Data Preservation in High Energy Physics.

Safeguarding the heritage of HEP data for the future

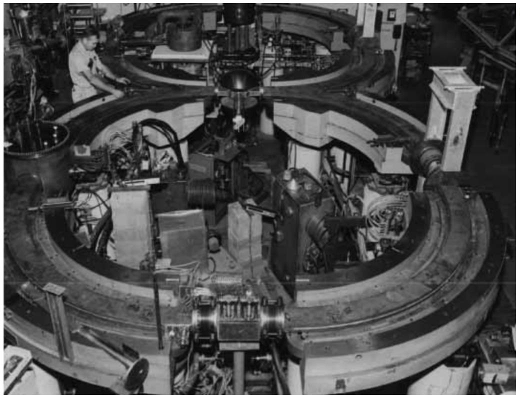
27th IEEE (MSST 2011) Symposium on Massive Storage Systems and Technologies and Co-located Events

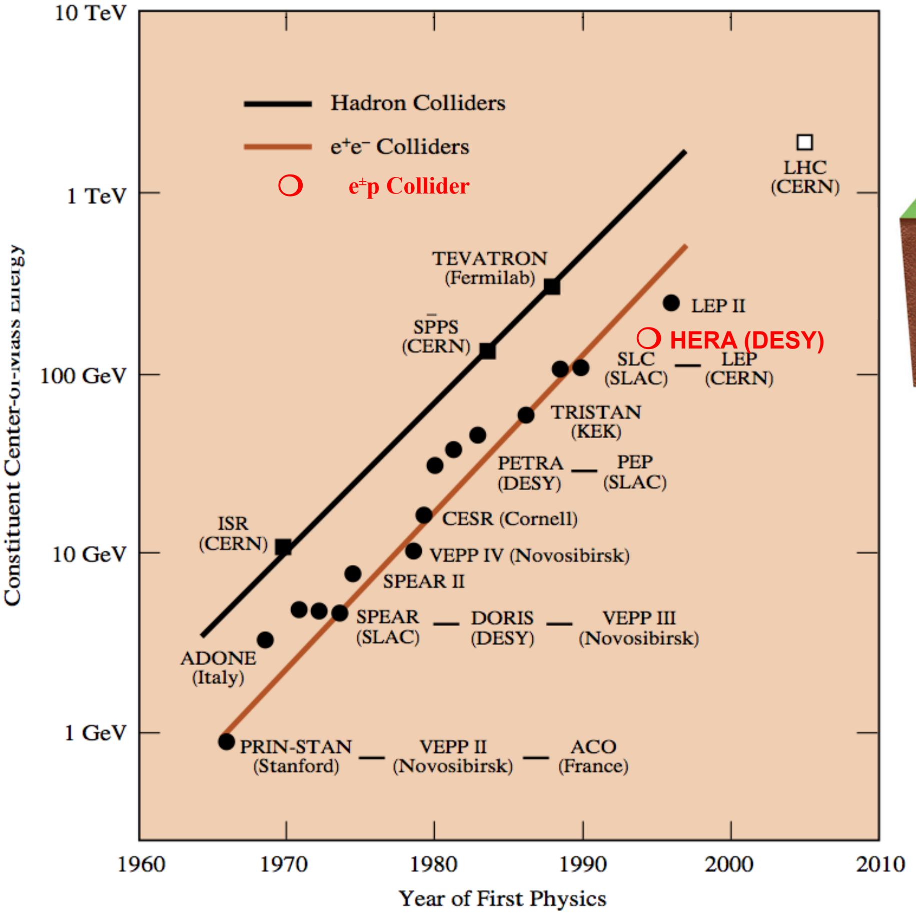
Dmitry Ozerov (DESY)
on behalf of the ICFA
DPHEP Study Group, dphep.org

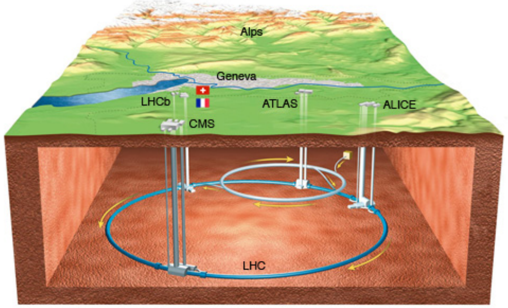
The Last 50 Years of High Energy Physics

PRIN-STAN,
built late 1950's

The first colliding-beam machine, a double-ring electron-electron collider, built by a small group of Princeton and Stanford physicists. (Courtesy Stanford University)



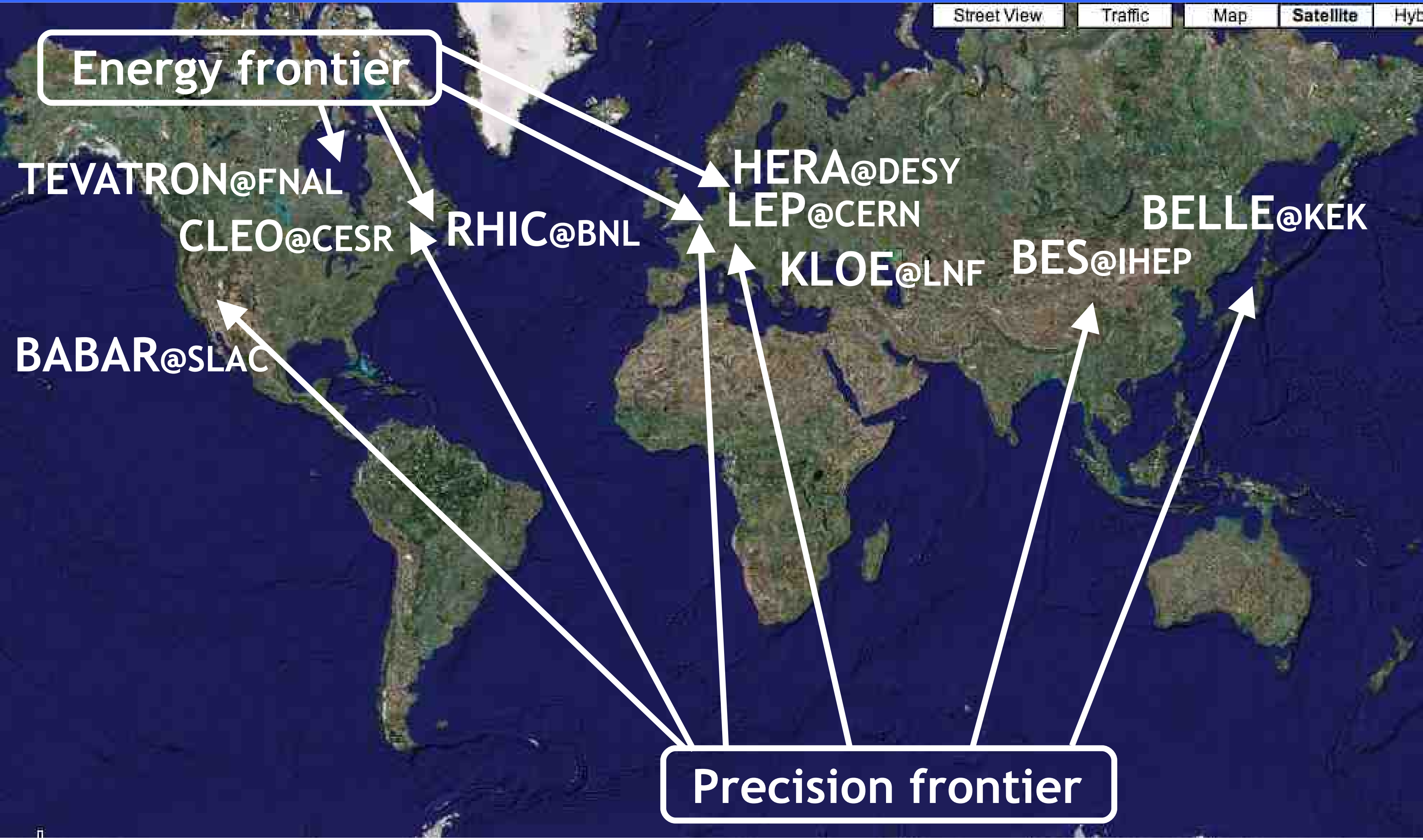




First collisions observed at the LHC in 2008; first data taking at 7 TeV now!

- Energy frontier probed with complex experimental installations
- New experiments normally supercede previous/similar ones - but not always..
- What is the present situation?

Active Experiments in the Pre-LHC Landscape



Energy frontier

TEVATRON@FNAL

CLEO@CESR

RHIC@BNL

HERA@DESY

LEP@CERN

KLOE@LNF

BES@IHEP

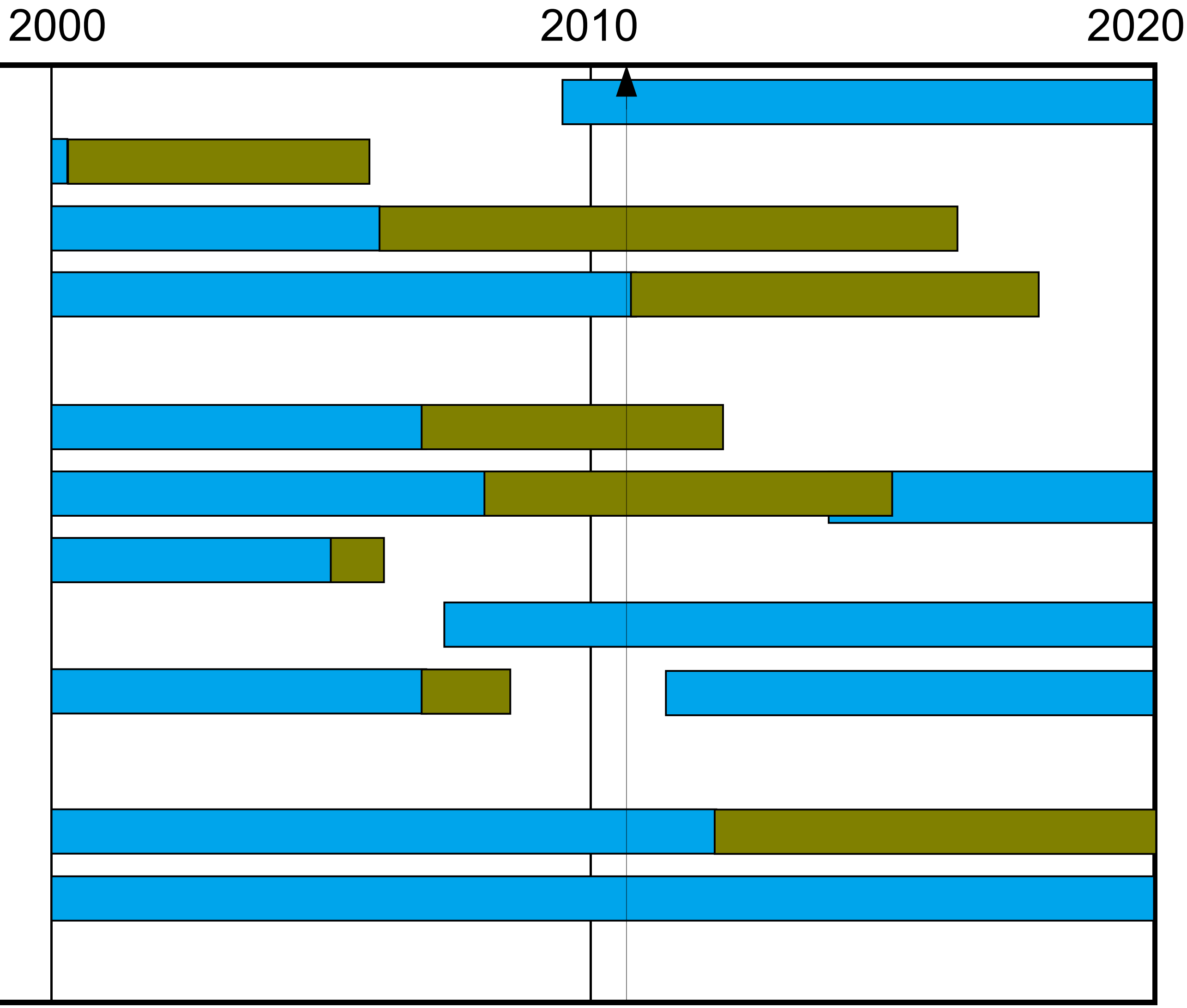
BELLE@KEK

BABAR@SLAC

Precision frontier

HEP Experimental Programmes in ± 10 Years

LHC	pp / ions	fb^{-1} ++
LEP	ee	0.9 fb^{-1}
HERA	ep	0.5 fb^{-1}
Tevatron	$p\bar{p}$	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



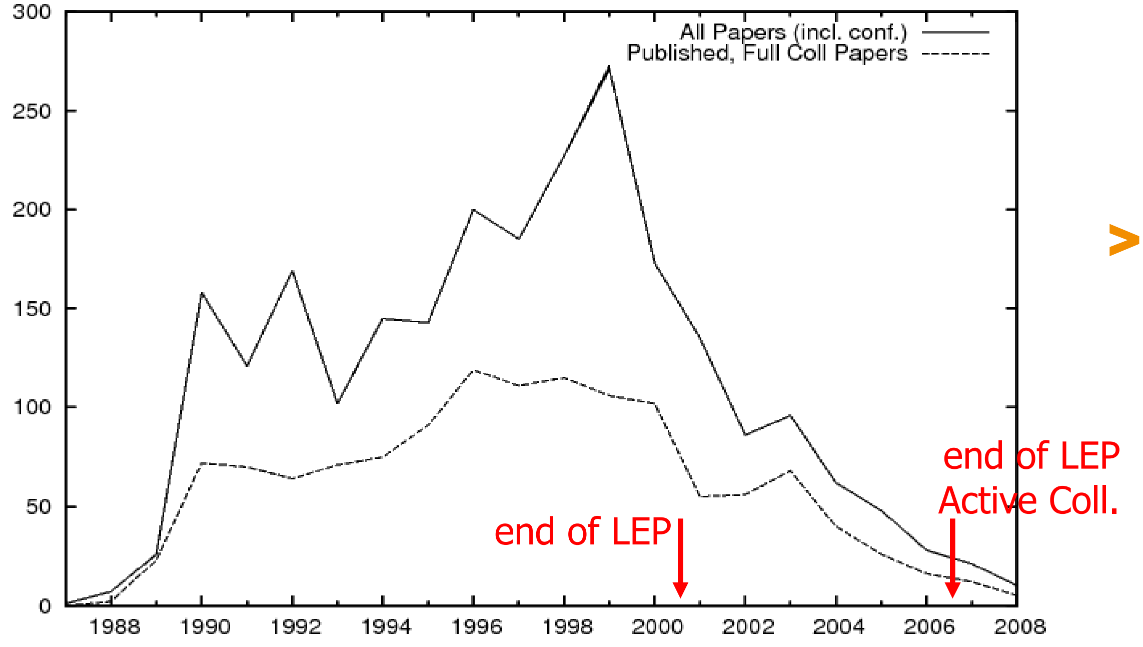
[not all programmes, dates are approximate, just to give the picture]

🟦 **-Data taking period**

🟫 **Post data taking, -Active Collaborations**

The Long Tail of LEP

Papers from all 4 LEP experiments (SPIRES Data)

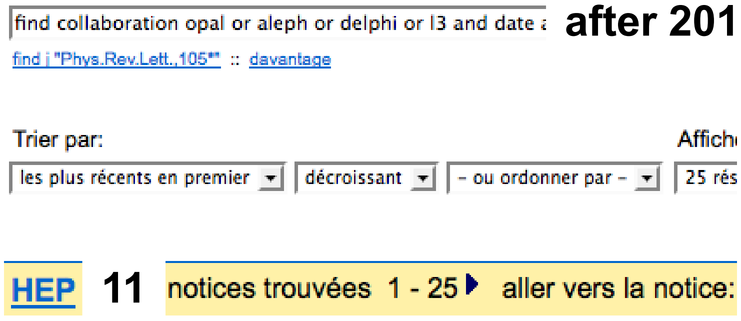


	All	ALEPH	DELPHI	L3	Opal
All physics	345	65	114	85	81
Electroweak	89	17	26	22	24
QCD	85	19	25	19	22
Higgs searches	37	6	14	8	9
SUSY searches	25	4	7	5	9
Exotica search	34	5	12	10	7
Flavor physics	30	6	15	4	5
Exclusive channels	21	3	8	8	2
Cosmo-LEP	12	3	3	6	-
Other	13	2	4	3	3

LEP Publications after 2004

S.Mele, P.Igo-Kemens

- Physics subjects are published after the end of collisions and/or collaborations
- 5-10% of the papers are finalized in the "archival mode"
  - Large number of publications well after data taking stopped
  - Large variety of topics
  - Legacy publications (full data, combined results) came later



find collaboration opal or aleph or delphi or l3 and date a

find j "Phys.Rev.Lett.,105" :: davantage

Trier par: les plus récents en premier ▾ décroissant ▾ - ou ordonner par - ▾

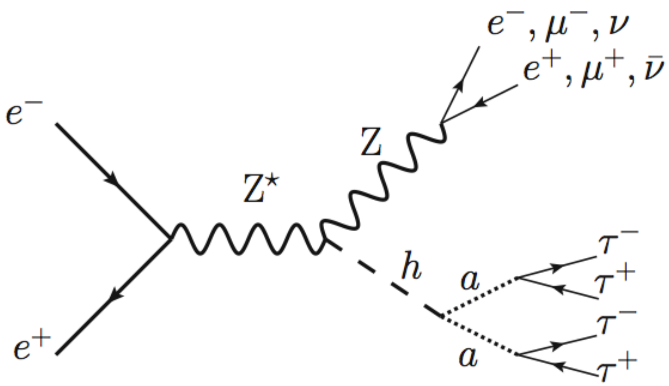
Affich 25 rés

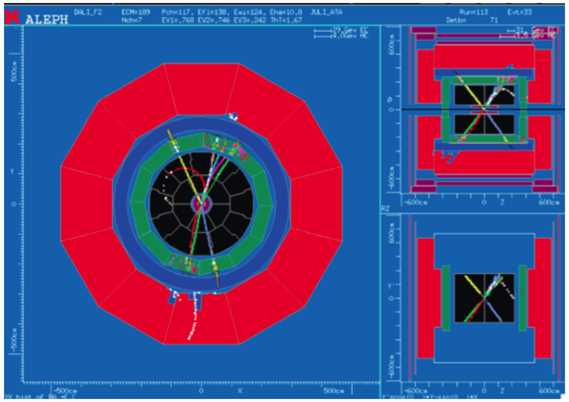
HEP 11 notices trouvées 1 - 25 ▸ aller vers la notice:

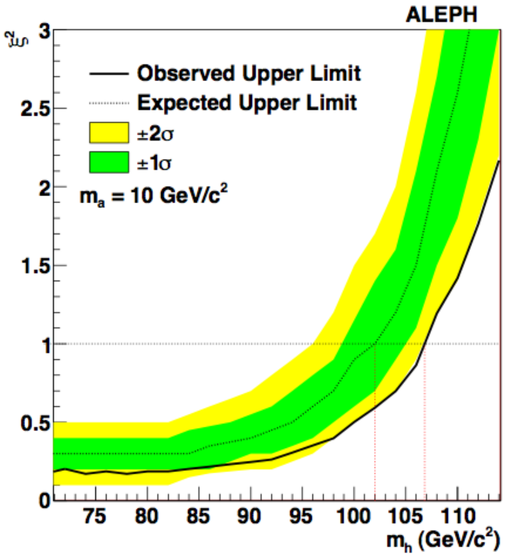
after 2010

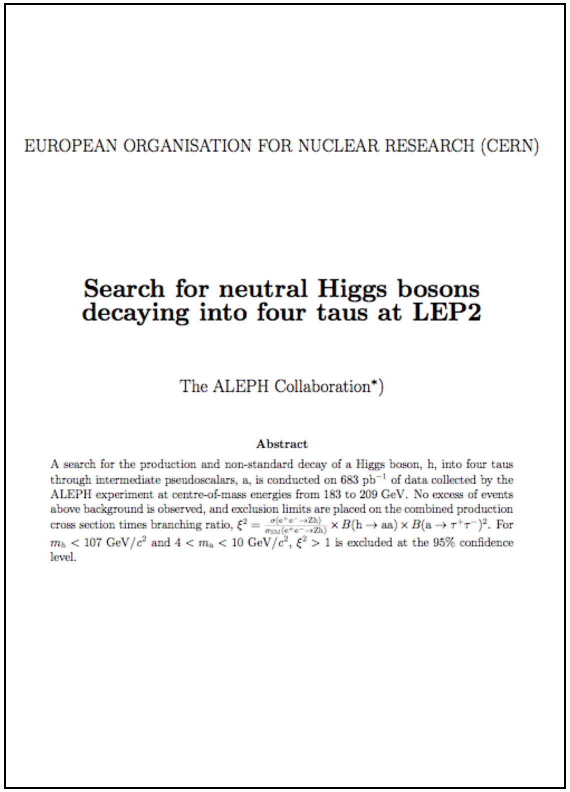
Searches still possible

- Theory and "common sense" evolve
- Unique physics case analysed 10 years after the end of collisions (and 5 years after the official end of the collaboration)









EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

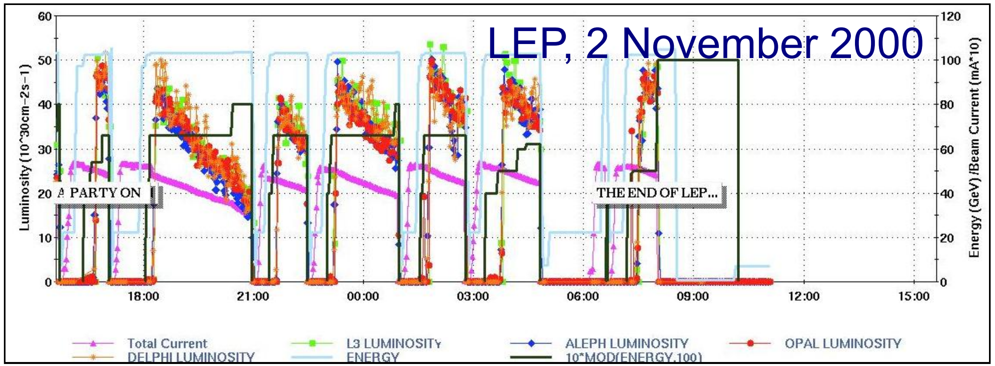
Search for neutral Higgs bosons decaying into four taus at LEP2

The ALEPH Collaboration*)

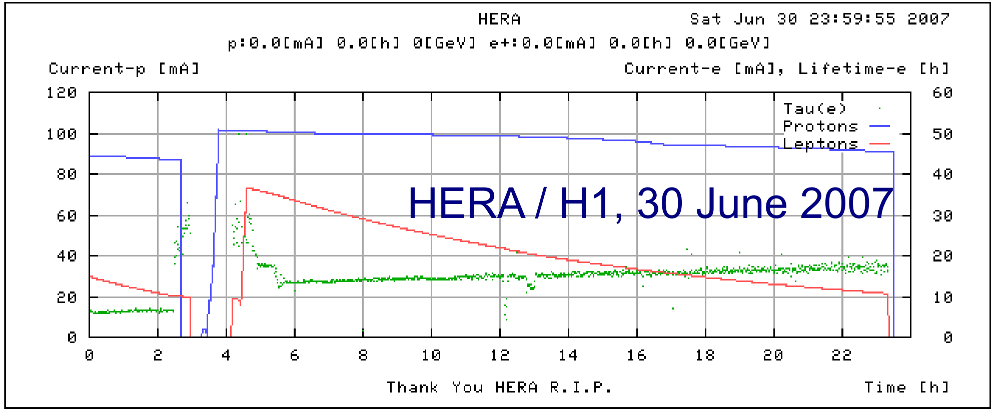
Abstract

A search for the production and non-standard decay of a Higgs boson, h, into four taus through intermediate pseudoscalars, a, is conducted on 683 pb^{-1} of data collected by the ALEPH experiment at centre-of-mass energies from 183 to 209 GeV. No excess of events above background is observed, and exclusion limits are placed on the combined production cross section times branching ratio, $\xi^2 = \frac{\sigma(e^+e^- \to Zh)}{\sigma_{SM}(e^+e^- \to Zh)} \times B(h \to aa) \times B(a \to \tau^+\tau^-)^2$. For $m_h < 107$ GeV$/c^2$ and $4 < m_a < 10$ GeV$/c^2$, $\xi^2 > 1$ is excluded at the 95% confidence level.

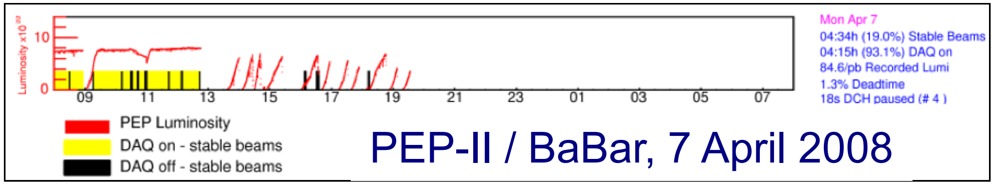
After the End of Data Taking



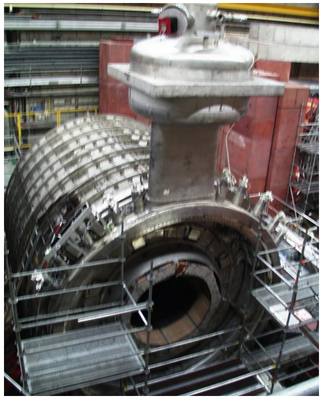
LEP, 2 November 2000

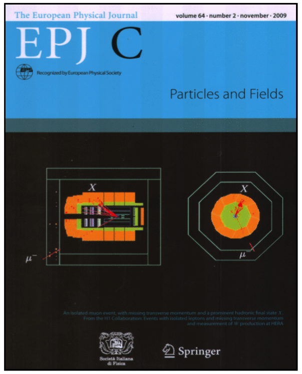


HERA / H1, 30 June 2007



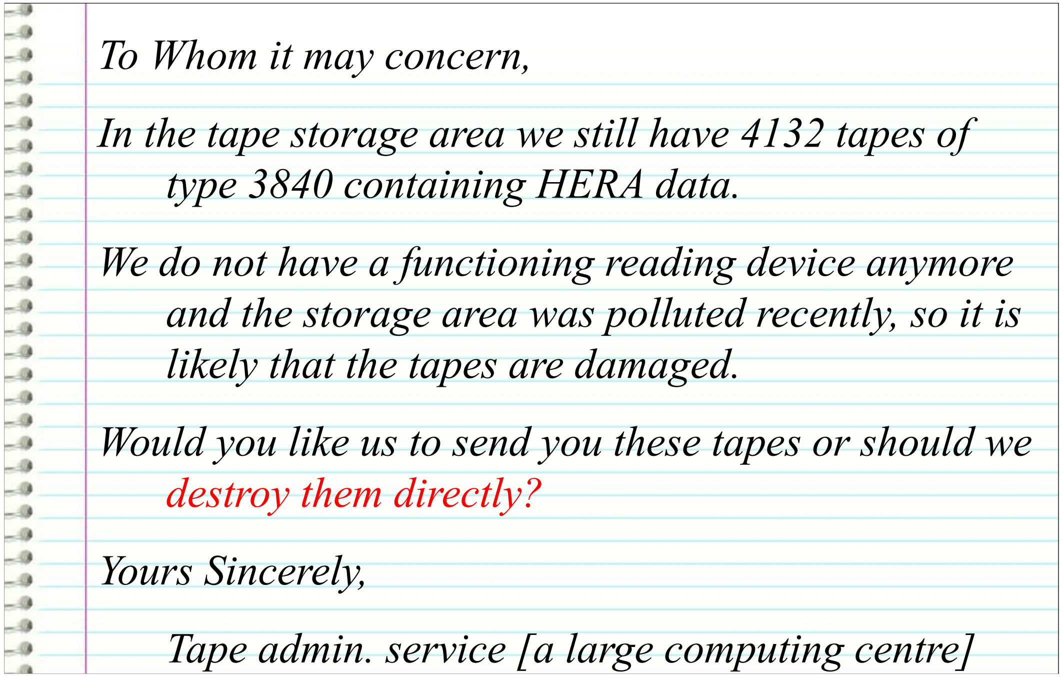
PEP-II / BaBar, 7 April 2008




The European Physical Journal
volume 64 · number 2 · november · 2009
EPJ C
Particles and Fields
Springer

- Have an end of run party, dismantle the detector, finalize the analyses,.. *all in all about 5 years*
- *And then what do you do with the data?*

A Few Communiqués Suggest a Common Problem…



To Whom it may concern,

In the tape storage area we still have 4132 tapes of type 3840 containing HERA data.

We do not have a functioning reading device anymore and the storage area was polluted recently, so it is likely that the tapes are damaged.

Would you like us to send you these tapes or should we destroy them directly?

Yours Sincerely,

Tape admin. service [a large computing centre]

- Some other choice quotes:

"We cannot ensure data is stored in file formats appropriate for long term preservation.

"We cannot ensure those data are still usable. The software for exploiting those data is under the control of the experiments.

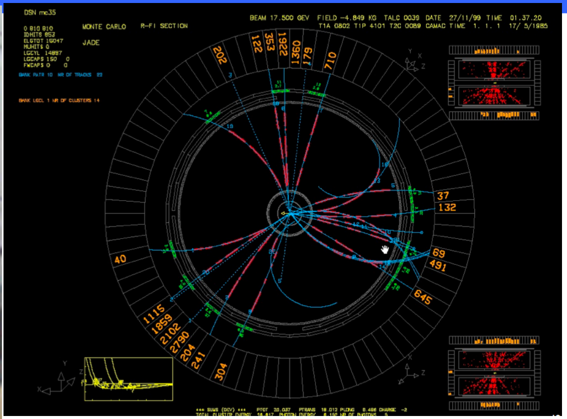
"We are sure most of the data are (not easily) accessible!"

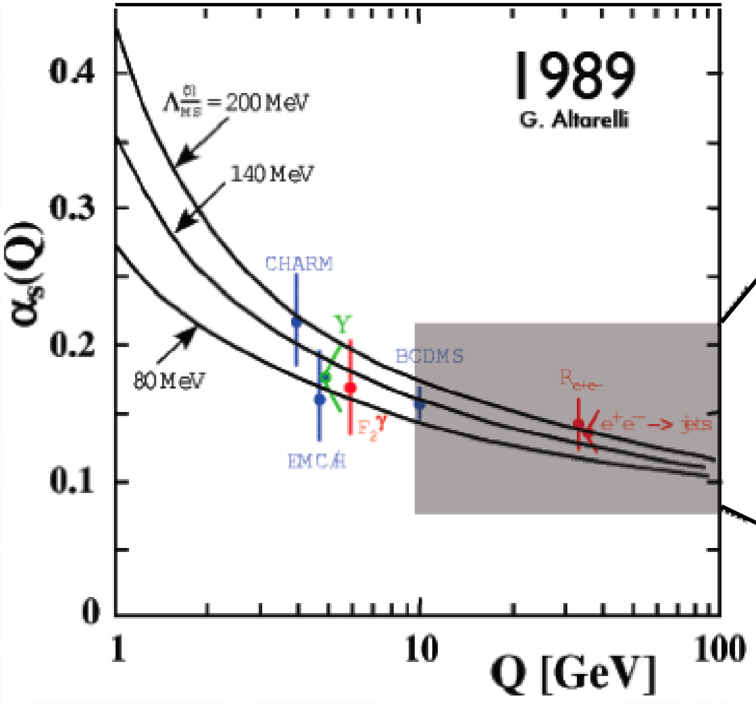
Past Experiences of Data Preservation in HEP

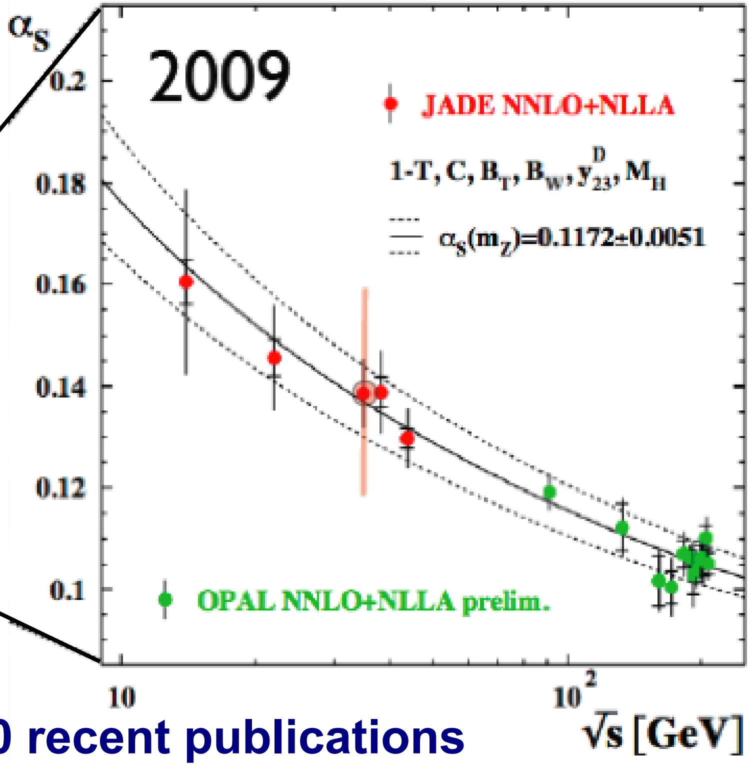
- No tradition, no model
- Data is lost or practically unavailable after a few years
- DP Not part of the planning, software design or budget of a HEP experiment
- Preservation examples are so far individual initiatives

Successful Resurrection of JADE Data Analysis

- Required full raw data preservation, software revitalisation, needed many individual initiatives…





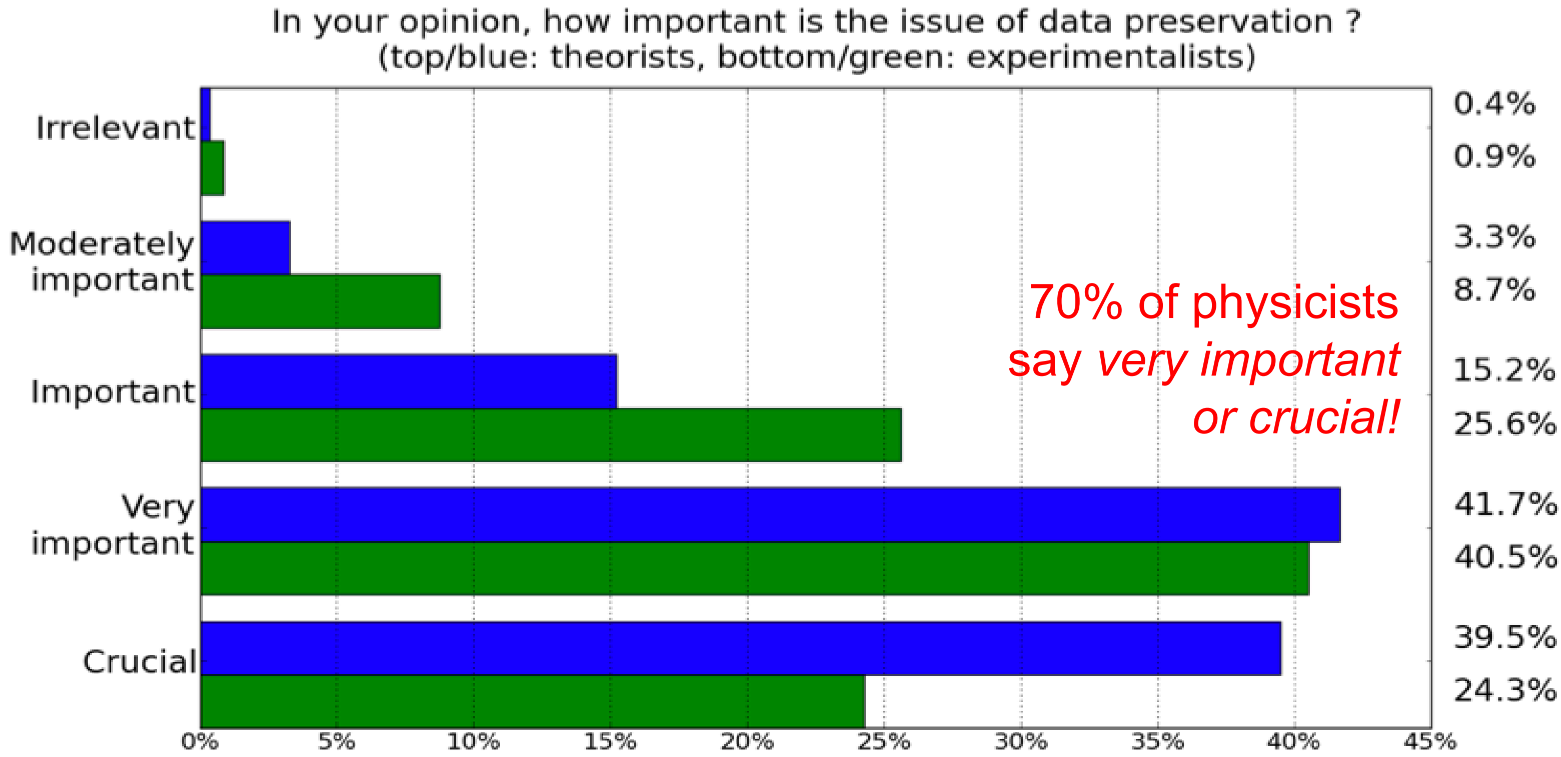


10 recent publications

PARSE.Insight: Support in the HEP Community

PARSE.Insight is financed by the European Commission and run at CERN

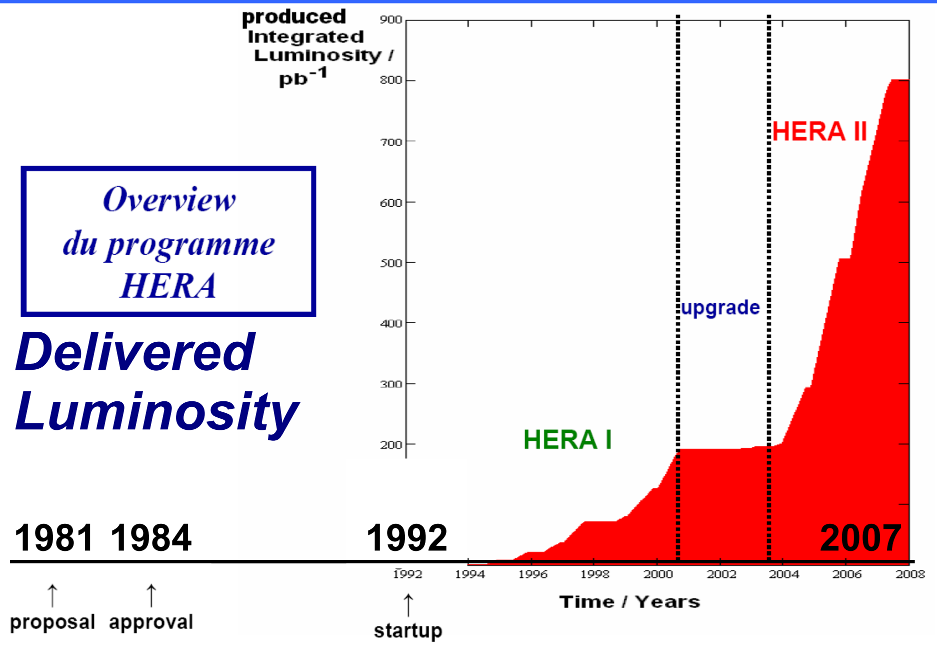
arXiv:0906.0485



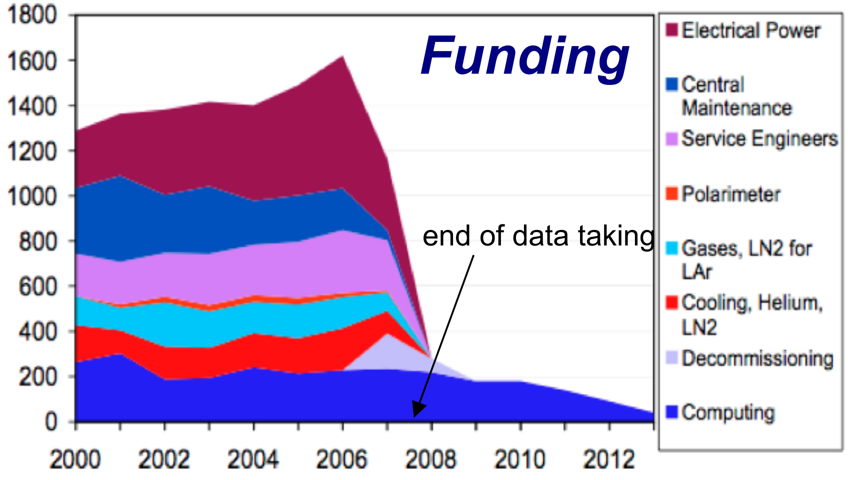
70% of physicists say *very important or crucial!*

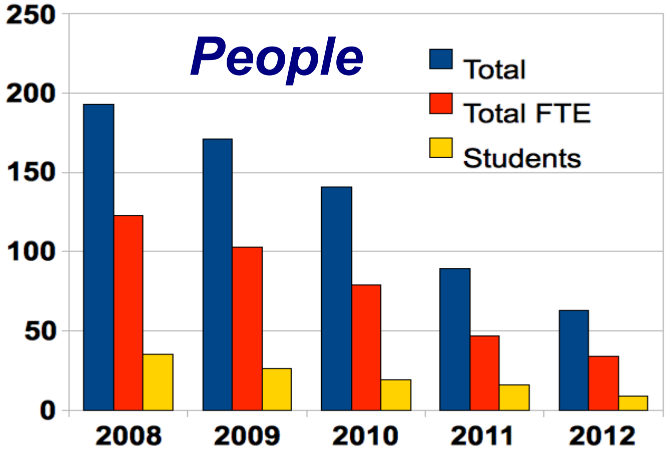
- However, no coherent strategy exists: in general, HEP data are lost
- The task in hand is to provide a coherent set of guidelines for future experiments to ensure the longevity of our data

Why is it Difficult to Preserve HEP Data?



- Good data taking period is towards the end of running
- The existing resources (funding and expertise) then decrease when the data taking stops





DPHEP: International Study Group on Data Preservation

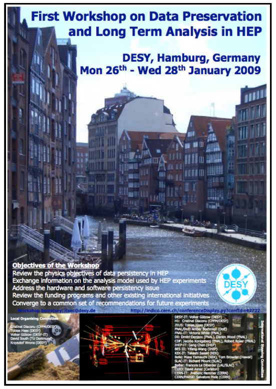
DPHEP Study Group for Data Preservation and Long Term Analysis in High Energy Physics

- Group has grown since 2008 to over 100 contact persons
- Endorsed by ICFA summer 2009
- **LHC** experiments joined in 2011

- Chair: Cristinel Diaconu (DESY/CPPM)
- Working Groups
  - Physics Cases: François Le Diberder (SLAC/LAL)
  - Preservation Models: D. South (DESY), Homer Neal (SLAC)
  - Technologies: Stephen Wolbers (FNAL), Yves Kemp (DESY)
  - Governance: Salvatore Mele (CERN)
- International Steering Committee
  - Participants from ee, ep and pp collider experiments
  - Associated computing centres at the labs
  - Some funding agencies
- International Advisory Committee
  - Chairs: Jonathan Dorfan (SLAC), Siegfried Bethke (MPIM)
  - Advisers: Gigi Rolandi (CERN), Michael Peskin (SLAC), Dominique Boutigny (IN2P3), Young-Kee Kim (FNAL), Hiroaki Aihara (IPMU/Tokyo), Alex Szalay (JHU)

DPHEP Activities

- First contacts established in September 2008
- Series of DPHEP workshops held since 2009
  - Jan2009: DESY May 2009: SLAC Dec 2009: CERN Jul 2010: KEK May 2011: Fermilab



First Workshop on Data Preservation and Long Term Analysis in HEP

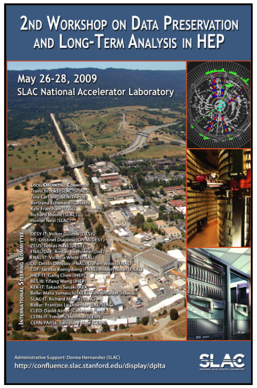
DESY, Hamburg, Germany
Mon 26th - Wed 28th January 2009

Objectives of the Workshop
Review the physics objectives of data persistency in HEP
Exchange information on the analysis model used by HEP experiments
Address the hardware and software persistency issue
Review the funding programs and other existing international initiatives
Converge to a common set of recommendations for future experiments

Workshop Secretary: itsec@desy.de http://indico.cern.ch/conferenceDisplay.py?confId=42722

Local Organising Committee
Cristinel Diaconu (CPPM/DESY)
Tobias Haas (DESY)
Volker Gülzow (DESY-IT)
David South (TU Dortmund)
Krzysztof Wrona (DESY)

International Steering Committee
DESY-IT: Volker Gülzow (DESY)
H1: Cristinel Diaconu (CPPM/DESY)
ZEUS: Tobias Haas (DESY)
FNAL/DoE: Amber Boehnlein (DoE)
FNAL-IT: Victoria White (FNAL)
D0: Dmitri Denisov (FNAL), Darien Wood (FNAL)
CDF: Jacobo Konigsberg (FNAL), Robert Roser (FNAL)
IHEP-IT: Gang Chen (IHEP)
BES III: Yifang Wang (IHEP)
KEK-IT: Takashi Sasaki (KEK)
Belle: Masa Yamauchi (KEK), Tom Browder (Hawaii)
SLAC-IT: Richard Mount (SLAC)
BaBar: Francois Le Diberder (LAL/SLAC)
CLEO: David Asner (Carleton)
CERN-IT: Frederic Hemmer (CERN)
CERN/PARSE: Salvatore Mele (CERN)



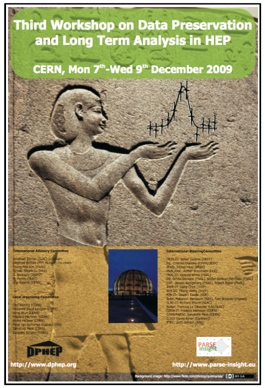
2nd Workshop on Data Preservation and Long-Term Analysis in HEP

May 26-28, 2009
SLAC National Accelerator Laboratory

Local Organizing Committee:
Travis Brooks (SLAC SPIRES)
Tina Cartaro (INFN/Trieste)
Bertrand Echenard (Caltech)
Kyle Fransham (UVic)
Richard Mount (SLAC)
Homer Neal (SLAC)

International Steering Committee
DESY-IT: Volker Gülzow (DESY)
H1: Cristinel Diaconu (CPPM/DESY)
ZEUS: Tobias Haas (DESY)
FNAL/DoE: Amber Boehnlein (DoE)
FNAL-IT: Victoria White (FNAL)
D0: Dmitri Denisov (FNAL), Darien Wood (FNAL)
CDF: Jacobo Konigsberg (FNAL), Robert Roser (FNAL)
IHEP-IT: Gang Chen (IHEP)
BES III: Yifang Wang (IHEP)
KEK-IT: Takashi Sasaki (KEK)
Belle: Masa Yamauchi (KEK), Tom Browder (Hawaii)
SLAC-IT: Richard Mount (SLAC)
BaBar: Francois Le Diberder (LAL/SLAC)
CLEO: David Asner (Carleton)
CERN-IT: Frederic Hemmer (CERN)
CERN/PARSE: Salvatore Mele (CERN)

Administrative Support: Donna Hernandez (SLAC)
http://confluence.slac.stanford.edu/display/dplta



Third Workshop on Data Preservation and Long Term Analysis in HEP

CERN, Mon 7th-Wed 9th December 2009

International Advisory Committee
Jonathan Dorfan (SLAC) (co-chair)
Siegfried Bethke (MPI Munich) (co-chair)
Young-Kee Kim (FNAL)
Hiroaki Aihara (U. Tokio)
D. Boutigny (IN2P3)
M. Peskin (SLAC)
Gigi Rolandi (CERN)

Local organising Committee
Olof Barring (CERN)
Fabienne Baud-Lavigne (CERN)
René Brun (CERN)
Frédéric Hemmer (CERN)
André Holzner (CERN)
Peter Igo-Kemenes (Gjøvik/CERN)
Salvatore Mele (CERN)
Susanne Schäfer (CERN)

International SteeringCommittee
DESY-IT: Volker Gülzow (DESY)
H1: Cristinel Diaconu (CPPM/DESY)
ZEUS: Tobias Haas (DESY)
FNAL/DoE: Amber Boehnlein (DoE)
FNAL-IT: Victoria White (FNAL)
D0: Dmitri Denisov (FNAL), Stefan Söldner-Rembold (FNAL)
CDF: Jacobo Konigsberg (FNAL), Robert Roser (FNAL)
IHEP-IT: Gang Chen (IHEP)
BES III: Yifang Wang (IHEP)
KEK-IT: Takashi Sasaki (KEK)
Belle: Masanori Yamauchi (KEK), Tom Browder (Hawaii)
SLAC-IT: Richard Mount (SLAC)
BaBar: Francois Le Diberder (LAL/SLAC)
CERN-IT: Frederic Hemmer (CERN)
CERN/PARSE: Salvatore Mele (CERN)
CLEO: David Asner (Carleton)
STFC: John Gordon (RAL)

http://www.dphep.org http://www.parse-insight.eu

Background image: http://www.flickr.com/photos/quinnanya/ BY-SA

4th Workshop on Data Preservation in High Energy Physics

8-10 July 2010 @ KEK

Local organising Committee
Nobu Katayama
Takanori Hara
Mitsuaki Nozaki
Yoshihide Sakai
Takashi Sasaki

International Advisory Committee
Jonathan Dorfan (SLAC) (co-chair)
Siegfried Bethke (MPI Munich) (co-chair)
Young-Kee Kim (FNAL)
Hiroaki Aihara (U.Tokyo)
D. Boutigny (IN2P3)
M. Peskin (SLAC)
Gigi Rolandi (CERN)

International SteeringCommittee
DESY-IT: Volker Gülzow (DESY)
H1: Cristinel Diaconu (CPPM/DESY)
ZEUS: Tobias Haas (DESY)
FNAL/DoE: Amber Boehnlein (DoE)
FNAL-IT: Victoria White (FNAL)
D0: Dmitri Denisov (FNAL), Stefan Söldner-Rembold (FNAL)
CDF: Jacobo Konigsberg (FNAL), Robert Rose (FNAL)
IHEP-IT: Gang Chen (IHEP)
BES III: Yifang Wang (IHEP)
KEK-IT: Takashi Sasaki (KEK)
Belle: Masanori Yamauchi (KEK), Tom Browd (Hawaii)
SLAC-IT: Richard Mount (SLAC)
BaBar: Francois Le Diberder (LAL/SLAC)
CERN-IT: Frederic Hemmer (CERN)
CERN/PARSE: Salvatore Mele (CERN)
CLEO: David Asner (Carleton)
STFC: John Gordon (RAL)

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

5th Workshop on Data Preservation and Long Term Analysis in HEP

May 16-18, 2011
Fermilab, Batavia, Illinois, USA

www.dphep.org

Objectives
- Review the physics objectives of data preservation
- Identify common features in analysis models
- Understand the status of data preservation efforts
- Review funding programs and similar international initiatives
- Develop and document a plan for the future

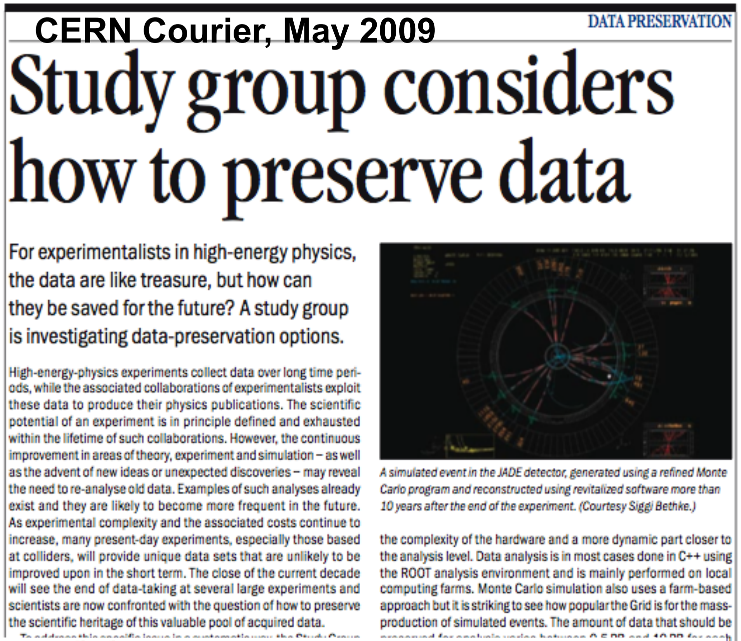
International Steering Committee
DESY-IT: Volker Gülzow (DESY)
H1: Cristinel Diaconu (CPPM/DESY)
ZEUS: Tobias Haas (DESY)
FNAL-IT: Victoria White (FNAL)
D0: Dmitri Denisov (FNAL), Stefan Söldner-Rembold (FNAL)
CDF: Giovanni Punzi (INFN Pisa), Robert Roser (FNAL)
IHEP-IT: Gang Chen (IHEP)
BES III: Yifang Wang (IHEP)
KEK-IT: Takashi Sasaki (KEK)
Belle: Masanori Yamauchi (KEK), Tom Browder (Hawaii)
SLAC-IT: Richard Mount (SLAC)
BaBar: Michael Roney (SLAC/Victoria)
SLAC: Amber Boehnlein (SLAC)
CERN-IT: Frederic Hemmer (CERN)
CERN/PARSE: Salvatore Mele (CERN)
CLEO: David Asner (Carleton)
STFC: John Gordon (RAL)
JLAB: Graham Heyes (JLAB)

International Advisory Committee
Jonathan Dorfan (SLAC) (co-chair)
Siegfried Bethke (MPI Munich) (co-chair)
Young-Kee Kim (FNAL)
Hiroaki Aihara (U Tokyo)
Dominique Boutigny (IN2P3)
Michael Peskin (SLAC)
Gigi Rolandi (CERN)
Alex Szalay (JHU)

Local Organizing Committee:
Stu Fuess (FNAL)
Qizhong Li (FNAL)
Carla Lloyd (FNAL)
Gene Oleynik (FNAL)
Ruth Pordes (FNAL)
Rick Snider (FNAL)
Margaret Votava (FNAL)
Steve Wolbers (FNAL) (Chair)

- Confront data models, clarify the concepts, set a common language, investigate technical aspects, compare with other fields such as astrophysics and others handling large data sets
- With the ultimate aim of providing a set of recommendations concerning data preservation for past, present and future HEP experiments

DPHEP Visibility



CERN Courier, May 2009

DATA PRESERVATION

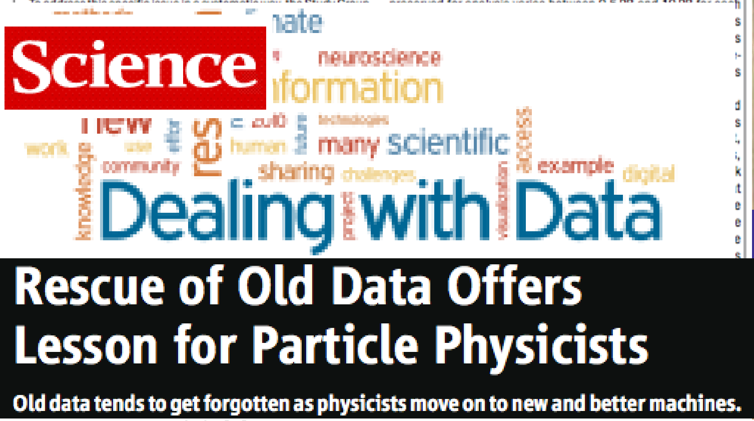
Study group considers how to preserve data

For experimentalists in high-energy physics, the data are like treasure, but how can they be saved for the future? A study group is investigating data-preservation options.

High-energy-physics experiments collect data over long time periods, while the associated collaborations of experimentalists exploit these data to produce their physics publications. The scientific potential of an experiment is in principle defined and exhausted within the lifetime of such collaborations. However, the continuous improvement in areas of theory, experiment and simulation – as well as the advent of new ideas or unexpected discoveries – may reveal the need to re-analyse old data. Examples of such analyses already exist and they are likely to become more frequent in the future. As experimental complexity and the associated costs continue to increase, many present-day experiments, especially those based at colliders, will provide unique data sets that are unlikely to be improved upon in the short term. The close of the current decade will see the end of data-taking at several large experiments and scientists are now confronted with the question of how to preserve the scientific heritage of this valuable pool of acquired data.

A simulated event in the JADE detector, generated using a refined Monte Carlo program and reconstructed using revitalized software more than 10 years after the end of the experiment. (Courtesy Siggi Bethke.)

the complexity of the hardware and a more dynamic part closer to the analysis level. Data analysis is in most cases done in C++ using the ROOT analysis environment and is mainly performed on local computing farms. Monte Carlo simulation also uses a farm-based approach but it is striking to see how popular the Grid is for the mass-production of simulated events. The amount of data that should be



Science

neuroscience
information
many scientific
sharing
example
digital
community
human
access
knowledge
visualization
res
new
Dealing with Data

Rescue of Old Data Offers Lesson for Particle Physicists

Old data tends to get forgotten as physicists move on to new and better machines.

February 2011

Canning, pickling, drying, freezing–physicists wish there were an easy way to preserve their hard-won data so future generations of scientists, armed with more powerful tools, can take advantage of it. They've launched an international search for solutions.

By Nicholas Bock

ZEUS
H1
JADE
CLEO
BaBar
Fermi
OPAL
ALEPH
TeV
L3
DELPHI

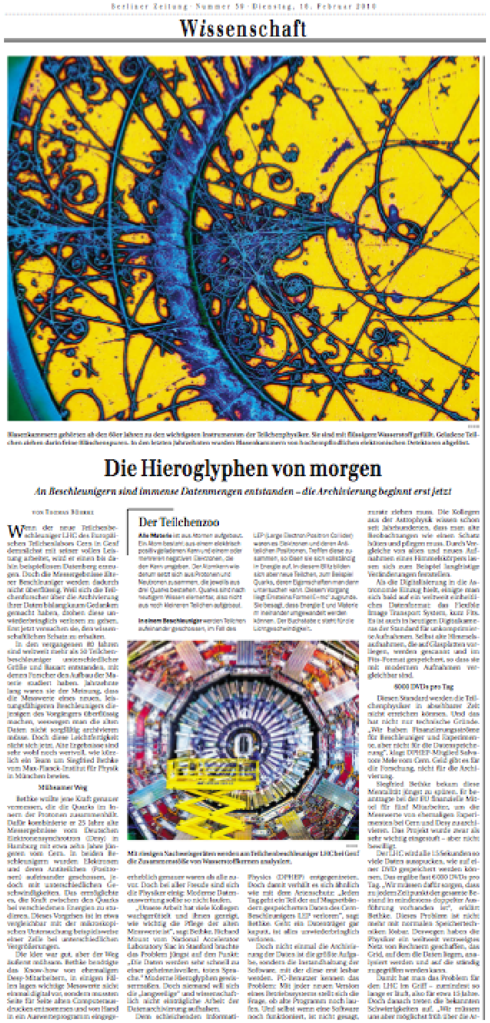
symmetry

dimensions of particle physics

A joint Fermilab/SLAC publication

VOLUME 06 ISSUE 06 DECEMBER 09

Symmetry, December 2009



Berliner Zeitung · Nummer 39 · Dienstag, 16. Februar 2010

Wissenschaft

Die Hieroglyphen von morgen

An Beschleunigern sind immense Datenmengen entstanden – die Archivierung beginnt erst jetzt

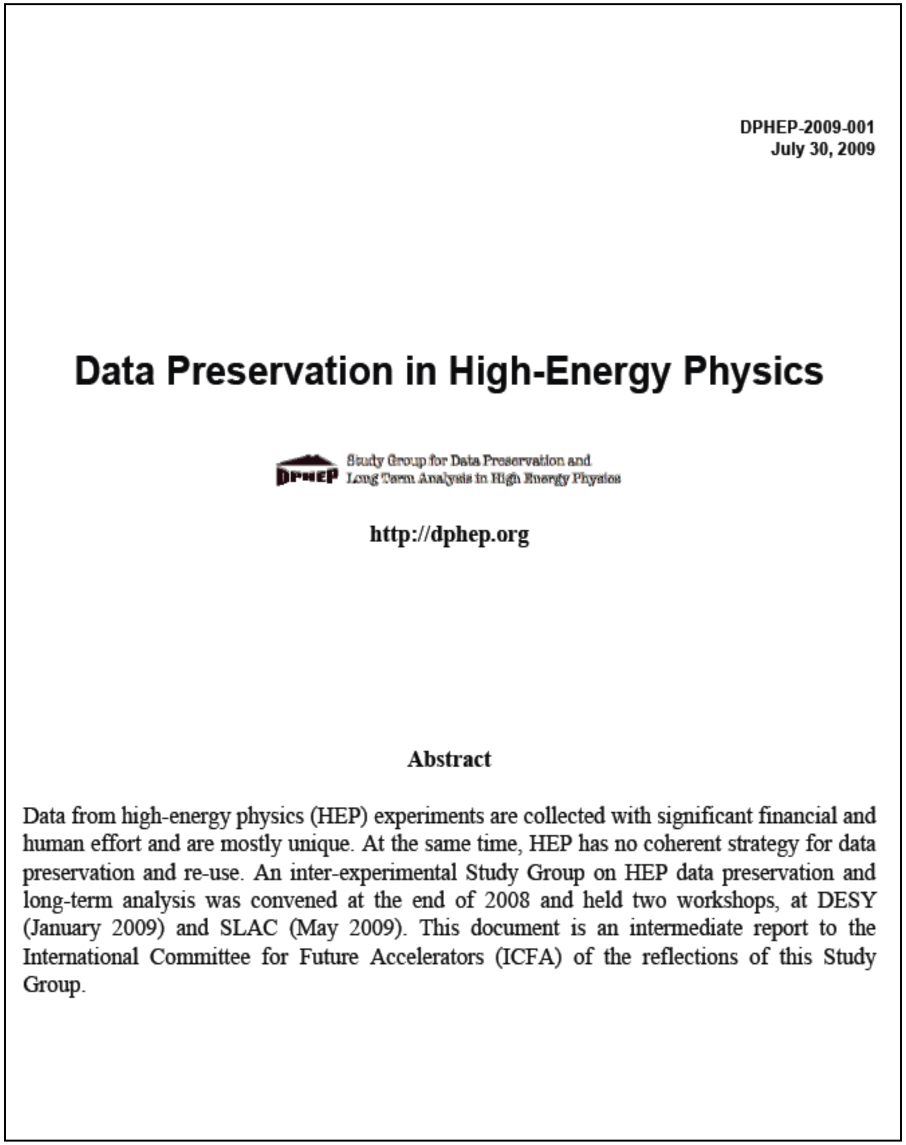
Der Teilchenzoo

Mühsamer Weg

6000 DVDs pro Tag

Berliner Zeitung and Frankfurter Rundschau, February 2010

Intermediate DPHEP Report Released Nov 2009



DPHEP-2009-001
July 30, 2009

Data Preservation in High-Energy Physics

DPHEP 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.

- First recommendations of the group published November 2009 ***arXiv:0912.0255***

- The report covers the four key areas
  - Physics Case for Data Preservation
  - Preservation Models
  - Technologies
  - Governance

Governance

- HEP Collaborations function as international bodies with well defined policies over a few decades
  - A long term data management plan must include a solid governance solution
- Management of the preservation project
  - Scientific supervision of the preserved data sets
  - Authorship and Access to data
  - Channels to outreach and education
  - Endorsement of the project from the experiment, host laboratory and funding agencies
  - HEP global solutions: common policy and standards

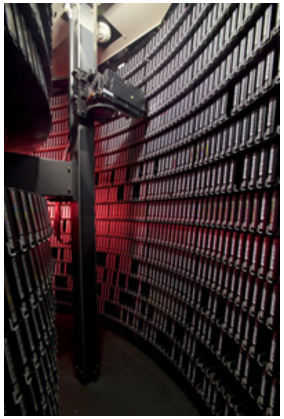
Physics case: Why would we want to re-use old HEP Data?

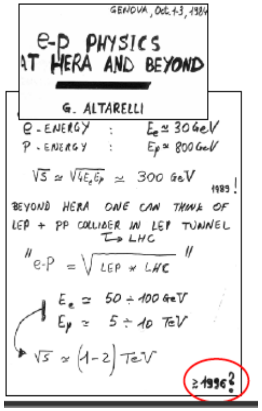
- We may want to re-do previous measurements
  - Increased precision, reduced systematics
  - New and improved theoretical calculations / MC models
  - Newly developed analysis techniques

- We may want to perform new measurements
  - At energies and processes where no other data are available (or will become available in the future)
  - Particularly relevant to HERA $e^{\pm}p$ data (and also Tevatron)

- Investigate if new phenomena found today
  - Go back and check in the old data

What is “HEP Data” anyway?





GENOVA, Oct. 1-3, 1984

e-p PHYSICS
AT HERA AND BEYOND

G. ALTARELLI

e - ENERGY : $E_e \simeq 30\,\text{GeV}$
P - ENERGY : $E_p \simeq 800\,\text{GeV}$

$\sqrt{s} \simeq \sqrt{4E_eE_p} \simeq 300\ \text{GeV}$ 1989 !

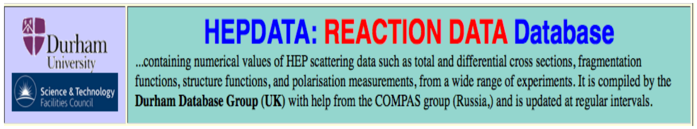
BEYOND HERA ONE CAN THINK OF
LEP + PP COLLIDER IN LEP TUNNEL
→ LHC

"e-P = $\sqrt{\text{LEP} * \text{LHC}}$ "

$E_e \simeq 50 \div 100\ \text{GeV}$
$E_p \simeq 5 \div 10\ \text{TeV}$

$\sqrt{s} \simeq (1-2)\ \text{TeV}$

≥ 1996 ?



Durham University

Science & Technology Facilities Council

HEPDATA: REACTION DATA Database

...containing numerical values of HEP scattering data such as total and differential cross sections, fragmentation functions, structure functions, and polarisation measurements, from a wide range of experiments. It is compiled by the **Durham Database Group (UK)** with help from the COMPAS group (Russia,) and is updated at regular intervals.

Journal of High Energy Physics

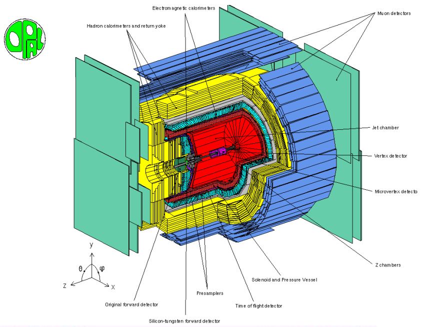
A refereed journal, written, run and distributed by electronic means

SPIRES

EPJ C

Particles and Fields

Springer


Electromagnetic calorimeters
Hadron calorimeters and return yoke
Muon detectors
Jet chamber
Vertex detector
Microvertex detecto
Z chambers
Solenoid and Pressure Vessel
Presamplers
Time of flight detector
Original forward detector
Silicon-tungsten forward detector
y
θ
φ
z
x

> Digital information: Data event files, database

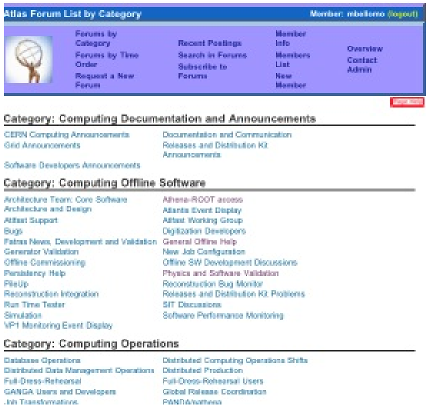
> Software: Simulation, reconstruction, analysis, user

> Publications: Journals, arXiv, Spires/INSPIRE, HEPDATA

> Documentation: Publications, notes, manuals, slides

> Meta information: Hyper-news, messages, wikis, forums

> Expertise (people): Often the hardest to secure



Category: Computing Documentation and Announcements

CERN Computing Announcements
Grid Announcements
Software Developers Announcements
Documentation and Communication
Releases and Distribution Kit Announcements

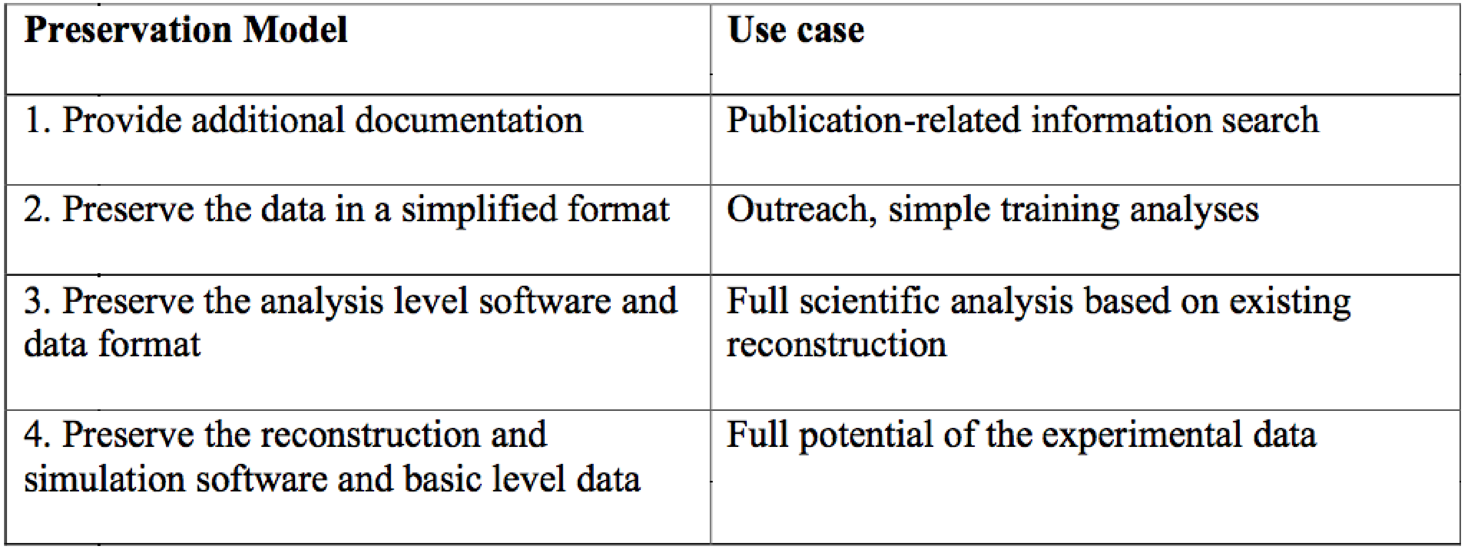
Category: Computing Offline Software

Architecture Team: Core Software
Architecture and Design
Atlfast Support
Bugs
Fatras News, Development and Validation
Generator Validation
Offline Commissioning
Persistency Help
PileUp
Reconstruction Integration
Run Time Tester
Simulation
VP1 Monitoring Event Display
Athena-ROOT access
Atlantis Event Display
Atlfast Working Group
Digitization Developers
General Offline Help
New Job Configuration
Offline SW Development Discussions
Physics and Software Validation
Reconstruction Bug Monitor
Releases and Distribution Kit Problems
SIT Discussions
Software Performance Monitoring

Category: Computing Operations

Database Operations
Distributed Data Management Operations
Full-Dress-Rehearsal
GANGA Users and Developers
Job Transformations
Distributed Computing Operations Shifts
Distributed Production
Full-Dress-Rehearsal Users
Global Release Coordination
PANDA/pathena

Data Preservation Models identified by DPHEP



Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

Cost, complexity, benefits

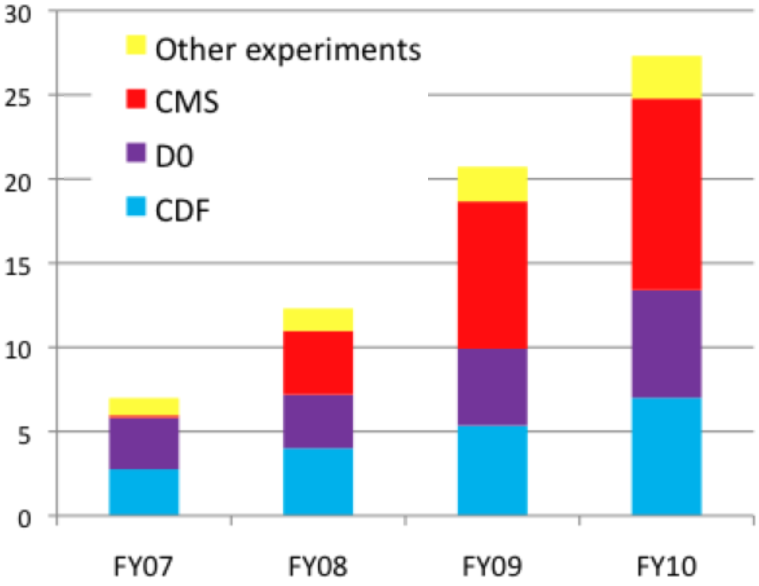
- Only with the full flexibility does the full potential of the data remain
  - Level 4 type programme was required by the JADE
- BaBar, H1, HERMES aim for DPHEP level 4, ZEUS between levels 3 and 4
  - Still some different approaches, can benefit from each other’s experiences
- Even with levels 1 and 2 preservation models one can publish new results

  (LEP analysis now, old data vs new theory)

How much Data are we talking about?

- Discussions in DPHEP lead to a number of around 0.5 to a few PB
  - Depending on preservation model
- Computing centres are, at least by volume arguments, able to store the data
  - Data preservation is not only about the data!
- Regular migration of the data to latest technologies should be considered and carefully planned
- However, currently employed storage systems may not be suited for archival storage
  - Regular integrity checks of the full sample
- Any **archival system** should be able to absorb future technological evolutions

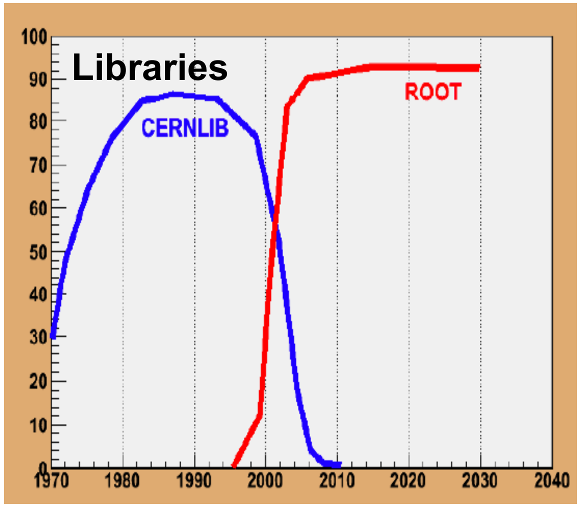
PB on tape at Fermilab at the end of each FY (1st October)

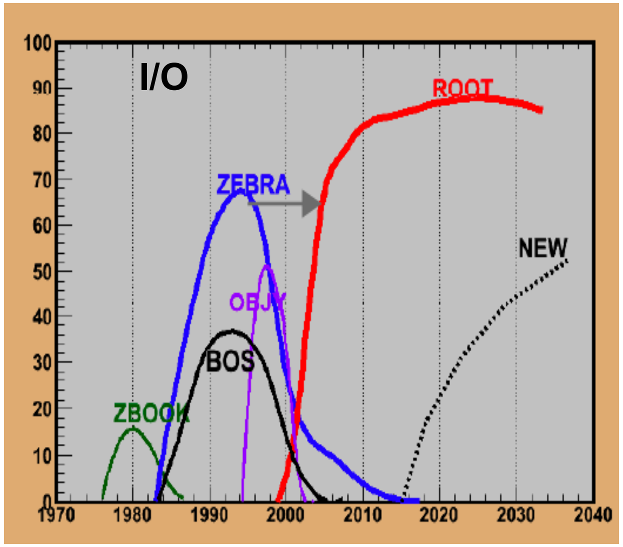


- Copies of the data
  - Different technologies (cost)
  - Geo-distributed (infrastructure to verify consistency, manage access, authentithication/authorisation)
- Standard protocols to access data

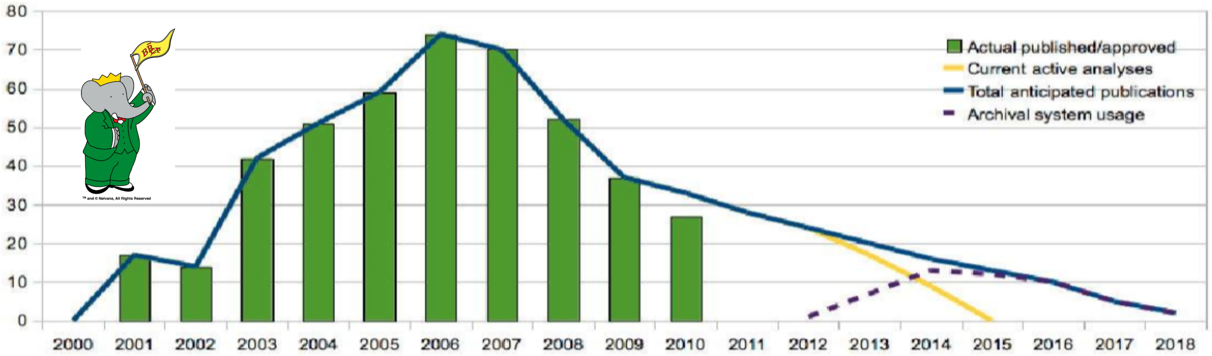
A serious issue: the software maintenance

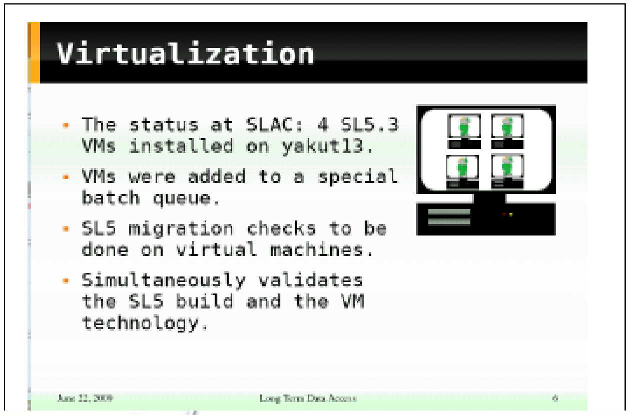
- Freezing: Technology preservation
  - Virtualisation techniques provide the software environment, freeze the hardware
  - Preparation step is not saved, lifetime limited as well
- Better: Continuous migration
  - Follow technology changes, external software, new OS, redesign, recompile etc
  - Virtualisation can help here too
- Preparation is not trivial
  - New operational model
  - Dependencies etc.
- Supervision is needed for both data and software
  - Data archivist position





Data Preservation at BaBar

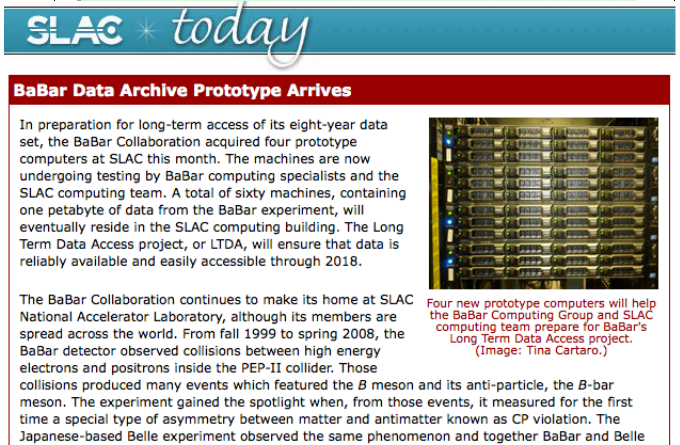




Virtualization

- The status at SLAC: 4 SL5.3 VMs installed on yakut13.
- VMs were added to a special batch queue.
- SL5 migration checks to be done on virtual machines.
- Simultaneously validates the SL5 build and the VM technology.

June 22, 2009 Long Term Data Access 6



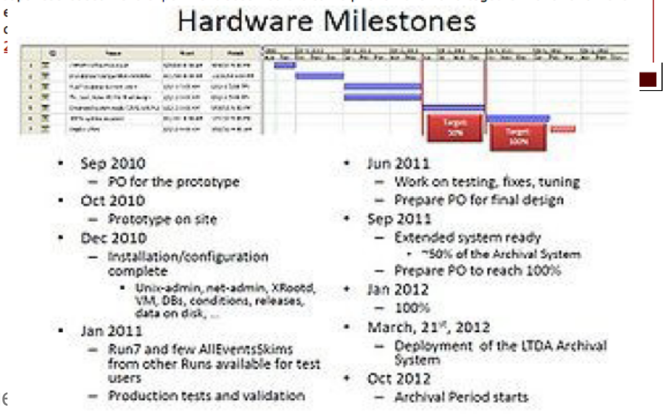
SLAC today

BaBar Data Archive Prototype Arrives

In preparation for long-term access of its eight-year data set, the BaBar Collaboration acquired four prototype computers at SLAC this month. The machines are now undergoing testing by BaBar computing specialists and the SLAC computing team. A total of sixty machines, containing one petabyte of data from the BaBar experiment, will eventually reside in the SLAC computing building. The Long Term Data Access project, or LTDA, will ensure that data is reliably available and easily accessible through 2018.

Four new prototype computers will help the BaBar Computing Group and SLAC computing team prepare for BaBar's Long Term Data Access project. (Image: Tina Cartaro.)

The BaBar Collaboration continues to make its home at SLAC National Accelerator Laboratory, although its members are spread across the world. From fall 1999 to spring 2008, the BaBar detector observed collisions between high energy electrons and positrons inside the PEP-II collider. Those collisions produced many events which featured the *B* meson and its anti-particle, the *B*-bar meson. The experiment gained the spotlight when, from those events, it measured for the first time a special type of asymmetry between matter and antimatter known as CP violation. The Japanese-based Belle experiment observed the same phenomenon and together BaBar and Belle



Hardware Milestones

- Sep 2010
  - PO for the prototype
- Oct 2010
  - Prototype on site
- Dec 2010
  - Installation/configuration complete
    - Unix-admin, net-admin, XRootd, VM, DBs, conditions, releases, data on disk, ...
- Jan 2011
  - Run7 and few AllEventsSkims from other Runs available for test users
  - Production tests and validation
- Jun 2011
  - Work on testing, fixes, tuning
  - Prepare PO for final design
- Sep 2011
  - Extended system ready
    - ~50% of the Archival System
  - Prepare PO to reach 100%
- Jan 2012
  - 100%
- March, 21st, 2012
  - Deployment of the LTDA Archival System
- Oct 2012
  - Archival Period starts

> BaBar moving to an "Archival Mode", preserving analysis ability beyond 2012

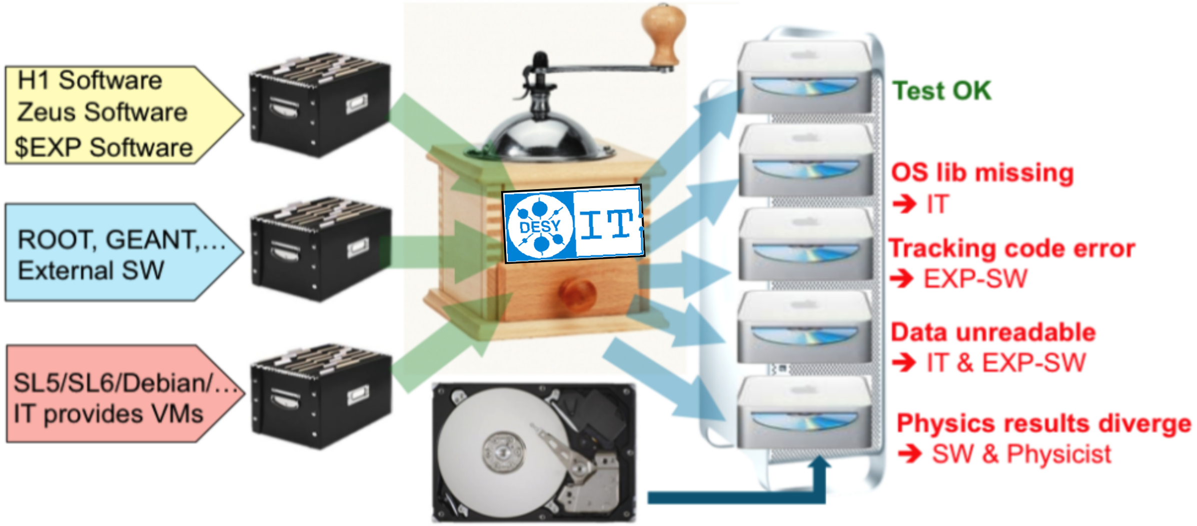
- In a very advanced state

> Use of virtualisation and cloud computing

Resources for projects at BaBar taken into account in funding model during analysis phase !

Towards a Generic Solution at DESY-IT

> Validation of experimental software using a virtual environment



H1 Software
Zeus Software
$EXP Software

ROOT, GEANT,…
External SW

SL5/SL6/Debian/…
IT provides VMs

DESY IT

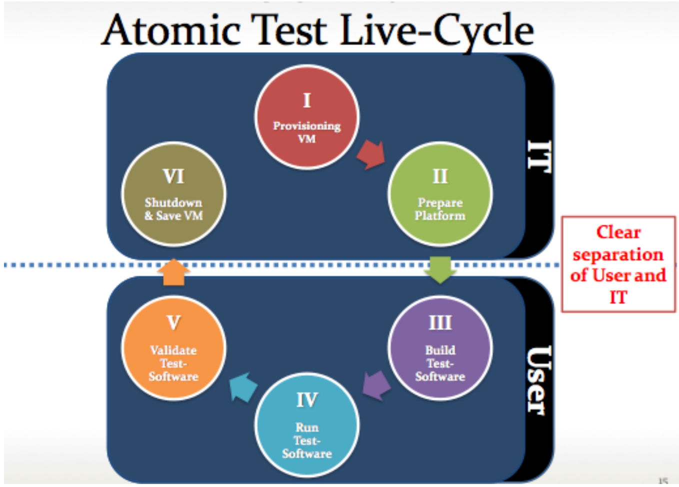
Test OK

OS lib missing
➔ IT

Tracking code error
➔ EXP-SW

Data unreadable
➔ IT & EXP-SW

Physics results diverge
➔ SW & Physicist



Atomic Test Live-Cycle

I Provisioning VM

II Prepare Platform

III Build Test-Software

IV Run Test-Software

V Validate Test-Software

VI Shutdown & Save VM

IT

User

Clear separation of User and IT

Clear separation between providers of input.

Automated VM image generator provided centrally.

Tests defined by $EXP.
Test data store provided by IT.

Different VMs run SW and tests. Depending on results, different action needed.

> Generic solution, for all HERA experiments: **validate the whole analysis chain**

- Useful collaboration for future OS, external software transitions
- Successful pilot project implemented, full project (people and $) now secured
- Should be useful for other experiments

Setting the Timescale

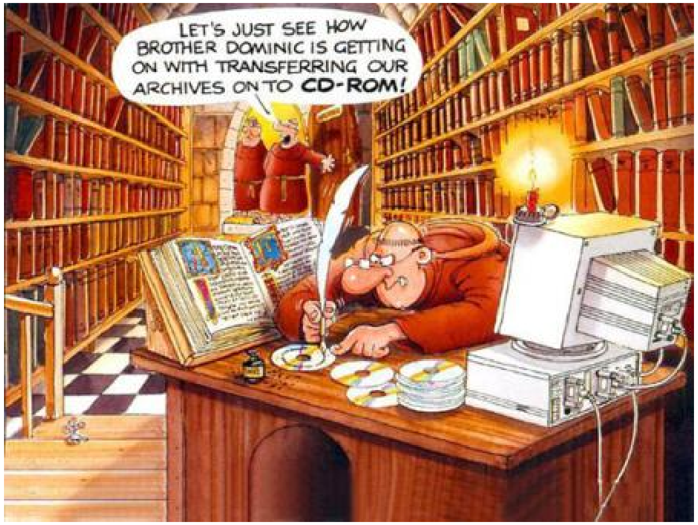
For how long HEP data should be saved?

- In case of the successor experiment (BELLE/BaBar) – 5-7 years
- HERA : > few decades
- LHC Run I (2008-2012) – minimum for 30 years
  (lifetime of LHC Collaborations)
  (different energy, less pile-up)

Archival System

Requirements for the archival system to cover (at least) few decades:

- Regular check for the files integrity
- Several consistent (automatic replication in case of failure) copies
- Ability to handle copies geographically separated (KEK, March-Apr. 2011)
- Cope with the copies on different storage types
- Fast and free-hands migration to new technology
- Fast access to the data when needed (active archive) via defined protocols


LET'S JUST SEE HOW BROTHER DOMINIC IS GETTING ON WITH TRANSFERRING OUR ARCHIVES ONTO CD-ROM!

http://treybig.org/Humor/CDmonk.jpeg

Conclusion and Outlook

- HEP data are mostly unique and have true scientific potential

- Data preservation in HEP is important because:
  - Relevant physics cases for future use can be made
  - It is timely, given the current experimental situation and plans
  - It may enhance the return on the initial investment in the experimental facilities
  - It provides additional research at particularly low cost

- It requires a strategy and well-identified resources

- International cooperation is the best way to proceed
  - Unique opportunity to build a coherent structure for the future: DPHEP

- Blueprint for Data Preservation is on the way
  - Skeleton for local, national and international proposals, for past and future experiments