Data Preservation in High Energy Physics

Cristinel DIACONU
CPPM/CNRS/Aix-Marseille University

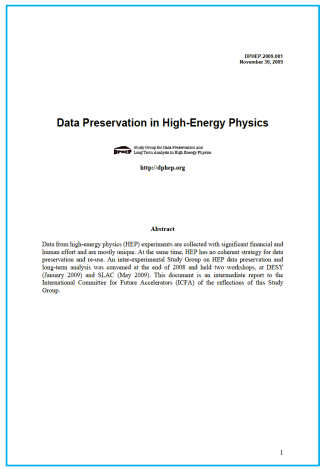
- What is “data”?
  - not (only) : “files”
  - but : “every digitally encoded information that was created as a result of planning, running and exploiting an experiment”
- What is “preservation”?
  - not: a freezer, a herbarium, a museum, an album, a cellar
  - but: the **process** of transforming a "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)

DPHEP Collaboration/ICFA Panel

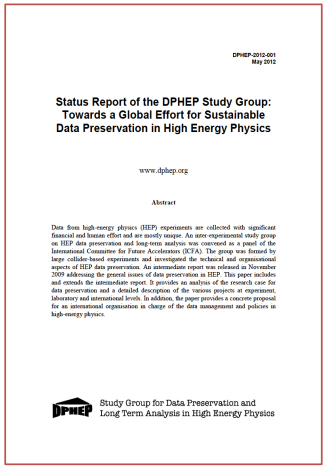
DPHEP Data Preservation in High Energy Physics
@DPHEPColl

http://dphep.org

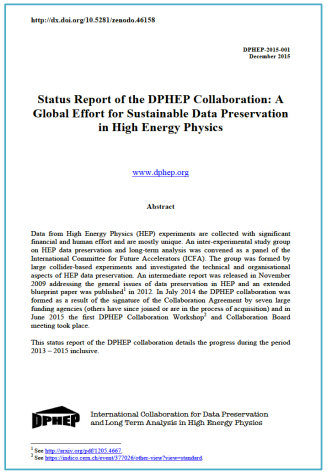
2009
LoI
arXiv:0912.0255



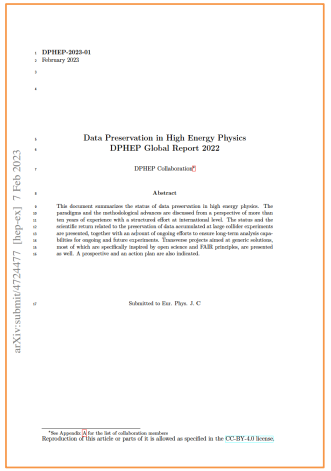
2012
Blueprint
arXiv:1205.4667



2015
Collaboration MoU
arXiv: 1512.02019



2023
Decade report
arXiv: 2302.03583

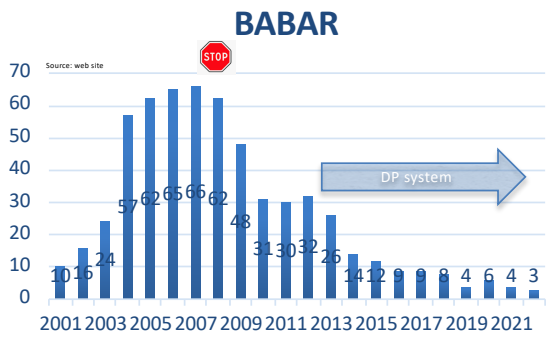


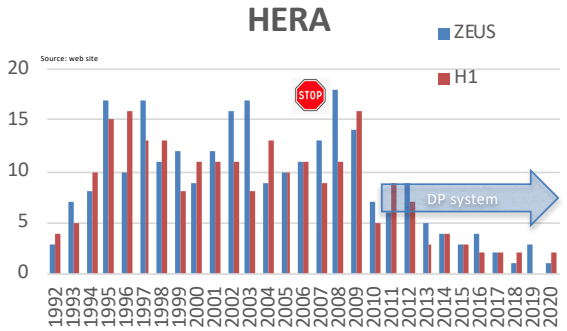
Data Preservation (DP) is a mandatory specification for any present and future experimental facility

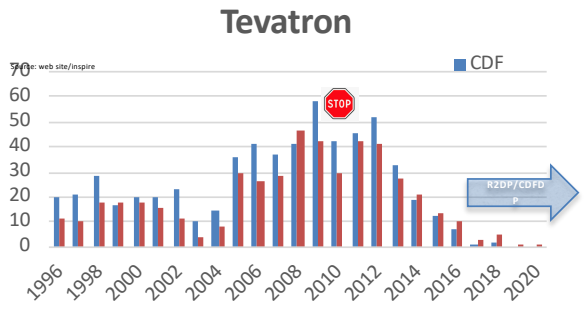
Enhancing the scientific output

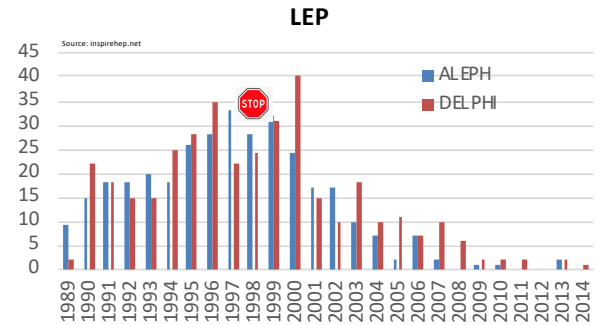
DP is a **cost-effective way of doing fundamental research** by exploiting unique data sets in the light of the increasing theoretical understanding.

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









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

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.

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**

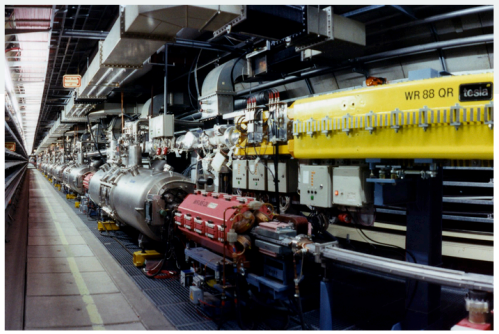
ARTICLE • MYSTERIES OF MATTER

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/