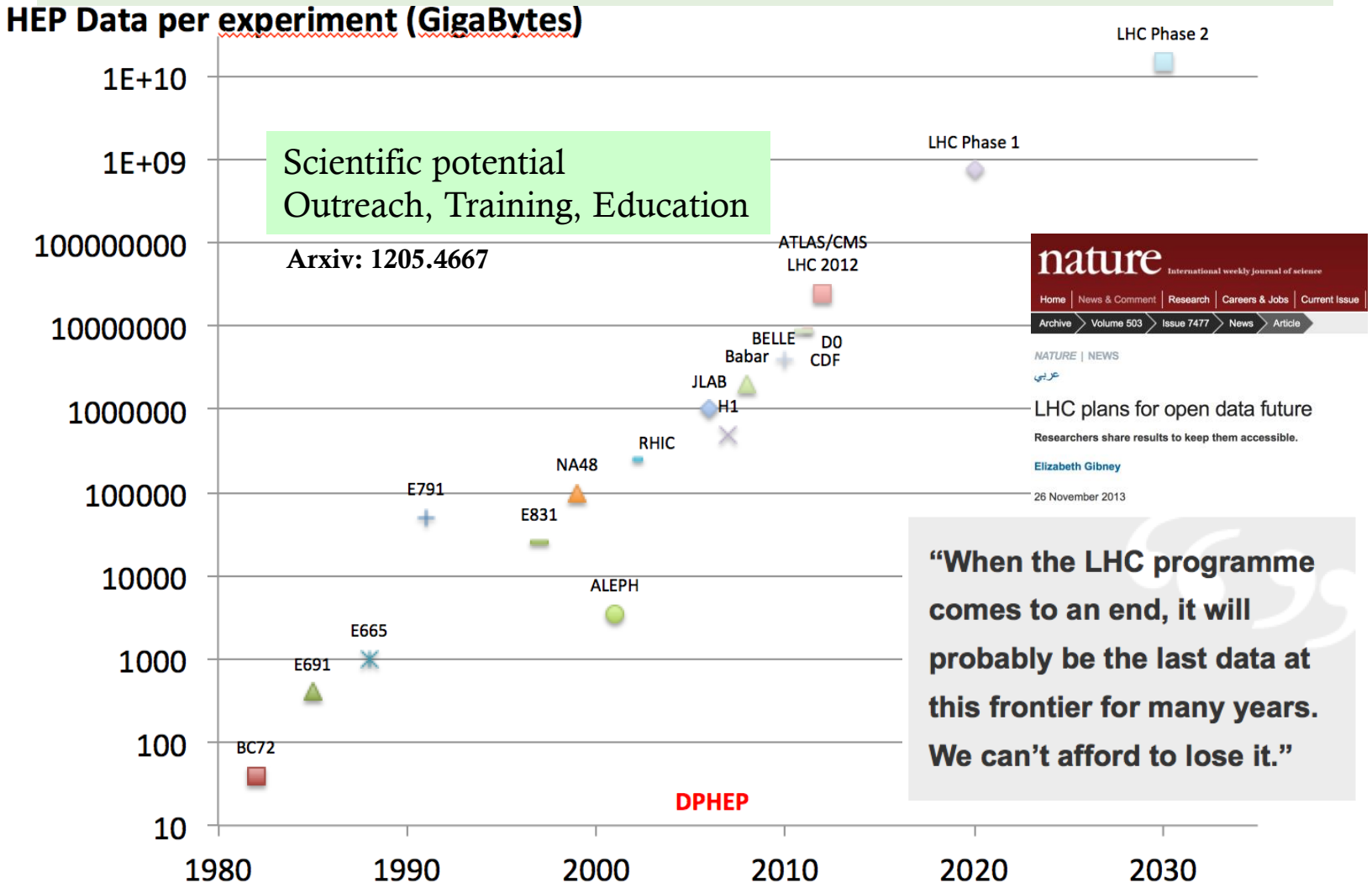
Data Preservation in High Energy Physics
The road to DPHEP

http://dphep.org

Data Preservation in HEP



Once upon a time (end 2008)

- Lots of HEP data collected, and lots of data to come
  - PEP2,HERA,Tevatron shutdown
  - The LHC start-up
- Some experiments wondered what's going to happen to their data
  - Is it going to be relevant?
  - Was it prepared?
  - Costs? Personpower?
  - Coordination?
- Study group convened: contacts to many labs and experiments, decide to stay around large colliders
  - Issue at ICFA meeting in SLAC in Nov 2008

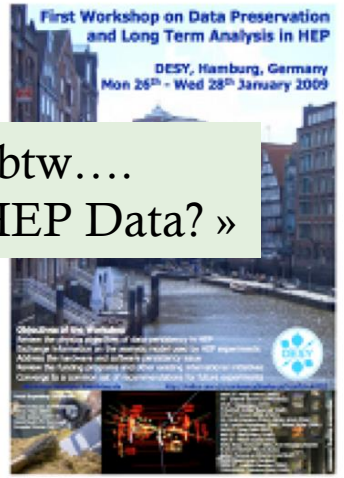
Study Group 2008

- **C**ommon 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.

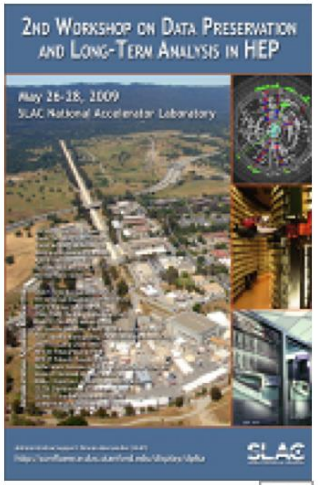
- Review and document the physics objectives of the data persistency in HEP.
- Exchange information concerning the analysis model: abstraction, software, documentation etc. and identify coherence points.
- Address the harware and software persistency status.
- Review possible fundings programs and other related international initiatives.
- Converge to a common set of specifications in a document that will constitute the basis **for future collaborations.**

Workshops 2009: the start-up

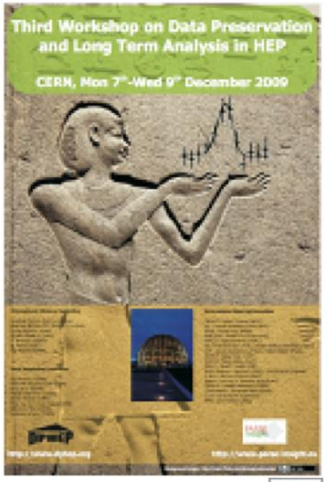
« … and btw….
what is HEP Data? »



DESY January 2009



SLAC May 2009



CERN December 2009

« It is clear that the issue is quite fresh in the community, in other words: **not defined.** »

WG1: Physics Case
WG2: Models
WG3: Governance
WG4: Technologies

First specifications Document **arXiv:0912.0255**

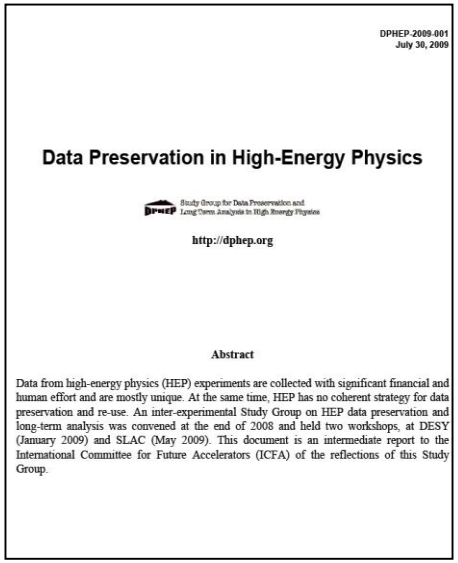
July 2009: DPHEP SG becomes ICFA Panel

Open Symposium

Start Planning Blueprint

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

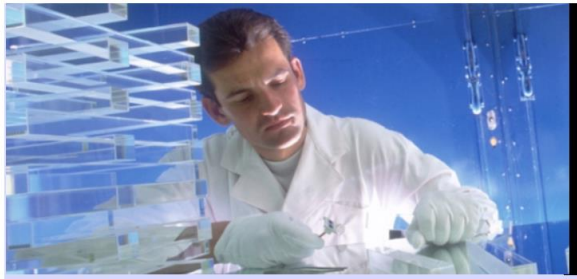
http://dphep.org

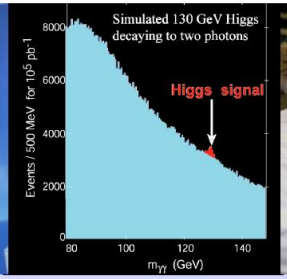
Abstract

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
  - 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

CERN DG Talk in Dec. 2009





After long preparation times and exciting physics:
Data preservation should be prepared as a part of the experimental programs

- **Need a strategy: coherent action, global initiative**
- **Need academic incentives and financial stimulus**

Workshops 2010-2012

KEK July 2010

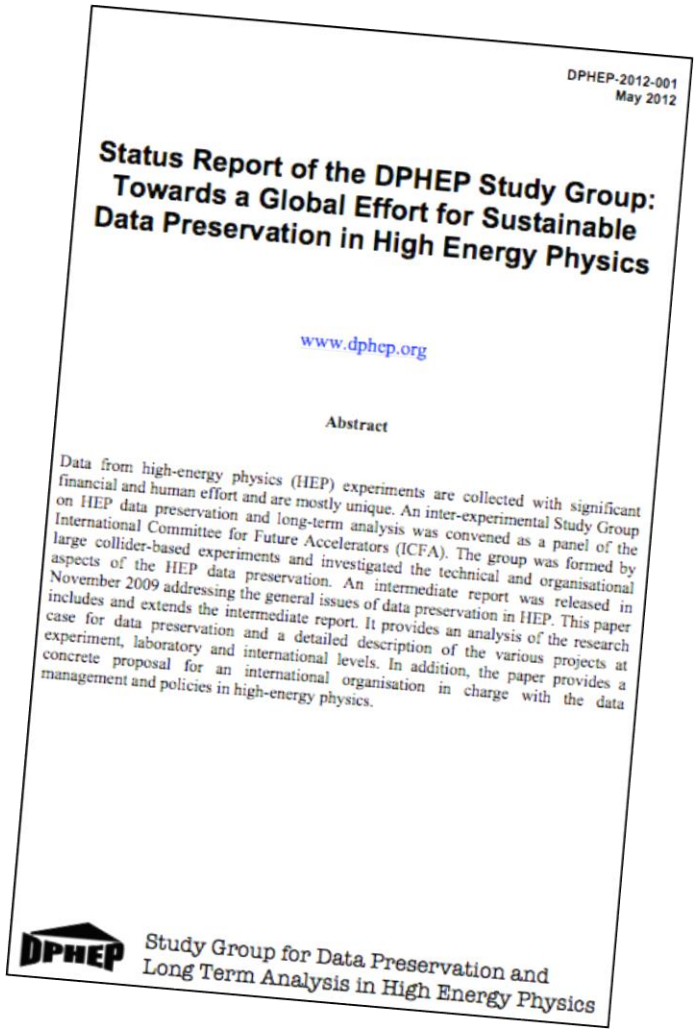
Fermilab May 2011

CPPM November 2012

- Exploring phase: building the community and working towards the blueprint:
  - Support of large labs essential
  - Continue to report to ICFA
- LHC experiments joined in 2011
  - Harmonisation and policy advances
- LEP data re-resurection discussed.
- Connections to multi-disciplinary projects, DASPOS

DPHEP Blueprint May 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



arXiv:1205.4667

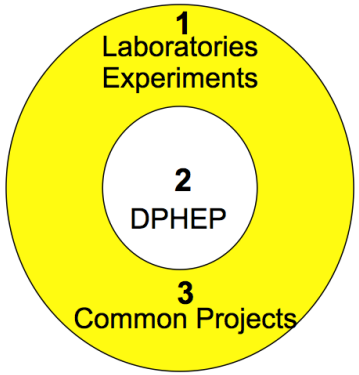
DPHEP 2008-2012

Year	2007	2008	2009	2010	2011	2012
HEP	**HERA stops**	**Babar stops**	**LHC starts**	**Belle I stops**	**Tevatron stops**	
DPHEP Meetings			DESY, SLAC, CERN 1st doc.	KEK	FNAL	CPPM
DPHEP Group		1st contacts	Endorsed by ICFA		**LHC exp. joined**	Confirmed by ICFA
DPHEP Docs			**White Paper**			**Status Report**
DP Projects within expts.			**Babar starts**	**HERA starts**		**CDF/D0**

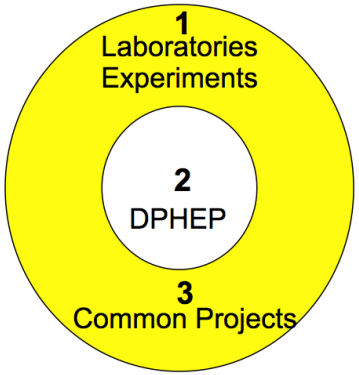
Aggregate person power - preliminary estimates

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.

Next step towards DPHEP consolidation



Projects and PP estimates
In the released document
Table 8



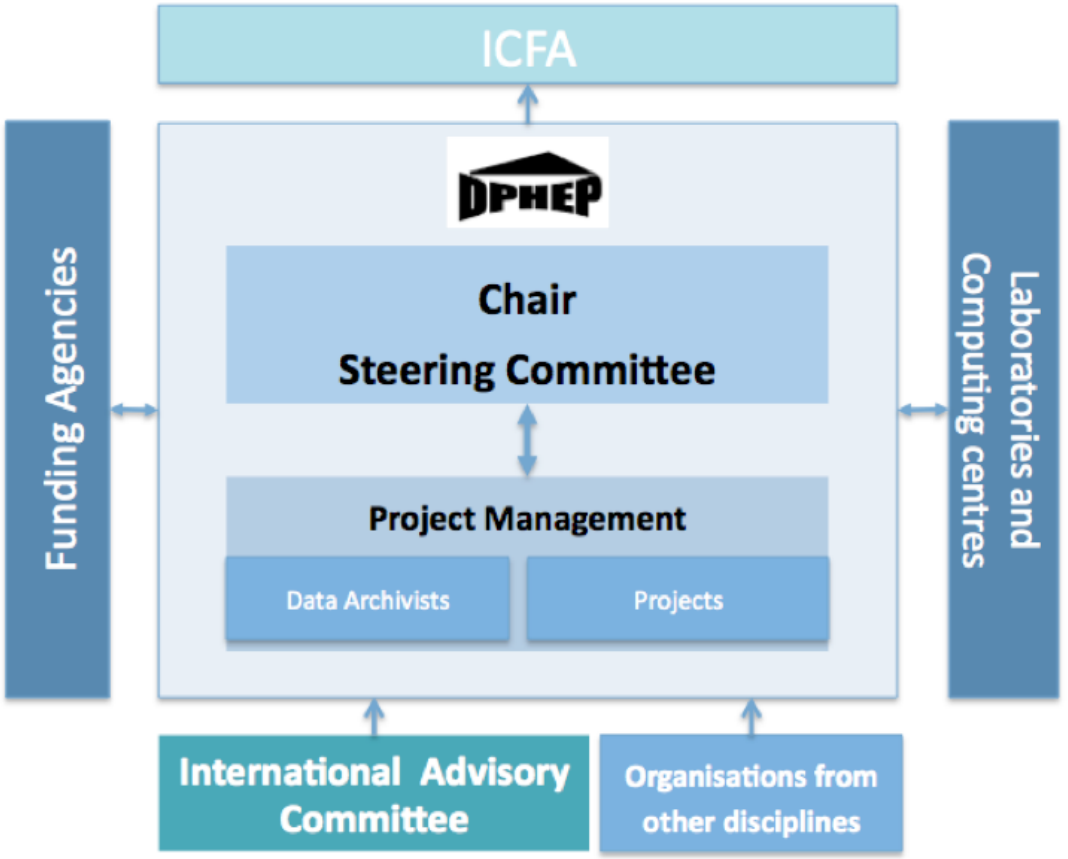
	Project	Goals and deliverables	Resources and timelines	Location, possible funding source, DPHEP allocation
Experiment and laboratory *Priority: 1*	Experimental Data Preservation Task Force	Install an experiment data preservation task force to define and implement data preservation goals.	1 FTE installed 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.

DPHEP international organisation

A “local success” is undefined
DP must be a global enterprise
or it will dissapear

There is a clear and urgent
need for **a project manager**



From the BluePrint:

There is a need for much more

- **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.

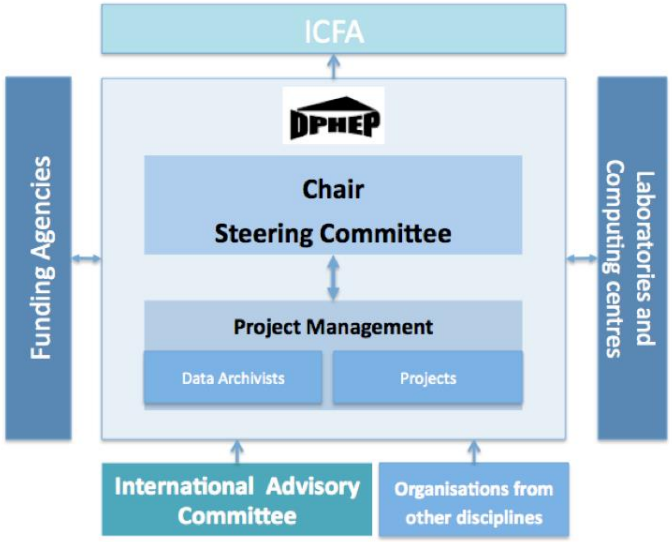
- **More Cooperation:** Cooperation with other fields in data management: access, mining, analysis and preservation; appears to be unavoidable and will also dramatically change the management of HEP data in the future.

Communication essential at the level of funding agencies, example: HEPAP

- **Report on current policies and practices of the High Energy Physics program for disseminating research results**
  - June 3, 2011
- "To date **no HEP experiment** has provided large-scale **open access** to its raw form digital data, although limited access to processed data has sometimes been granted upon request. The size and complexity of these datasets present **significant** technological, governance, and support **challenges.** "
- "**DPHEP** Study Group is an international effort working to **develop solutions** to these challenges and to provide **common guidelines** for use by future collaborations. "
- "The preservation of HEP data and its dissemination requires **organized action** from the experimental collaborations, the participating laboratories, and the funding agencies."
  - **May 2011: NSF initiates a funding line for data preservation in HEP: proposal accepted (DASPOS)**

The DPHEP Collaboration

- October, 2012: CERN endorse 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
- The collaboration agreements signed in 2013



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 maturation phase: 2012-2014

- Central activity intensified (Project Manager)
- Implementation board meetings (every 6-8 weeks)
- Several topical workshops: (costs, technologies)
- Ramp-up reflection and activties at LHC
- DPHEP visible in interdisciplinary initiatives

Data Preservation at present

- Data preservation is discussed widely in HEP
  - Dedicated projects in SLAC, DESY, FNAL,…
    - Transition from R&D to service is critical
  - New projects: MPI (OPAL-JADE-HERA), ALEPH
- LHC experiments
  - Data preservation is a « spec », included in the computing models and plans for Phase I/II upgrades
  - Most experiments have prepared Data Preservation and Open Access policies, concrete implementations starting

A note on the physics case

- Many hypothesis and concrete examples discussed in the past workshops: re-analysis, re-cast, combinations etc.
- Did all this continue to happen?
  - Do we have now continued evidence for the physics case of preserved data?
- Did data preservation initiatives within experiments played any role in enhancing the physics output flow for the ending experiments?
- Does it play any role in the running of the present experiments?
- … and in the planning for new experiments?

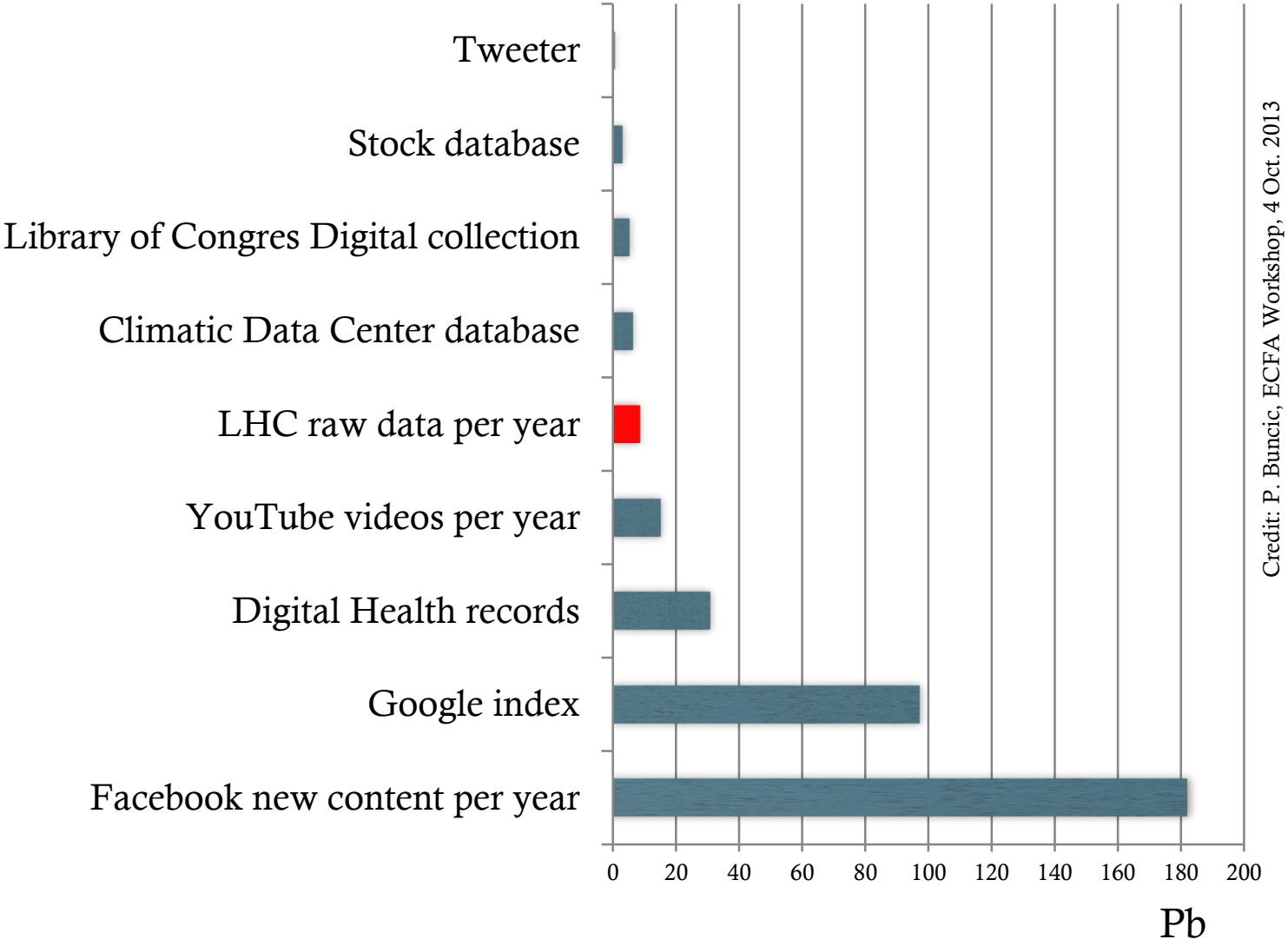
Costs

- Estimates in the blueprint (2012)
- Workshop « Costs of curation » at CERN in Jan. 2014
- Discuss concrete examples with (by now) more realistic estimations of costs
  - Material costs decrease
  - Personnel costs remain constant
  - Critical steps and associated costs
- Document with models and possibly business plan proposals/variants.

From common projects to services

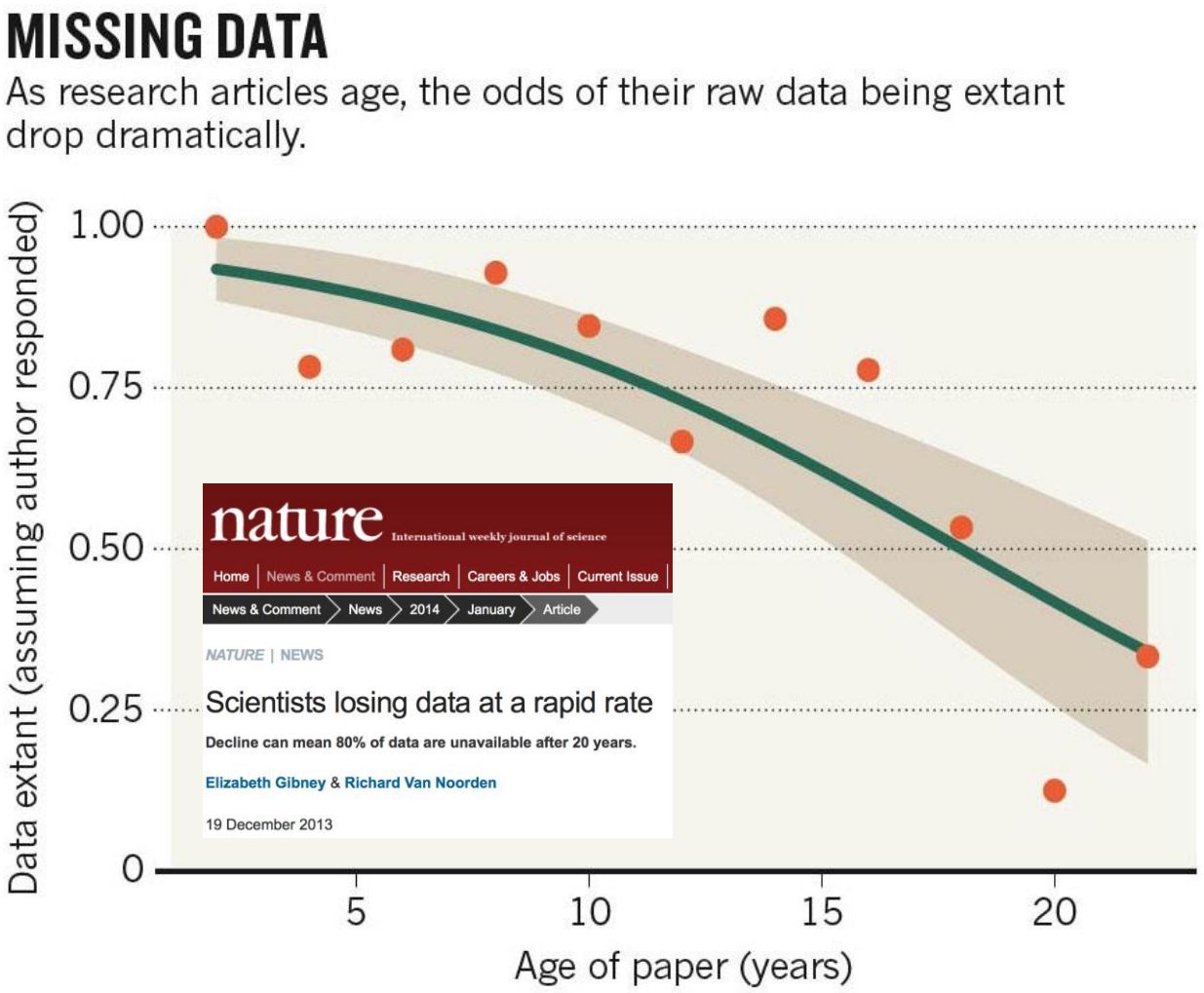
- Projects proposed in the blueprint (with costs estimation)
- Concrete work plan for 2014/2015:
  - DPHEP portal
  - Pub/high-level data projects (INSPIRE)
  - Virtualisation (sp-desy, slac, cernvm)
  - Bit-level preservation (HEPiX WG)
  - Open Data formats
  - Document « Costs » (workshop in 2014)
  - More? (Less?)
- Significant progress expected in 2015/2016
  - Expertise sharing, new opportunities
  - Person-power optimisation

A context of « big data »



And their dissapearance…

Study over 516 ecology papers published between 1991 and 2011.



Interdisciplinary approaches

- HEP connected to global efforts and potentially to common future funding programs
  - EU-4C, APA, SCIDIP-ES, EUDAT,…
  - RDA (Research Data Alliance) - WG on Data Preservation
- National initiatives with strong/leading HEP component emerging:
  - DASPOS (US)
  - PREDON (France)
  - Finland (educational pilot project with CMS data)
  - … more to come

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:

Annex 1: Partners of the DPHEP Project and contact persons

Initial DPHEP Partner	Location	Contact person
European Organization for Nuclear Research, **CERN**	Switzerland	J. Shiers
Deutsches Elektronen-Synchrotron, **DESY**	Germany	D. South
Helsinki Institute of Physics, **HIP**	Finland	K. Lassila-Perini
Institute of High Energy Physics, **IHEP**	China	G. Chen
Institut national de physique nucléaire et de physique des particules, **IN2P3**	France	G. Lamanna
Institute of Particle and Nuclear Studies, High Energy Accelerator Research Organisation, **IPNS, KEK**	Japan	T. Hara
Max Planck Institut für Physik , **MPP**	Germany	S. Kluth

Following institutes are members of the DPHEP Study Group and intend to join formally the DPHEP Collaboration:

Brookhaven National Laboratory, **BNL**	USA	M. Ernst
CSC- IT Center for Science	Finland	N.N.
Fermi National Accelerator Laboratory, **FNAL**	USA	S. Wolbers
Institute of Particle Physics, **IPP**	Canada	R. Sobie
Istituto Nazionale di Fisica Nucleare, **INFN**	Italy	M. Maggi
SLAC National Accelerator Laboratory	USA	C. Cartaro
Science and Technology Facilities Council, **STFC**	UK	J. Bicarregui

DPHEP Collaboration

- The Collaboration Agreement are signed, DPHEP Collaboration exists
  - Give a clear sign of the will of all labs to co-operate and collaborate in this common challenge
- New members will join:
  - IOP Canada (CB)
  - More laboratories, funding agencies?
- First official DPHEP Collaboration Meeting:
  - Discuss the objectives and the long term future of the collaboration
  - Use this meeting to have a new impulse
    - « Data Preservation in HEP Vol. 3 » ?

The goals of this workshop are:

- Establish the motivation for long-term data preservation in HEP in terms of succinct Use Cases
  - Are there a common set of Use Cases, such as those that were recently agreed for the 4 main LHC experiments but in a more global scope?
- Review the existing areas of "Common Projects"
  - Can these be extended (similarly) from their current scope - often LHC - to become more global?
- Perform a site-experiment round-table to capture the current situation HEP-wide
  - >5 years experience in what is (still) possible/feasible in HEP
  - Report back to the community on our most recent findings

First DPHEP Collaboration Board on Wednesday June 10, 9h30

- Proposal: Inaugural Open CB
- Agenda:

  - MoU review (short summary)
  - Discussion on Collaboration functioning
  - Next 2 years plans, elections
  - Person power and continuation of the project management
  - New partners (HEP FAs or labs)
  - International cooperation with similar projects

BACKUP

	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.

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

Data Longevity: > 10 ans