Live-Open-Preserved Quo Vadis?

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

DPHEP Data Preservation in High Energy Physics
@DPHEPColl

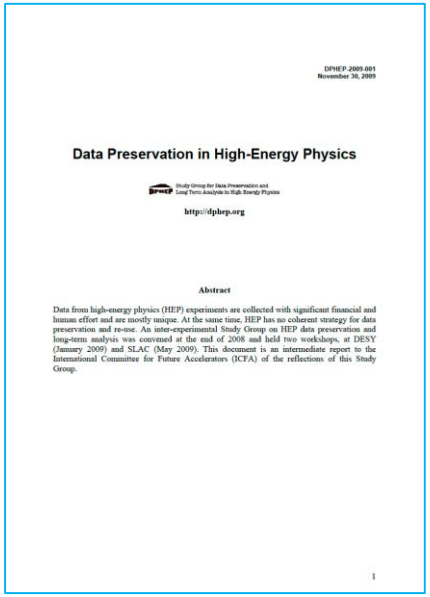
http://dphep.or

Cristinel DIACONU
CPPM/CNRS/Aix-Marseille University

EN 2023, LE CPPM CÉLÈBRE SES 40 ANS
40
CENTRE DE PHYSIQUE DES PARTICULES DE MARSEILLE
CPPM
AU CŒUR DE LA MATIÈRE ET DE L'UNIVERS

DPHEP

> arXiv:0912.0255



DPHEP-2009-001
November 30, 2009

Data Preservation in High-Energy Physics

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

http://dphep.org

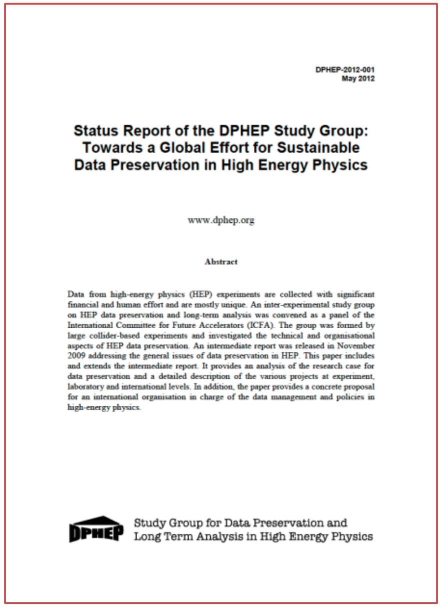
Abstract

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

LoI

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

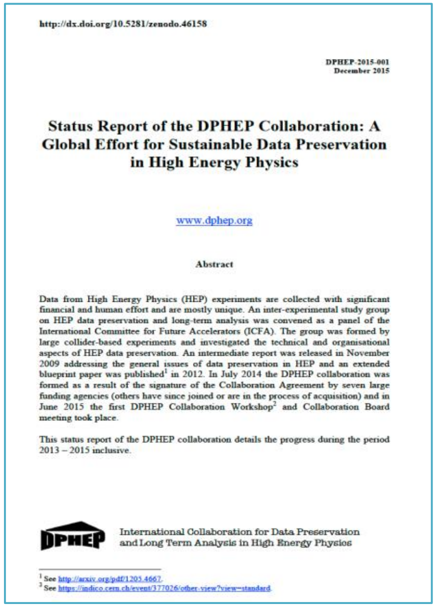
Abstract

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

Blueprint

- 1512.02019 [hep-ex]



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.

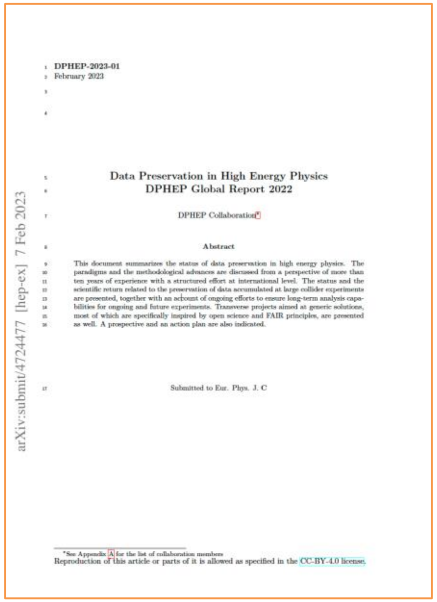
This status report of the DPHEP collaboration details the progress during the period 2013 – 2015 inclusive.

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.

Collaboration

- 2302.03583 [hep-ex]



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

DPHEP-2023-01
February 2023

Data Preservation in High Energy Physics
DPHEP Global Report 2022

DPHEP Collaboration[*]

Abstract

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

Submitted to Eur. Phys. J. C

[*] 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.

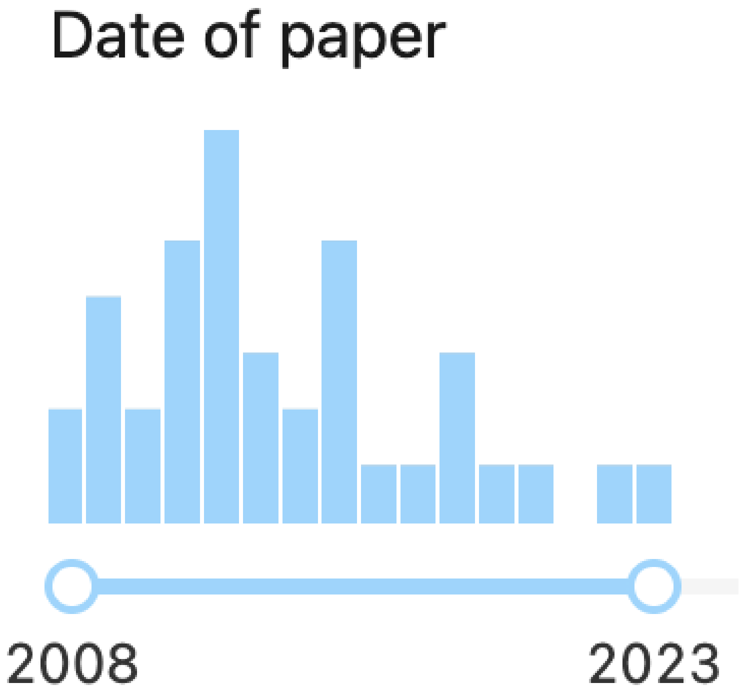
10 y

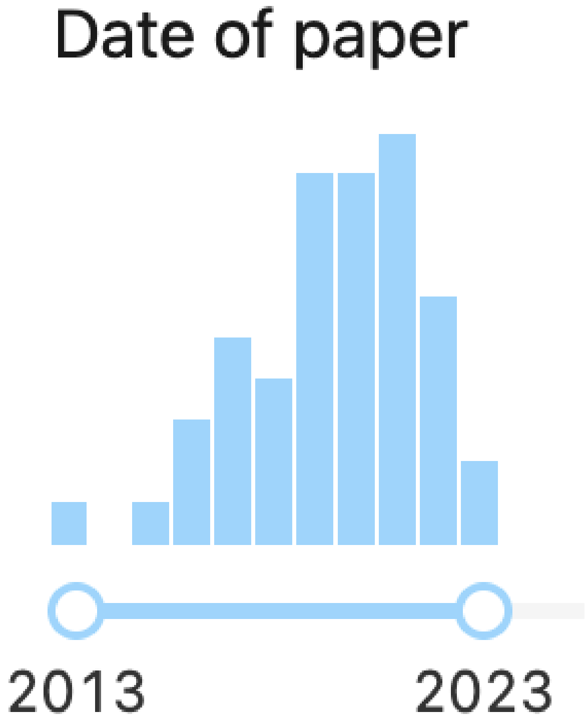
See Maxim's DPHEP talk FAIROS on Wed Dec 8

Open and preserved

find t **data preservation**
39 results

find t **open data**
50 results





2009 Report

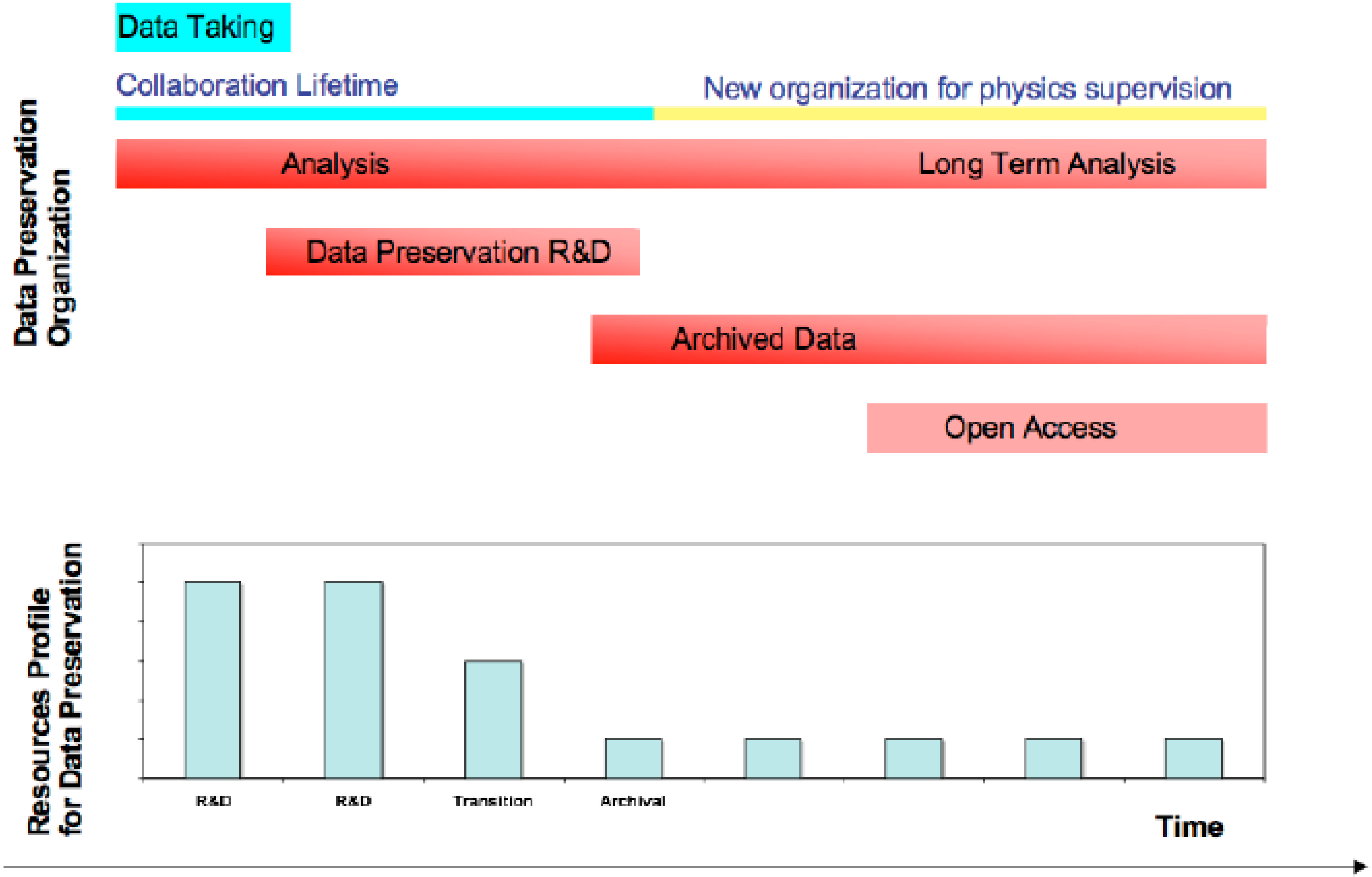
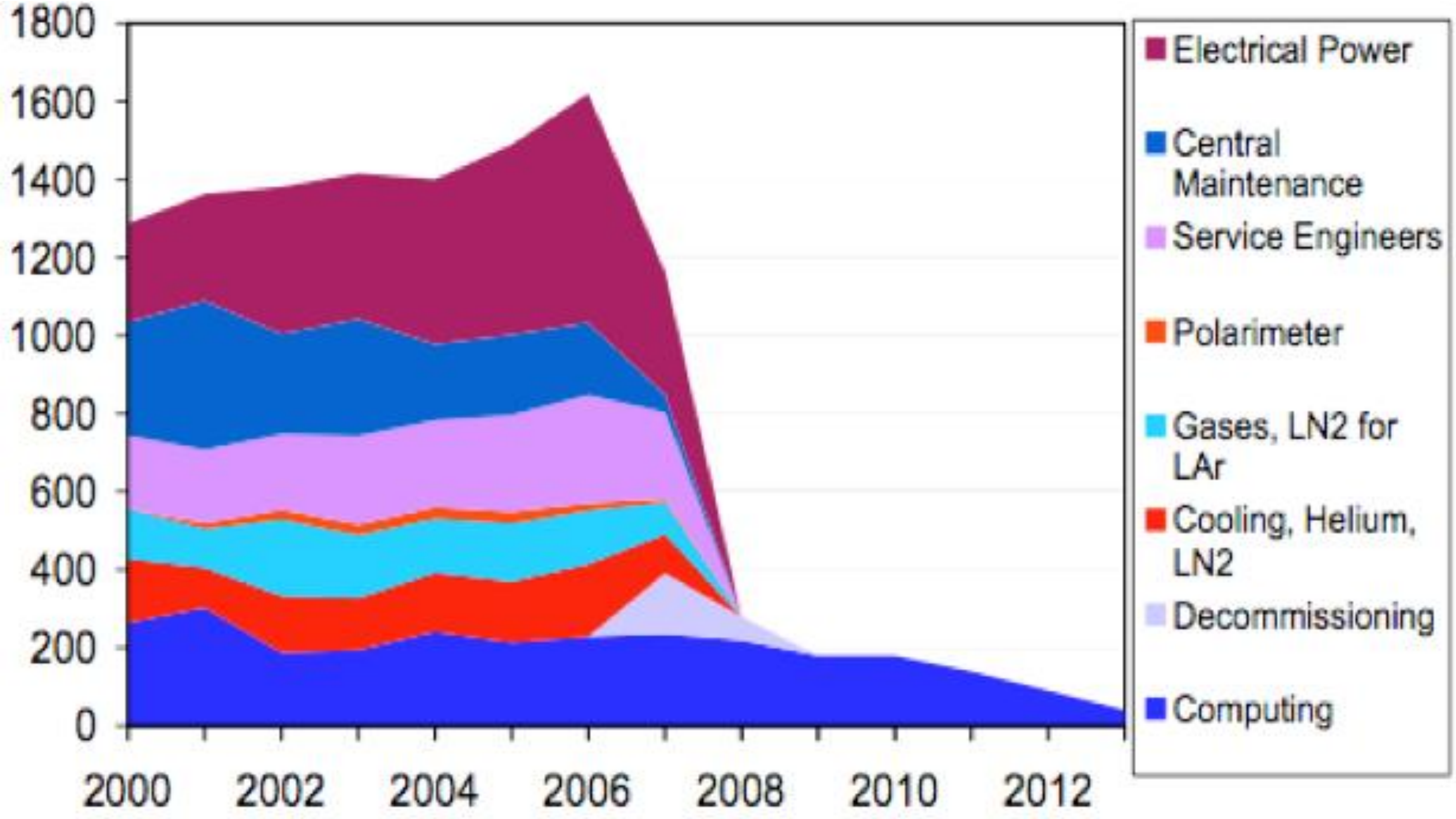


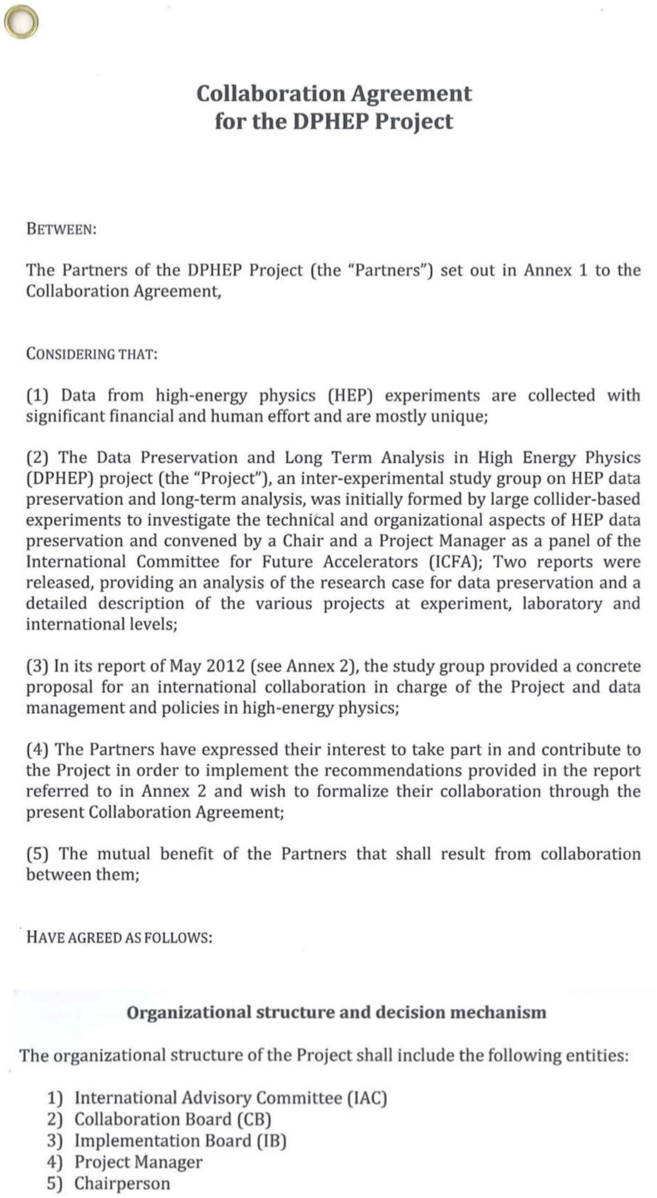
Figure 1: A possible model for data preservation organisation and resources presented as the milestones of the organisation and the resources evolution as a function of time.

When it stops taking data



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 prolongued to 2024
- DPHEP Self-Preservation?



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:

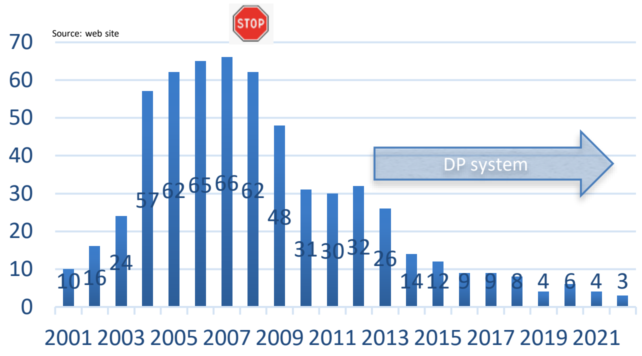
1) International Advisory Committee (IAC)
2) Collaboration Board (CB)
3) Implementation Board (IB)
4) Project Manager
5) Chairperson

The DPHEP 2020 Vision

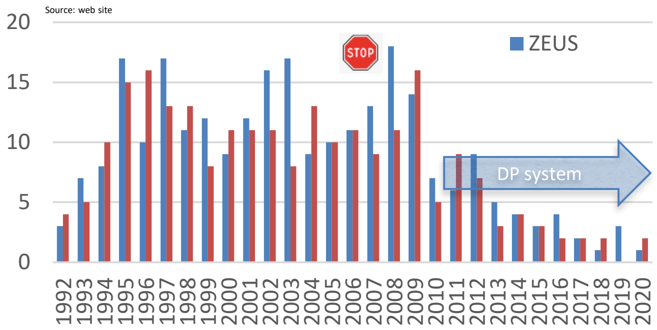
- *The "vision" for DPHEP – first presented to ICFA in* ***February 2013*** *– a consists of the following key points:*
  - By 2020, all **archived data** – e.g. that described in DPHEP Blueprint, including LHC data – should be easily **findable** and fully **usable** by the **designated communities** 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.

Scientific output from preserved data

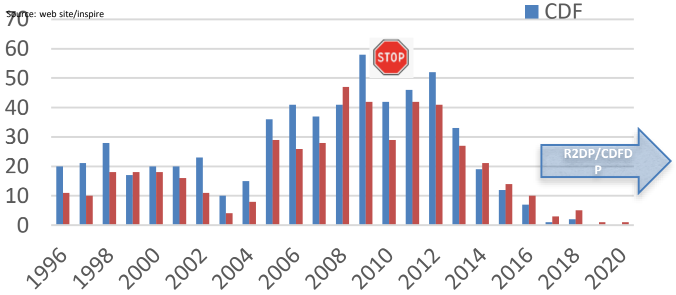
BABAR



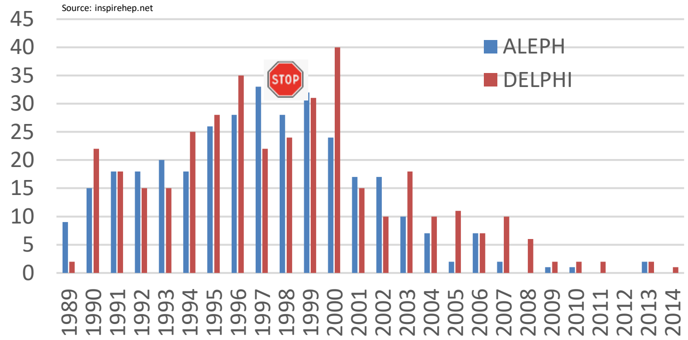
HERA



Tevatron



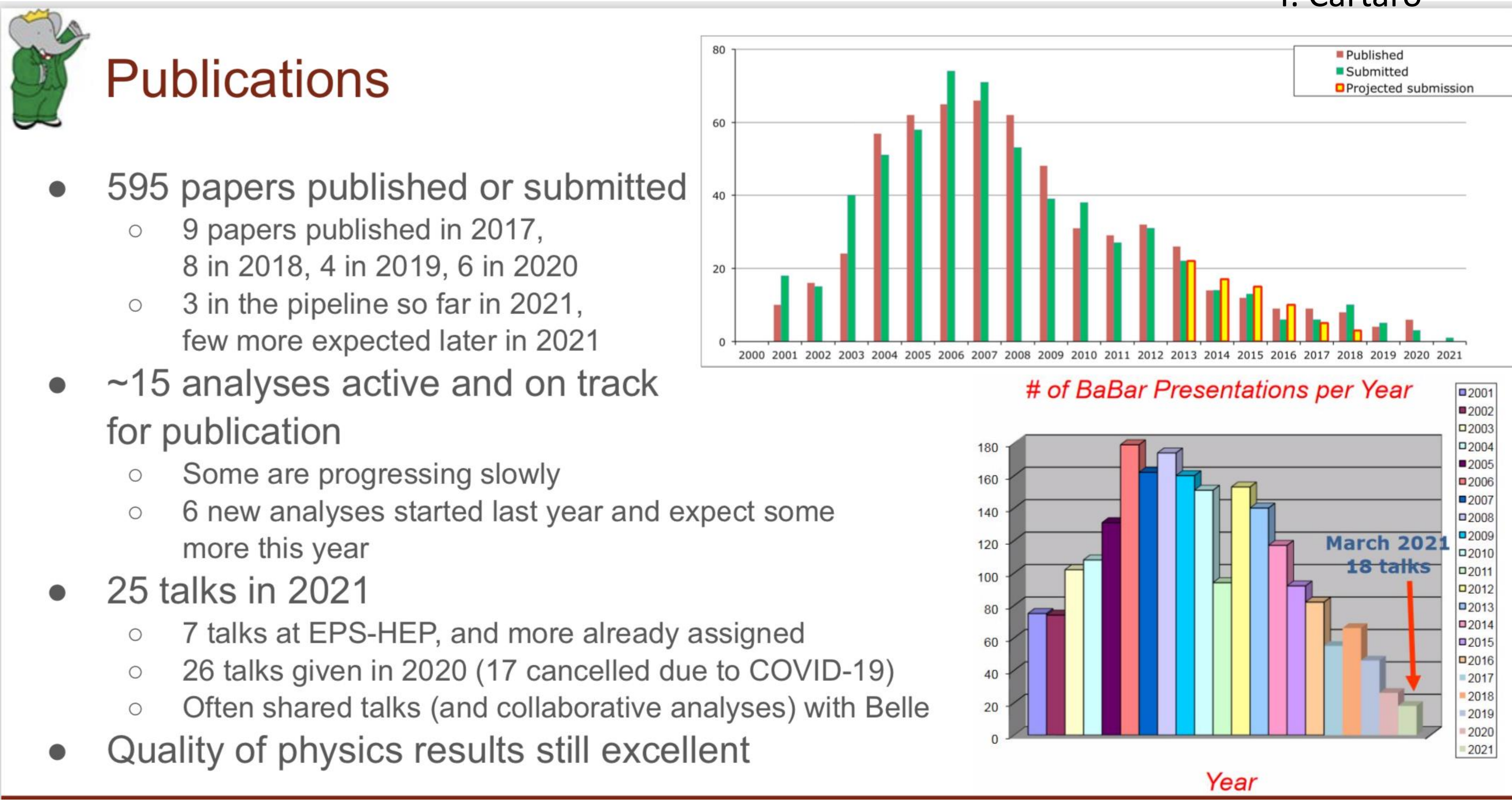
LEP



	before 2012	after 2012	%
Babar	471	154	32,70%
H1+ZEUS	436	62	14,22%

Babar today

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

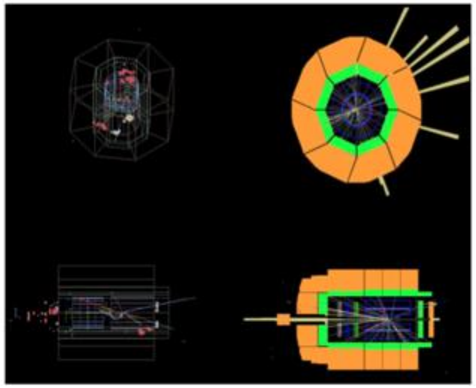
Open Data decided

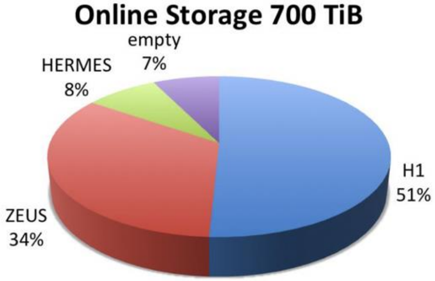
HERA: succesful DP, towards open data

- 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

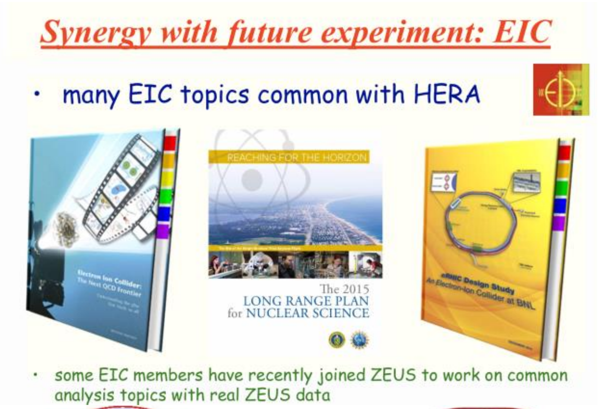
'H1Red' for simulated Pythia8.3 event





– New topics/collaborators (EIC)

A. Geiser, D. Britzger, D. South

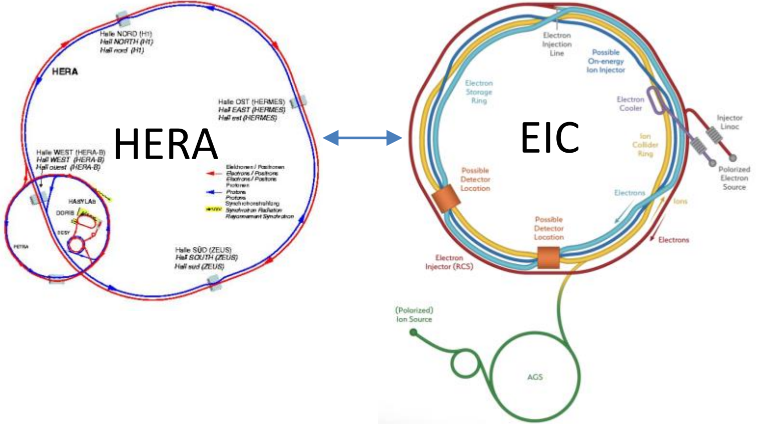


Synergy with future experiment: EIC

- many EIC topics common with HERA

Electron Ion Collider:
The Next QCD Frontier
REACHING FOR THE HORIZON
The 2015
LONG RANGE PLAN
for NUCLEAR SCIENCE
eRHIC Design Study
An Electron-Ion Collider at BNL

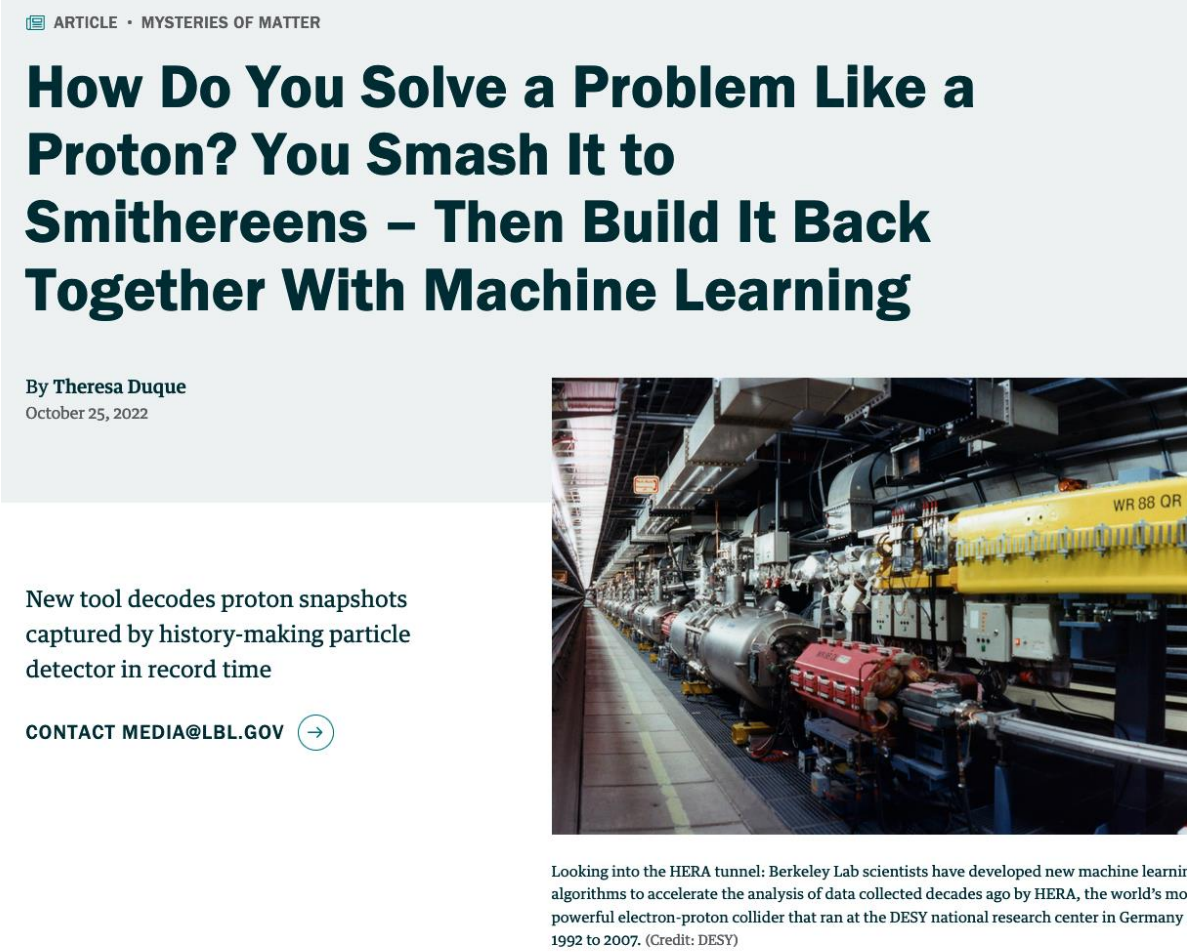
- some EIC members have recently joined ZEUS to work on common analysis topics with real ZEUS data



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). `

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



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 learnir algorithms to accelerate the analysis of data collected decades ago by HERA, the world's mo powerful electron-proton collider that ran at the DESY national research center in Germany 1992 to 2007. (Credit: DESY)

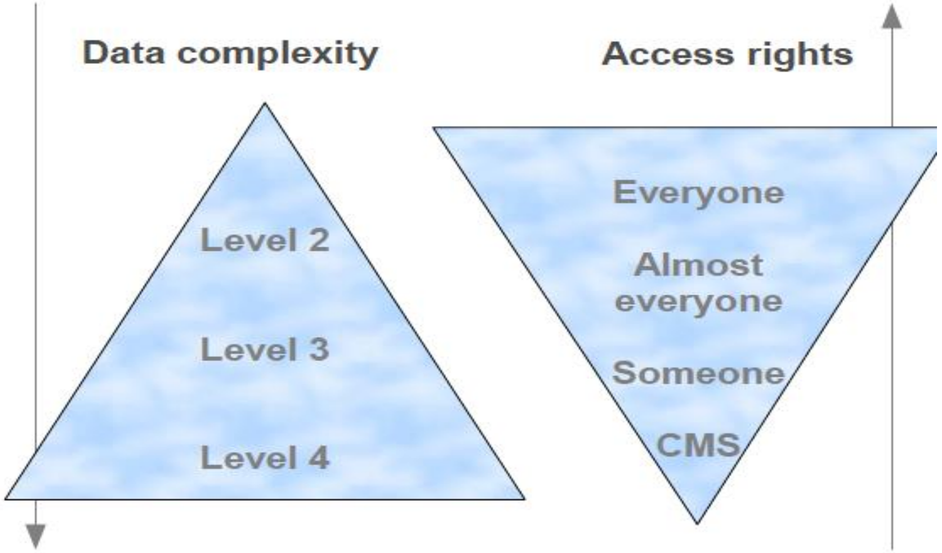
Live, preserved, open

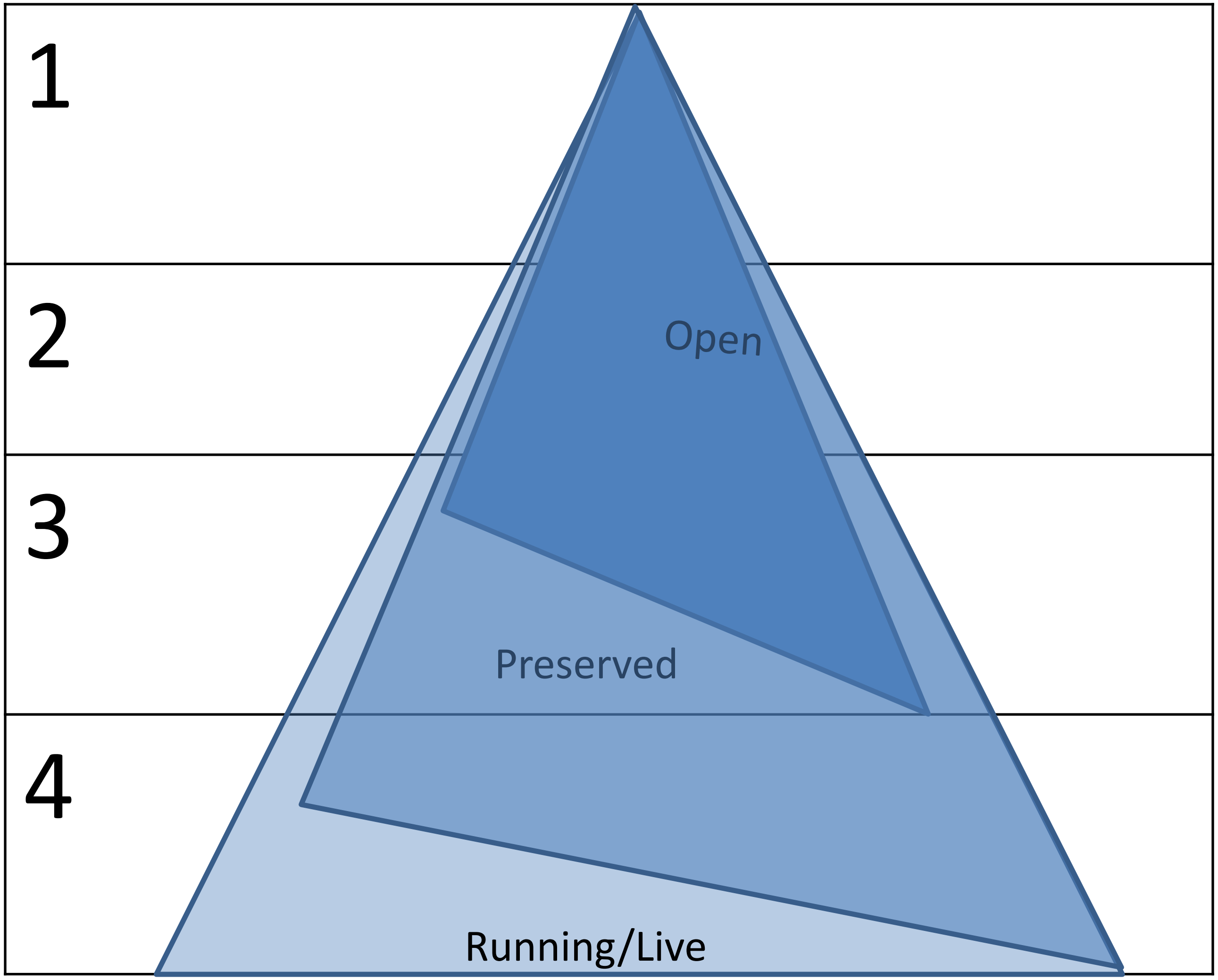
- “Complementary” requirements
  - **Live** computing: full force, speed, productivity, cooperation
    - within “limited” resources
  - **Preservation** (DPHEP): target as close as possible to 4
    - within “constrainted” resources ➔ 0
  - **Open**: is not *chmod a+r* applied to the preserved nor to running
    - Resources by projects; a global view?

Guidance into complexity/sharing

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




1
2
3
4
Open
Preserved
Running/Live

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 dedicatedDP 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

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 OD

Open Questions for DP/ how much they apply for OD/OS?

- 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. What global resources were used 5 and 10 years past the end of the experiment to keep systems alive and publish?
- 6. Are the DP requirements compatible with the running experiments conditions? How much extra investments are needed to make ”fresh” data suitable for a long term preservation and how those investments can be optimised further when considering **open data and open science aspects?**
- 7. How are future projects supporting, stimulating and shaping data preservation projects and how are the cost and benefits of this transfer of knowledge accounted for?

Discussion incentives

- Preservation and sharing:
  - 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