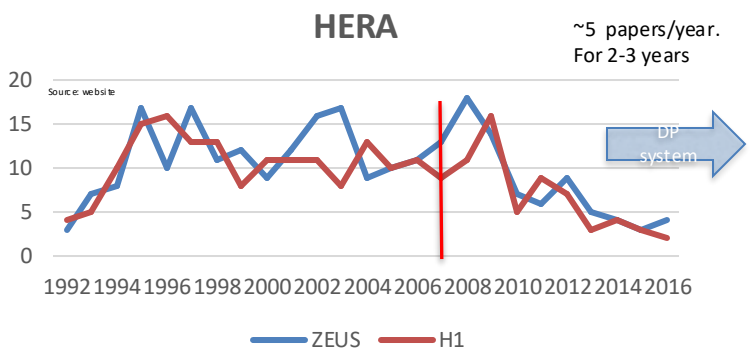
2018 status

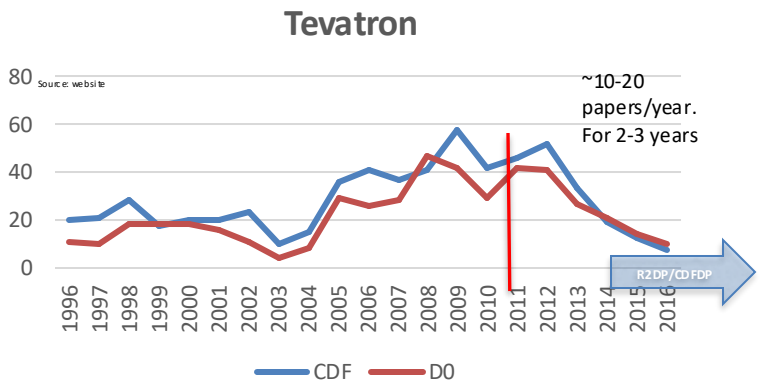
DPHEP timelines

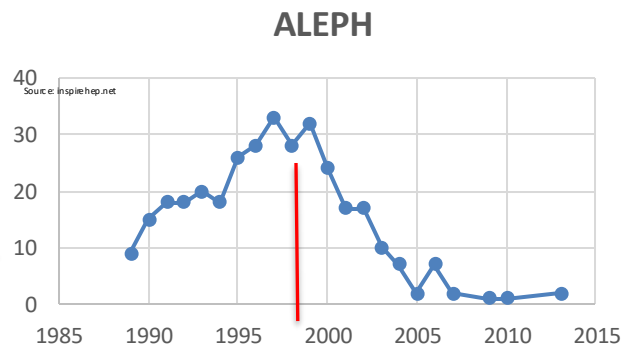
Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
	Start-up					Consolidation			DPHEP Collaboration		
HEP	HERA stops	Babar stops	LHC starts	Belle I stops	Tevatron stops				LHC Run 2		
DPHEP Group			ICFA Panel		LHC exp. joined	DPHEP Manger appointed at CERN		DPHEP Collaboration Agreements signed	1st DPHEP Collaboration Meeting		2nd 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/D0 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

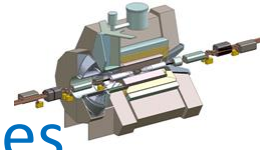








BABAR Highlights and Press Releases



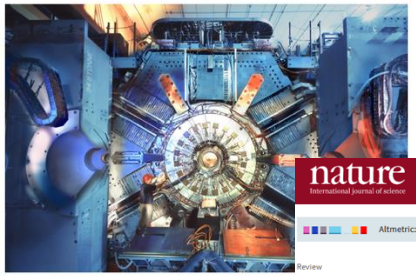
BERKELEY LAB
Bringing Science Solutions to the World

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

November 2017



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

INTERACTIONS.ORG
PARTICLE PHYSICS NEWS AND RESOURCES

New study: Scientists narrow down the sea... photons using decade-old particle collider d...

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

Physics ABOUT BROWSE PRESS COLLECTIONS

Viewpoint: New Light Shed on Dark Photons

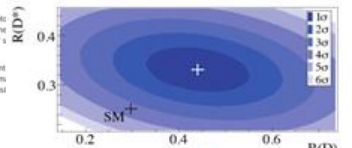
Douglas Bryman, University of British Columbia, Vancouver, British Columbia V6T2A3, Canada

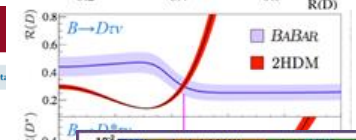
A search for a photonlike particle that could be related to dark matter has come up empty, putting new constraints on models that imagine a dark form of electromagnetism.

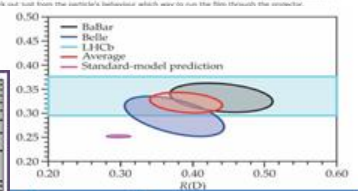
Search for a Dark Photon in e^+e^- Collisions at *BaBar*
J. P. Lees *et al.* (BaBar Collaboration)
Phys. Rev. Lett. 113, 201801 (2014)
Published November 10, 2014

Features
Q&A: A Condensed Matter Theorist Embraces AI
Juan Carrasquilla gave himself a crash course on machine learning and found a new way of approaching condensed-matter theory.
Meetings: Interplanetary GPS

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







nature
International Journal of science

Altmetric: 512 Citations: 2

Review

A challenge to lepton universality in B-meson decays

Gregory Ciezarek, Manuel Franco Sevilla, Brian Hamilton, Robert Kowalewski, Thomas Kuhr, Vera Lüth & Yutaro Sato

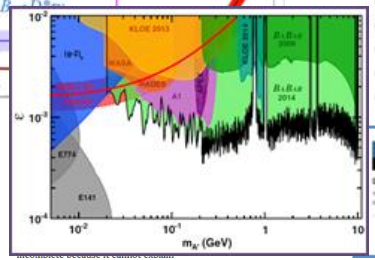
Nature 546, 227–233 (08 June 2017)
doi:10.1038/nature2...
Download Citation

Received: 15 December 2016
...d: 20 March 2017
...d online: 07 June 2017

June 2017

Experimental particle physics | Phenomenology | Theoretical particle physics

Abstract



physicstoday

Democracy suffers a blow—in particle physics

Dataset:
Y(4S): 433/fb
Y(3S): 30/fb
Y(2S): 14/fb
Off resonance: 10%
Y(1S) accessed via
$Y(2S,3S) \to Y(1S)\,\pi^+\pi^-$

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 Y(3S) resonance** (no plan at the moment to run at the Y(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**