La préservation des données scientifiques dans la physique des hautes énergies

C. Diaconu
CPPM

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

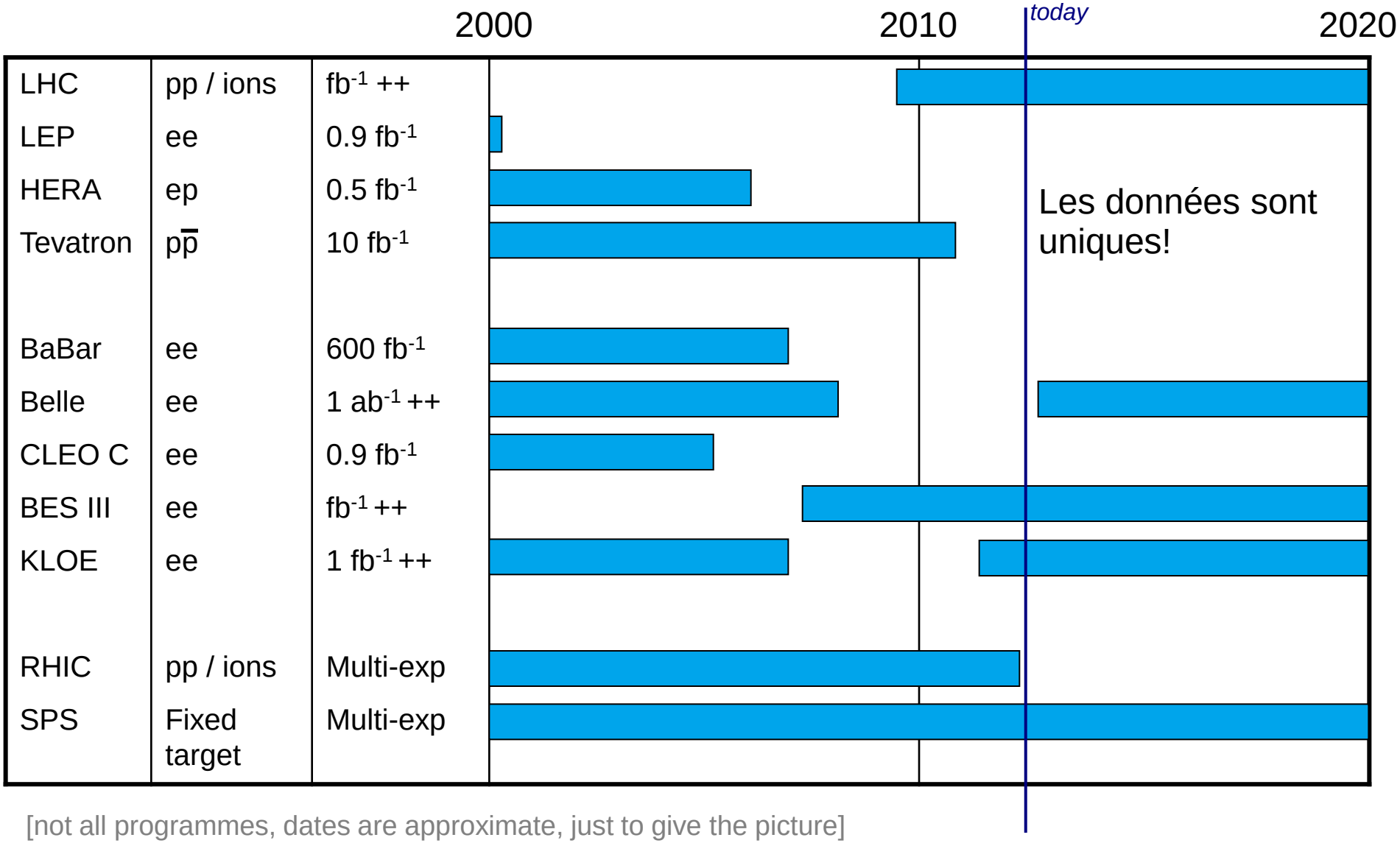
PREDON

A project for scientific data preservation in France

High Energy Physics

- Exemple: Physique des Hautes énergies
- Large Hadron Collider (27 Km, 13 TeV, 40MHz)

Exemple: Programme experimental de la physique des hautes énergies HEP ± 10 ans

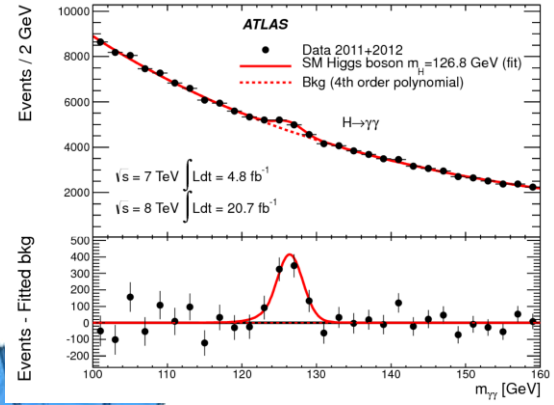


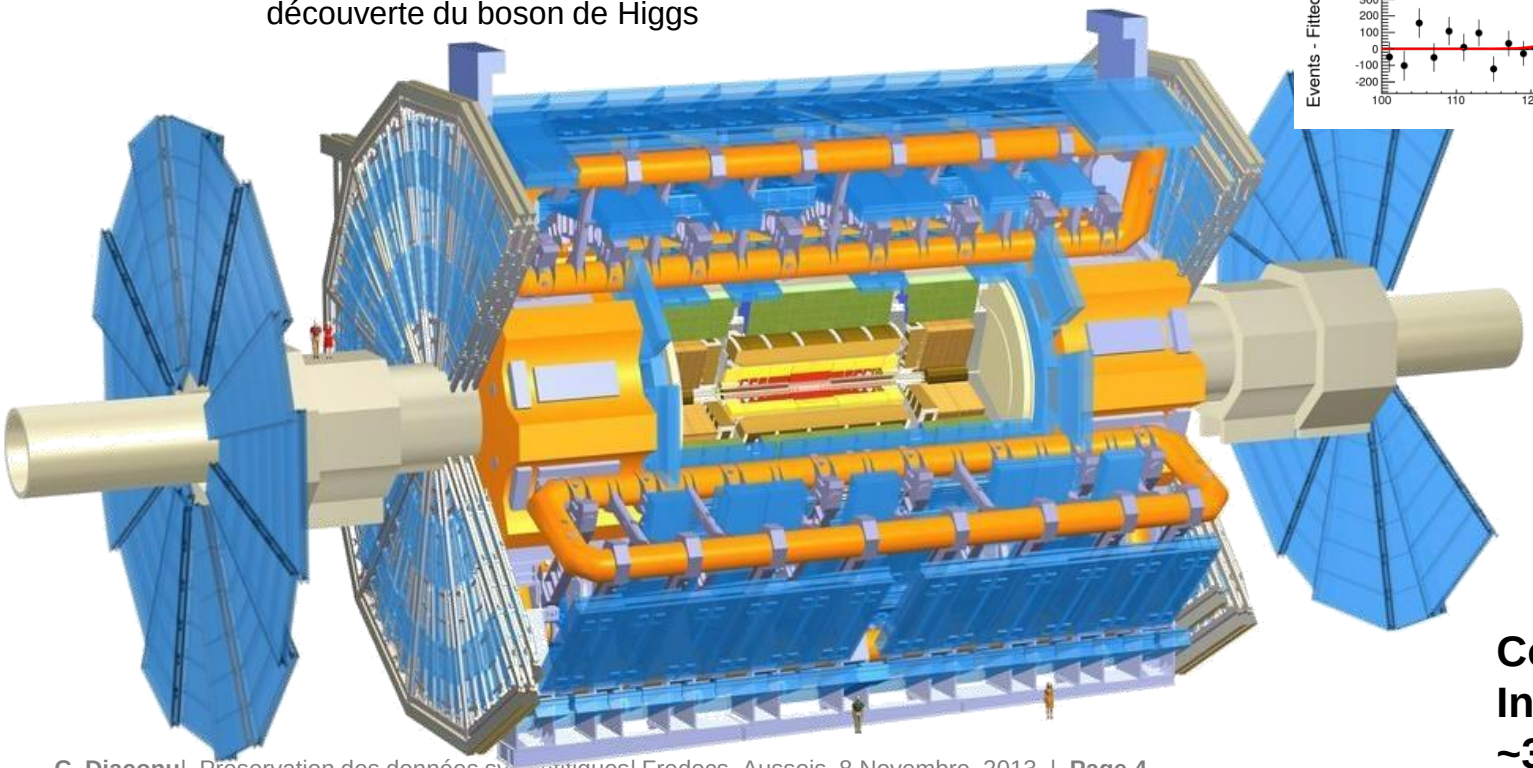
[not all programmes, dates are approximate, just to give the picture]

Des instruments gigantesques

- ATLAS: L'équivalent d'une camera avec 25Gpixels (avec une cinquantaine de technologies différentes) et 40 000 000 000 « photos » par seconde (100Pb)
- manips LHC:
  - 1000 articles scientifiques en 2 ans

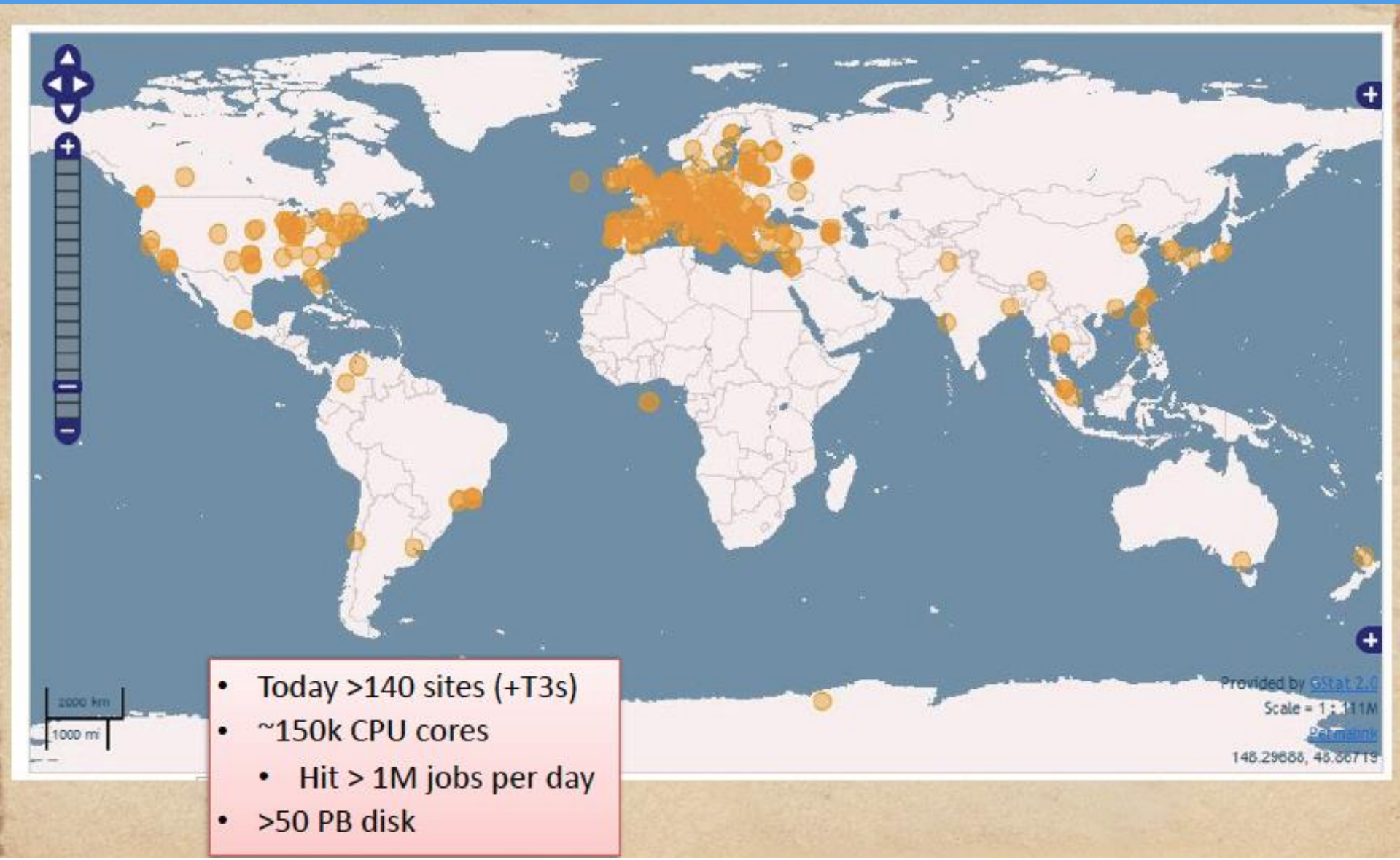
découverte du boson de Higgs





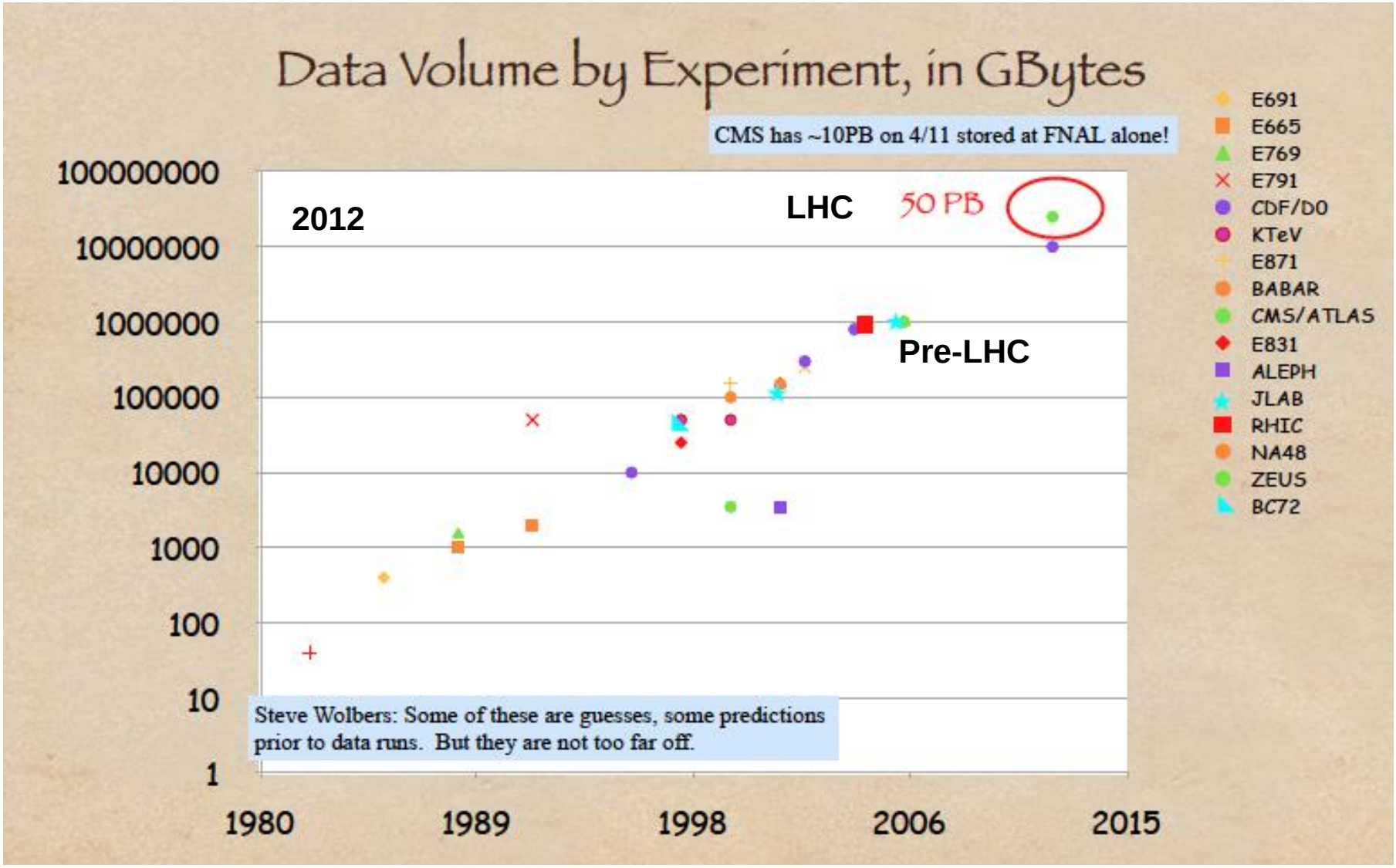
Collaboration Internationale ~3000 chercheurs

LHC computing

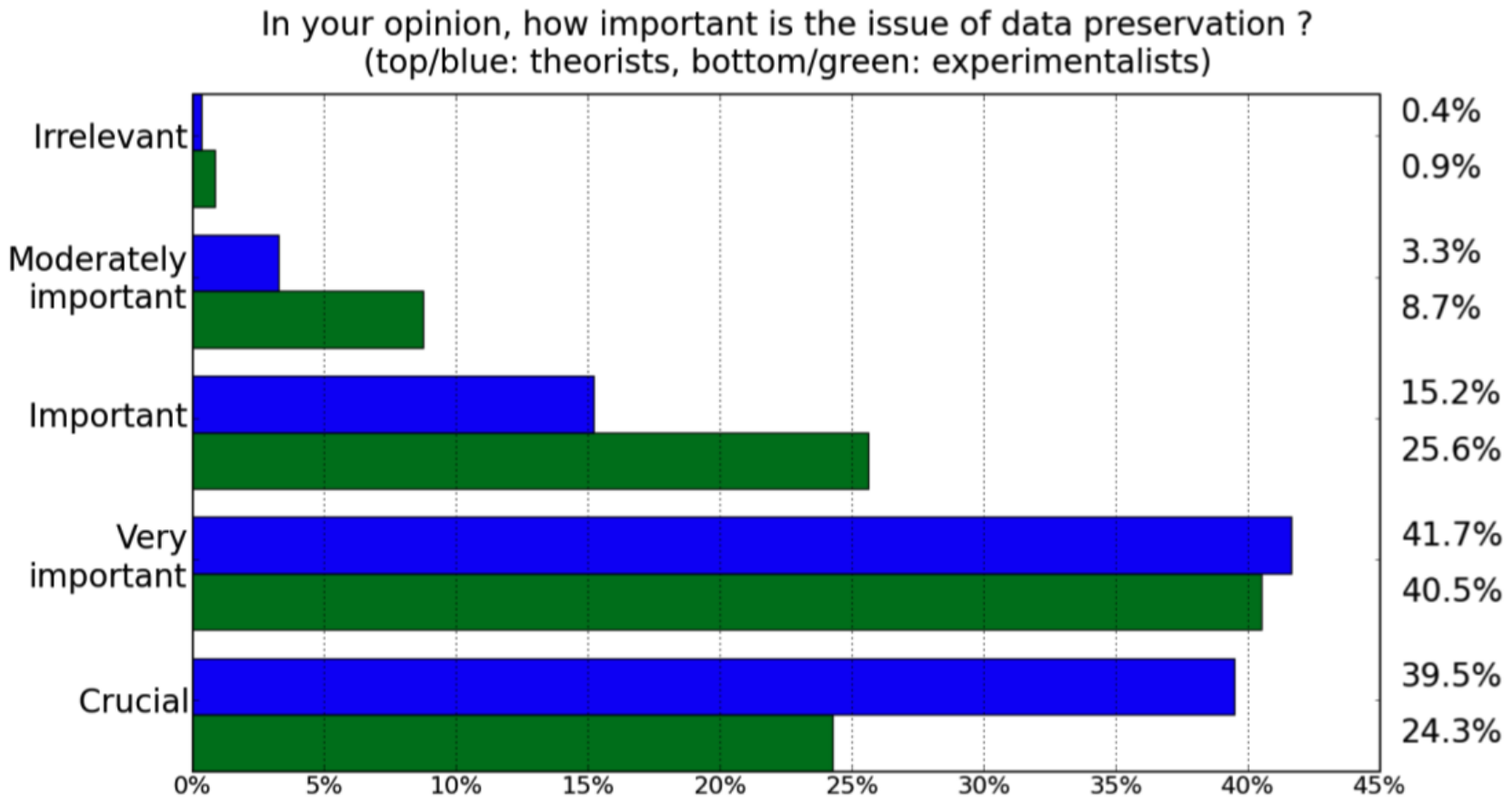


- Today >140 sites (+T3s)
- ~150k CPU cores
  - Hit > 1M jobs per day
- >50 PB disk

Quantité de données: 1PB -> 100PB->1EB



L’opinion de la communauté scientifique



PARSE insight
Permanent Access to the Records of Science in Europe

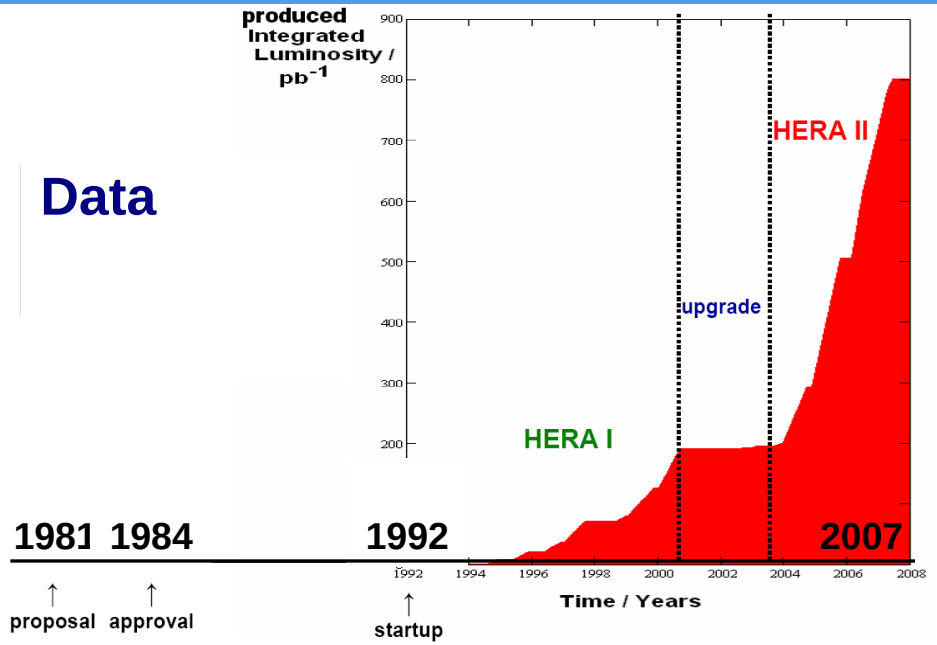
e-infrastructure

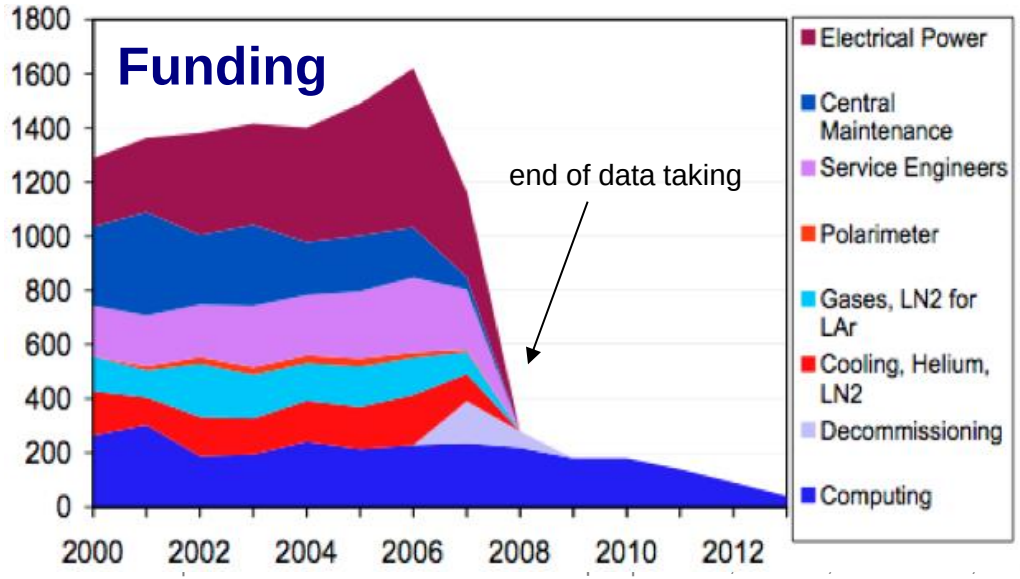
SEVENTH FRAMEWORK PROGRAMME

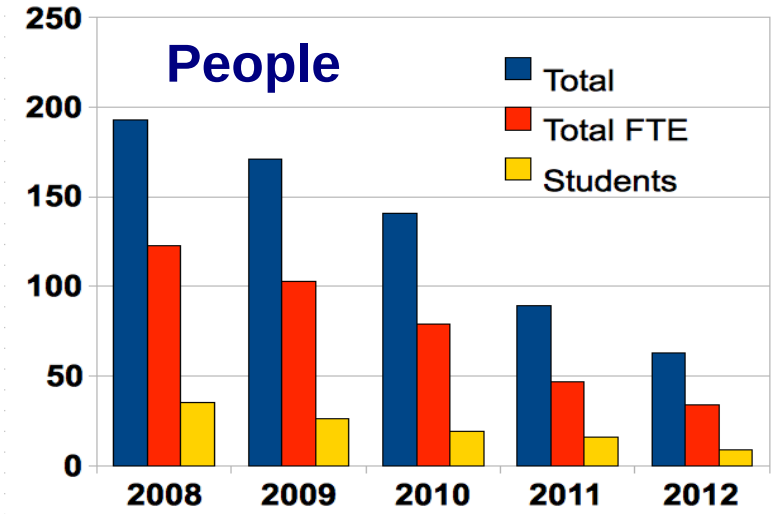
arXiv:0906.0485

Est-il difficile de préserver les données?

- Les programmes accumulent la plupart des données vers la fin du programme
- Les ressources (financières et humaines) décroissent rapidement après la fin des manips
- En absence d'une planification rigoureuse les conditions pour des nouveau projets ne sont pas idéales







What is HEP “data”?

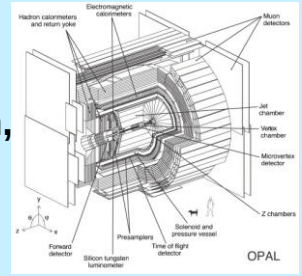
Digital information

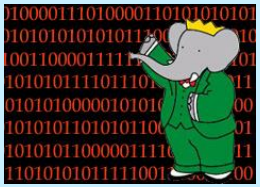
The data themselves, volume estimates for preservation data of the order of a few to 10 PB

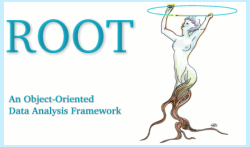
Other digital sources such as databases to also be considered

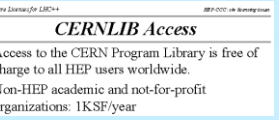
Software

Simulation, reconstruction, analysis, user, in addition to any external dependencies









Meta information

Hyper-news, messages, wikis, user forums..

Publications

arXiv.org

INSPIRE

Durham University — Science & Technology Facilities Council

HEPDATA: REACTION DATA Database

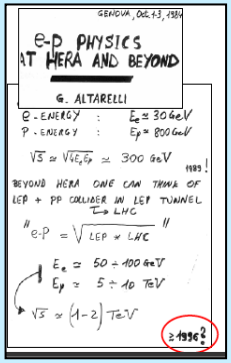
...containing numerical values of HEP scattering data such as total and differential cross sections, fragmentation functions, structure functions, and polarisation measurements, from a wide range of experiments. It is compiled by the **Durham Database Group (UK)** with help from the COMPAS group (Russia,)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

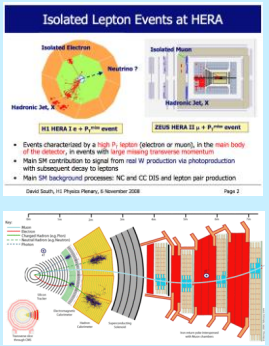
Journal of High Energy Physics

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

Documentation

Internal publications, notes, manuals, slides



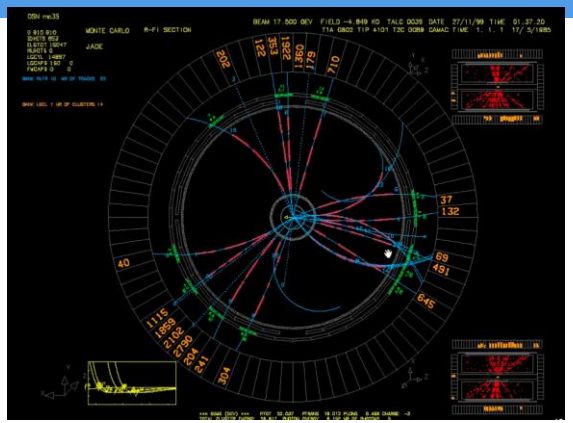


Expertise and people

Un exemple typique

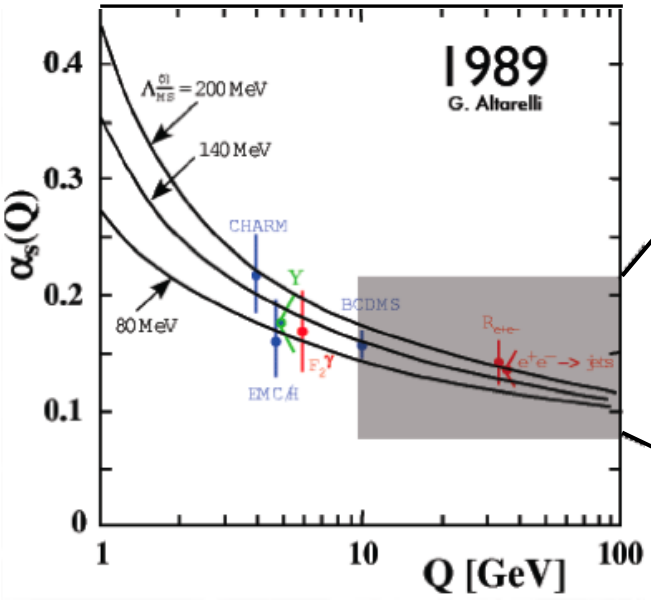
- Experience JADE
  - Données sauvées par hasard

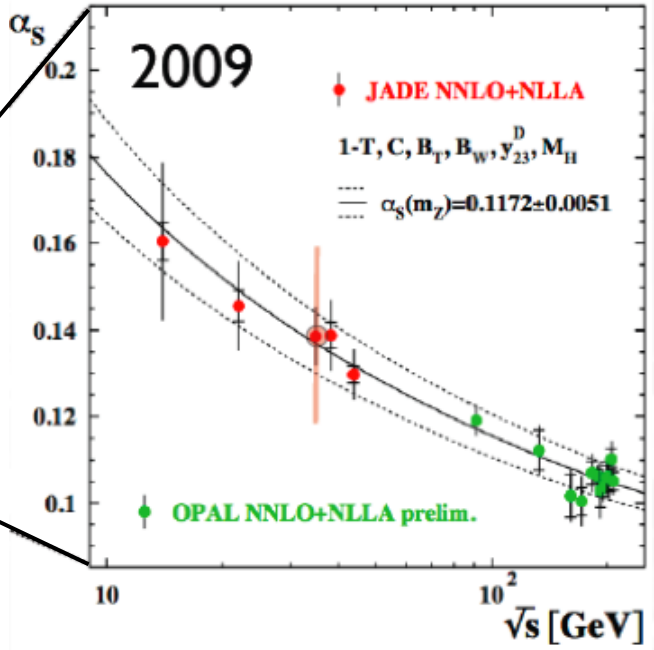
    Nom de code: “la valise”
  - Ré-anlayse après 20 ans



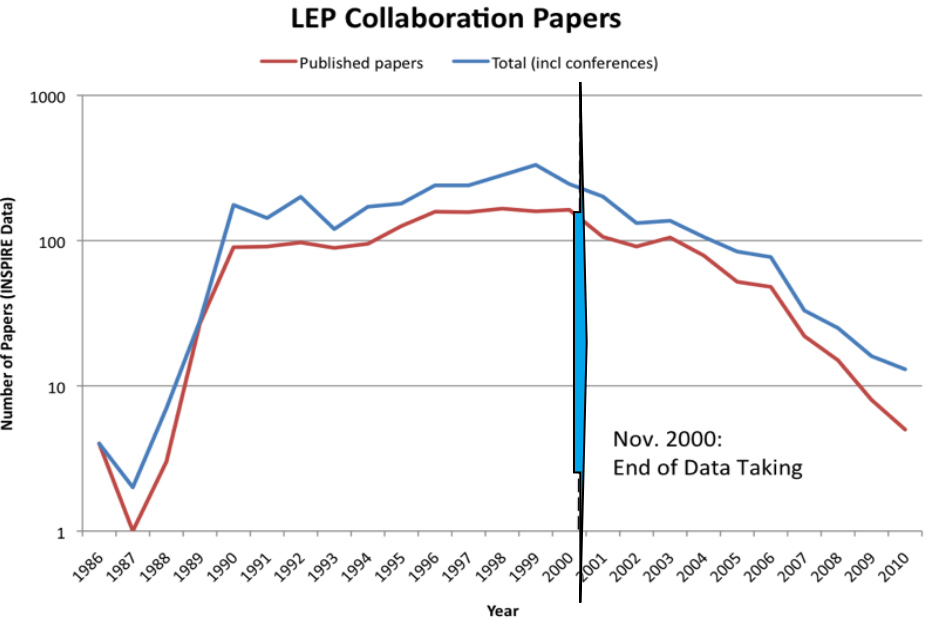
10 publications

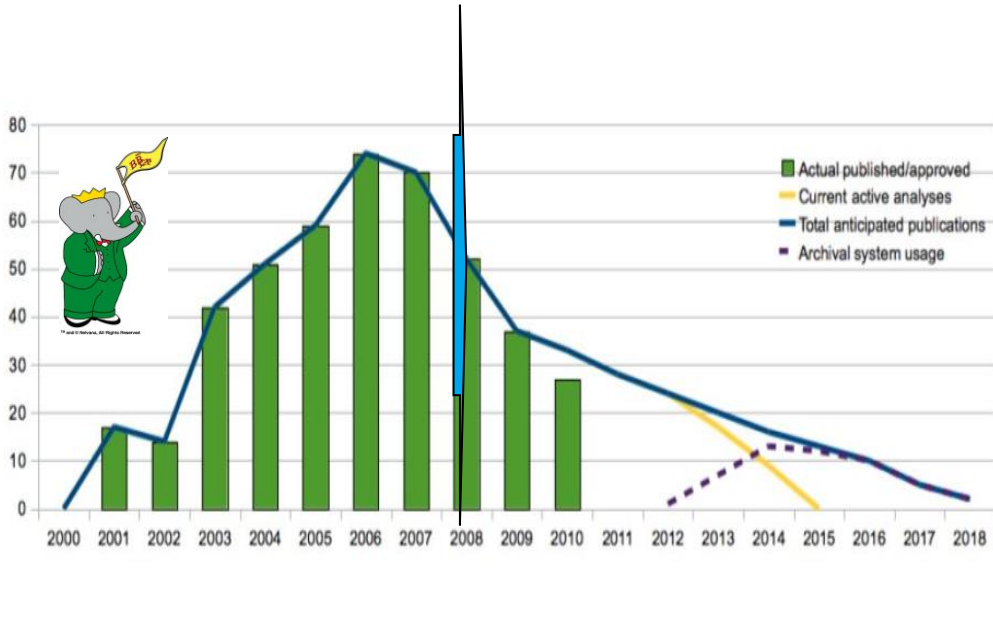
2011





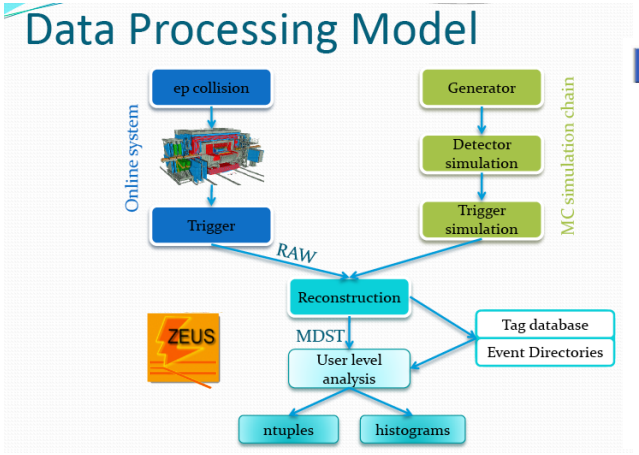
Publications à long terme

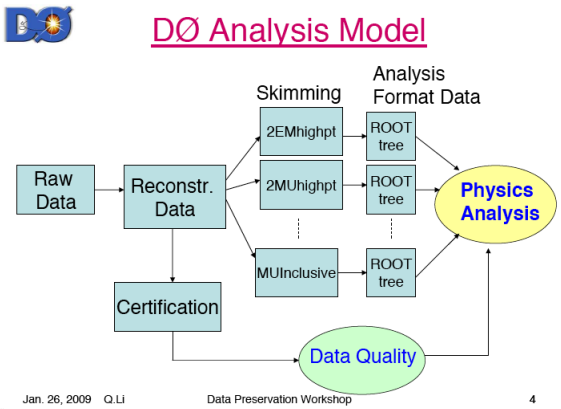


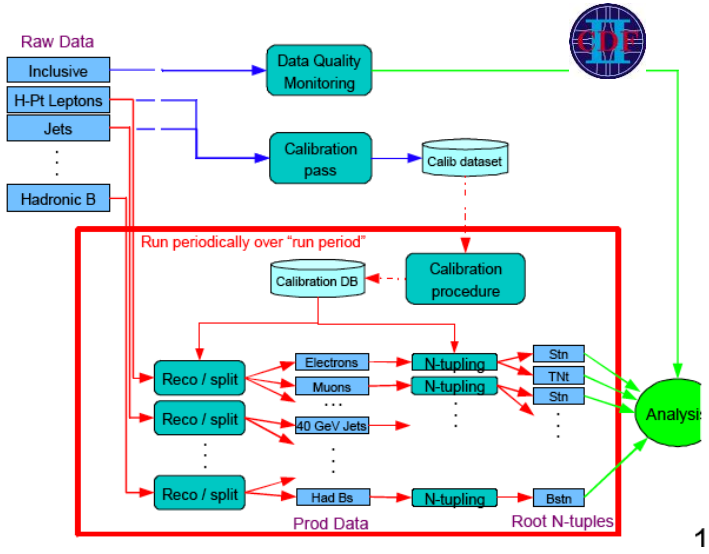


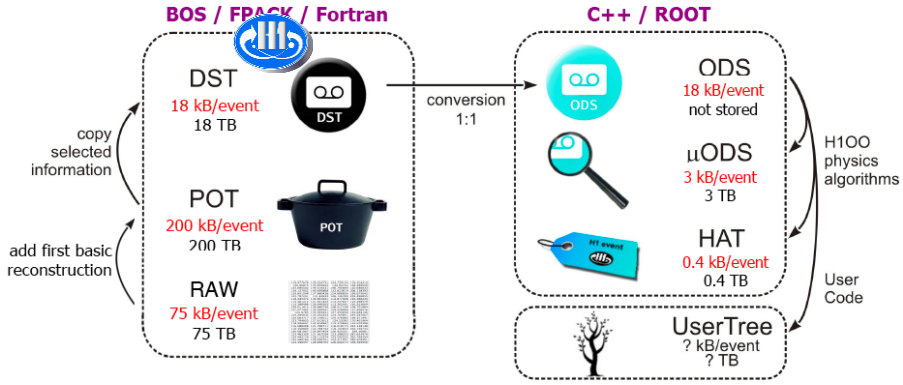
- LEP: 1989-2000
  - mais 300 publications produites apres 2000 et une centaine après 2005
- C'est systématique: les publications continuent long temps après la fin de l'experience
  - Nouvelles idées, théories etc.

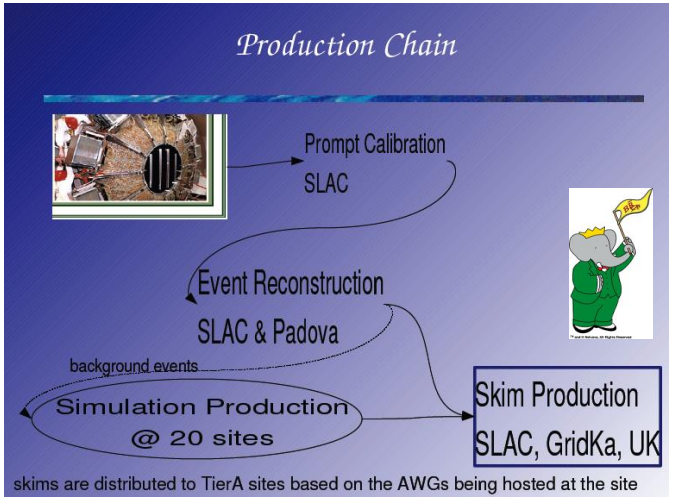
Modèles de données











Réduction et abstraction

- RAW (→ POT)→ DST → ntuple → article
- 1.5Mb->15Mb ----------→ 30 Kb (par "événement")
- Pas de format standardisé

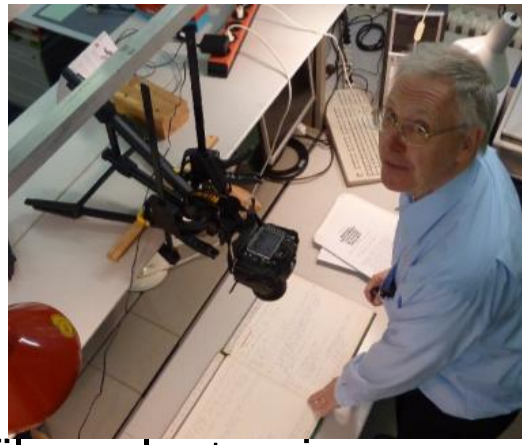
DPHEP : définition des niveaux de préservation

> En progression de la complexité et les couts

Preservation Model		Use Case	
1	Provide additional documentation	Publication related info search	**Documentation**
2	Preserve the data in a simplified format	Outreach, simple training analyses	**Outreach**
3	Preserve the analysis level software and data format	Full scientific analysis, based on the existing reconstruction	**Technical Preservation Projects**
4	Preserve the reconstruction and simulation software as well as the basic level data	Retain the full potential of the experimental data	

Level 1: Documentation

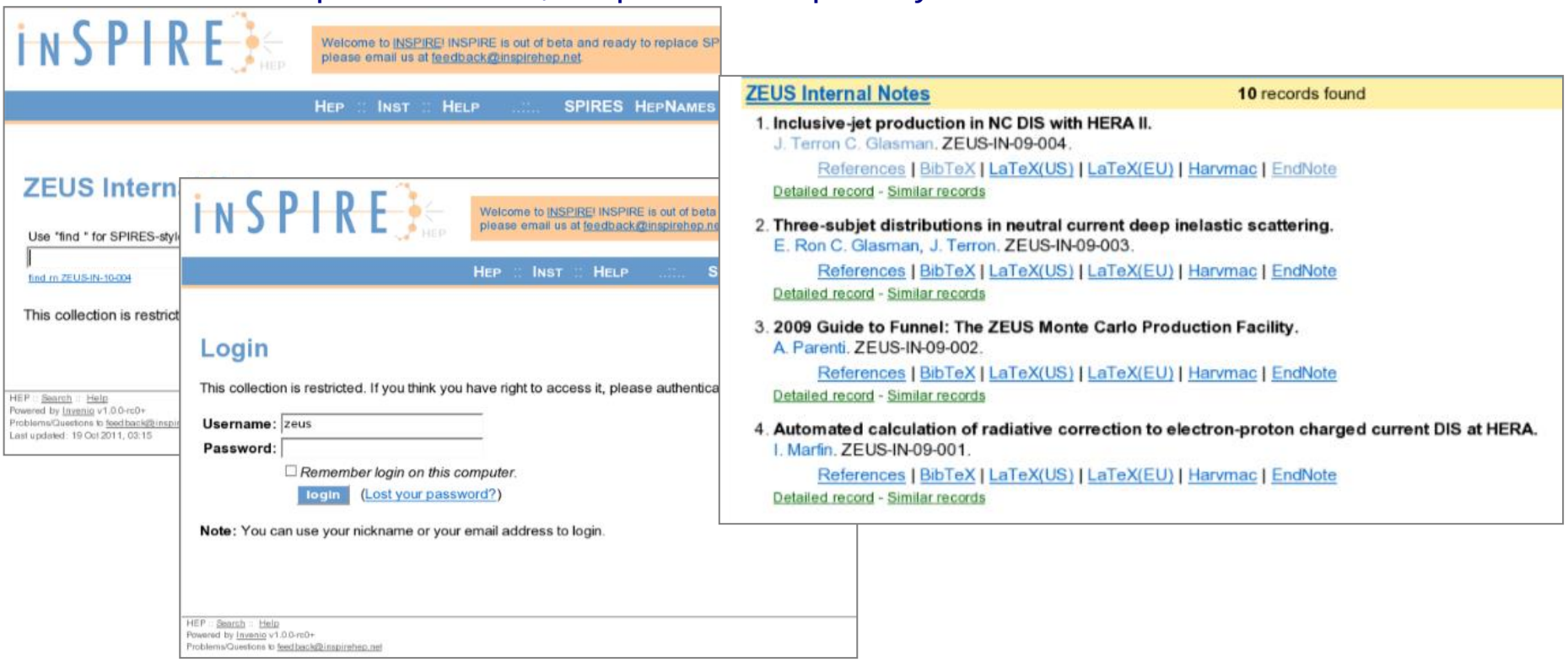
- Une tache considérable: des groupes de travail dédiées



- Non-digital: Cataloguing, organisation, scanning or photographing of appropriate of papers, notes, drawings, talks from pre-web days, detector schematics, blueprints, logbooks, ...
  - *Virtual Archives* established by the experiments
- Digital: Old online shift tools, detector configuration files, electronic logbooks, detailed run information, web content from out-dated servers with dead links, various wikis, meetings, talks, ...
  - Replacement of old web servers by VMs, hosted by the computer centres
  - Replacement of old pages to newer technologies such as wikis (use of (T)wikis much more prevalent in the LHC era)
  - Use of external services for hosting collaboration material

Documentation projects with INSPIRE

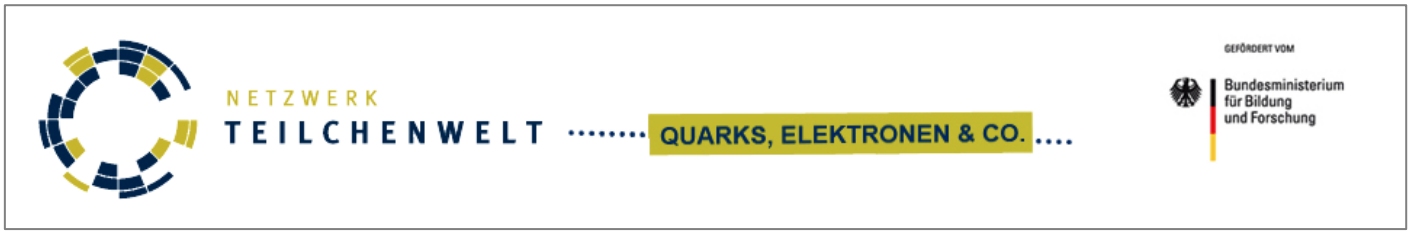
- Internal notes now available on INSPIRE
  - Password protected now, simple to make publicly available in the future

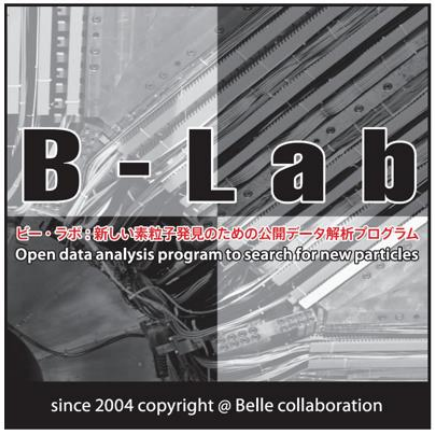


- The ingestion of other documents is under discussion, including theses, preliminary results, conference talks and proceedings, paper drafts, ...
- More on InSpire: reduced data?

HEP outreach initiatives

- Many initiatives promoting outreach efforts and to improve the public understanding of science in general





Outreach

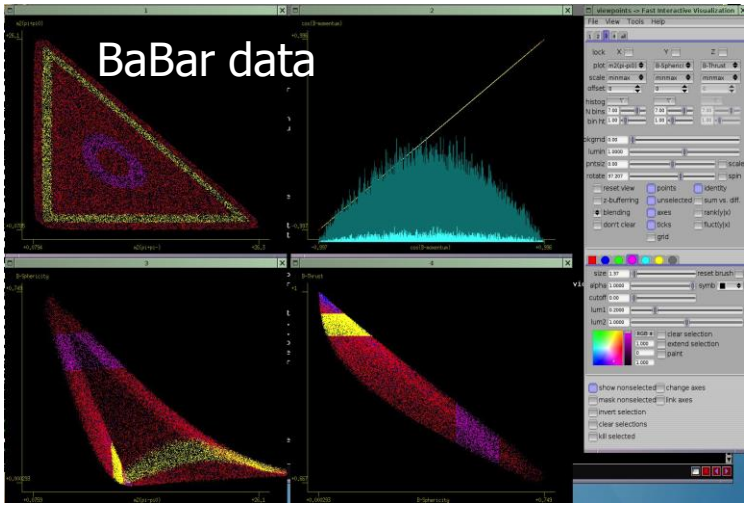
- Use **real and preserved** data to enhance scientific education worldwide
- Simple data format: input using text file of kinematics of HEP events

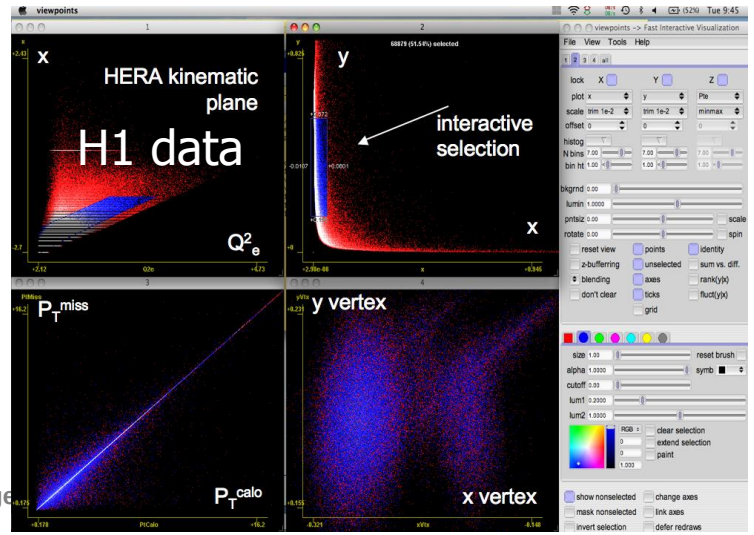
HepEdu

- Discussions within DPHEP: format for outreach.
- BaBar, Belle, H1.
- First order attempt: text. Number of particles/tracks in the event
  index PID E px py pz for particle 0
  index PID E px py pz for particle 1
  index PID E px py pz for particle 2
  index PID E px py pz for particle 3

```
6
0    -13  1.313  0.407  1.241  0.075
1    -11  2.010 -1.813  0.039 -0.865
2   -211  0.474 -0.134  0.304 -0.308
3    211  0.480 -0.353 -0.112 -0.273
4   -211  1.003 -0.905 -0.369  0.176
5     22  0.212 -0.108  0.147  0.108
3
0    211  1.316 -0.414 -1.239  0.075
1   -211  2.014 -1.802  0.204 -0.865
2    211  0.474 -0.307  0.127 -0.308
```

Viewpoints (NASA)





- Discussions about common formats ongoing
  - B-lab (KEK) example considered
  - Experience at LHC
  - Connect to existing projects (master classes etc.)

Summary of information from the (pre-LHC) experiments

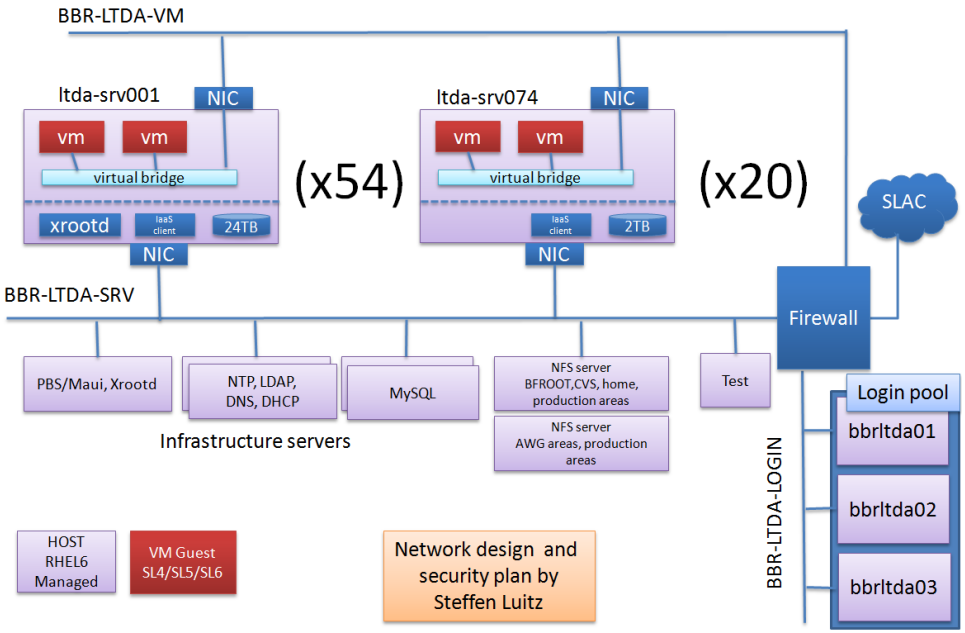
	BaBar	H1	ZEUS	HERMES	Belle	BESIII	CDF	DØ
End of data taking	07.04.08	30.06.07	30.06.07	30.06.07	30.06.10	2017	30.09.11	30.09.11
Type of data to be preserved	RAW data Sim/rec level Data skims in ROOT	RAW data Sim/rec level Analysis level ROOT data	Flat ROOT based ntuples	RAW data Sim/rec level Analysis level ROOT data	RAW data Sim/rec level	RAW data Sim/rec level ROOT data	RAW data Rec. level ROOT files (data+MC)	Raw data Rec. level ROOT files (data+MC)
Data Volume	2 PB	0.5 PB	0.2 PB	0.5 PB	4 PB	6 PB	9 PB	8.5 PB
Desired longevity of long term analysis	Unlimited	At least 10 years	At least 20 years	5-10 years	5 years	15 years	Unlimited	10 years
Current operating system	SL/RHEL3 SL/RHEL 5	SL5	SL5	SL3 SL5	SL5/RHEL5	SL5	SL5 SL6	SL5
Languages	C++ Java Python	C C++ Fortran Python	C++	C C++ Fortran Python	C C++ Fortran	C++	C C++ Python	C++
Simulation	GEANT 4	GEANT 3	GEANT 3	GEANT 3	GEANT 3	GEANT 4	GEANT 3	GEANT 3
External dependencies	ACE CERNLIB CLHEP CMLOG Flex GNU Bison MySQL Oracle ROOT TCL XRootD	CERNLIB FastJet NeuroBayes Oracle ROOT	ROOT	ADAMO CERNLIB ROOT	Boost CERNLIB NeuroBayes PostgresQL ROOT	CASTPR CERNLIB CLHEP HepMC ROOT	CERNLIB NeuroBayes Oracle ROOT	Oracle ROOT

Longévité recherchée: > 10 ans

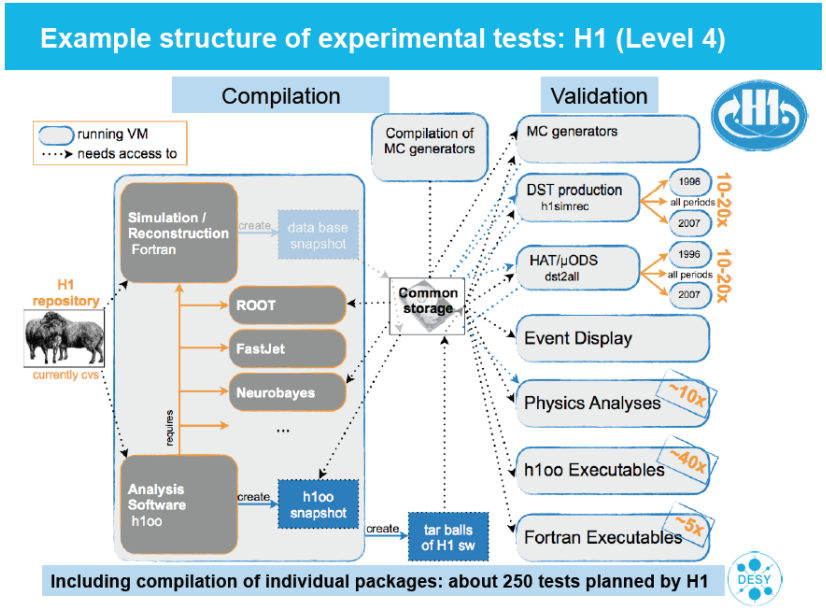
Exemples projets

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

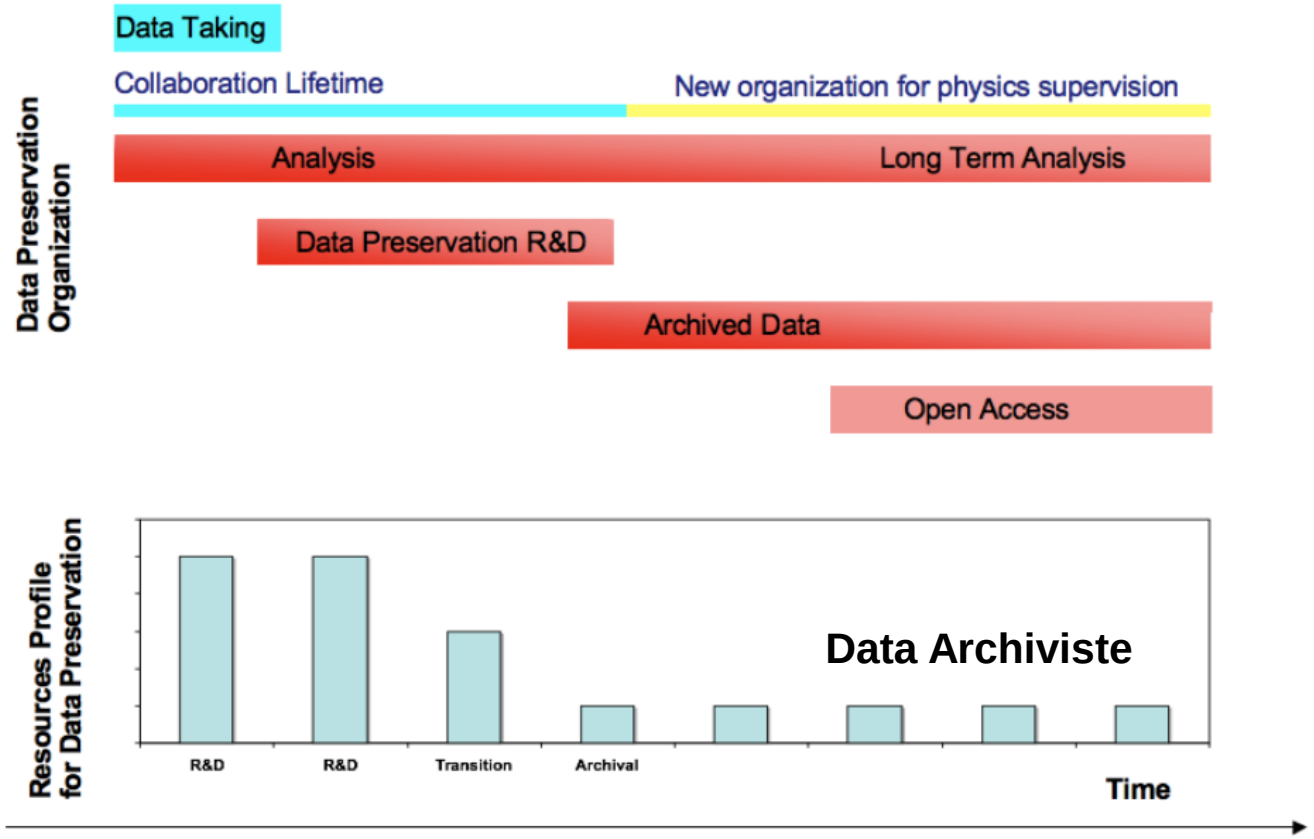
Préservation d'un système d'accès et calcul à des données complexes (SLAC/Stanford USA)



Système de préservation et migration Virtualisation, validation intensive (DESY, Hambourg, Allemagne)



Plan de sauvetage, vers un modèle économique



Estimation du cout du projet:

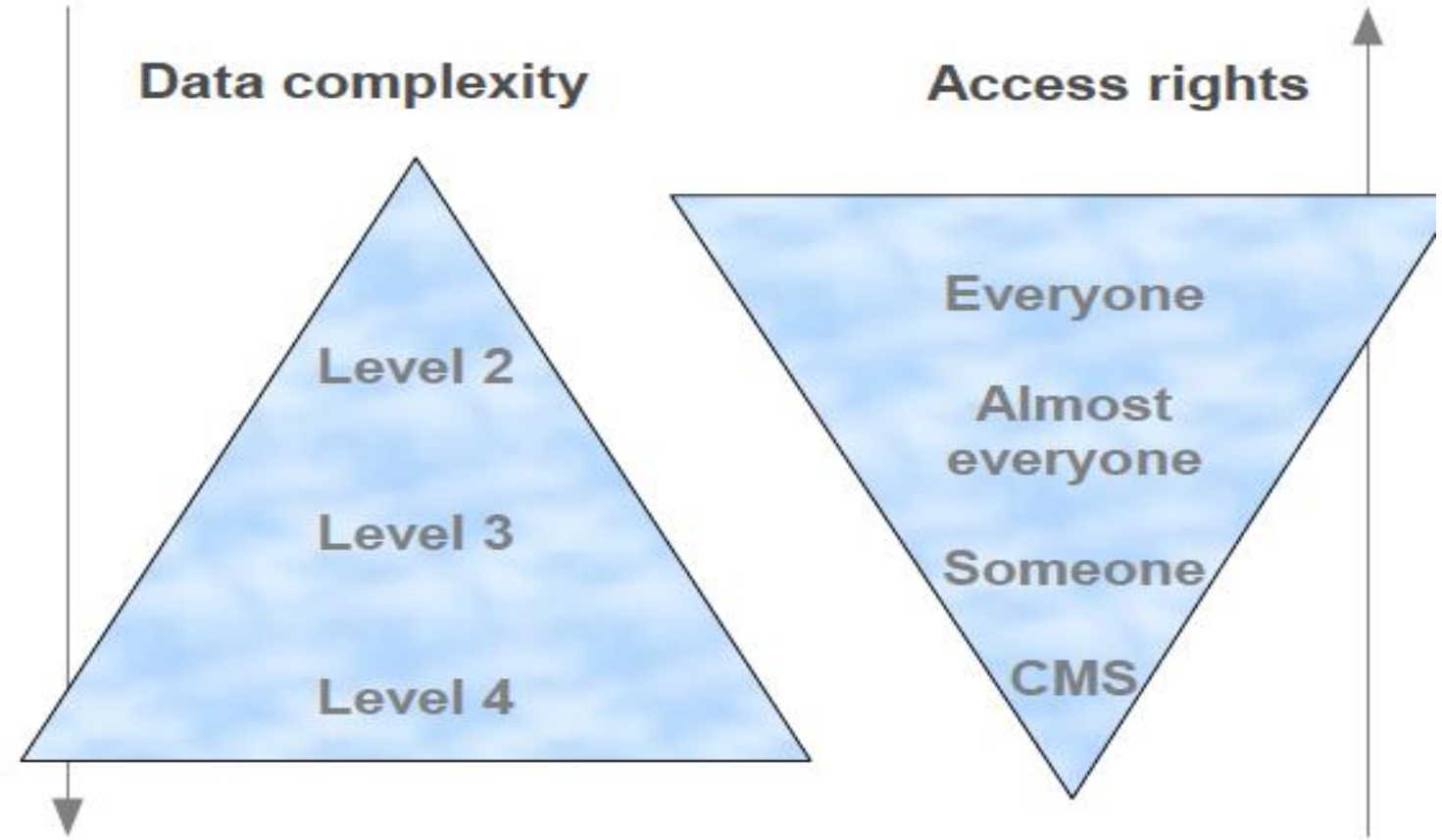
- **>1% du cout total pour une production scientifique augmentée de 10%**

Des réflexions en cours sur:

- **Systèmes de stockage robustes, basse consomation etc.**
- **Couts de la préservation de données à long terme**

Data Preservation and Open Access at the LHC

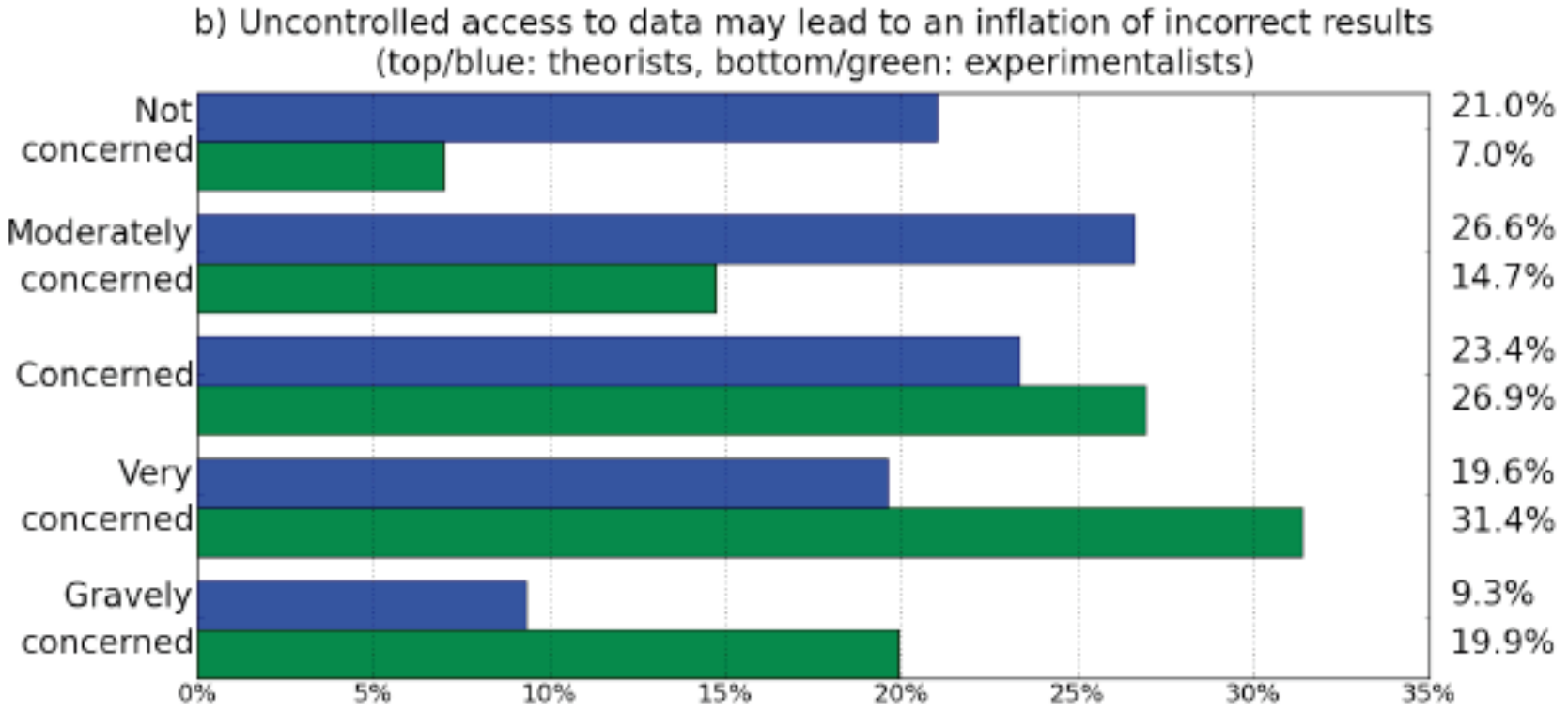
- Reflection just started in ATLAS, ALICE, CMS, LHCb
  - Common understanding that starting earlier will consolidate the long term future
  - Strong wish to develop a common policy at CERN and within DPHEP



Est-ce que la réutilisation des données est risquée?

"Errors using inadequate data are much less than those using no data at all."
Charles Babbage

Parse.insight

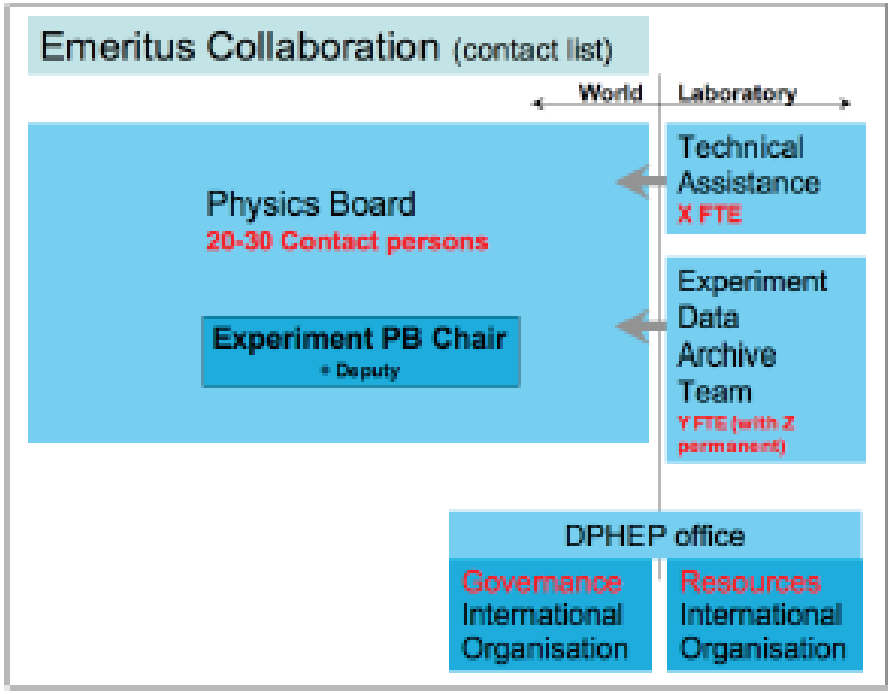


Governance issues are very important to support data usage

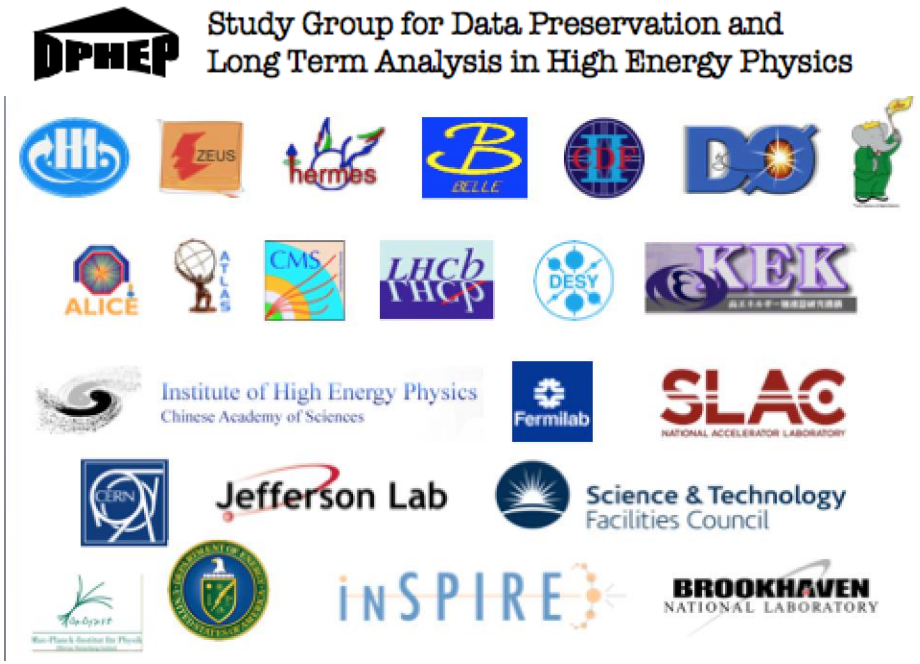
Long term organisation

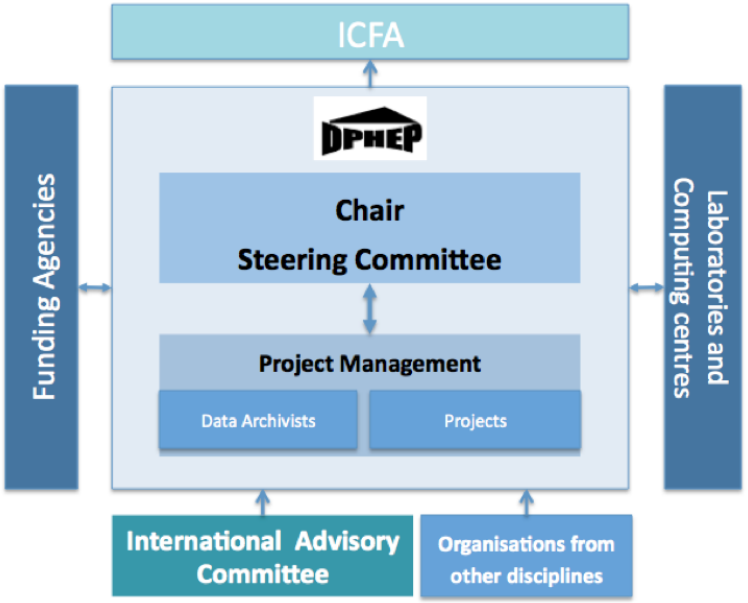
Preservation project make sense if the scientific supervision is ensured

- Future structure of the collaboration should also be considered by HEP experiments
  - Experimental organisation risks being left in an undefined state
  - Transition should also be planned in advance of the projected end date



DPHEP: Groupe d'études et organisation internationale





- Study Group DPHEP:
  - Participation des grands laboratoires (CERN, DESY, FERMILAB, SLAC, KEK,IHEP et experiences
- Organisation internationale en cours de mise en place
  - 100 contact personnes de contact
  - Chair: D. Diaconu Project Manager: Jamie Shiers (CERN)

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.

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.)

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 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- ed events. The amount of data that should be is varies between 0.5 PB and 10 PB for each ot huge by today's standards but nonetheless legree of preparation for long-term data varies but it is obvious that no preparation was fore- of the programs; any conservation initiatives llel with the end of the data analysis.

Science

February 2011

Rescue of Old Data Offers Lesson for Particle Physicists

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

nature news

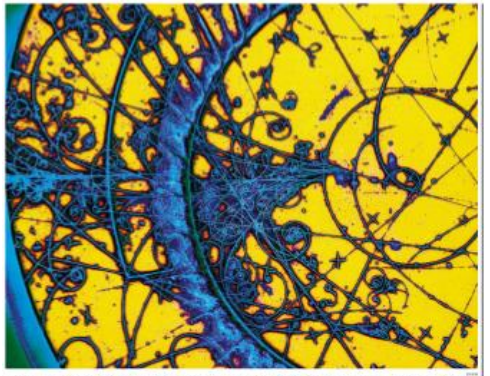
May 2011

symmetry dimensions of particle physics
A joint Fermilab/SLAC publication
VOLUME 06 ISSUE 06 DECEMBER 09

Symmetry, December 2009

Berliner Zeitung · Nummer 50 · Dienstag, 16. Februar 2010

Wissenschaft



Die Hieroglyphen von morgen

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

Der Teilchenzoo

Berliner Zeitung and Frankfurter Rundschau, February 2010

PDG particle data group

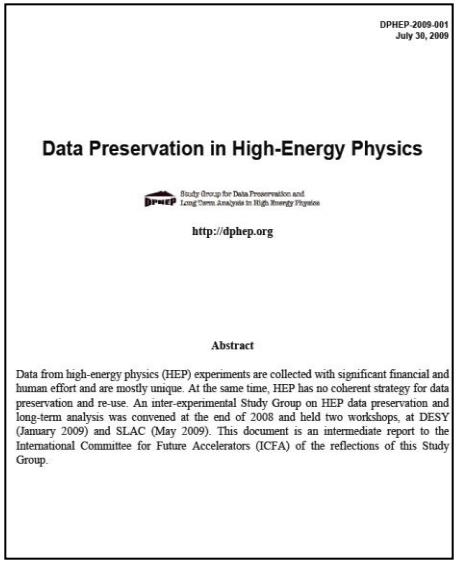
Data Preservation

- ICFA Study Group on Data Preservation and Long Term Analysis in High Energy Physics. High Energy Physics experiments initiated with this Study Group a common reflection on data persistency and long term analysis in order to get a common vision on these issues and create a multi-experiment dynamics for further reference:

https://www.dphep.org/

DPHEP Intermediate Recommendations (end 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

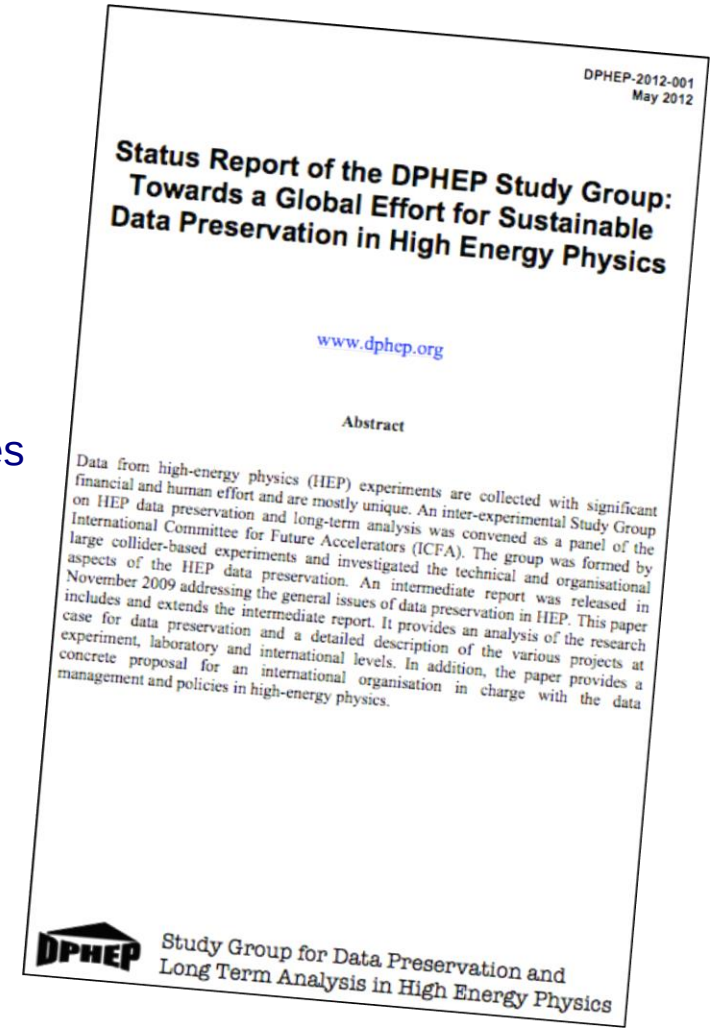
Abstract

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
  - Many 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 publication 2012

- Full status report of the activities of the DPHEP study group, including:
  - Tour of data preservation activities in other fields
  - 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



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

ICFA Statement on LTDP

ICFA: International Committee for Future Accelerators

- *The International Committee for Future Accelerators (ICFA) supports the efforts of the Data Preservation in High Energy Physics (DPHEP) study group on long-term data preservation and welcomes its transition to an active international collaboration with a full-time project manager.* ***It encourages laboratories, institutes and experiments to review the draft DPHEP Collaboration Agreement with a view to joining by mid- to late-2013.***

- *ICFA notes the lack of effort available to pursue these activities in the short-term and the possible consequences on data preservation in the medium to long-term.* ***We further note the opportunities in this area for international collaboration with other disciplines and encourage the DPHEP Collaboration to vigorously pursue its activities****. In particular, the effort required to prepare project proposals must be prioritized, in addition to supporting on-going data preservation activities.*

- ***ICFA notes the important benefits of long-term data preservation to exploit the full scientific potential of the, often unique, datasets****. This potential includes not only future scientific publications but also educational outreach purposes, and the Open Access policies emerging from the funding agencies.*

- 15 March 2013

Data Preservation in a multidisciplinary context

- **More Coordination**: The organisation should be brought to a long-term perspective by solid, commensurate and courageous decisions of the funding and coordination bodies responsible for the wealth of HEP experimental data produced so far.

- **More Standards** An increased standardisation will increase the overall efficiency of HEP computing systems and it will also be beneficial in securing long-term data preservation.

- **More Technology:** These new techniques (virtualisation etc.) seem to fit well within the context of large scale and long-term data preservation and access.

- **More Experiments**: The expansion of the DPHEP organisation to include more experiments is one of the goals of the next period.

- **More Cooperation: Cooperation with other fields in data management: access, mining, analysis and preservation; appears to be unavoidable and will also dramatically change the management of HEP data in the future.**

Conclusions HEP

- Les données scientifiques ont un potentiel qui dépasse le cadre de recherche initial et qui doit être exploité à long terme
- La préservation de données scientifique est économiquement avantageuse:
  - Recherche à bas cout
- Une technologies de frontière est nécessaire
  - Virtualisation, cloud computing, workflows….
  - Expertise IST essentielle
- La collaboration internationale est essentielle
- La collaboration multi-disciplinaire est nécessaire
  - Méthodes, technologies, approches (PREDON)