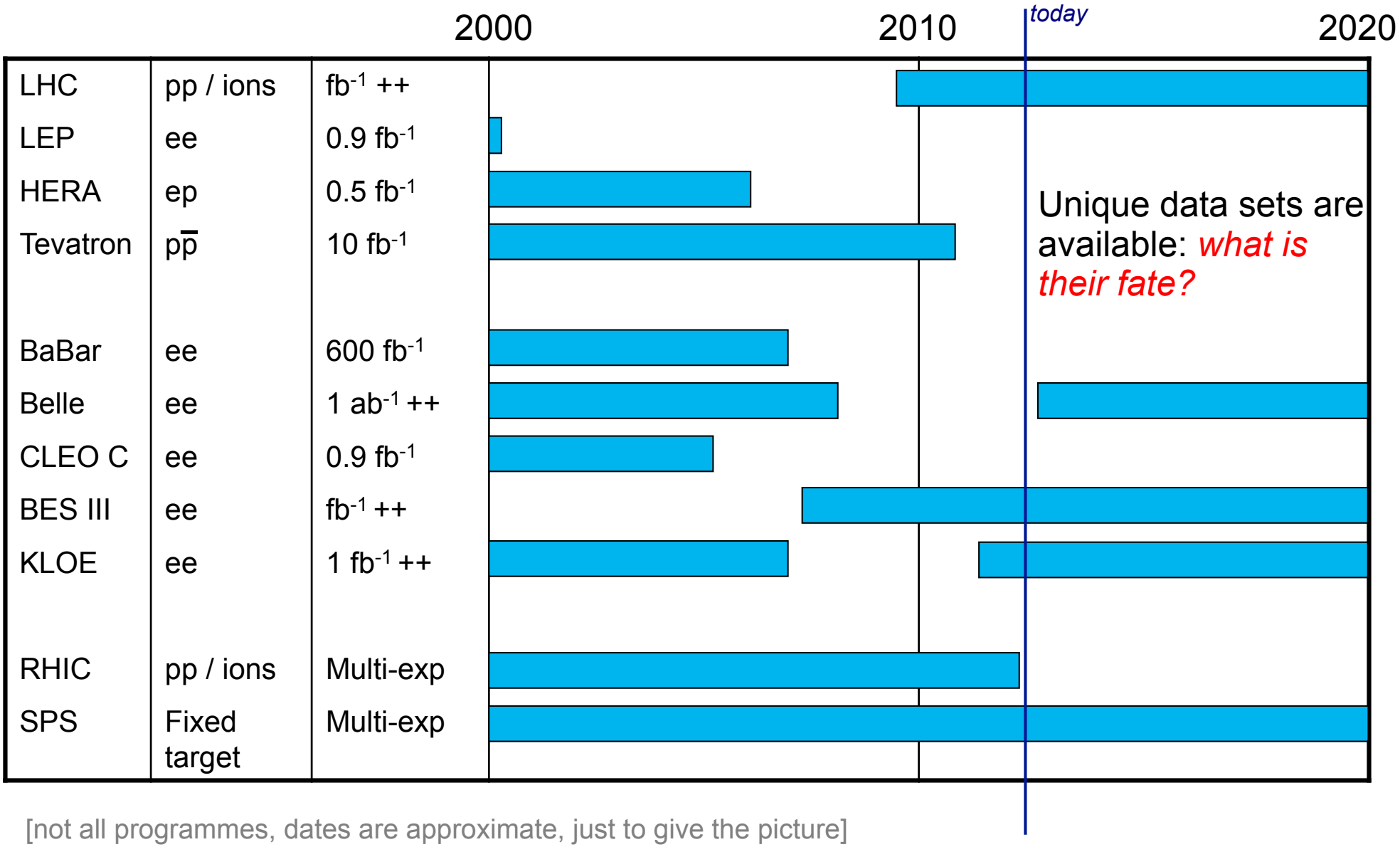
Scientific Data Preservation: from HEP to multi-disciplinary

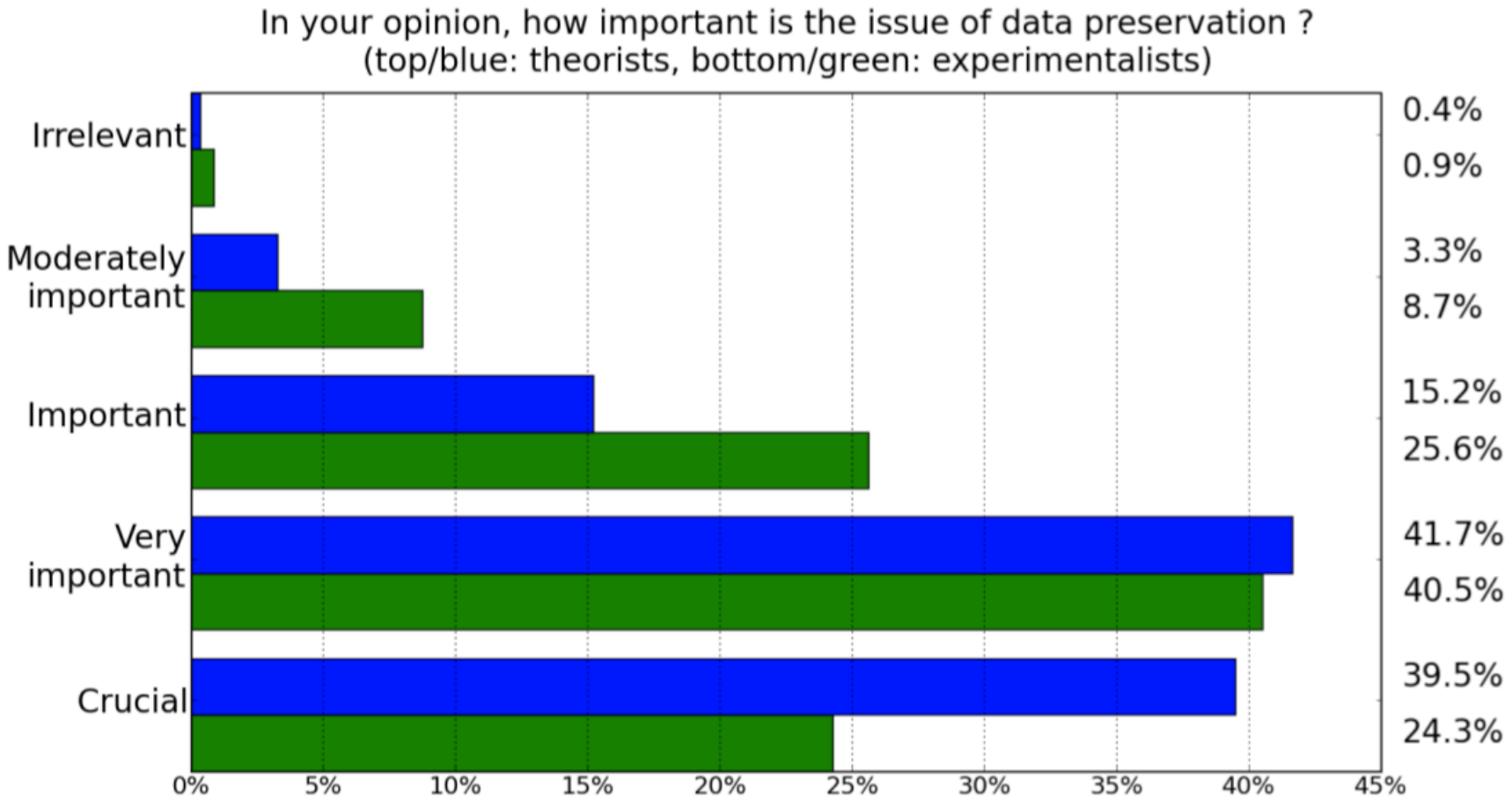
C. Diaconu
CPPM

HEP experimental programmes ± 10 years



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

Support for data preservation in the HEP community



PARSE insight
Permanent Access to the Records of Science in Europe

e-infrastructure

SEVENTH FRAMEWORK PROGRAMME

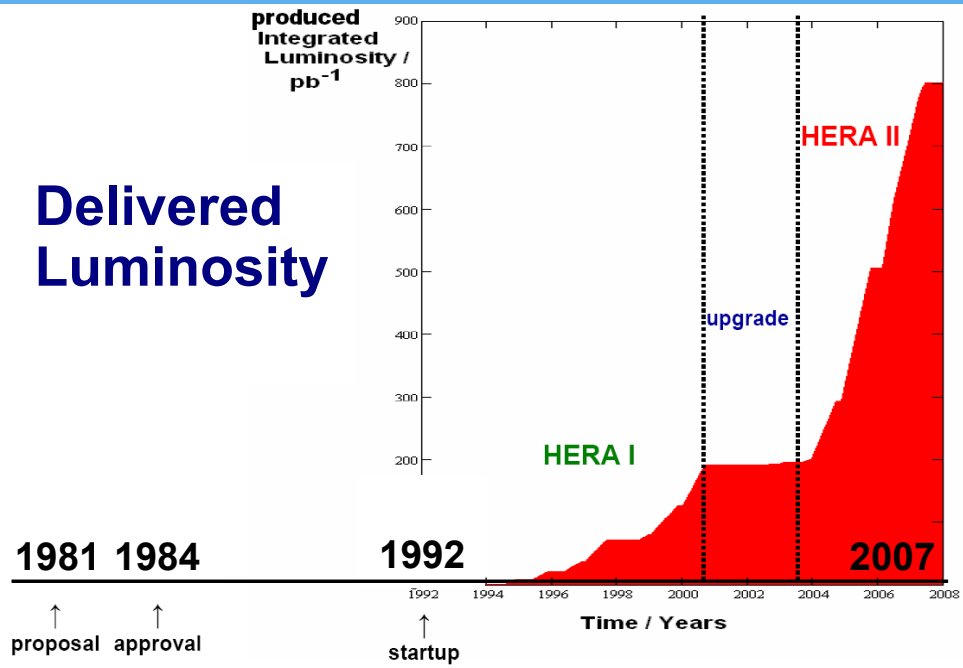
CERN

arXiv:0906.0485

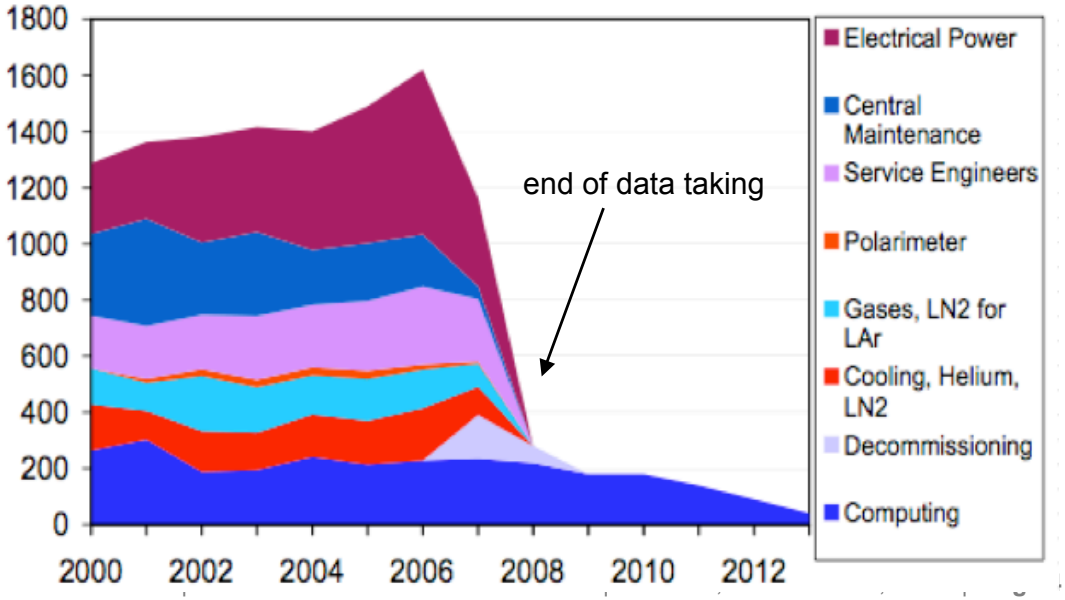
Why is it Difficult to Preserve HEP Data?

- Lots of data available to analyse at the end of collisions
- The existing resources (funding and expertise) then decrease when the data taking stops

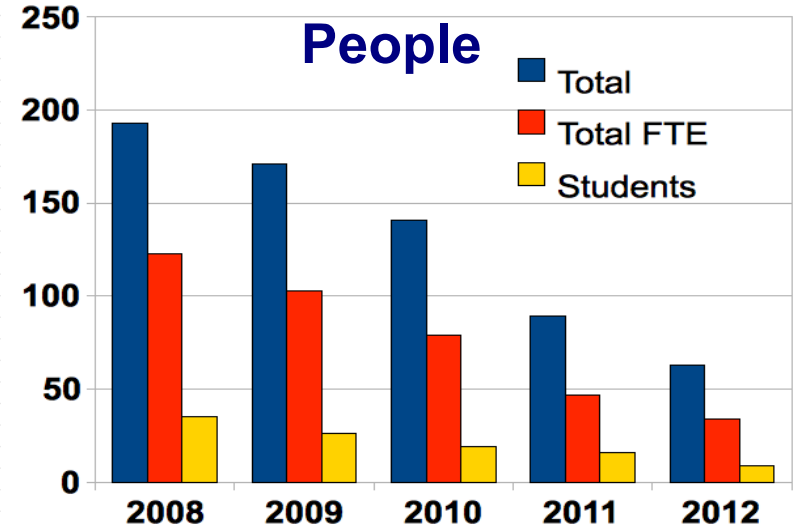
Delivered Luminosity



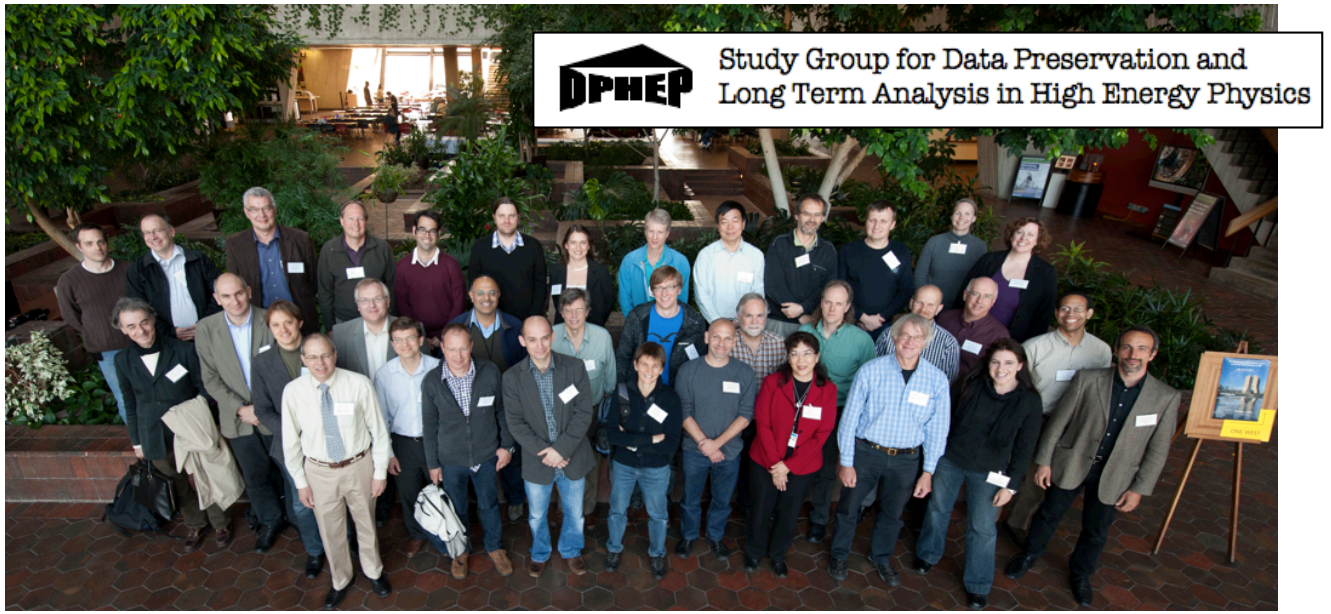
Funding



People

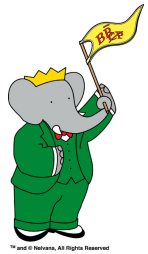


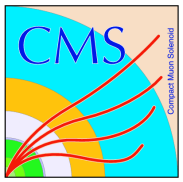
DPHEP: An international study group on data preservation



- First contacts established in September 2008
  - Group since grown to over 100 contact persons (chair : CD)
  - Endorsed as an ICFA panel summer 2009
  - *All 4 LHC experiments joined in 2011*
- Steering Committee: representatives from all members
- International Advisory Committee:
  - Jonathan Dorfan (Chair, SLAC), Siegfried Bethke (Chair, MPIM), Gigi Rolandi (CERN), Michael Peskin (SLAC) Dominique Boutigny (IN2P3), Young-Kee Kim (FNAL), Hiroaki Aihara (IPMU/Tokyo), Alex Szalay (JHU)

DPHEP: An international study group on data preservation



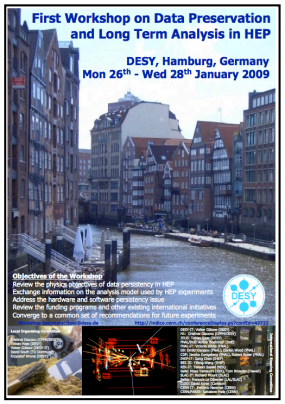

Institute of High Energy Physics
Chinese Academy of Sciences



Science & Technology
Facilities Council

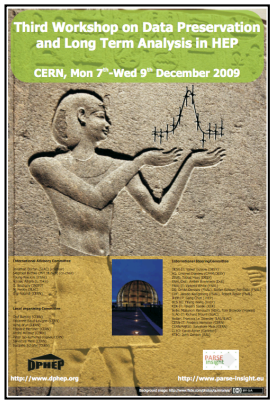

DPHEP: An international study group on data preservation

- Series of DPHEP workshops held since 2009



Jan 2009: DESY

May 2009: SLAC



Dec 2009: CERN

Jul 2010: KEK

May 2011: Fermilab

- The first task of the group was to establish the working directions
  - "To confront data models, clarify the concepts, set a common language, investigate technical aspects, compare with other fields handling large data."

- Initial findings published in an interim report December 2009
  - Focus on four key areas of the study group: Physics Case for Data Preservation, Preservation Models, Technologies, Governance

arXiv:0912.0255

DPHEP Visibility

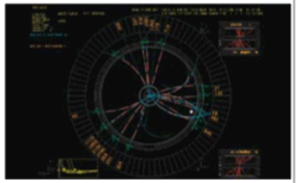
CERN Courier, May 2009

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



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-...ed events. The amount of data that should be ...is varies between 0.5 PB and 10 PB for each ...not huge by today's standards but nonetheless ...degree 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**

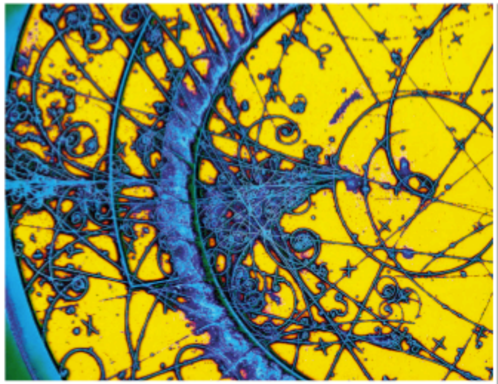
Rescue of Old Data Offers Lesson for Particle Physicists

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

nature news **May 2011**

Symmetry, December 2009

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

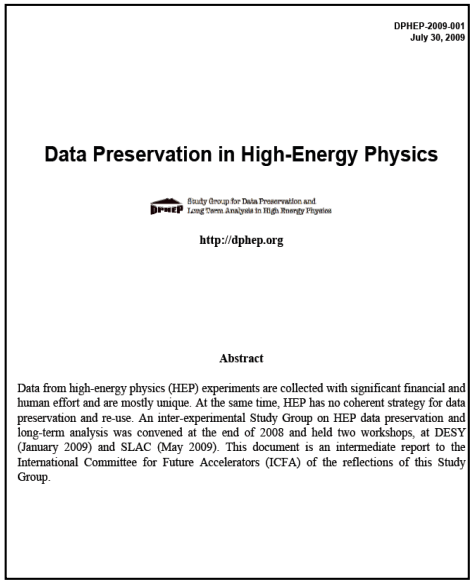
Data Preservation

- 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

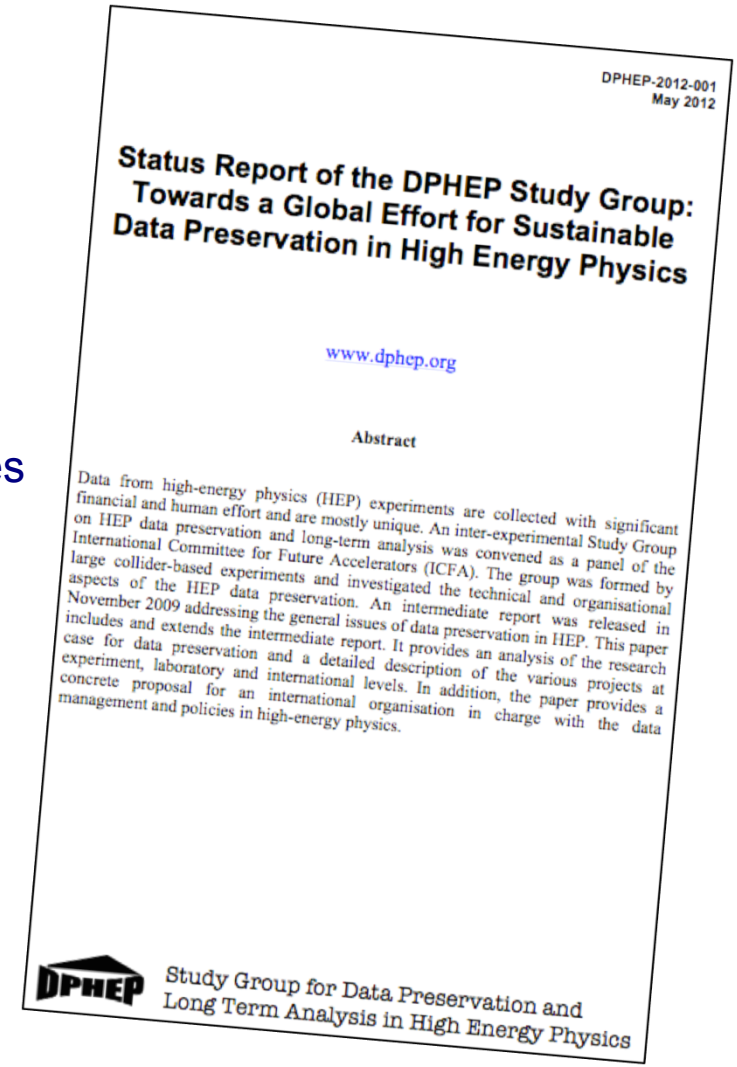
Abstract

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

New DPHEP publication

- 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

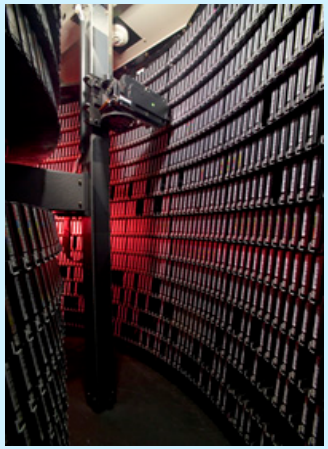
Abstract

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

What is HEP "data"?



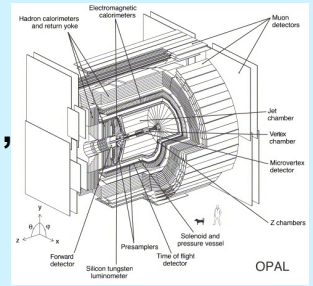
Digital information

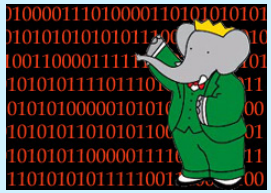
The data themselves, volume estimates for preservation data of the order of a few to 10 PB

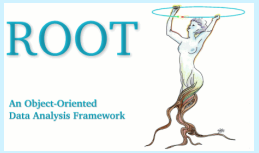
Other digital sources such as databases to also be considered

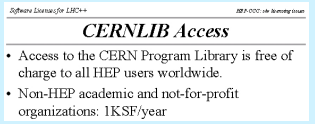
Software

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



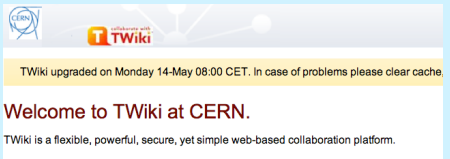


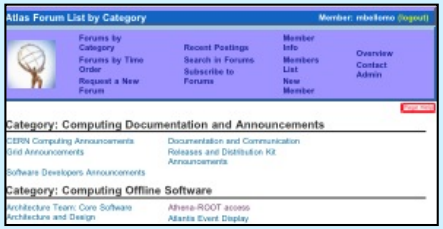




Meta information

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





Publications

arXiv.org

INSPIRE

Durham University — Science & Technology Facilities Council

HEPDATA: REACTION DATA Database

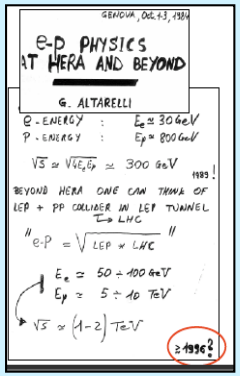
...containing numerical values of HEP scattering data such as total and differential cross sections, fragmentation functions, structure functions, and polarisation measurements, from a wide range of experiments. It is compiled by the **Durham Database Group (UK)** with help from the COMPAS group (Russia,)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

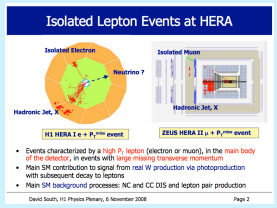
Journal of High Energy Physics

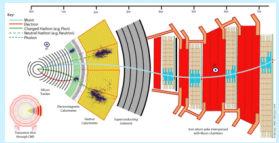
A refereed journal, written, run and distributed by electronic means

Documentation

Internal publications, notes, manuals, slides





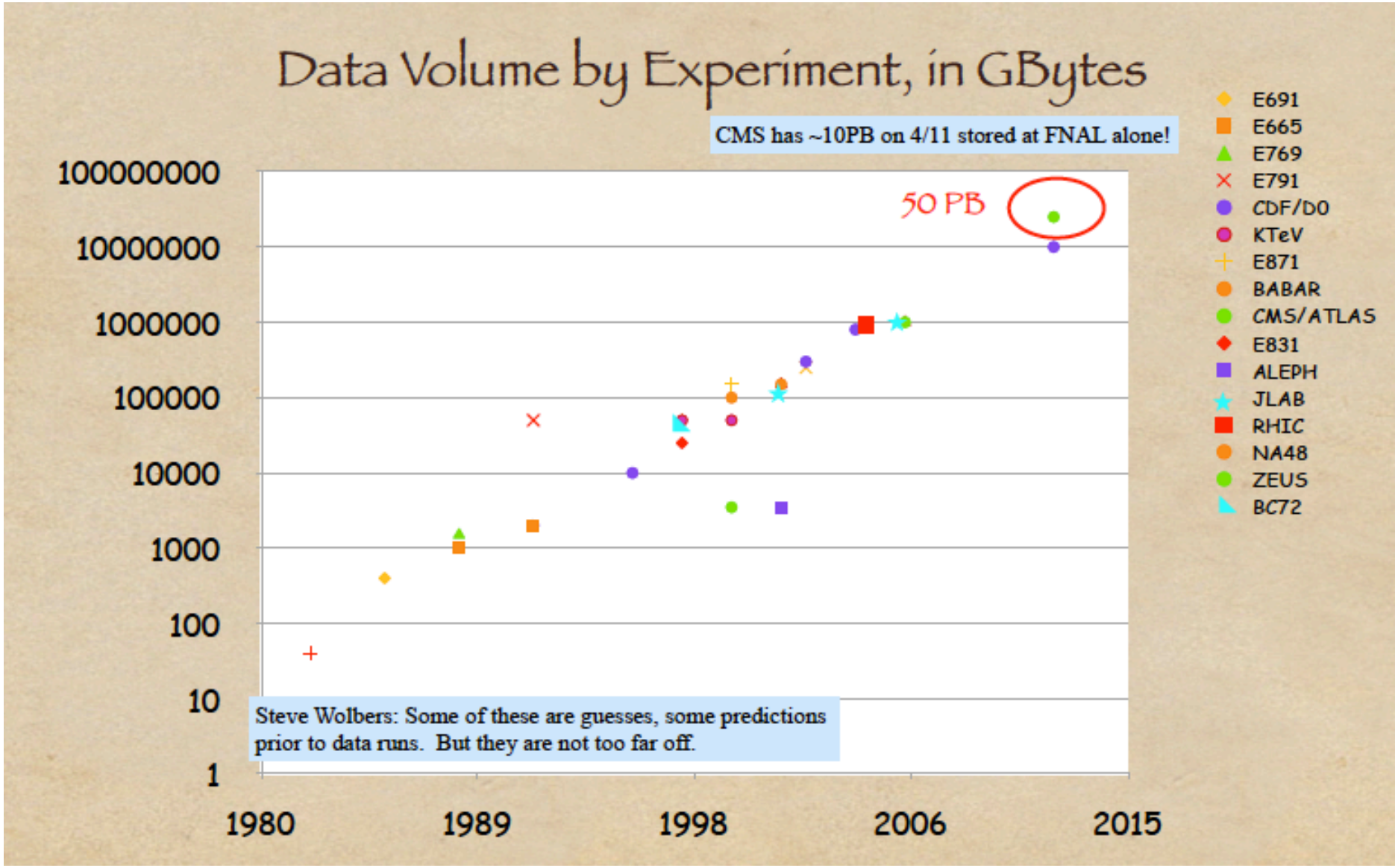


Expertise and people

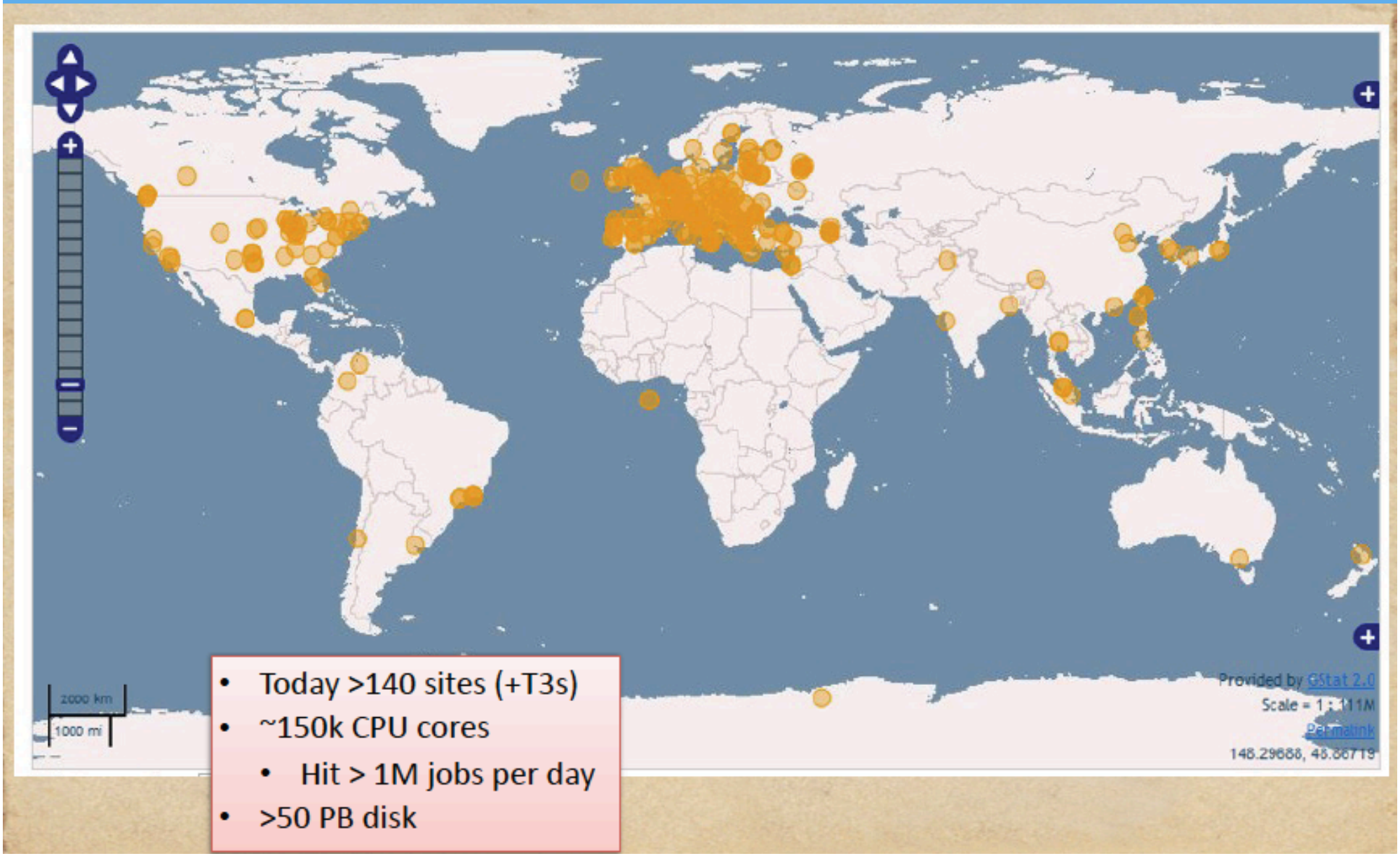
Summary of information from the (pre-LHC) experiments

	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

LHC: a different scale….



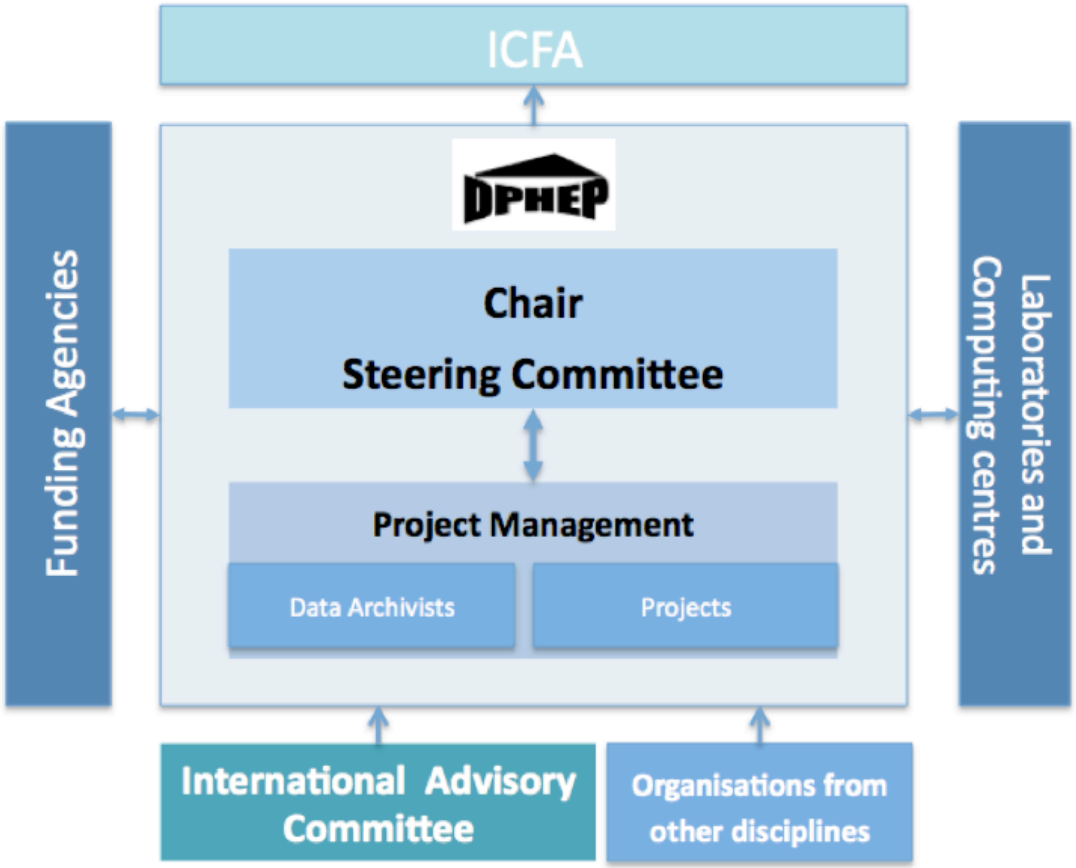
LHC computing



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

The DPHEP Organisation

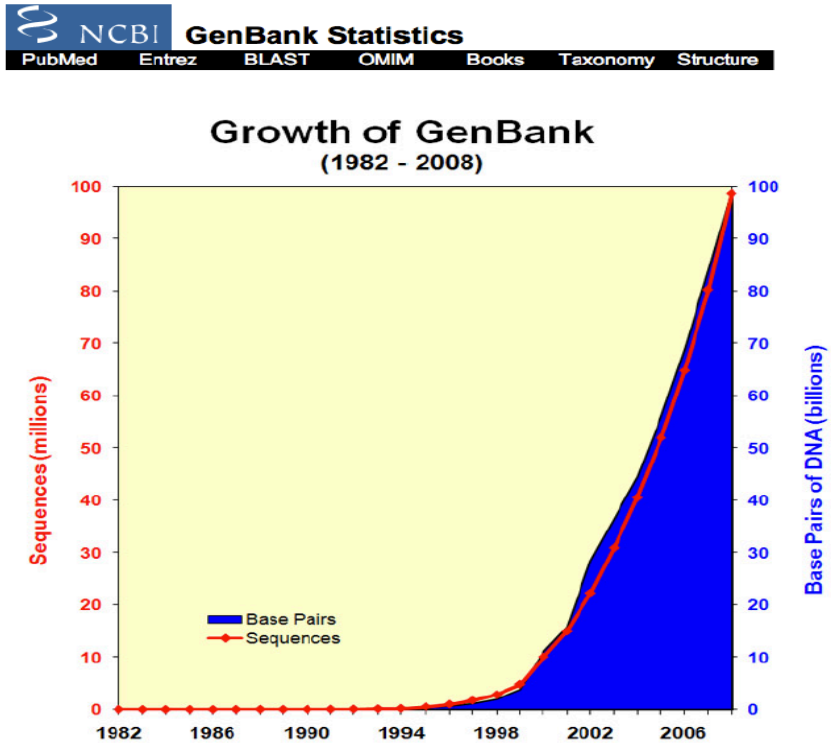
- Retain the basic structure of the Study Group, with links to the host experiments, labs, funding agencies, ICFA
- Installation of a full time DPHEP Project Manager, who acts as the main operational coordinator
- The DPHEP Chair (appointed by ICFA) coordinates the steering committee and represents DPHEP in relations with other bodies



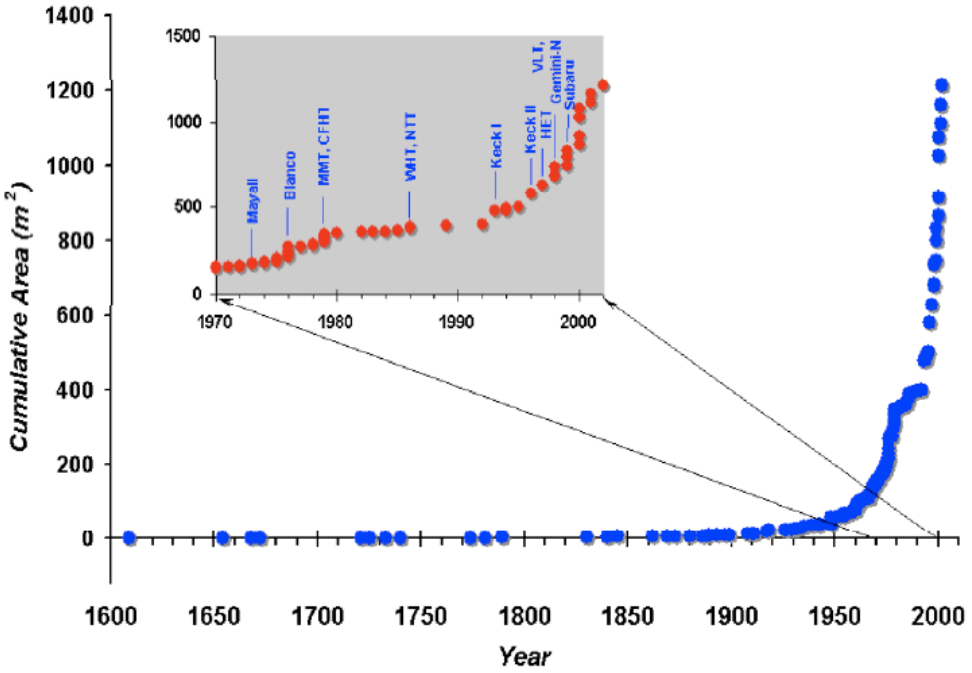
October 11, 2012: CERN endorse the blueprint and appoints the DPHEP Project Manager

We are not alone….

- Other fields observe a dramatic increase in data and are questioning the long term future of this data



Telescope Collecting Area



Generic arguments

- Task forces already in place to address this issue in a generic way (standards)
  - e.g. Blue Ribbon, APA, DPC, eSciDir, …

http://www.alliancepermanentaccess.eu
http://brtf.sdsc.edu
(intermediate report and references)

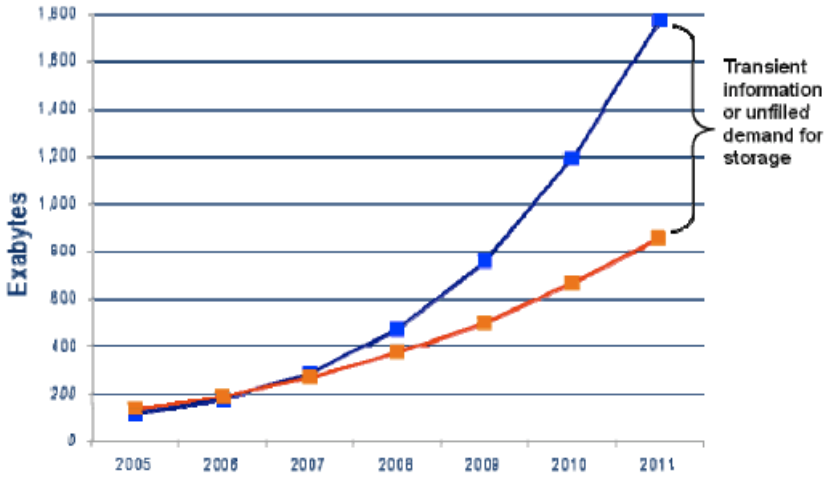


FIGURE 1.3: **Information and Storage**
Source: J. Gantz January 2008 (revised). Used with permission.

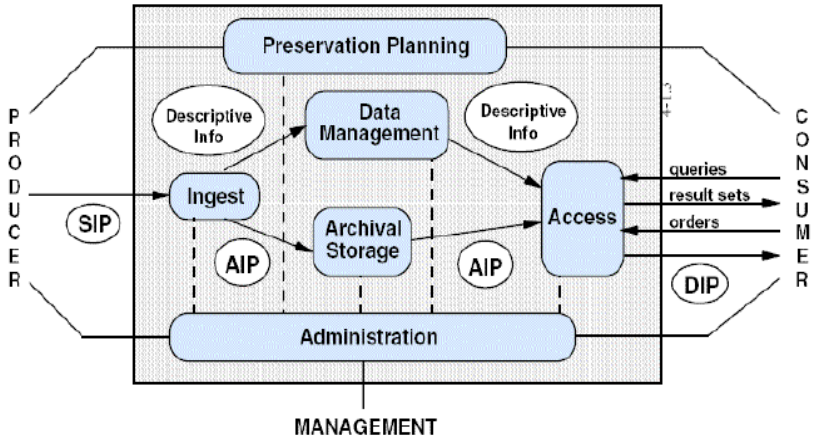
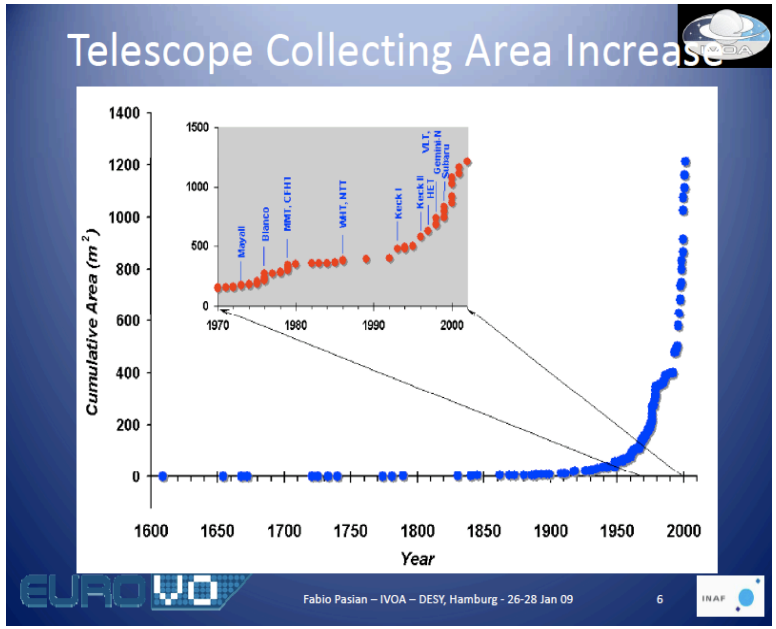
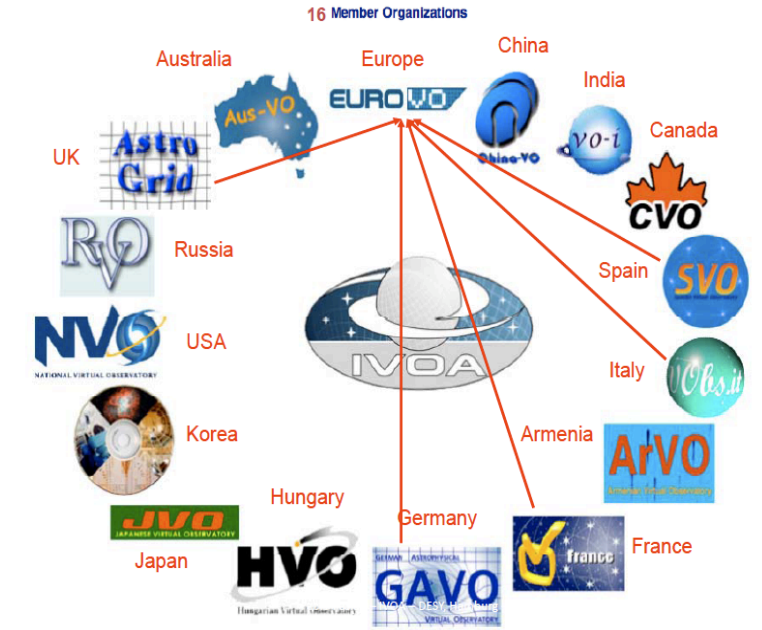


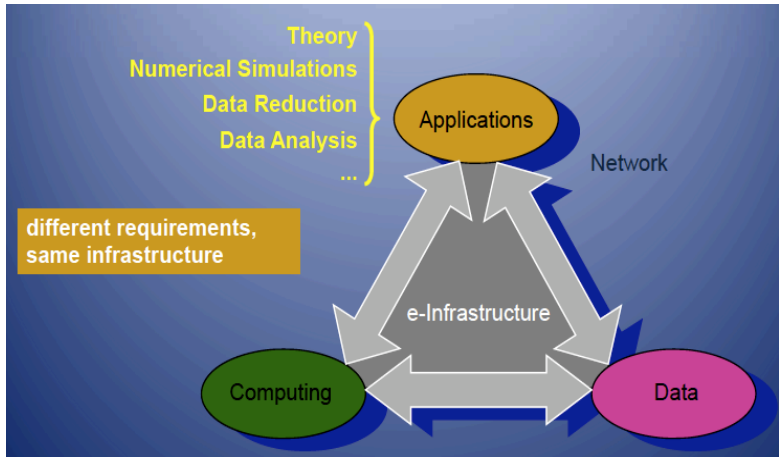
FIGURE 2.1: **The OAIS Reference Model**
http://public.ccsds.org/publications/archive/650x0b1.pdf, Page 4-1.
Source: Consultative Committee for Space Data Systems January 2002.

- Scientific Data is a major component of the ongoing efforts (complexity)
- Some scientific fields are well advanced : astrophysics

Virtual Observatories in Astrophysics





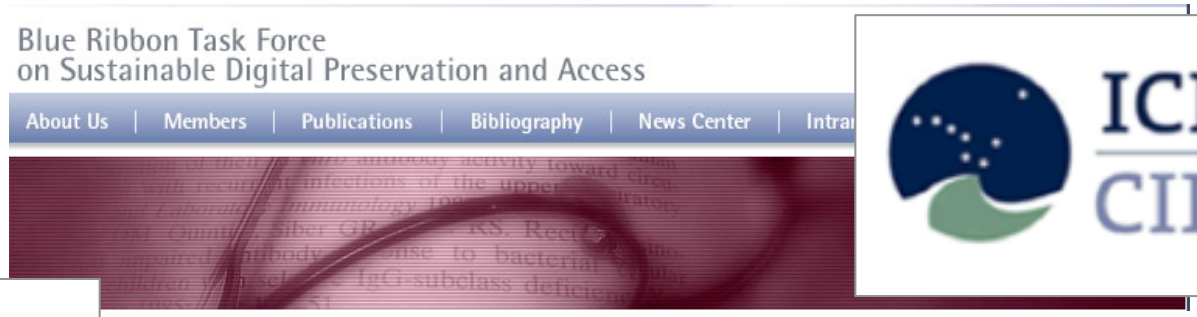


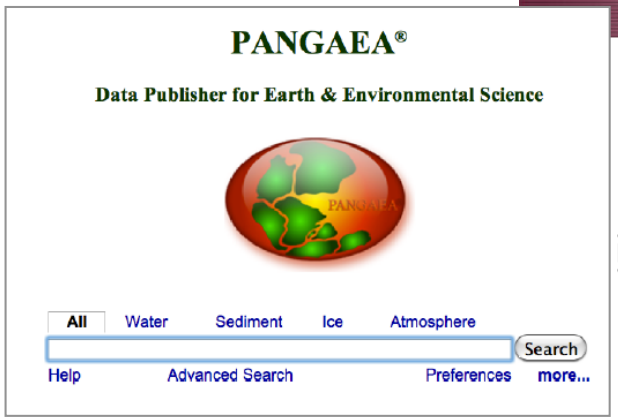
- Data Archives Inter-operable
- Work on standards and access to
  - Data, simulation, mining techniques
- International, multi-experiment
- Agregated Person-power: about 100FTE

F.Pasian

Initiatives in other fields

- Data preservation and in particular open access and data sharing are present in other fields such as:
  - Astrophysics, molecular biology, earth sciences, humanities and social sciences
  - Towards Data-scopes (Alex Szalay)





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.**

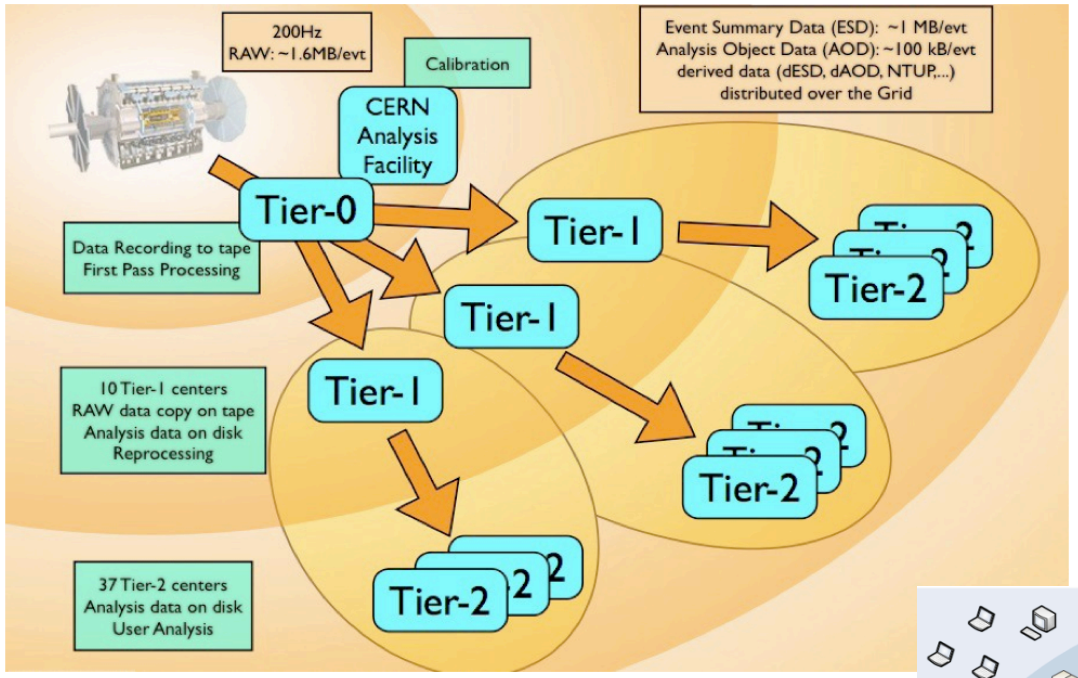
In France…

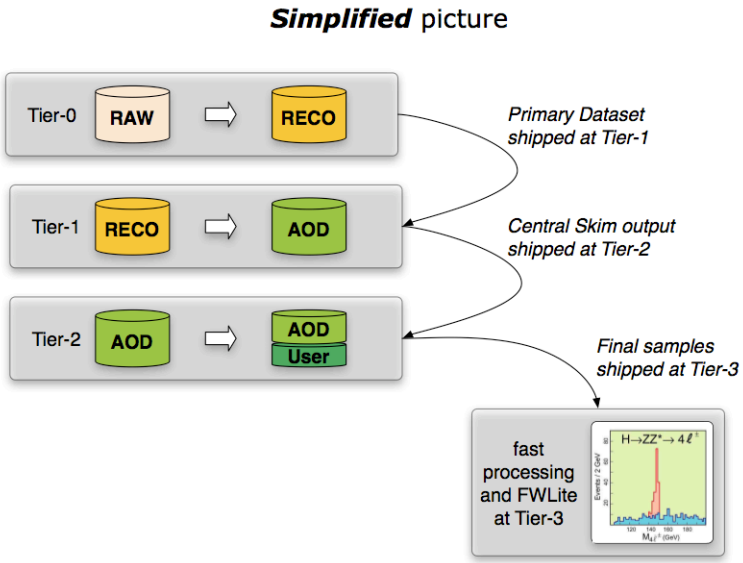
- Mastodons:
  - La Mission Interdisciplinarité (MI) du CNRS lance **un défi sur la gestion, l'analyse et l'exploitation des très grandes masses de données scientifiques** (MASTODONS).
  - Projet: PREDON C. Diaconu (CPPM), G.Lammana (LAPP). S. Kraml (LPSC)

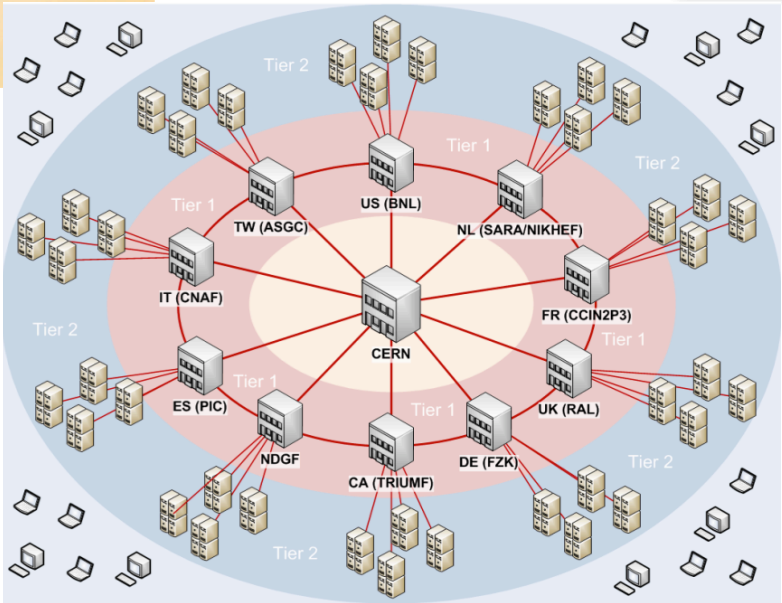
    le projet PREDON propose une approche nouvelle qui mélange les capacités scientifique, technique et organisationnelle des grandes collaborations en physique des particules et astrophysique pour définir et construire un system robuste de stockage et analyse des donnés à long terme.

  - But pour 2012: montrer qu'il existe un interêt a travers les disciplines et les instituts du CNRS, Workshop in Marseille November, 19-21, 2012
  - Initiatives similaires MPI (Allemagne), INFN(Italie), STFC(UK)
  - http://indico.cern.ch/conferenceDisplay.py?confId=209688

Data analysis models in HEP in the LHC era







- More skims - *yes*
- More distribution - *certainly*
- More complexity - *perhaps..*
- Data placement is key, but analysis-wise it's still very similar to what we had before