The past and future of Data Preservation in High Energy Physics

Collider data ageing: the bottle and the cellar

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

http://dphep.org

DPHEP Data Preservation in High Energy Physics
@DPHEPColl

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

10 years ago.....

DPHEP

Episode X

THE ADVENT OF PRESERVED DATA ANALYSIS

At the end of the first decade of the XXI century, many HEP experiments were close to the data taking.

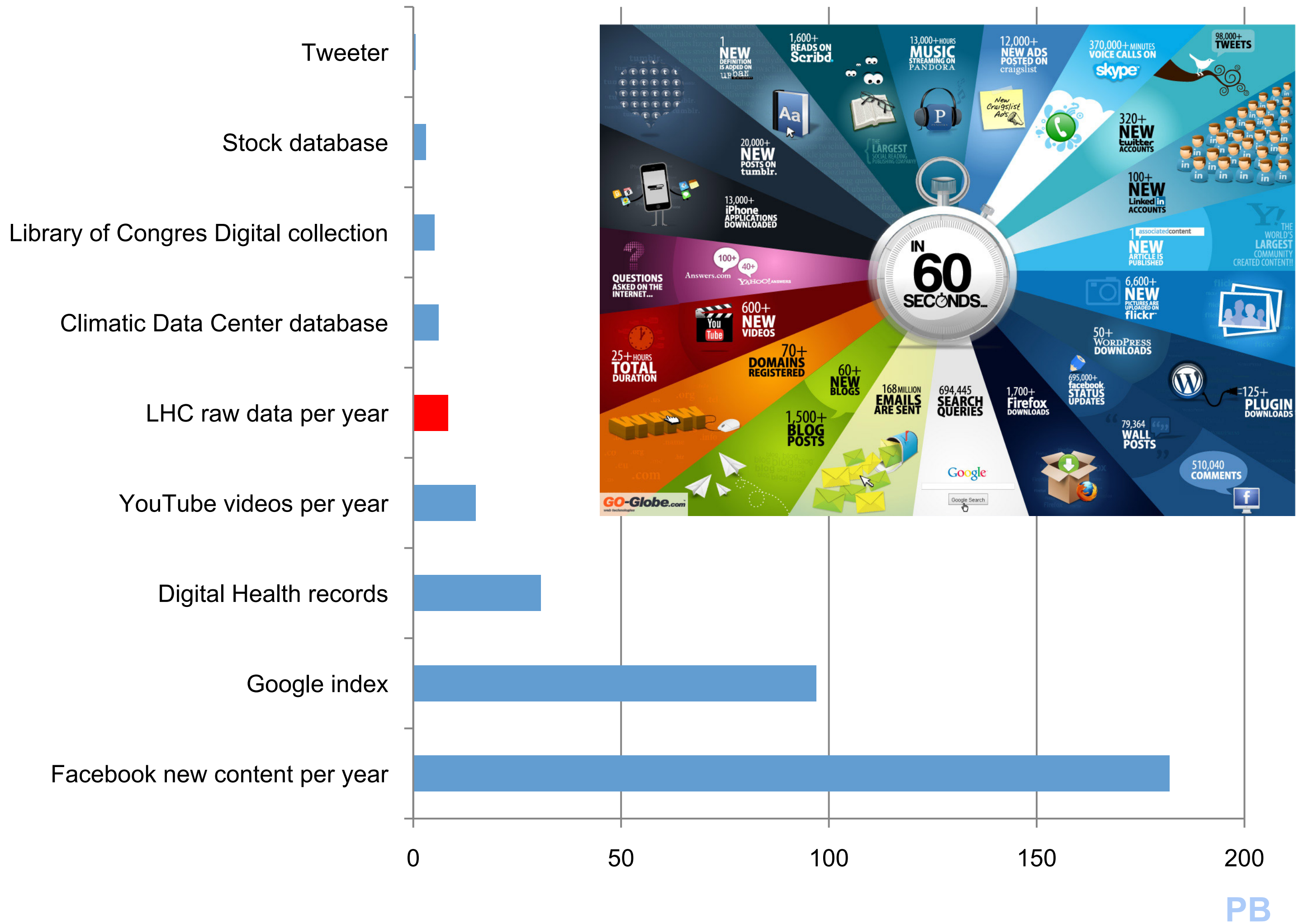
The world faced the rise of the LHC experiments.

A big data deluge was threatening the HEP computing centers.

The DPHEP collaboration federates the data preservation initiatives, monitor and report the worldwide status of preserved data.

But the fight is not over, data deletion forces are still alive….

Data collection accellerates



Digital data are fragile

- Storage capacity is physically exceeded
- Unattended/orpahned data vanishes quickly

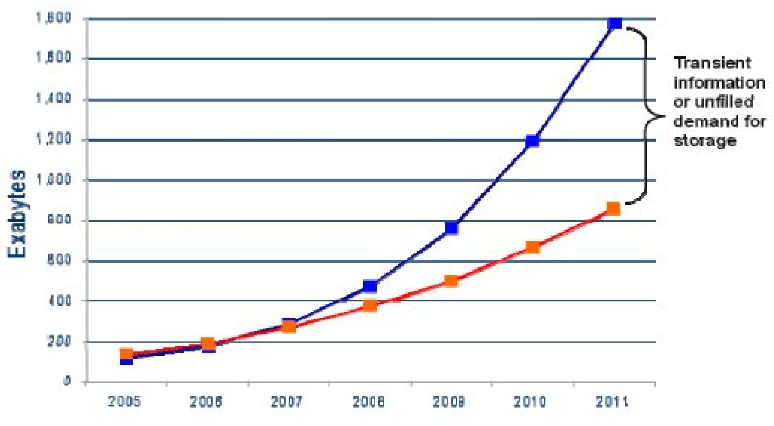
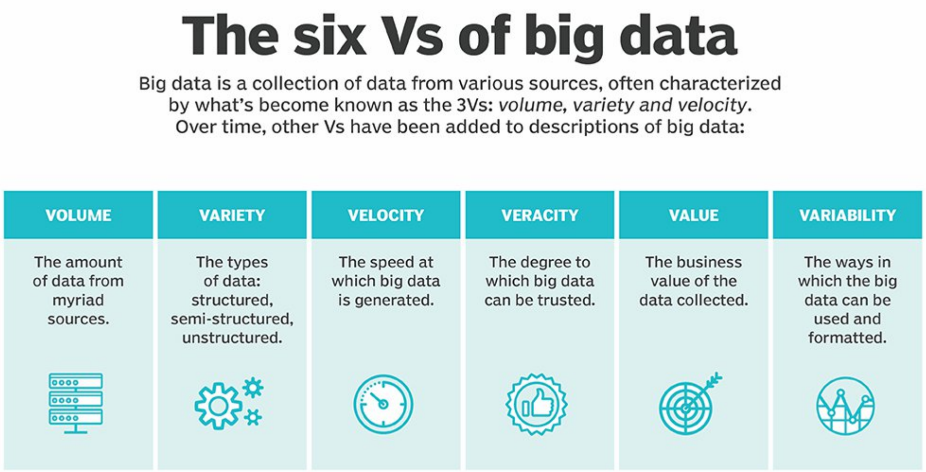


FIGURE 1.3: **Information and Storage**
Source: J. Gantz January 2008 (revised). Used with permission.

In Blue Ribbon Task Force report



https://www.techtarget.com/searchdatamanagement/definition/big-data

Vanish-ability?

Models of data preservation and acces

- Collaborations addressed this issue in a generic way
  - e.g. Blue Ribbon, APA, DPC, eSciDir, RDA …

http://www.alliancepermanentaccess.eu
http://brtf.sdsc.edu

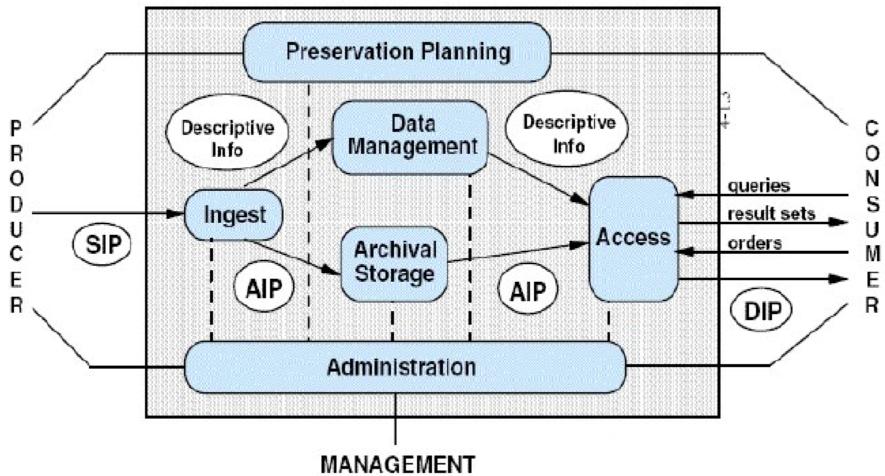
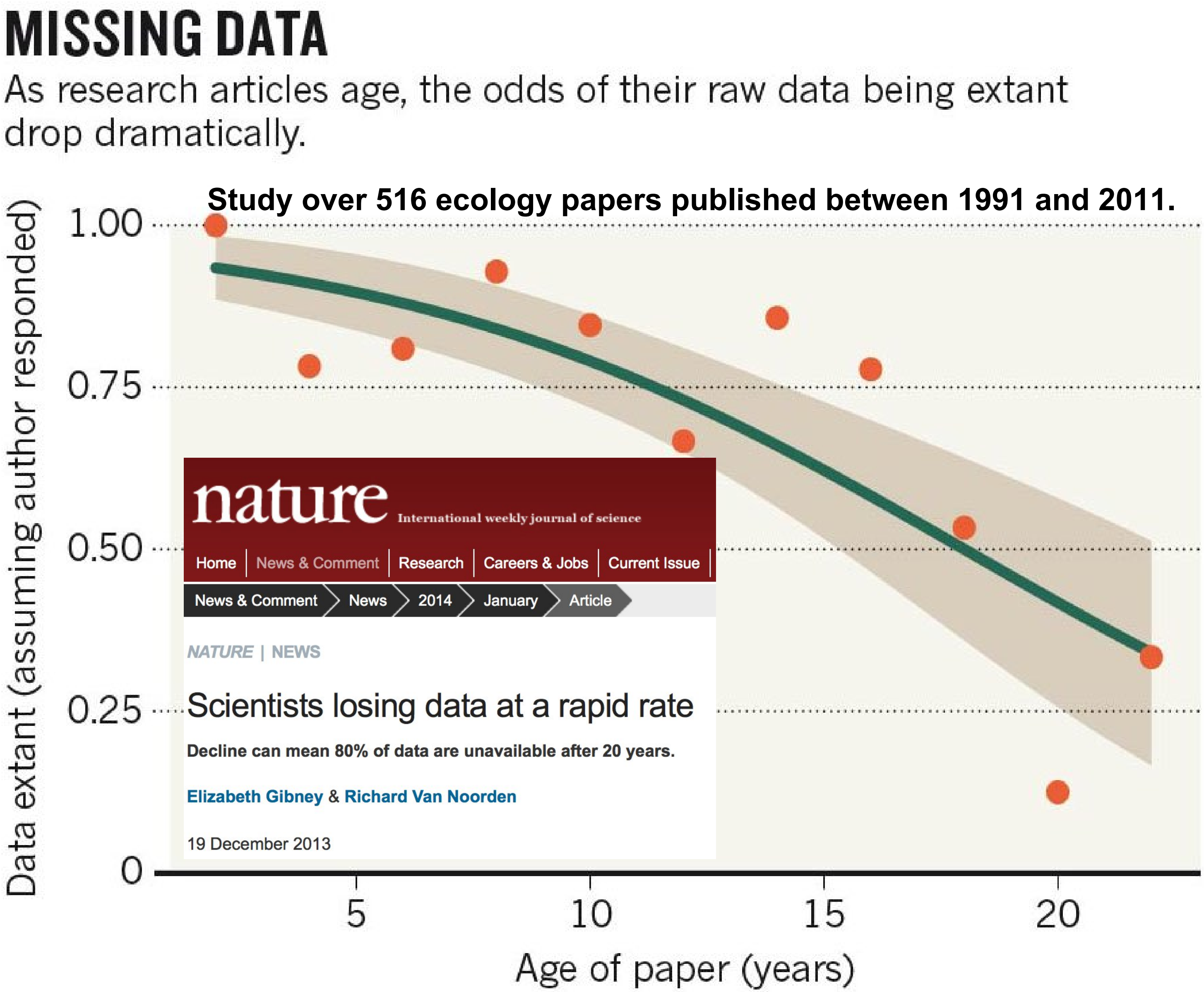


FIGURE 2.1: **The OAIS Reference Model**
http://public.ccsds.org/publications/archive/650x0b1.pdf, Page 4-1.
Source: Consultative Committee for Space Data Systems January 2002.

- Scientific … fforts
  - comple

Where is your data?

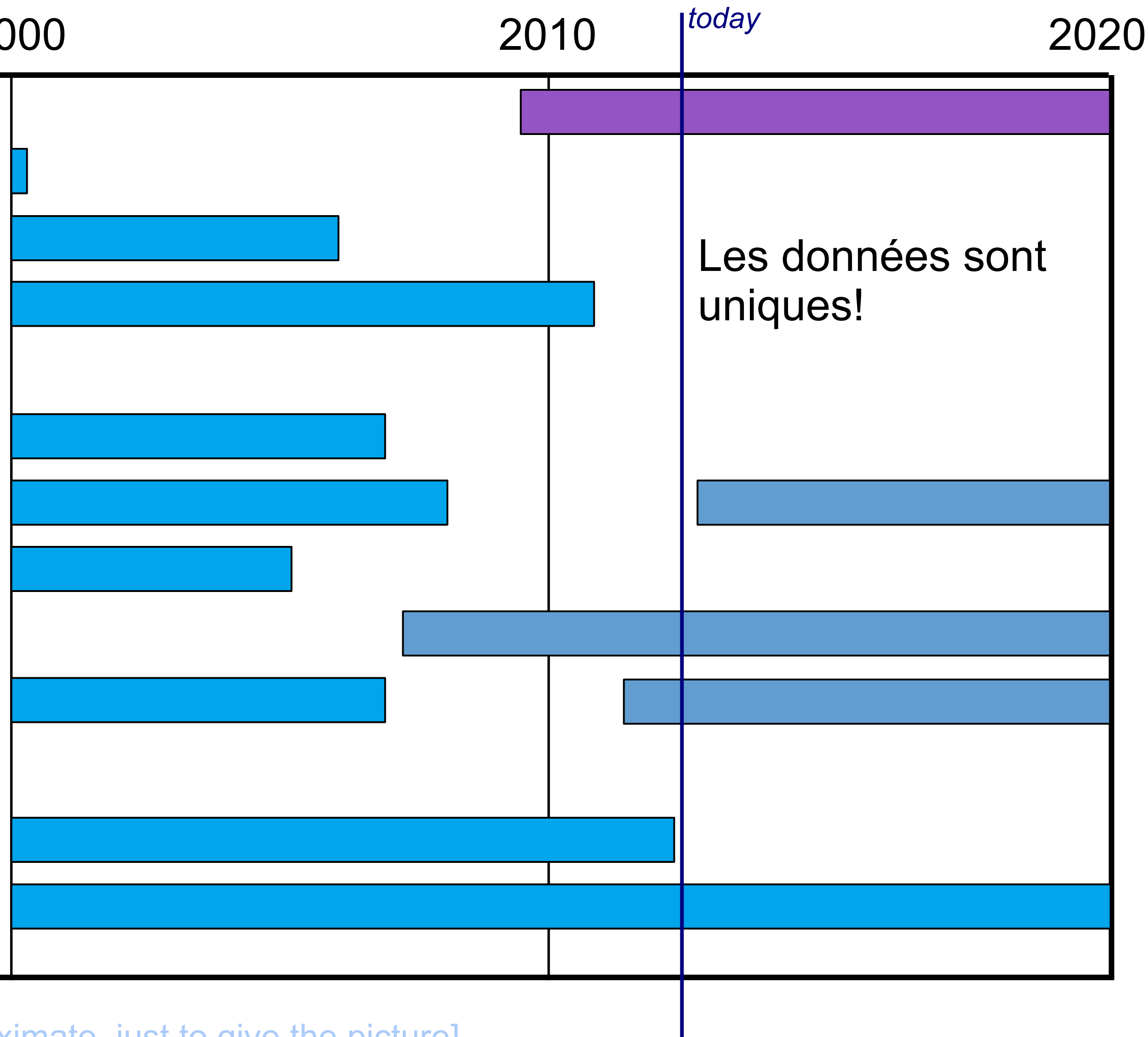


Scientific Data

- Structured following a scientific plan
- Diverse sources
- Large and expensive projects
  - Not easy to repeat
- Contain unique knowledge, objective/subjective rigidity in the long term
  - « time stamped »
  - « technology stamped »
  - « common knowledge stamped » ….
- Data Observatories
  - SD usually contain more information than initially needed/intended
  - less than possible, more than intended

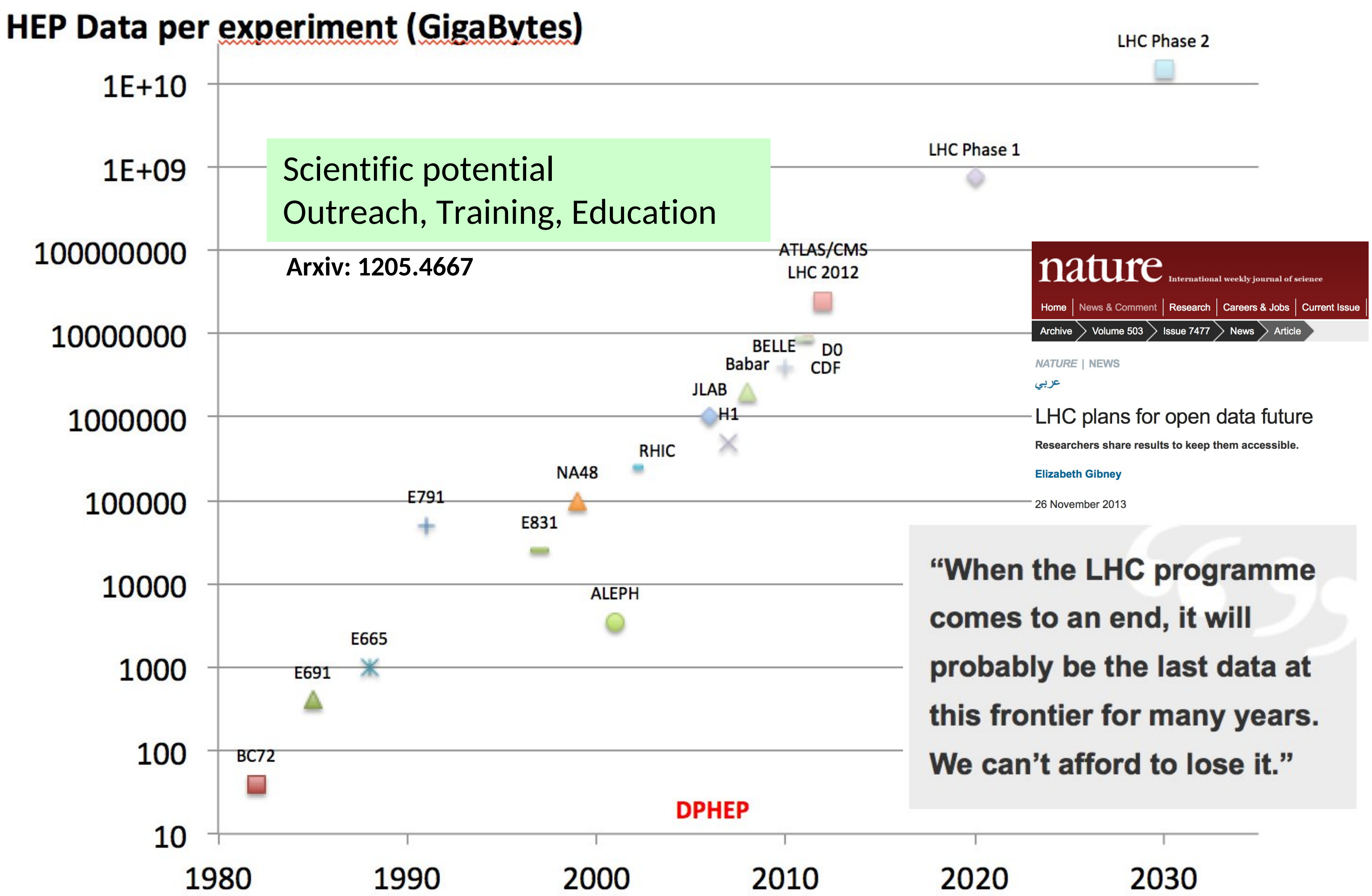
Exemple: HEP experiments in ± 10 ans

LHC	pp / ions	fb^{-1} ++
LEP	ee	0.9 fb^{-1}
HERA	ep	0.5 fb^{-1}
Tevatron	pp	10 fb^{-1}
BaBar	ee	600 fb^{-1}
Belle	ee	1 ab^{-1} ++
CLEO C	ee	0.9 fb^{-1}
BES III	ee	fb^{-1} ++
KLOE	ee	1 fb^{-1} ++
RHIC	pp / ions	Multi-exp
SPS	Fixed target	Multi-exp



[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
ATLAS/CMS
LHC 2012
BELLE
D0
Babar
CDF
JLAB
H1
RHIC
NA48
E791
E831
ALEPH
E665
E691
BC72
DPHEP
Scientific potential
Outreach, Training, Education
Arxiv: 1205.4667
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.”

What is “data”?

- The persistent confusion:
  - "data" = an operating system files, i.e. bits on a memory support such as disks, tapes etc.
    - This very superficial view, is useless for any running experiment and cannot exclusively apply to any useful thinking on the long term data preservation.

- Data = “every digitally encoded information that was created as a result of planning, running and exploiting an experiment”
  - digital data files: raw and processed, control/configuration, meta-data resuming environmental parameters, operational data etc.
  - software in all its forms (front-end, trigger, middleware, reconstruction, classification include machine learning setups, high-level analysis, visualisation etc.)
  - documentation files (internal/public notes, publications, manuals, workflows)
  - organisation and diffuse knowledge files (rules, minutes, meetings and slides, news, blogs, logbooks etc.)

Can HEP data “age”?

- The wine example (*)
  - Constituents
  - Bottle
  - Storage
  - Evolution (including “dumb phase”)

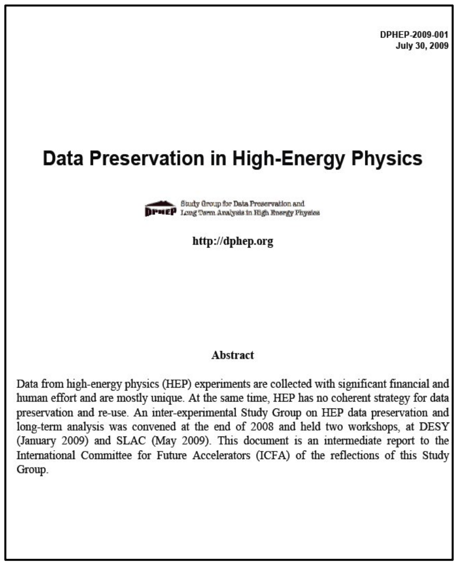
(*) l'abus d'alcool est dangereux pour la santé
à consommer avec modération
<u>alcohol abuse is dangerous for your health and should be consumed in moderation</u>

What is “preservation”?

- Not: a freezer, a herbarium, a museum, an album etc. … and not a cellar
  - And certainly not a “save it on tape, we will download it later if needed”
- Preservation: 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" (within its definition of above).
  - **ensure physical existence of dat**a from a digital point of view (see data definition above, all this has to be physically saved and secured at long term - and that includes software, of course) - note that this is the simplest and basically solved aspect of DP in HEP.
  - as an obvious (and relatively easy to solve) aspect of the previous item: **identify and provide computing and storage resources.**
  - **ensure the functionality** of the whole system, identify the potential risks and take appropriate measures as technology and community evolve. The level of complexity differ for the various aspects of the data. The simplest examples include the digital files, the documentation etc. that need only storage and access, i.e. rather standard operations independent of the experiment complexity in general. In contrast, specific experimental software and databases are much more difficult to keep functional across technological changes (hardware, operating systems etc.).
  - **define and identify the human resources** related to the research plan
  - **oversight and manage the collaborative work** and manage the preserved data analysis activity according to the DP design.
  - define and implement the **policies for data usage**, including **open**ing the access to data to new collaborators and/or releasing the data to larger (not identified apriori) communities.
  - observe and update the **physics case of the preserved data**. It should be noted that the technical solutions and the necessary choices on the information to be dismissed while designing a long term preservation system should decouple as much as possible from the epoch-related physics case. Indeed, the door should remain open for unexpected analyses (see below discussion on preservation levels).

DPHEP Study Group (2009)

- *arXiv:0912.0255*



DPHEP-2009-001
July 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

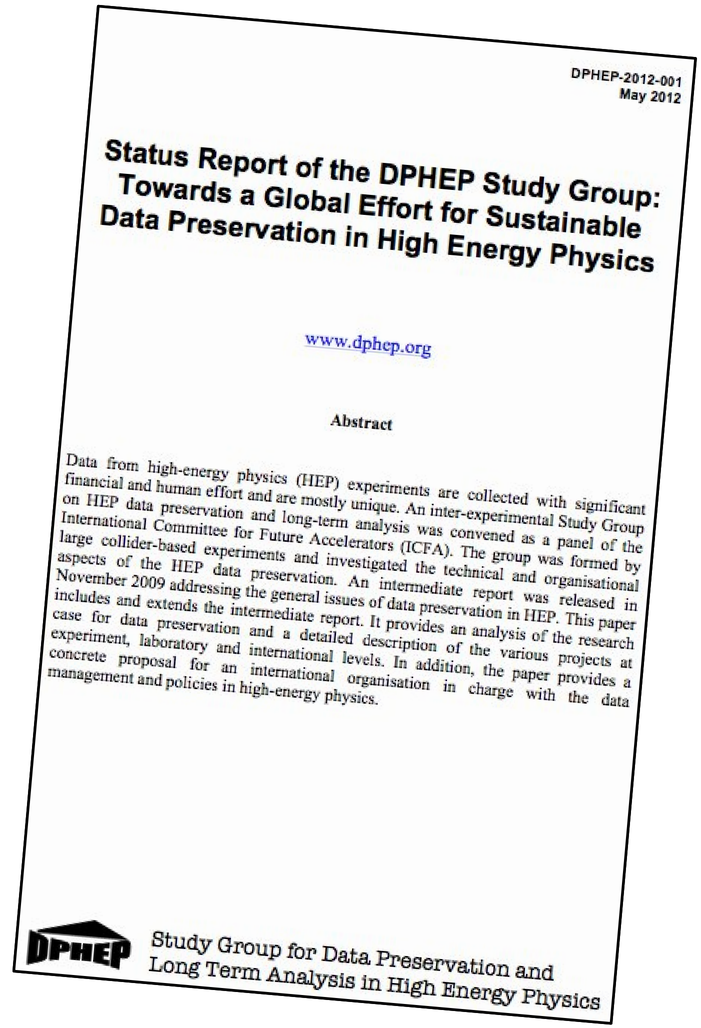
Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. At the same time, HEP has no coherent strategy for data preservation and re-use. An inter-experimental Study Group on HEP data preservation and long-term analysis was convened 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.

- 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

DPHEP Blueprint May 2012

- Full status report of the activities of the DPHEP study group, including:
  - An expanded description of the physics case
  - Defining and establishing data preservation principles
  - Updates from the experiments and joint projects
  - FTE estimates for these and future projects
  - Next steps to establish fully DPHEP in the field

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.


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

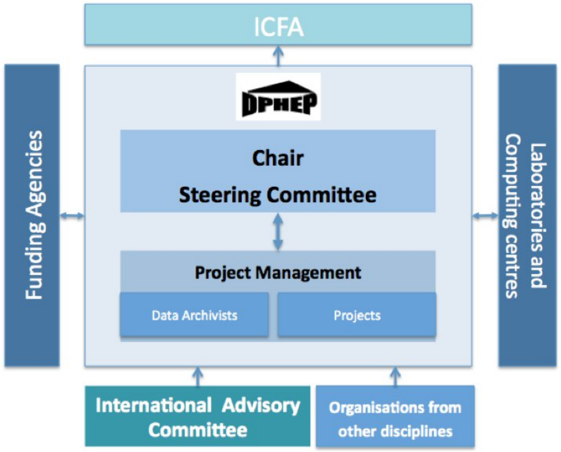
Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. An inter-experimental Study Group on HEP data preservation and long-term analysis was convened as a panel of the International Committee for Future Accelerators (ICFA). The group was formed by large collider-based experiments and investigated the technical and organisational aspects of the 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 with the data management and policies in high-energy physics.

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

arXiv:1205.4667

2012: Study Group ℗ Collaboration

- October, 2012: CERN endorses the blueprint and appoints the DPHEP Project Manager (Jamie Shiers)
- Retain the basic structure of the Study Group, with links to the host experiments, labs, funding agencies, ICFA



Dear Dr. Diaconu,

Following the delivery of the final DPHEP blueprint, various inputs received into the European Strategy for Particle Physics symposium earlier this week and after consultation with my colleagues, I would like to inform you that CERN offers to provide the role of the initial DPHEP project manager.

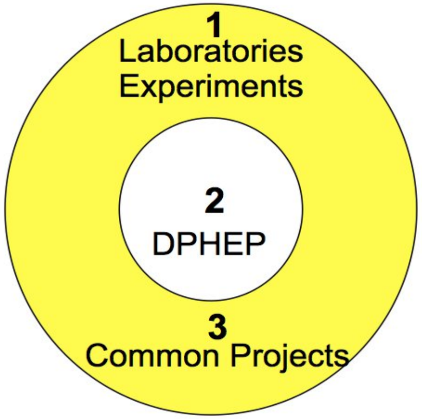
We would propose to appoint Jamie Shiers in this role for an initial period of 3 years starting 1 January 2013, after which the role may be assumed by another laboratory, as suggested in the blueprint.

We would anticipate that during this period the DPHEP organization will be launched (year 1) and that the initial deliverables defined in the blueprint would be achieved.

CERN would also foresee participation in the other activities described in the document in areas such as R&D into the use of virtual machine technology for data preservation purposes (PH-SFT input to ESPP) and into the management of very large data stores.

Yours sincerely,

Sergio Bertolucci
Director for Research and Computing



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
- DPHEP Activity
  - Global reports 2009(whitepaper), 2012 (blueprint), 2015, 2017 (global reports)
  - Collaboration meetings: 2015, 2017, 2021
  - Remote panel discussion March 2nd 2021
  - Reports to ICFA 2021 and 2022

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.

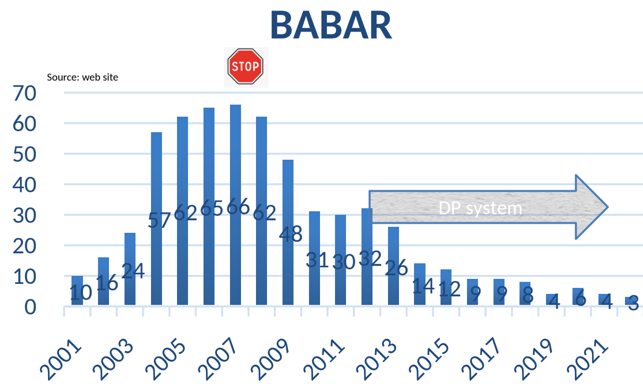
DPHEP recent activities

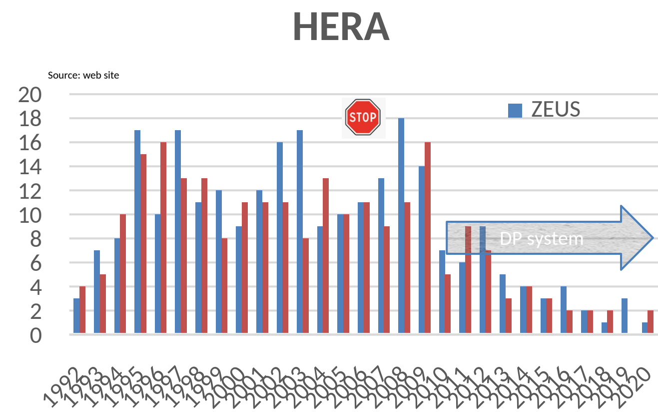
- Remote discussion March 2021
- ICFA Mandate prolongued 2021-2024
- 3rd Collaboration meeting (remote) June 21-23, 2021
  - https://indico.cern.ch/event/1043155/timetable/
  - 22 contributions: experiments, dedicated projects

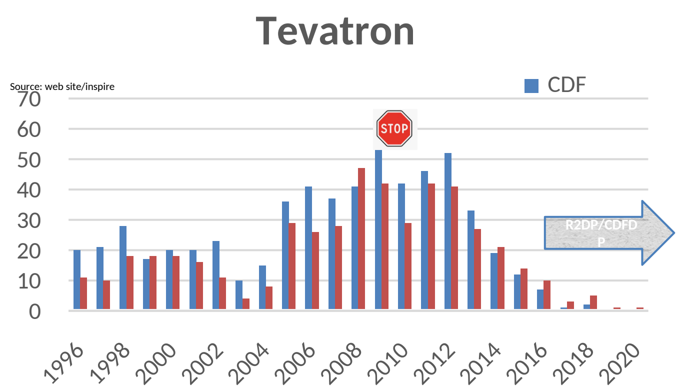
Panel remote discussion: March 2nd

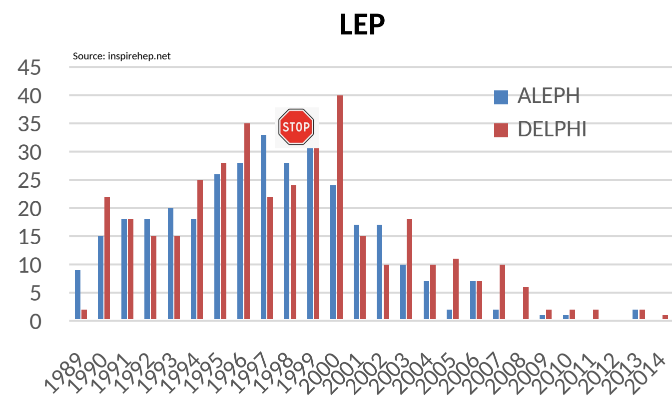
- A decade perspective
  - The 2009 recommendations were crucial and are still valid:
    - address DP asap through dedicated projects
    - make it global via DPHEP
    - develop technologies

Scientific output from preserved data









Data Preservation projects labs: recent update

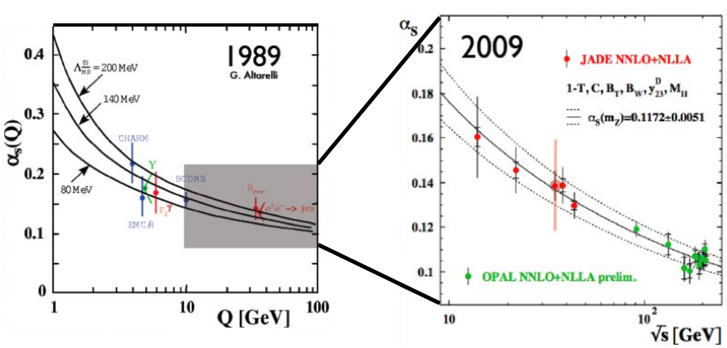
- **@DESY:** H1 (migration) and ZEUS (encapsulation) in great shape
  - successful transitions to the DP systems, publication plans continues and includes O(10) papers
  - objective: alive by 2030; New institutes joining (synergy with EIC)
- **@CERN: strong LHC activity**, LEP data/sw refreshed, OD/OS standards/technologies, DPHEP portal
  - Need for the continuation of the central management support
- **@MPI**: multi-experiment framework explored (JADE, HERA, OPAL)
  - JADE on a desktop

- **@KEK:** BELLE I data readable in Belle II framework ;
  - objective maintain Belle I data by 2023 (when the precision will be exceeded by the new data)
- **@IHEP/BES3:** The experiment is expected to stop data taking by 2022
  - Data to be preserved for 15 years
  - Strong support to DP national and international activities expressed
- **@BNL/JLAB**: DP activity ongoing (ATLAS, EIC), discussed with NPC
- **@Babar:** LTDA supported analysis since 2012. SLAC support ended in February. Data almost entirely copied to CERN/GridKa.
  - Data saved at CERN/GridKa: ~ 1.2 PB+ 0.5 PB (ongoing), Minimal user infrastructure for ongoing analyses and documentation hosted at U. of Victoria.
- **@FNAL**: (indirect news this time) transition to a DP system for both CDF (CDFDP) and D0 (R2DP)
  - Data stored/saved @FNAL+Italy, 500th paper from D0 in 2021

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

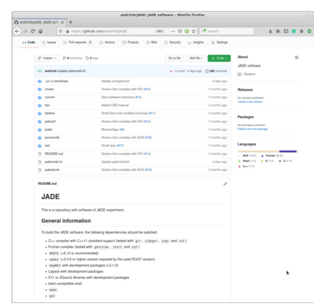


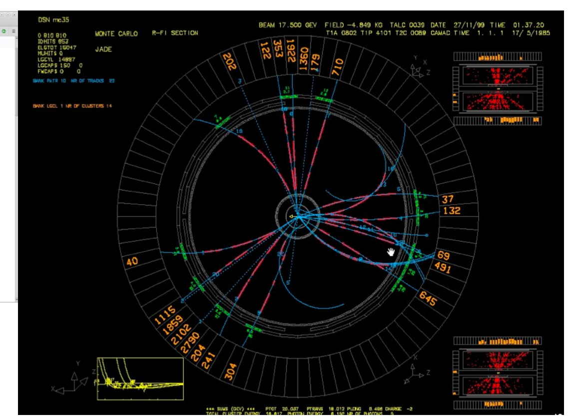
2021

JADE software: recent developments

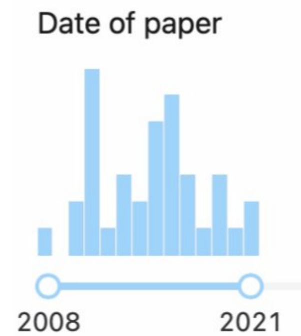
More portability, testing and documentation.

- 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



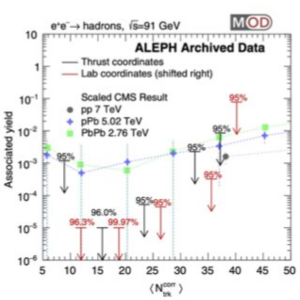
LTDP @LEP: Big Data Today – Peanuts Tomorrow

New physics with Archeodata

PHYSICAL REVIEW LETTERS 123, 212002 (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*



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

17 January 2020

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

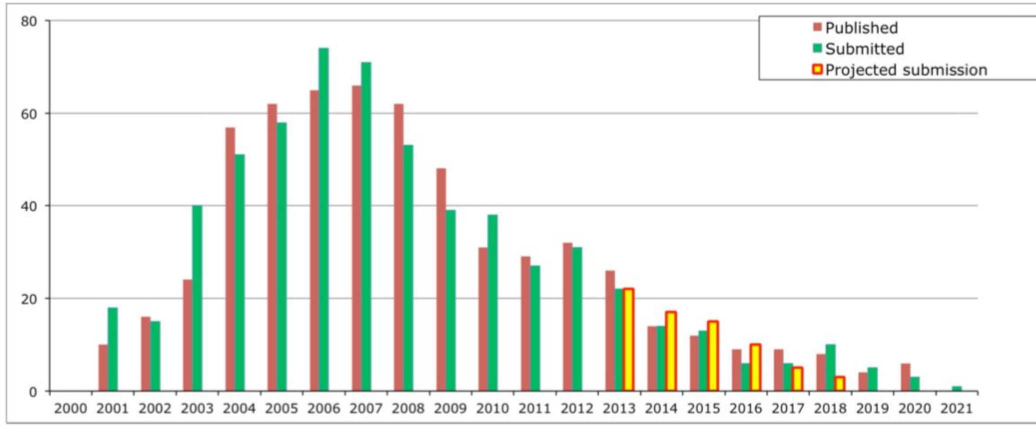
V. F. Perepelitsa
ITEP, Moscow
T. Ekelof
Department of Physics and Astronomy, Uppsala University
A. Ferrer
IFIC, Valencia University
B. R. French
bernardfrench@bluewin.ch

Ⓟ FCCee

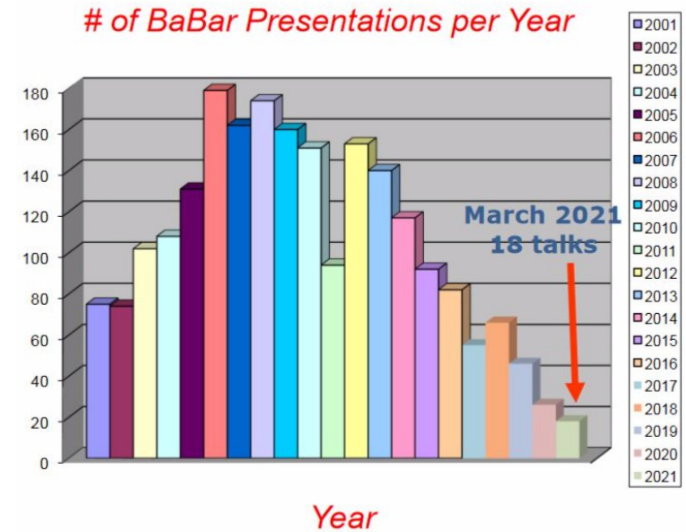
Babar today

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

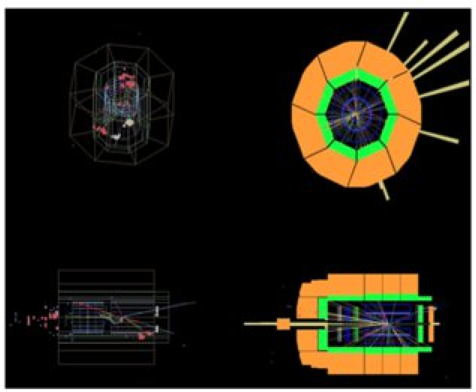
Open Data decided

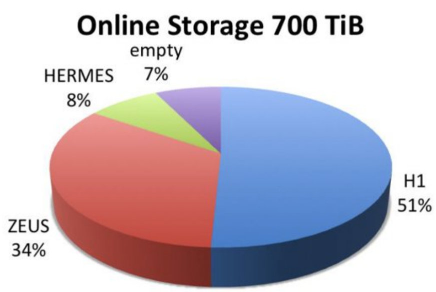
HERA: succesful DP, towards open data

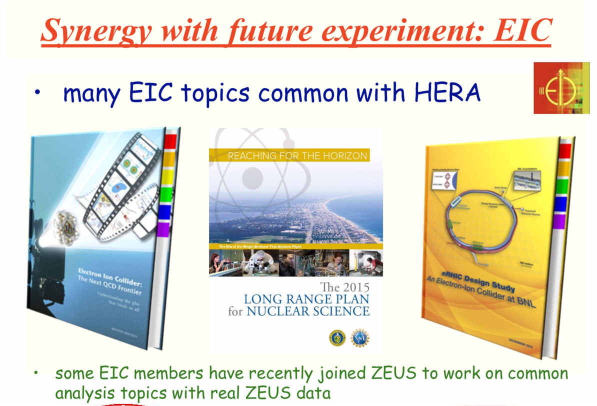
- H1: “Level 4” DPHEP strategy
  - All data, full migration, including regular recompilation/validation
  - Recent “technology jump” succesfull : in line with modern tools
    - “LHC”-like tools, ready for opendata

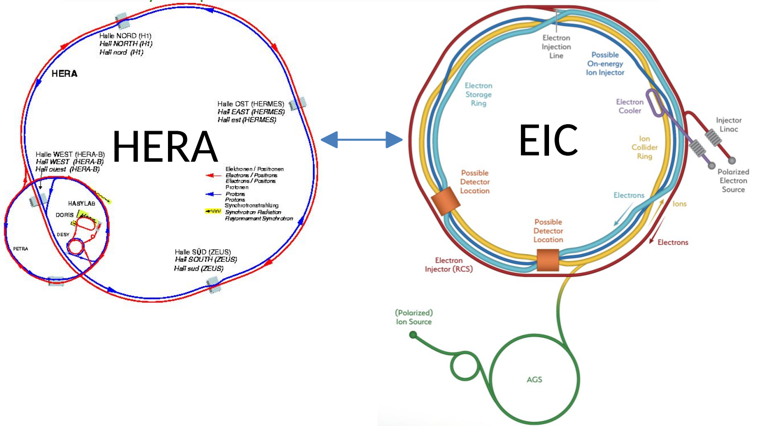
- ZEUS : “Level 3/4” DPHEP strategy
  - Root ntuples produced in the preparatory phase
  - easy to maintain/use/test/open

'H1Red' for simulated Pythia8.3 event









– New topics/collaborators (EIC)

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)](). `
- The EIC – to be built at Brookhaven National Laboratory in partnership with Thomas Jefferson National Accelerator Facility – will be a powerful and versatile new machine capable of colliding high-energy beams of polarized electrons with a wide range of ions (or charged atoms) across many energies, including polarized protons and some polarized ions.

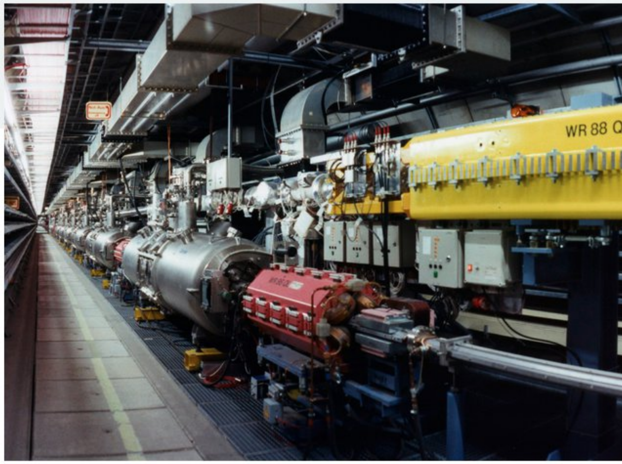
ARTICLE • MYSTERIES OF MATTER

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

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

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, réinterprétation, 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 ⌄

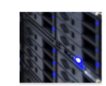
nature › nature physics › perspectives › article

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

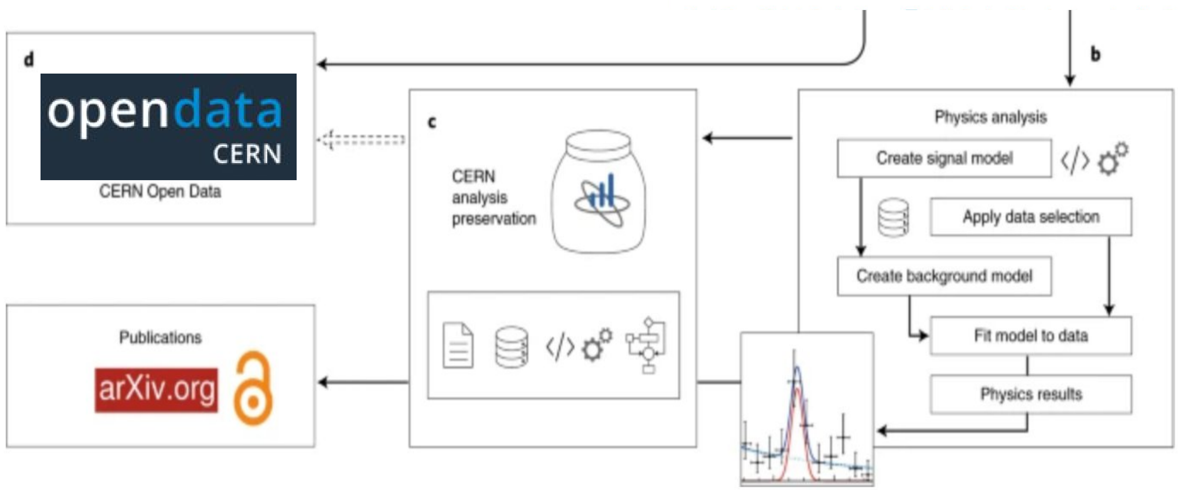
Perspective | Open Access | Published: 15 November 2018

Open is not enough

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



Filter by experiment	2017	2021
ALICE	15	15
ATLAS	101	109
CMS	878	1167
LHCb	3	4
OPERA		904



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

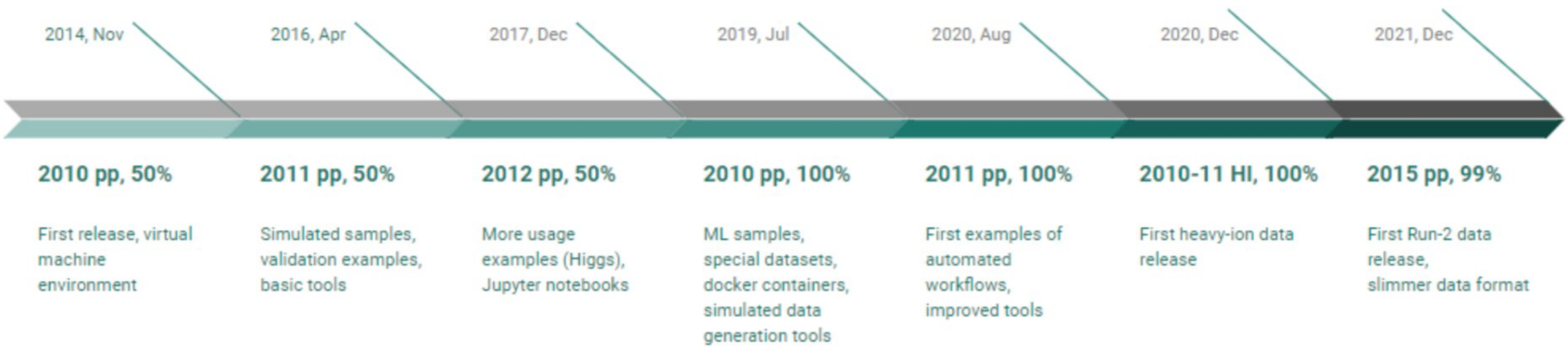
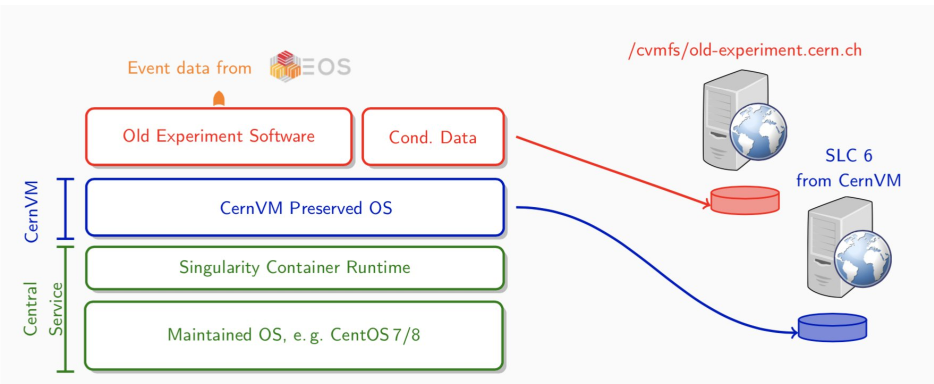


Figure 2: CMS data release timeline.

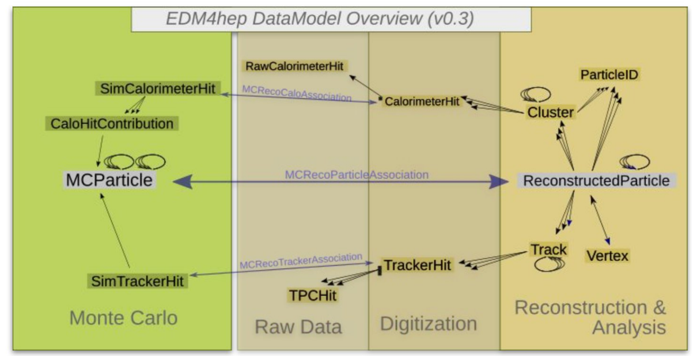
Towards more standards

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



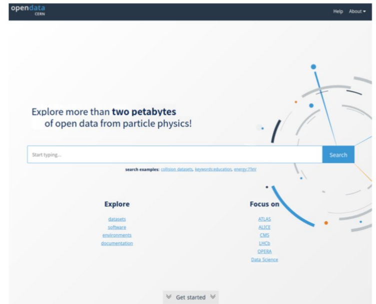
key4hep / EDM4hep and DPHEP?

- 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

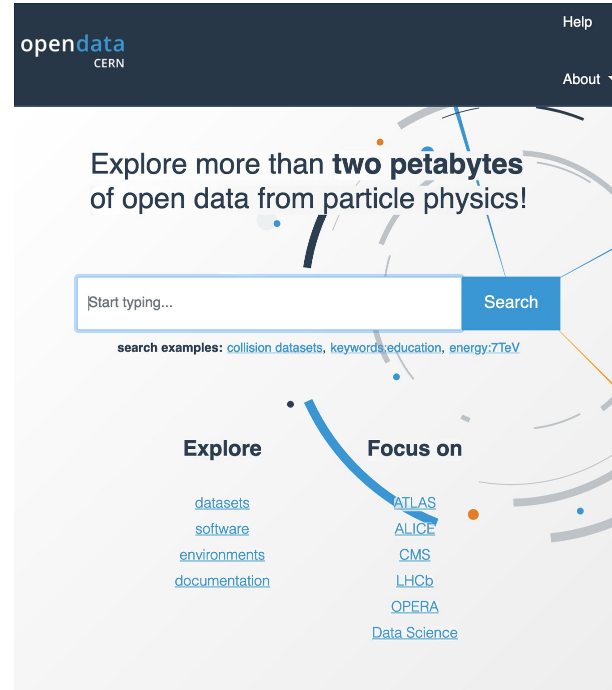
- CMS completed 2010–11 proton-proton data; released half of 2012 data; released 2010–11 heavy-ion data samples and corresponding pp reference datasets
- ATLAS released 13TeV educational samples
- Data science and Machine Learning (CMS, LHCb. . .)
- Non-LHC physics: OPERA neutrino physics data; interest 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:ff63668a, found adler32:ff63668a
```

Command-line client to ease data download



CERN Open Data portal: Plans

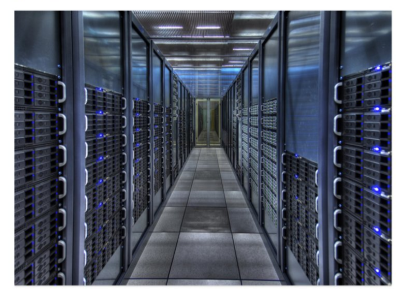
- December 2020: A common statement on the open data policy by CERN management and ATLAS, ALICE, CMS, LHCb and TOTEM experiments.

  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

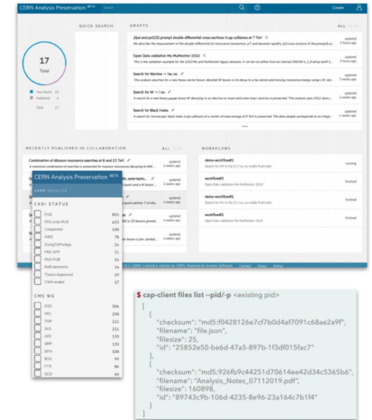
Xiaoli Chen, Sünje Dallmeier-Tiessen, Robin Dasler, Sebastian Feger, Pamfilos Fokianos, Jose Benito Gonzalez, Harri Hirvonsalo, Dinos Kousidis, Artemis Lavasa, Salvatore Mele, Diego Rodriguez Rodriguez, Tibor Šimko, Tim Smith, Ana Trisovic, Anna Trzcinska, Ioannis Tsanaktsidis, Markus Zimmermann, Kyle 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)



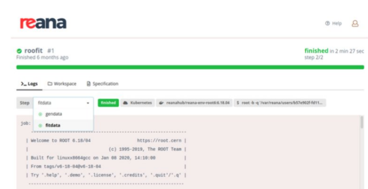
https://analysispreservation.cern.ch

@tiborsimko 2 / 3

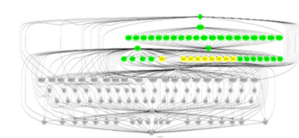
REANA reproducible analysis platform

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

A word on *FAIR*

- The DPHEP objectives (2012) intrinsically comply with what has became to be known as FAIR principles (2016)

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.

- 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

Situation and trends

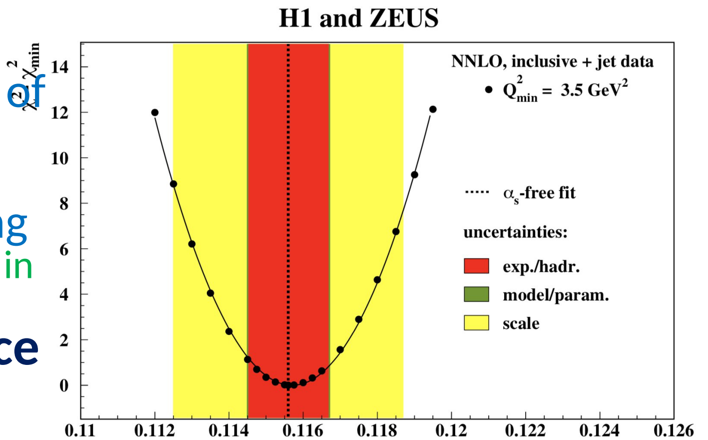
- Significant/measurable impact of dedicated DP projects @expts./labs
  - Production of high quality and unique scientific results at very low (non-zero) cost
    - 10% output for less than 1% investment: ✓
    - **Long term organisation proves to be productive**
  - Signs of re-vigorating collaborations 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
- LHC exps. very active in DP and **Open Data/Science**

- 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)
- Global status report
  - Several remote editorial meetings
  - Quite advanced, to be released by Summer 2022

Eur. Phys. J. C (2022) 82:243
https://doi.org/10.1140/epjc/s10052-022-10083-9
THE EUROPEAN PHYSICAL JOURNAL C
Regular Article - Experimental Physics

HERA; 2022

Impact of jet-production data on the next-to-next-to-leading-order determination of HERAPDF2.0 parton distributions



DESY 21-130, ISSN 0418-9833

Measurement of lepton-jet correlation in deep-inelastic scattering with the H1 detector using machine learning for unfolding

H1 Collaboration*

(To be submitted to Physical Review Letters)
(Dated: August 30, 2021)

H1; 2021

The first measurement of lepton-jet momentum imbalance and azimuthal correlation in lepton-proton scattering at high momentum transfer is presented. These data, taken with the H1 detector at HERA, are corrected for detector effects using an unbinned machine learning algorithm (OmniFold), which considers eight observables simultaneously in this first application. The unfolded cross sections are compared to calculations performed within the context of collinear or transverse-momentum-dependent (TMD) factorization in Quantum Chromodynamics (QCD) as well as Monte Carlo event generators. The measurement probes a wide range of QCD phenomena, including TMD parton distribution functions and their evolution with energy in so far unexplored kinematic regions.

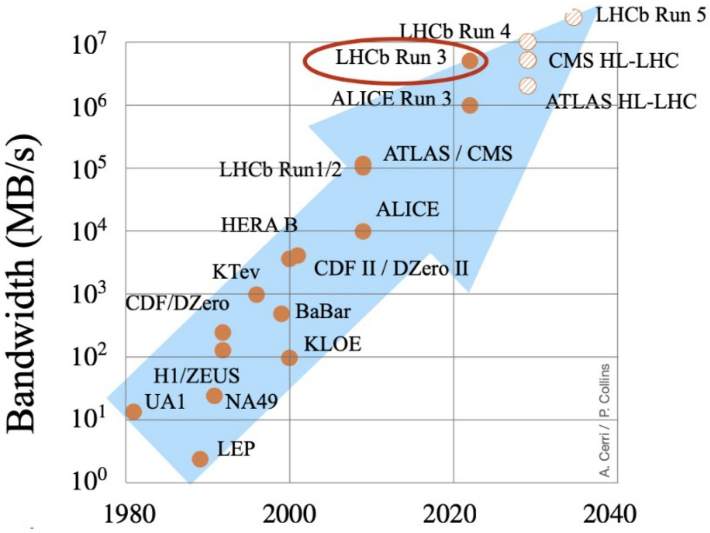
Data Preservation in High Energy Physics
Status Report, Perspectives, and Plans
DPHEP Global Report 2022

DPHEP Collaboration[1]

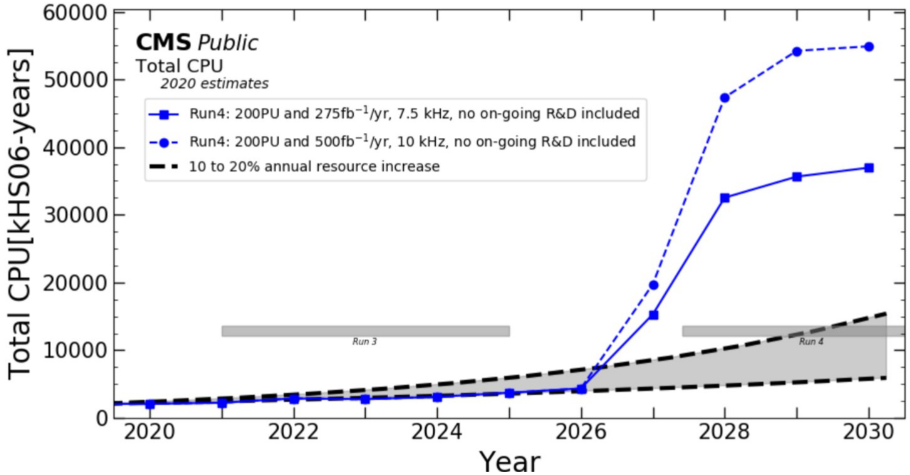
Abstract

This document contemplates more than ten years of experience and global efforts to pursue the preservation of data accumulated at large collider experiments.

Business as … (not) usual



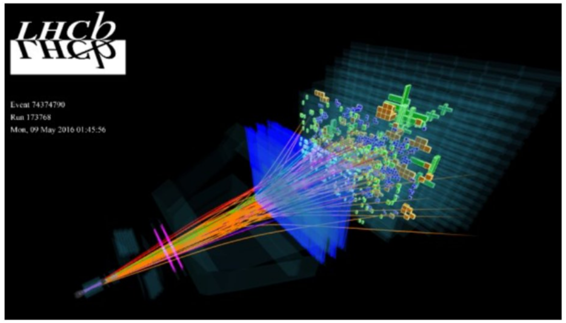
D. vom Bruch



https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults

Largest single internet exchange point:
14 Tbit/s

LHCb experiment @ CERN
40 Tbit/s



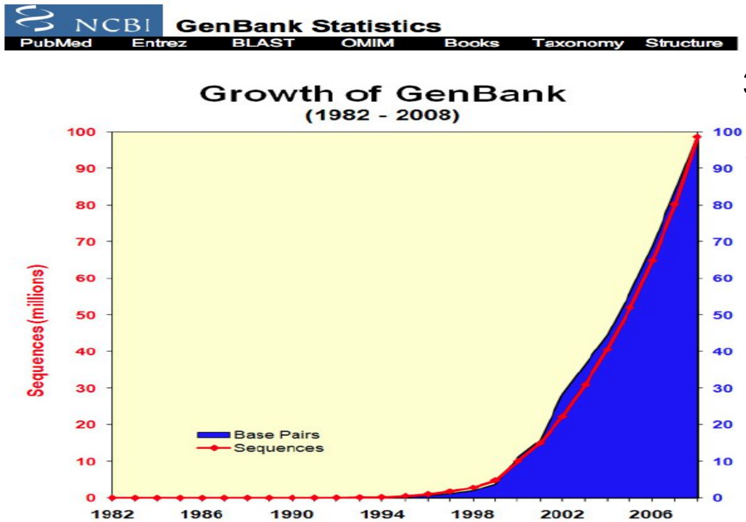
Next steps

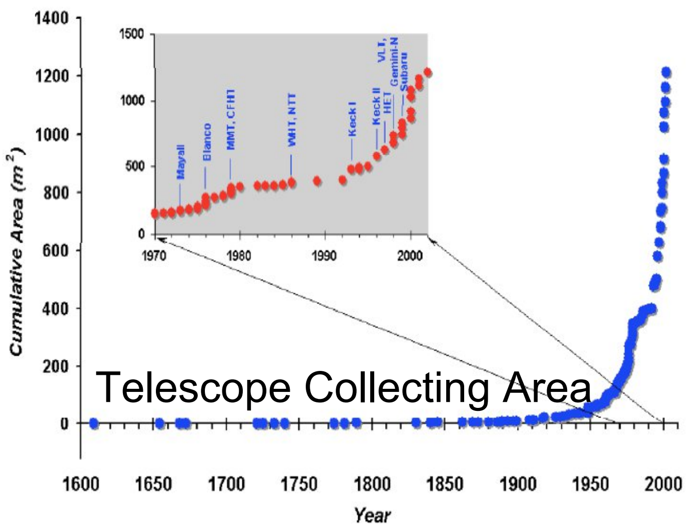
- Objectives 2021-2024:
  - improve the awareness and stimulate improvements in Data Preservation
    - Scientific motivation, organisation, technologies, standards, outreach and education
    - Organise Workshops / issue Global Reports, link to other communities
  - reinforce and support the ongoing laboratory/experiment-based projects and their cooperation
    - keep alive data sets that (can) still produce science, keep track on parked data sets
  - support/develop the DP aspects for future experiments and encourage the transfer of knowledge
  - encourage open data and open science as a way to preserve data and knowledge
- Plans
  - Réinforce Laboratory and FA contacts
  - Release GR2022 and discuss it widely
  - DPHEP Workshop : September 2022
    - “Data Preservation for the future HEP+neutrinos”
- CERN support needed: focal point of ongoing major experiments and computing standards
- What would be the role of the future (EIC, FCC, etc.) experiments?

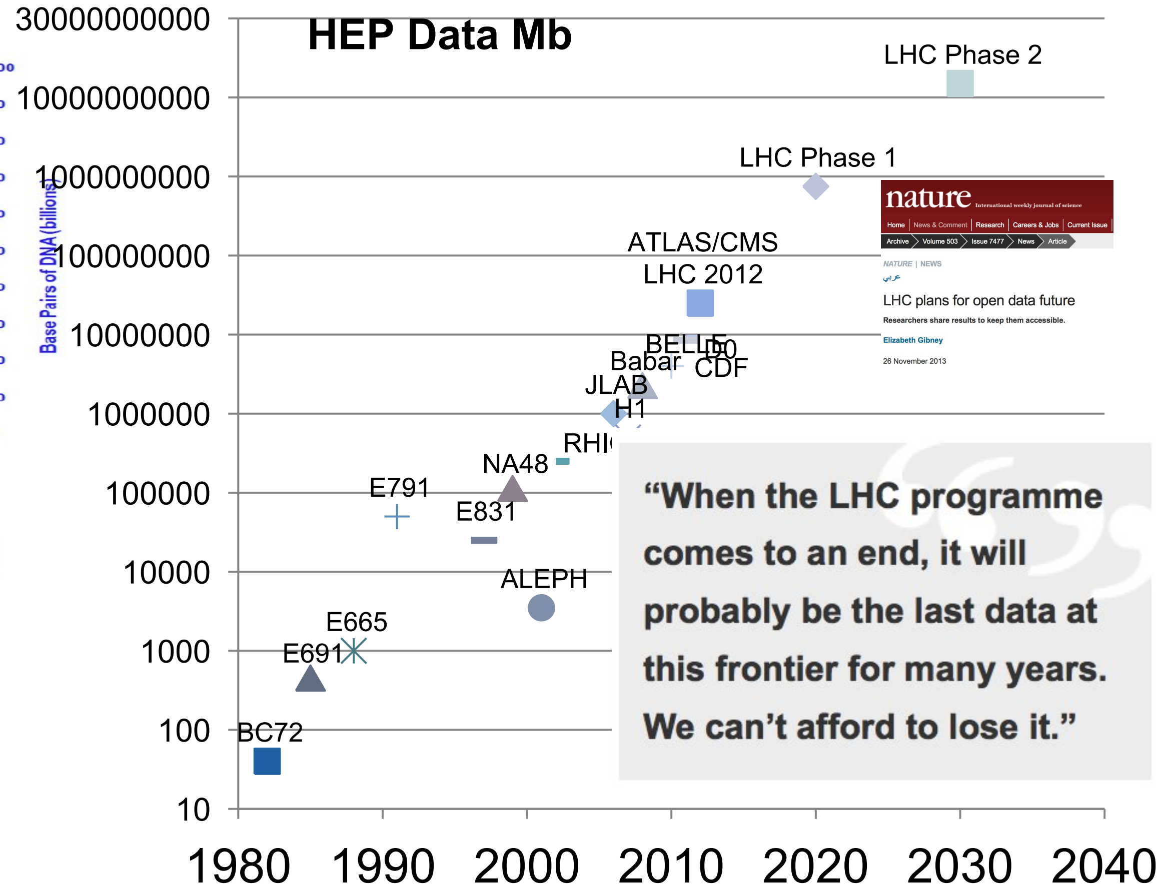
BACKUP

Big Scientific Data

- Scientific research observes a dramatic increase in data and are questioning the long term future of this data







Monday, 21 June

14:00 → 15:15 **The Landscape**

- 14:00 **Workshop Introduction and Goals**
  Speakers: Cristinel Diaconu (CPPM, Aix-Marseille Université, CNRS/IN2P3 (FR)) , Dirk Duellmann (CERN)
- 14:10 **An Open Data policy for LHC**
  Speaker: Jamie Boyd (CERN)
- 14:30 **Welcome from CERN Management**
  Speaker: Joachim Josef Mnich (CERN)
- 14:40 **From LEP to fcc-ee. A bridge too far?**
  Speaker: Jamie Shiers (CERN)
  DPHEP-status_repo... | DPHEP-status_repo... | FROM LEP to fcc-ee... | LEP to fcc-ee.pdf | LHC Computing – t...

15:15 → 15:35 **Coffee**

15:35 → 18:20 **Site / Experiment "Round table": Part 1**

- 15:35 **LHC session**
  - ATLAS activities & plans
    Speakers: Lukas Alexander Heinrich (CERN) , Marumi Kado (Sapienza Universita e INFN, Roma I (IT))
  - CMS activities & plans
    Speaker: Kati Lassila-Perini (Helsinki Institute of Physics (FI))
  - LHCb activities & plans
    Speaker: Adam Morris (University of Bonn (DE))
  - ALICE activities & plans
    Speakers: Jochen Klein (CERN) , Stefano Piano (INFN (IT))
- 16:55 **DESY / HERA**
  - H1 activities
    Speaker: Daniel Britzger (Max-Planck-Institut für Physik München)
  - ZEUS activities
    Speaker: Achim Geiser (Deutsches Elektronen-Synchrotron (DE))
- 17:35 **BaBar**
  Speaker: Concetta Cartaro (SLAC)
- 17:55 **Discussion**

Wednesday, 23 June

09:00 → 11:30 **Service / Project Updates: Service/Project updates**

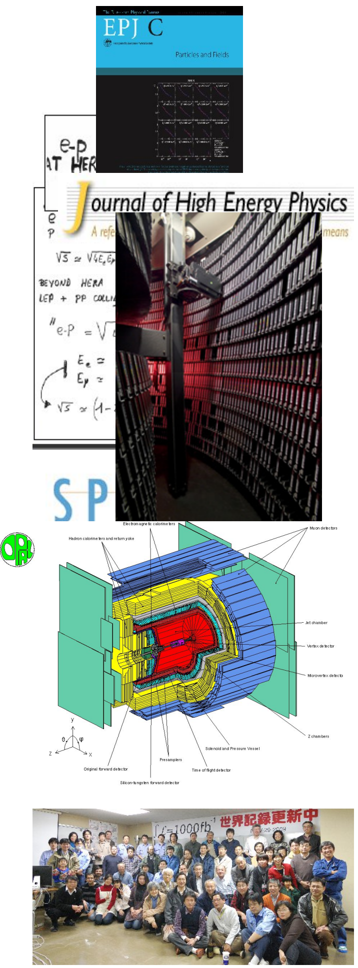
- 09:00 **CERN Open Data Portal**
  Speaker: Tibor Simko (CERN)
- 09:15 **CERN Analysis Preservation**
  Speaker: Pamfilos Fokianos (CERN)
- 09:30 **REANA Reproducible Analyses**
  Speaker: Tibor Simko (CERN)
- 09:45 **Bit Preservation**
  Speaker: Oliver Keeble (CERN)
- 10:00 **Coffee**
- 10:20 **"Software Preservation" - virtualisation et al.**
  Speaker: Jakob Blomer (CERN)
- 10:40 **Managed service migration/retirement**
- 11:00 **Discussion**

11:20 → 19:00 **Site / Experiment "Round table": Part 2**

- 11:20 **BES III**
  Speaker: Lu Wang (Computing Center,Institute of High Energy Physics, CAS)
- 11:40 **KEK / Belle I & II**
  Speaker: Takanori Hara (High Energy Accelerator Research Organization (JP))
- 12:00 **Lunch**
- 14:00 **LEP Session**
  - ALEPH, DELPHI and OPAL Status and Plans
    Speakers: Gerardo Ganis (CERN) , Marcello Maggi (Universita e INFN, Bari (IT)) , Matthias Schroeder (CERN) , Ulrich Schwickerath (CERN)
  - Opportunities offered by LEP data@edm4hep for future EW and Higgs factories
    Speaker: Gerardo Ganis (CERN)
  - Discussion
- 14:45 **CERNLIB**
- 15:15 **Coffee**
- 15:35 **FNAL / Tevatron**
- 15:55 **BNL / RHIC**
  Speaker: Maxim Potekhin (Brookhaven National Laboratory (US))
- 16:15 **Discussion on DPHEP futures**
- 17:45 **CB Meeting**

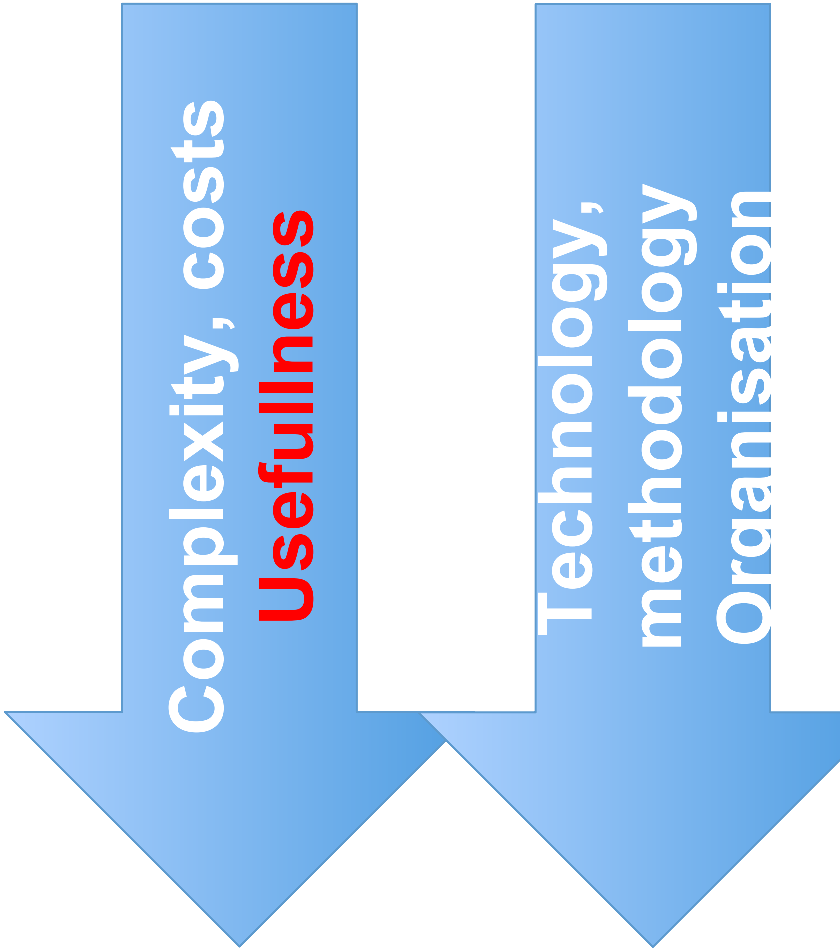
[3rd DPHEP Collab. Meeting](https://indico.cern.ch/event/1043155/)
https://indico.cern.ch/event/1043155/

Scientific Data: what is it?



- Publications
- Documentation
- Raw
- Processed data
- Meta-data
- Workflows
- **Software (all branches)**
- Diffuse knowledge

….more…



DPHEP ressources for DP

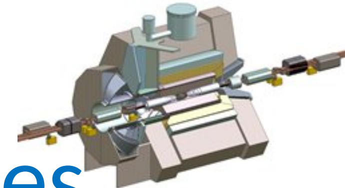
- 2012 Blueprint

	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.

2018 status

BABAR Highlights and Press Releases



BERKELEY LAB
Bringing Science Solutions to the World

About the Lab | Leadership/Organization | Calendar | News Cent

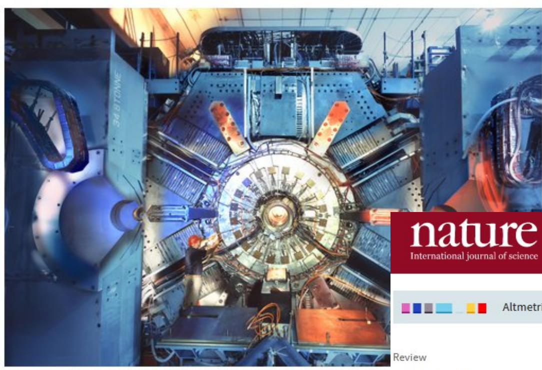
NEWS CENTER

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

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

News Release Glenn Roberts Jr. (510) 5

November 2017



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

INTERACTIONS.ORG
PARTICLE PHYSICS NEWS AND RESOURCES

Home | About | Particle People

A communication resource from the world's particle physics laboratorie

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

8 November 2017 - Lawrence Berkeley National Laboratory

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

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

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

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

"Although it does not rule out the existence of dark photo hide, and definitively rule out their explanation for anothe property of the subatomic particle known as the muon," s University of Victoria professor.

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

DATE ISS | November 8 | SOURCE | Lawrence B | CONTEN | Press Release | CONTACT: | y National Labo

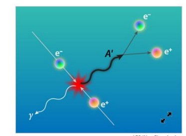
Physics | ABOUT | BROWSE | PRESS | COLLECTIONS

Viewpoint: New Light Shed on Dark Photons

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

November 10, 2014 • *Physics* 7, 115

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.



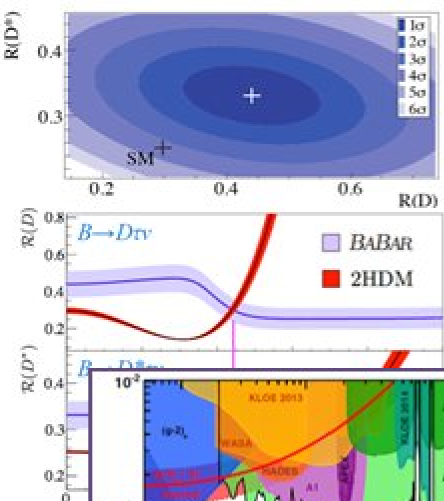
Search for a Dark Photon in e^+e^- Collisions at *BaBar*
J. P. Lees *et al.* (BaBar Collaboration)
Phys. Rev. Lett. 113, 201801 (2014)
Published November 10, 2014

Read PDF

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



nature
International journal of science

Altmetric: 512 | Citations: 2 | More det

Review

A challenge to lepton universality in B-meson decays

Gregory Ciezarek[1], Manuel Franco Sevilla[2], Brian Hamilton[3], Robert Kowalewski[4], Thomas Kuhr, Vera Lüth[6] & Yutaro Sato[7]

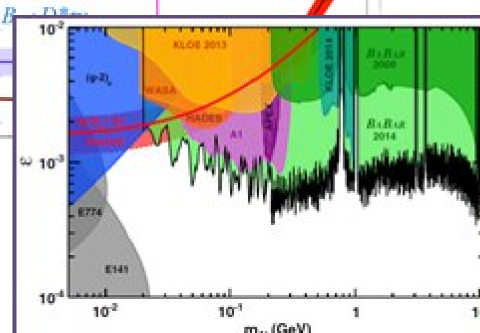
Nature 546, 227–233 (08 June 2017)
doi:10.1038/nature2
Download Citation

Received: 15 December 2016
arch 2017
e: 07 June 2017

June 2017

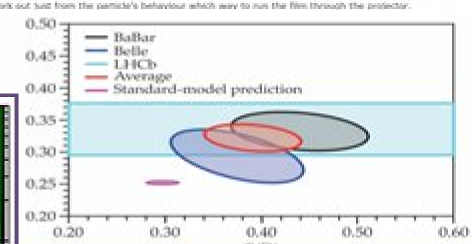
Experimental particle physics | Phenomenology | Theoretical particle physics

Abstract



The Economist

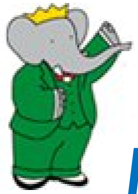
The arrow of time
Backward ran sentences...
To the relief of physicists, time really does have a preferred direction



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) → Y(1S) $\pi^+\pi^-$



2018 status

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

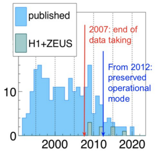
- 2PB of data on T10k-D tapes
  - raw, processed, Monte Carlo
  - **Unique dataset at the Y(3S) resonance** (no plan at the moment to run at the Y(3S) @ Belle II)
- Full environment enclosed in VMs (SL5,SL6)
- ~1TB of documentation, repositories, and dataset information (DBs, cvs, wiki, html)
  - Internal 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**

HERA: succesful DP, towards open data

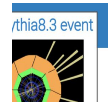
Preserved Operational Mode: 2012-2020

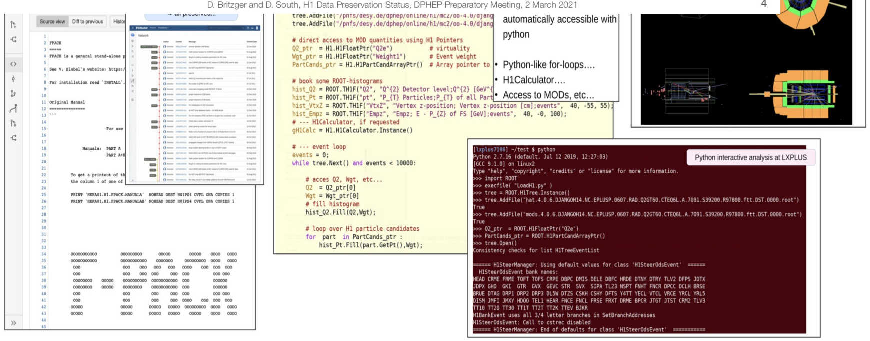
- Significant "level 1" effort made on documentation
- New collaboration model: H1 Physics Board
- H1 physics publications continued
  - 28% of 233 papers after data taking ended
  - 66 after 2007, including 18 after 2012



- 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, CVS compilers, GNU utilities, gmake, system libraries	CVS
External dependencies (selected fixed releases)	H1	fastjet, neurobayes–expert, MC generators	CERNLIB, GKS, GEANT3, ROOT5, LHAPDF5, MC generators

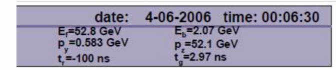




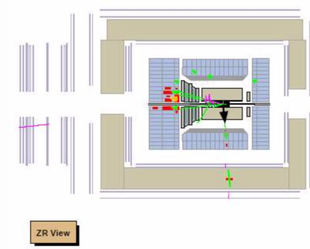
- H1 has unique data and continues to produce physics publications, long after data taking ended
- A recent software modernisation program has been performed to allow this to continue using modern analysis tools, recent programming languages and on state-of-the-art platforms

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, ...)



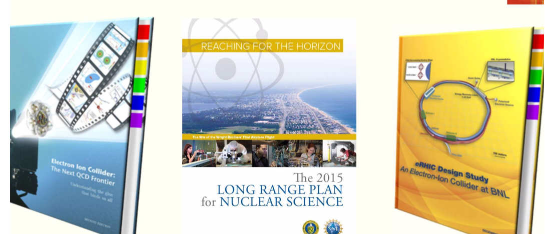
- **Well tested !**
  almost all recent ZEUS papers (24 out of 25) based on Common Ntuples
- **Easy to maintain**
  transition sl5 -> sl6 -> sl7 completely transparent
  (just use newer ROOT version)
- **"Easy" to use**
  most recent ZEUS papers based on results produced by master students, PhD students 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



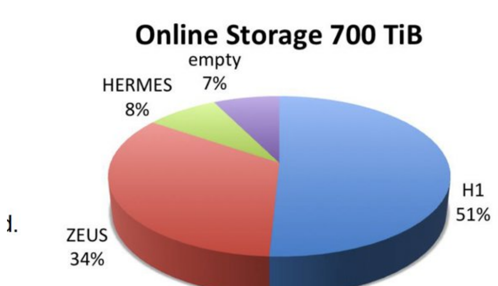
- **Low threshold for access to data by external groups**

...periment: EIC

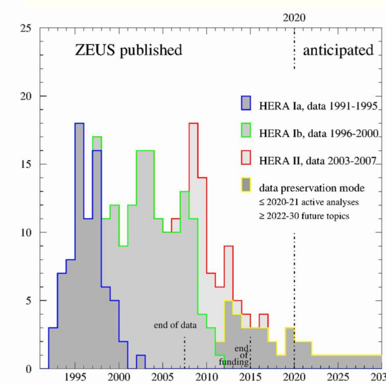
- many EIC topics common with HERA



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



ZEUS physics papers

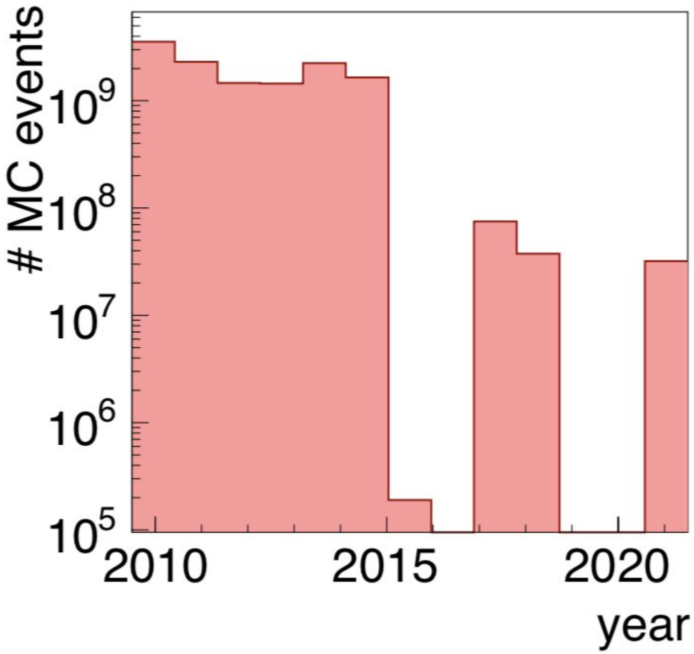


majority of papers produced in "data preservation mode" already since 2012 (25 papers)

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

H1 DP



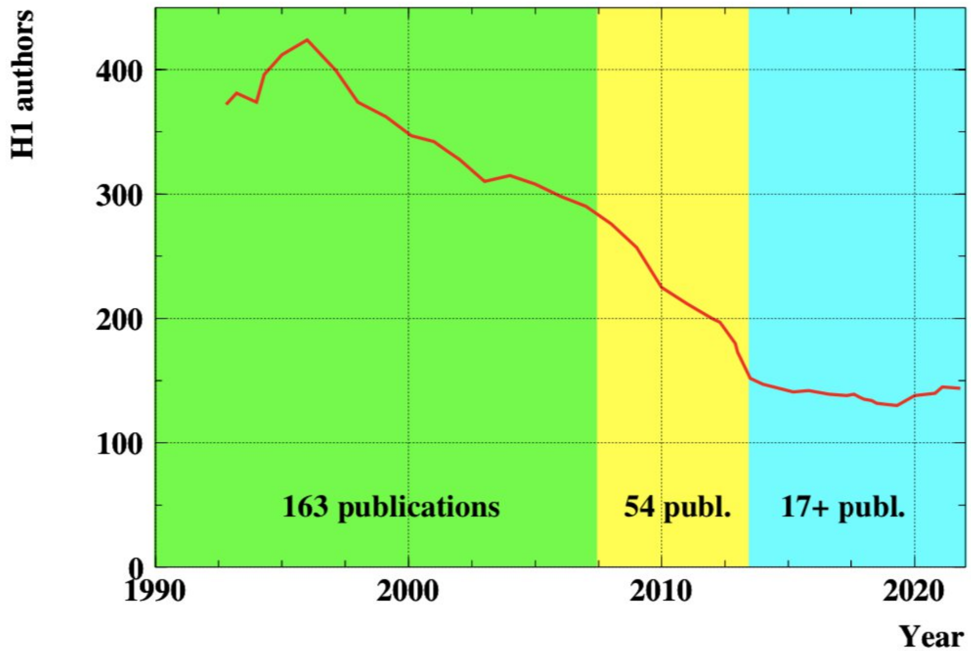


Figure 3: Left: Number of Monte Carlo events produced centrally by the H1 Collaboration. The years without MC production are related to a change of the computing environment, or no MC requests. Right: Number of H1 authors is increasing since 2019 due to retained analysis capabilities and new interest 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.

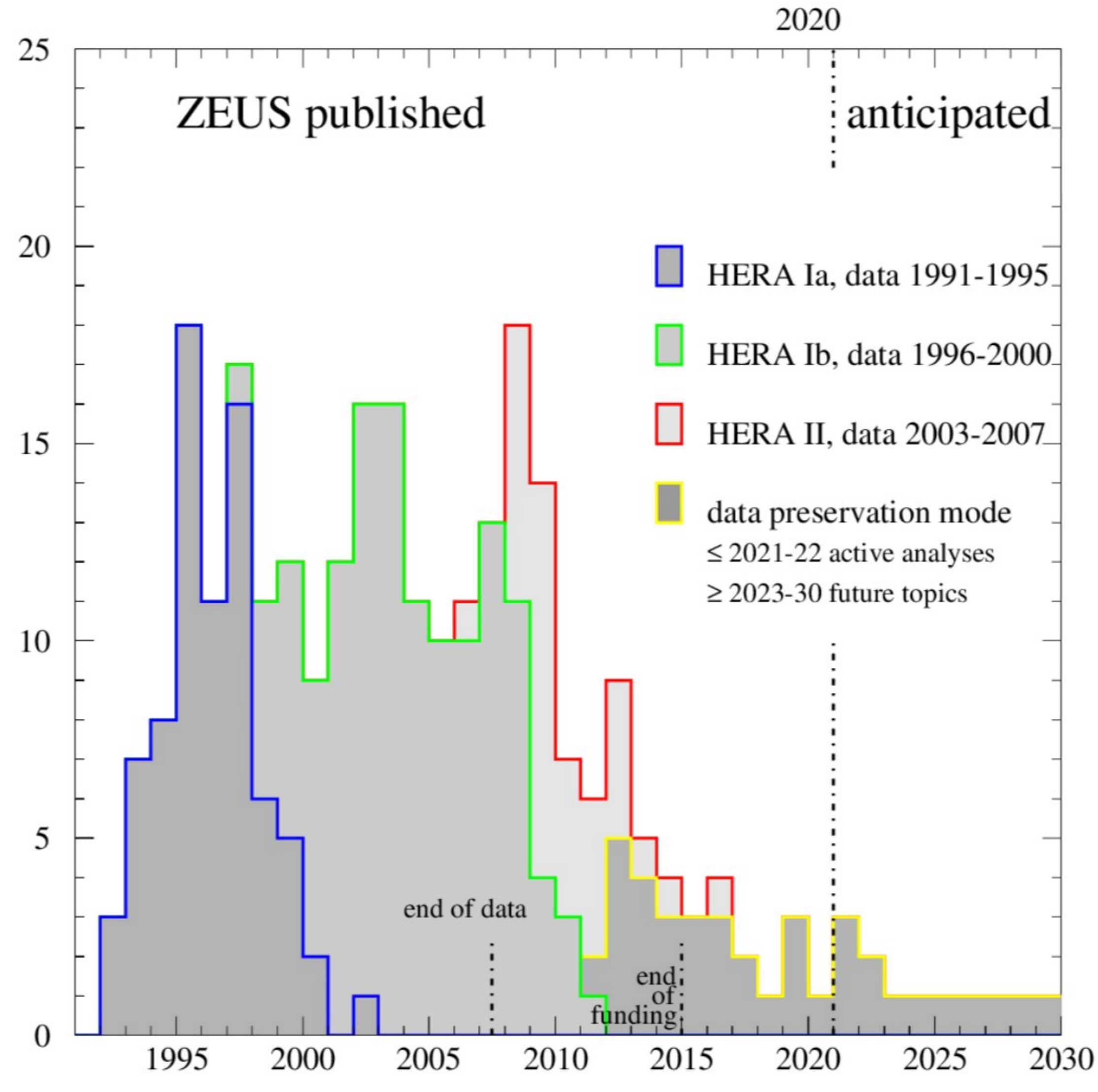


Figure 4: Number of ZEUS papers published and anticipated to be published per year (will be updated to 2022 version).

2018 status

DPHEP timelines

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	**2018**	**2019**	**2020**	**2021**	**2022**
HEP	**HERA tops**	**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 Collab. Meeting		2nd DPHEP Collab. Meeting				3rd DPHEP Collab. Meeting	
DPHEP Docs			**DPHEP White Paper**			**Blueprint Report**			DPHEP Status Report 2020 Vision		DPHEP 2017 Status Report				**ICFA mandate Renewed**	**DPHEP Status Report (in work)**
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					**CERN open science policy**	

Start-up | Consolidation | DPHEP Collaboration

Preservation: where is the problem?

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."**

Storing the data is not a problem: hard drives are cheap and getting cheaper. The challenge is preserving knowledge that is less commonly stored — the software, algorithms and reference plots specific to each experiment. These often degrade or disappear with time, says Cristinel Diaconu of the Marseilles Centre for Particle Physics in France, who is chair of the international Data Preservation in Long Term Analysis in High Energy Physics (DPHEP) study group. He worries that if the data continue to be stored in their current state, physicists trying to decipher them in 10 years' time will be unable to reconstruct the discovery of the Higgs boson. "When the LHC programme comes to an end, it will probably be the last data at this frontier for many years," he says. "We can't afford to lose it."