Data Preservation and Long Term Analysis in HEP.

Safeguarding the heritage of HEP data for the future

David South (DESY)
on behalf of the DPHEP Study Group
dphep.org

CHEP 2012, May 21-25
New York, USA

HELMHOLTZ ASSOCIATION

DPHEP

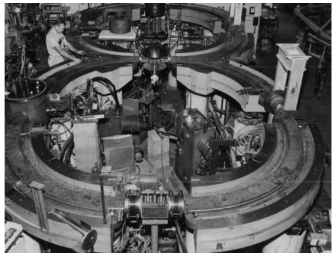
DESY

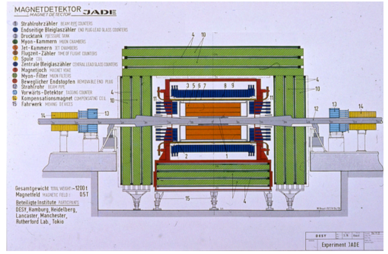
Outline

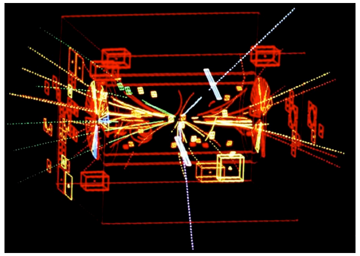
- Introduction
  - Data preservation in HEP
  - An international initiative: DPHEP
  - The scientific potential of HEP data
- DPHEP data preservation models
  - Current strategies of the experiments
  - Emerging projects in the DPHEP community
- Future working directions
  - Where we are now, where we need to go, and how that's going to happen

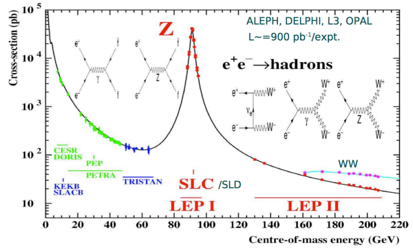
Experimental particle physics in the collider era

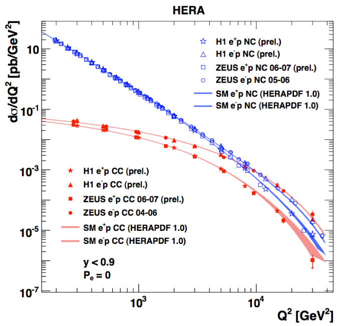
- A wide variety of physics results from many, often very different experiments
- Energy frontier probed with increasingly complex accelerator installations
  - New experiments typically supersede previous, similar ones - but not always
- Growth in size of the necessary international collaborations, as well as the diversity of the data management
- The age of the LHC has truly arrived
  - The Super-B factories and other projects such as the ILC or next e-p(A) collider are to come

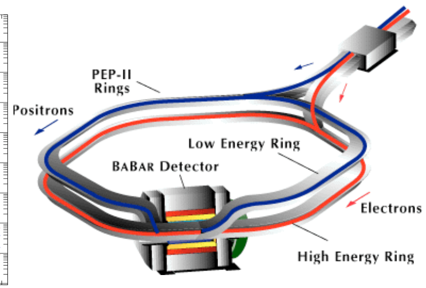


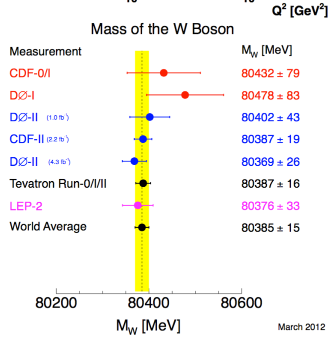


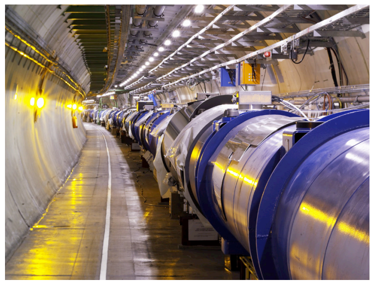




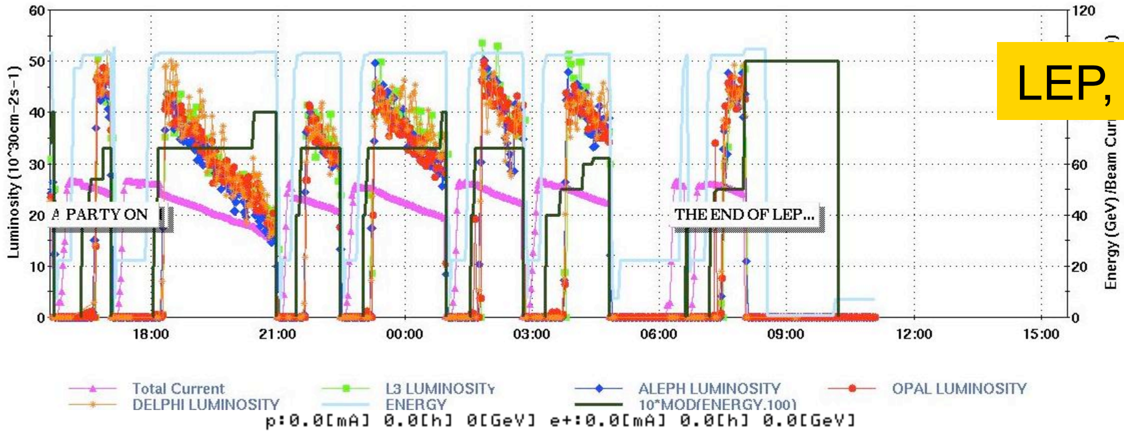




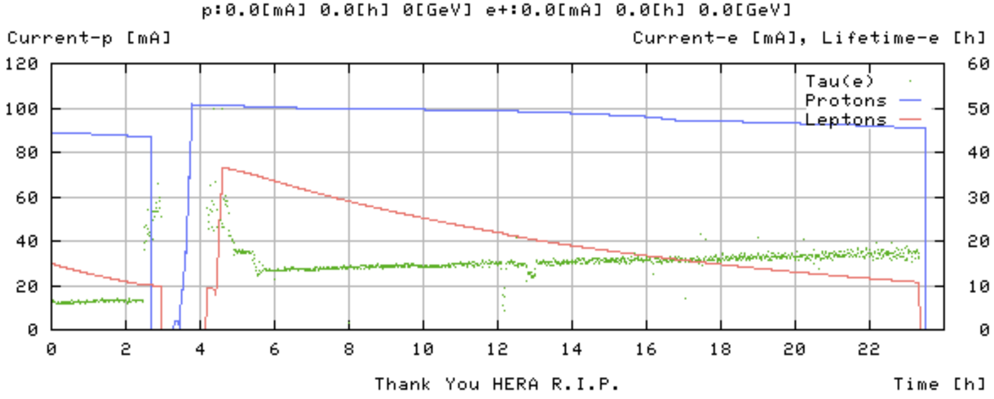




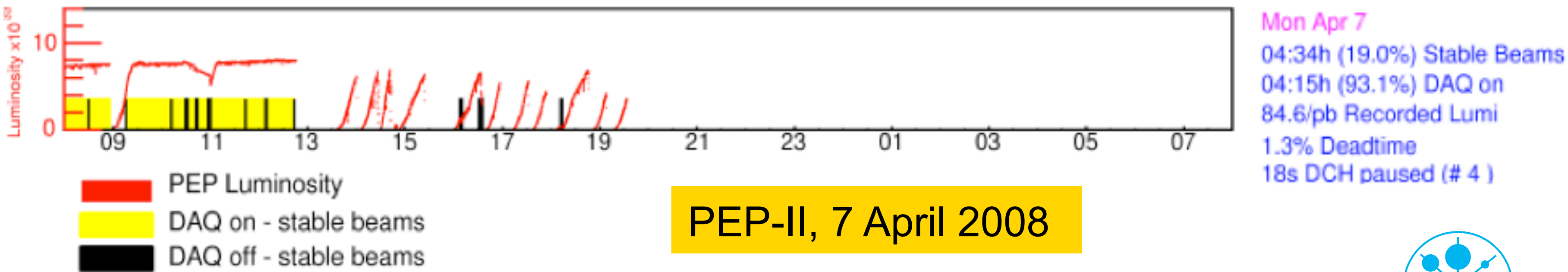
The last years have seen the end of several experiments



LEP, 2 November 2000

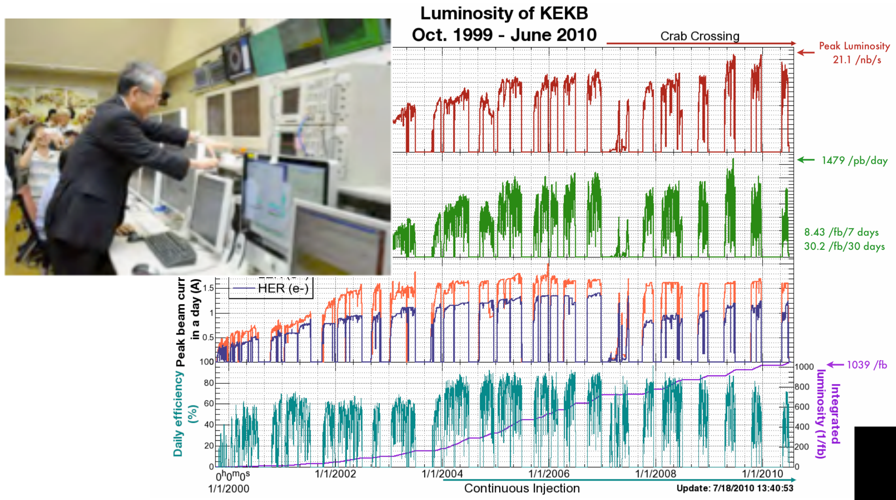


HERA, 30 June 2007



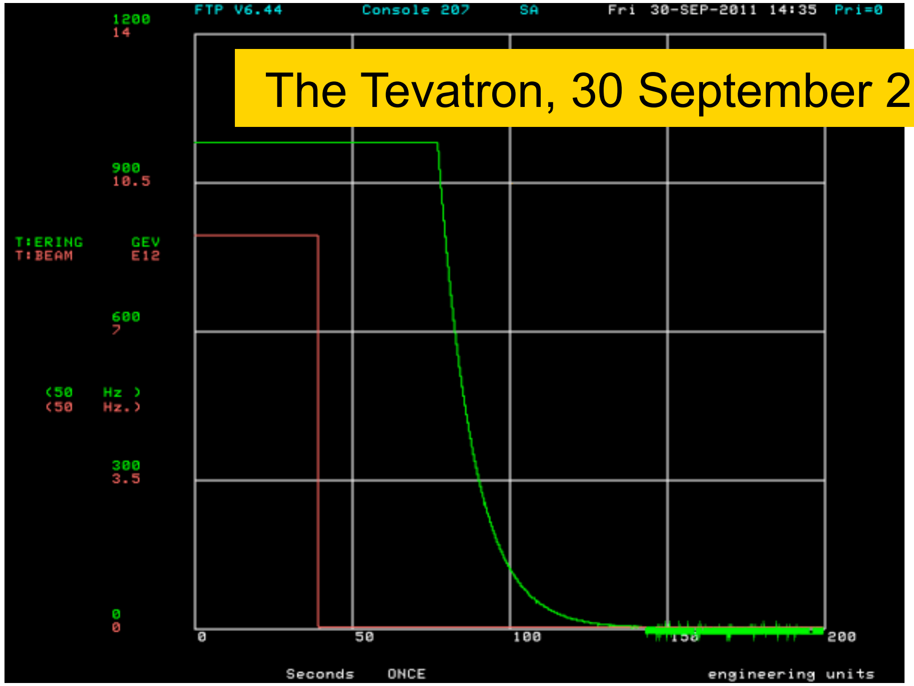
PEP-II, 7 April 2008

The last years have seen the end of several experiments



KEKB, 30 June 2010

The Tevatron, 30 September 2011



After the collisions have stopped

- Finish the analyses! But then what do you do with the data?
  - Until recently, there was no clear policy on this in the HEP community
  - It's possible that older HEP experiments have in fact simply lost the data
- Data preservation, including long term access, is generally not part of the planning, software design or budget of an experiment
  - So far, HEP data preservation initiatives have been in the main not planned by the original collaborations, but rather the effort a few knowledgeable people

- The conservation of tapes is not equivalent to data preservation!
  - *"We cannot ensure data is stored in file formats appropriate for long term preservation"*
  - *"The software for exploiting the data is under the control of the experiments"*
  - *"We are sure most of the data are not easily accessible!"*

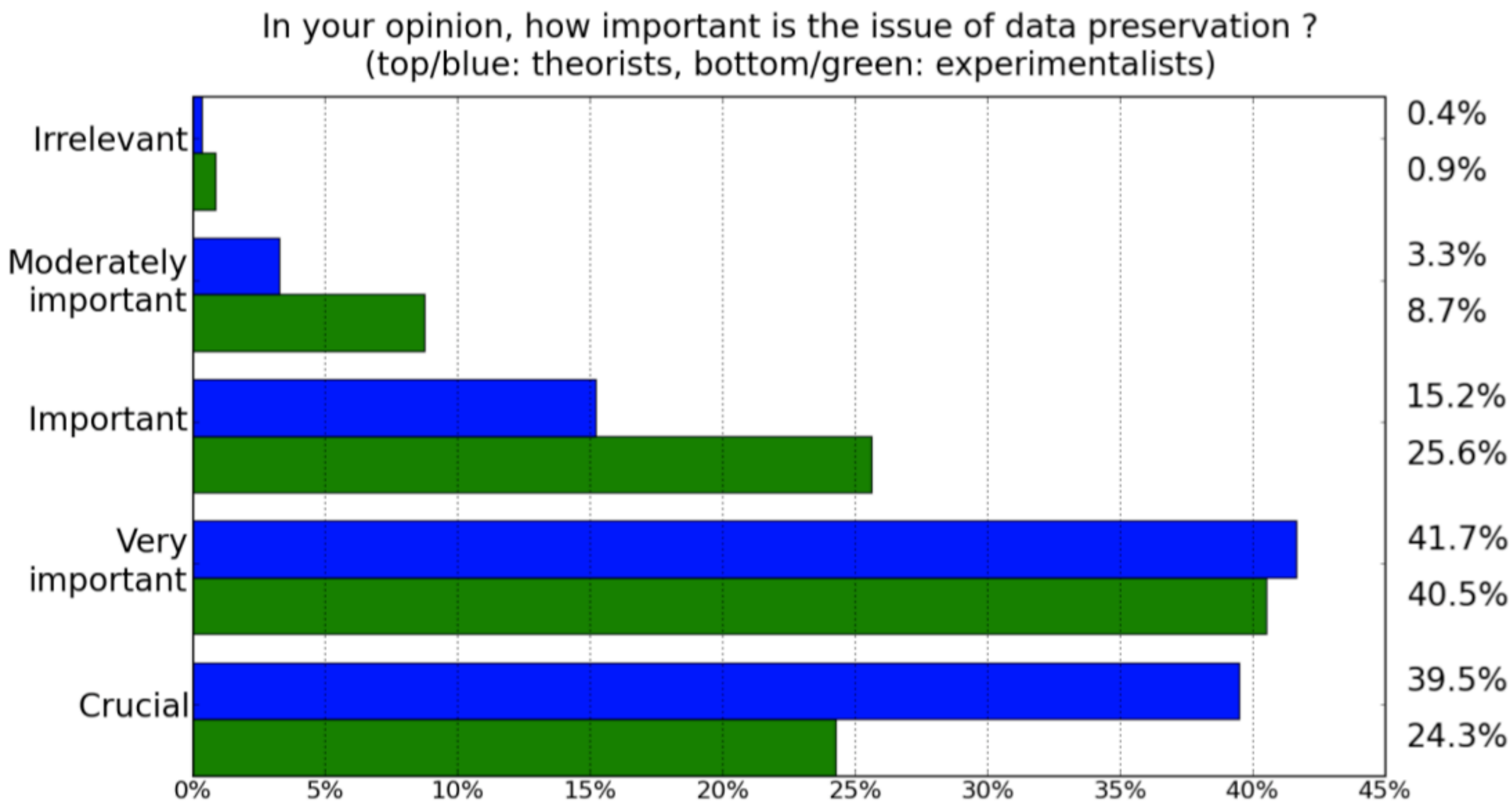
Initiatives in other fields

- 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

The difficulties of data preservation in HEP

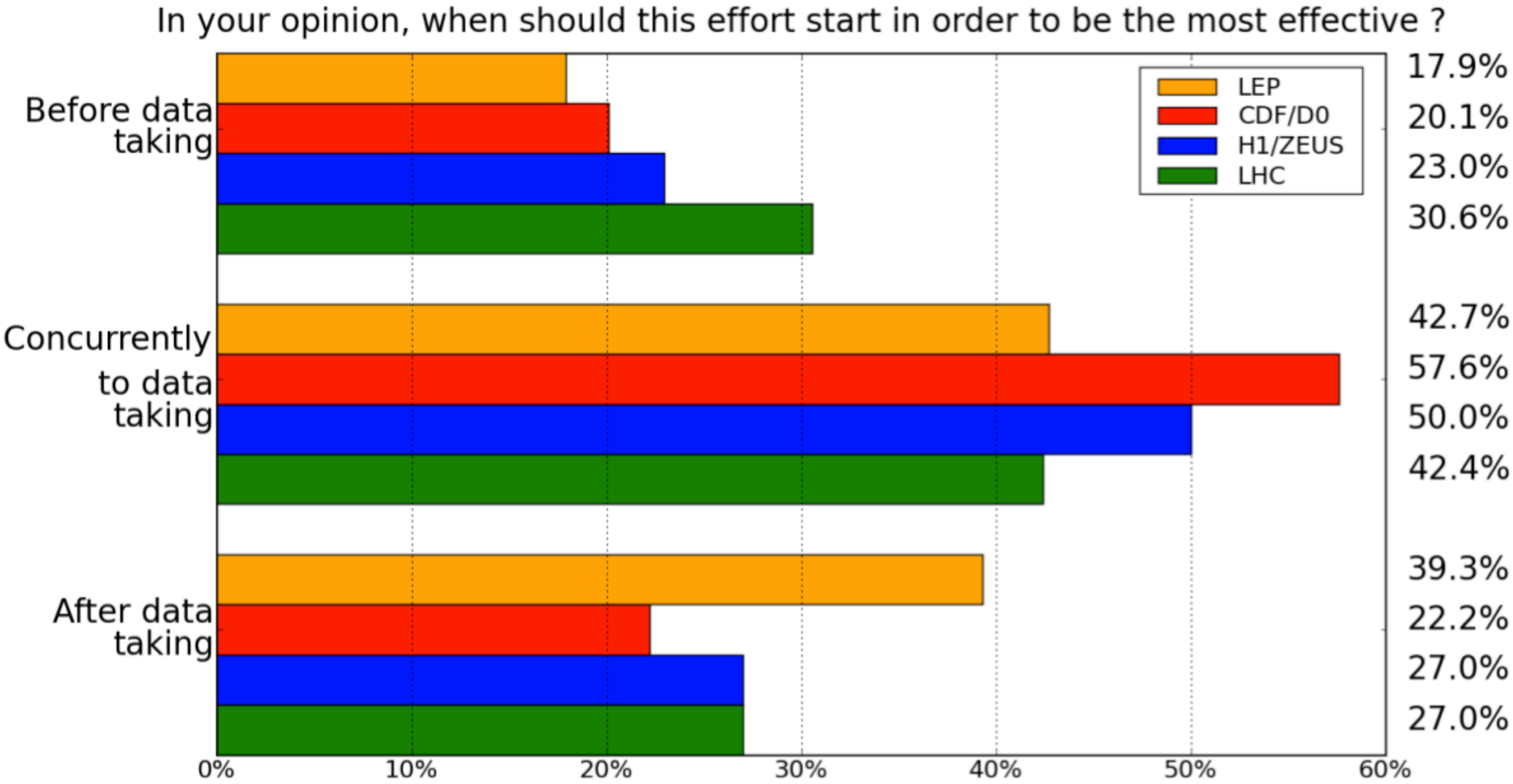
- Handling HEP data involves large scale traffic, storage and migration
  - The increasing scale of the distribution of HEP data can complicate the task
- Who is responsible? The experiments? The computing centres?
  - Problem of older, unreliable hardware: unreadable tapes after 2-3 years
  - The software for accessing the data is usually under the control of the experiments
- Key resources, both funding and person-power expertise, tend to decrease once the data taking stops
- And a rather key ingredient to all this is: *why do it?*
  - Can the relevant physics cases be made?
  - Who says we want to do this anyway?
  - Is the benefit of all this really worth the cost and effort?

Support for data preservation in the HEP community



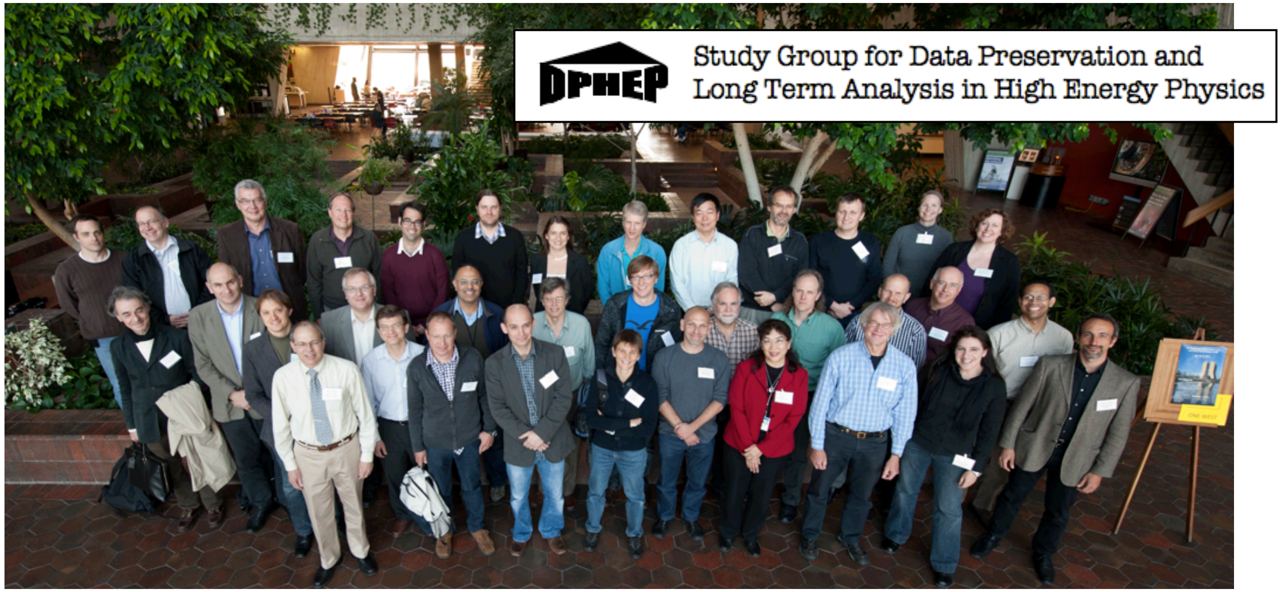
arXiv:0906.0485

Support for data preservation in the HEP community



arXiv:0906.0485

DPHEP: An international study group on data preservation



- First contacts established in September 2008
  - Group since grown to over 100 contact persons
  - 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

H1 | ZEUS | hermes | BELLE | CDF | DØ

ALICE | ATLAS | CMS | LHCb | DESY | KEK

Institute of High Energy Physics
Chinese Academy of Sciences

Fermilab

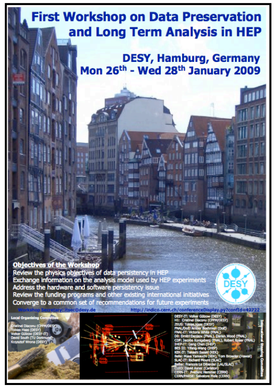
SLAC
NATIONAL ACCELERATOR LABORATORY

CERN | Jefferson Lab | Science & Technology Facilities Council

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA | INSPIRE | BROOKHAVEN NATIONAL LABORATORY

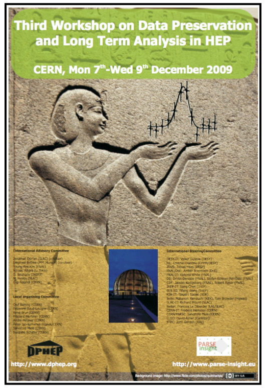
DPHEP: An international study group on data preservation

> 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

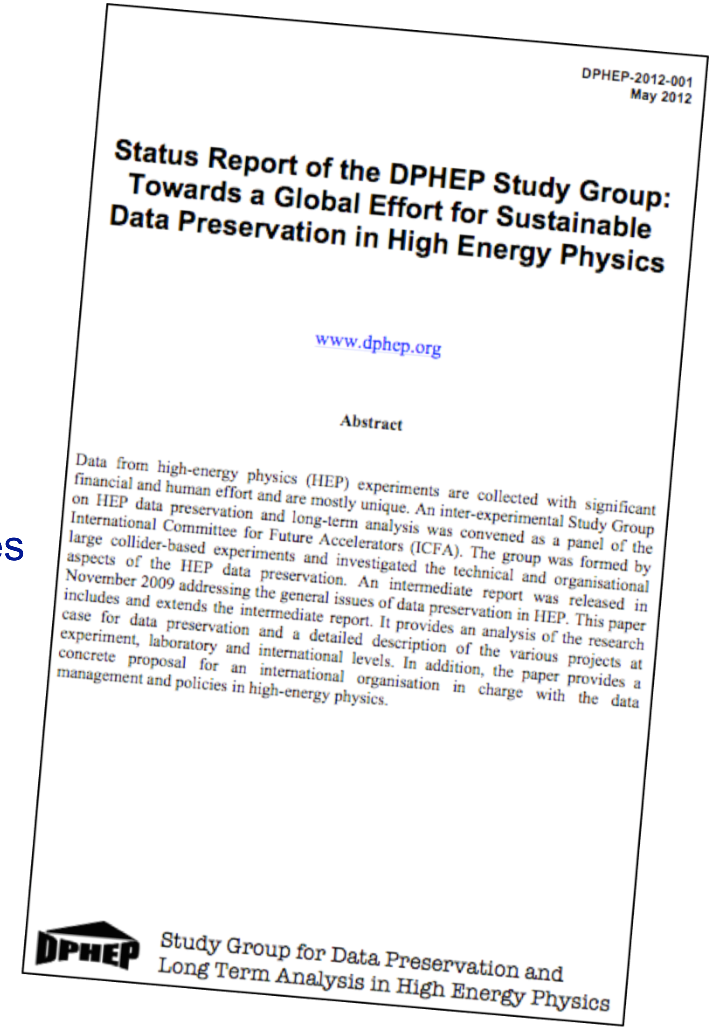
- Focus on four key areas of the study group: Physics Case for Data Preservation, Preservation Models, Technologies, Governance

arXiv:0912.0255

DESY

New DPHEP publication

- Available on arXiv since yesterday morning
- 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

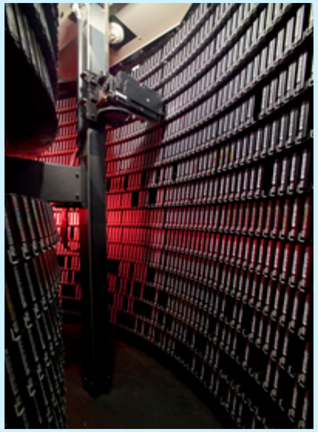
DESY

Building the physics case: Reasons to preserve HEP data

- Long term completion and extension of an existing physics program
  - Up to 10% of papers are finalised in the "archival mode"
  - Gain in scientific output of the experiments
- Cross-collaboration and combinations of physics results
  - During the active lifetime of similar experiments at one facility: LEP, HERA, TeVatron
  - And later across larger boundaries: Belle/BaBar, TeVatron/LHC
- Revisit old measurements or perform new ones
  - Access to newly developed techniques, comparisons to new theoretical models
  - Unique data sets available in terms of energy, initial states
- Use in scientific training, education, outreach

DESY

What is HEP data?



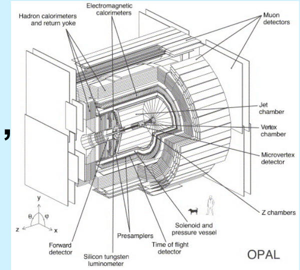
Digital information

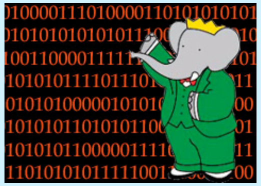
The data themselves, volume estimates for preservation data of the order of a few to 10 PB

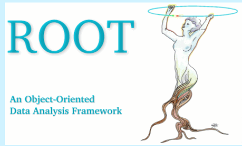
Other digital sources such as databases to also be considered

Software

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





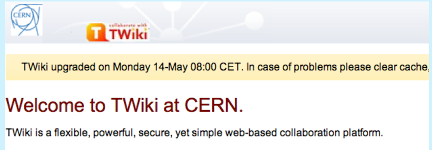


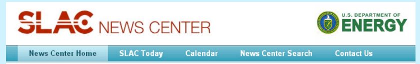
CERNLIB Access

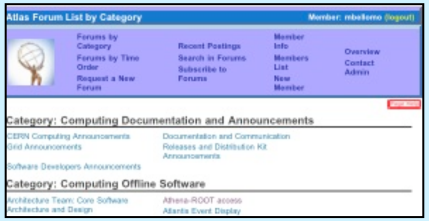
- Access to the CERN Program Library is free of charge to all HEP users worldwide.
- Non-HEP academic and not-for-profit organizations: 1KSF/year

Meta information

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







Publications

arXiv.org

iNSPIRE

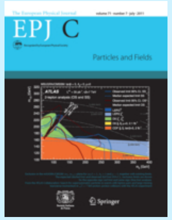
HEPDATA: REACTION DATA Database

Durham University

...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.

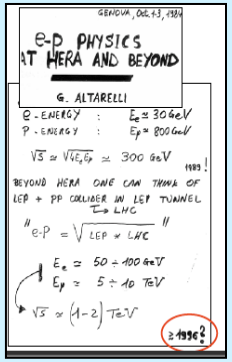
Journal of High Energy Physics

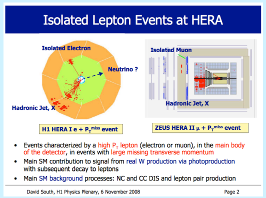
A refereed journal, written, run and distributed by electronic means

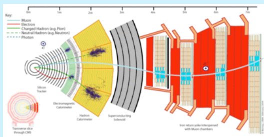


Documentation

Internal publications, notes, manuals, slides







Expertise and people

DESY

DPHEP models of HEP data preservation

	Preservation Model	Use Case
1	Provide additional documentation	Publication related info 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 the existing reconstruction
4	Preserve the reconstruction and simulation software as well as the basic level data	Retain the full potential of the experimental data

Increasing cost, complexity and benefits ↓

- These are the original definitions of DPHEP preservation levels from the 2009 publication
  - Still valid now, although interaction between the levels now better understood

DPHEP models of HEP data preservation

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	

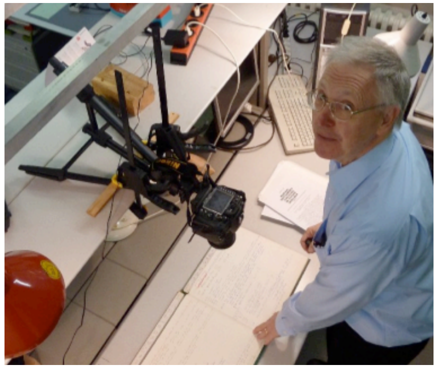
- These are the original definitions of DPHEP preservation levels from the 2009 publication
  - Still valid now, although interaction between the levels now better understood
- Originally idea was a progression, an inclusive level structure, but now seen as complementary initiatives
- Three levels representing three areas:
  - Documentation, Outreach and Technical Preservation Projects

Level 1: Documentation

> The organisation of documentation turns out to be quite a task

- Dedicated task forces set up by many of the experiments
- Much material from pre-web days, or using all kinds of web applications

> Non-digital: Cataloguing, organisation, scanning or photographing of appropriate of papers, notes, drawings, talks from pre-web days, detector schematics, blueprints, logbooks, ...



- New *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, ...

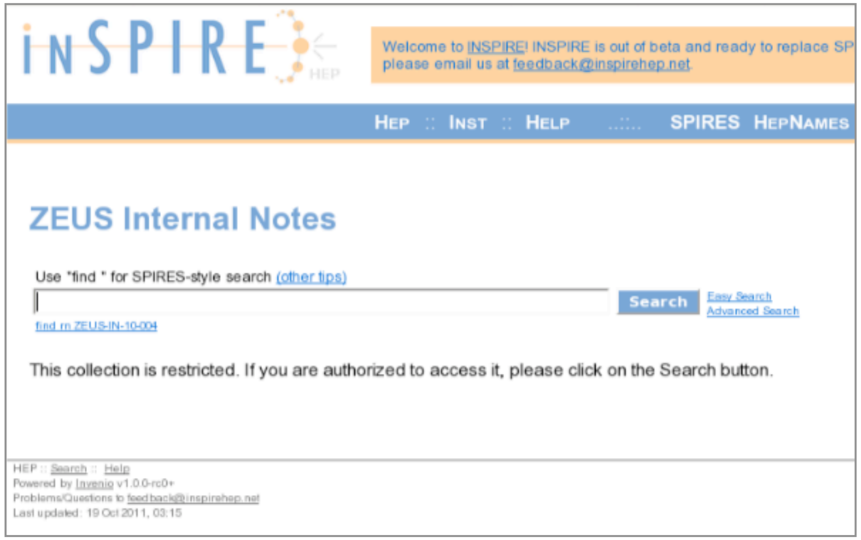
- Replacement of old web servers by VMs, hosted by the computer centres
- Replacement of old pages to newer technologies such as wikis (use of (T)wikis much more prevalent in the LHC era)
- Use of external services for hosting collaboration material

DESY

Documentation projects with INSPIRE

> Internal notes from all HERA experiments now available on INSPIRE

- Experiments no longer need to provide dedicated hardware for such things
- Password protected now, simple to make publicly available in the future



Documentation projects with INSPIRE

- Internal notes from all HERA experiments now available on INSPIRE
  - Experiments no longer need to provide dedicated hardware for such things
  - Password protected now, simple to make publicly available in the future

Documentation projects with INSPIRE

- Internal notes from all HERA experiments now available on INSPIRE
  - Experiments no longer need to provide dedicated hardware for such things
  - Password protected now, simple to make publicly available in the future

Documentation projects with INSPIRE

- Internal notes from all HERA experiments now available on INSPIRE
  - Experiments no longer need to provide dedicated hardware for such things
  - 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 experiments working with INSPIRE, including CDF, D0 as well as BaBar

HEP outreach initiatives

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

NETZWERK TEILCHENWELT ········ QUARKS, ELEKTRONEN & CO. ····

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

10 QuarkNet: The science connection you've been waiting for!

IPPOG International Particle Physics Outreach Group

INTERNATIONAL MASTERCLASSES hands on particle physics

Science Hack Day: Increasing the access to LHC data

http://cms.web.cern.ch/news/cms-public-data-activity-scoops-prize-nairobi

CMS public data activity scoops prize in Nairobi

+1 4 Tweet 11 Like 27

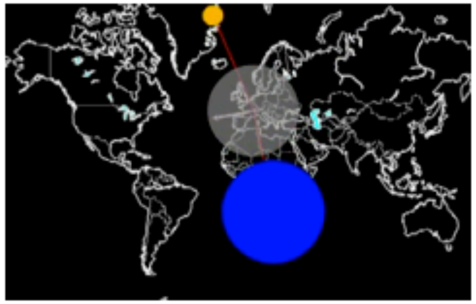
An application using real event data from CMS has won "Best Science" prize in a public "Science Hack Day" held in Nairobi between 13th and 15th April 2012. Science Hack days bring together a wide range of enthusiastic members of the public to create something completely new using existing scientific systems or data.

CMS use of public data in a "Science Hack" event in Nairobi. Photo credit: Matt Biddulph, via Flickr

The winning application visualized real CMS di-muon events from the 2011 LHC run, which are made public for use in various educational programmes, such as the IPPOG Masterclasses, Quarknet and I2U2. The application showed an animation of muons produced in CMS superimposed on a map of the world, showing where they would go if they were to continue without stopping (which they don't in reality).

Other prizes were awarded to Leah Atieno, a 15-year-old high-school student, for a voice-controlled walking robot and Denis Munene for a crowd-mapping platform to help promote the fight against malaria.

The Nairobi event, involving 240 developers, is part of broader series of Science Hack Day events. CMS data previously featured in another very successful event in San Francisco.

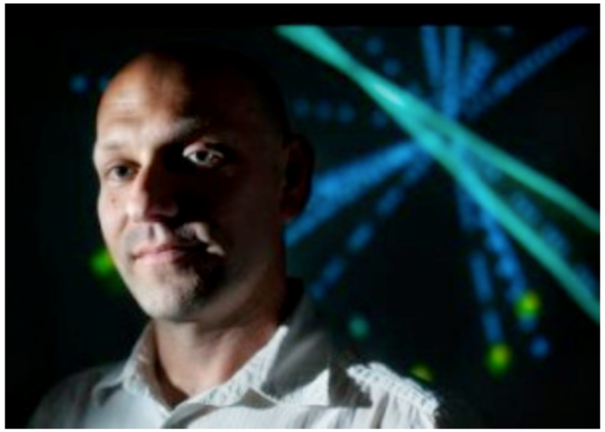


Application developed to visualise where muons from CMS would go if they continued forever

News article by Gythan Munga, HumanIpo
See photos of the event
Youtube film
Link to more Science hack events
2012-04-20, by Lucas Taylor

Level 2: Simplified formats for outreach

- Within DPHEP and the member collaborations there are generic ideas, such as common formats and user interfaces
  - In terms formats, much can be learned from other fields such as astrophysics or life sciences



- Such outreach formats in HEP are typically based on ROOT, containing particle 4-vectors and simple event information
  - Composite-particle reconstruction, finding signals
  - Initiatives in place at BaBar, Belle and LHC experiments

- A multi-experimental project is desirable, coordinated via DPHEP, and based in several locations (CERN, FNAL, DESY..)
  - To include associated tutorials linked to preserved HEP data from several sources

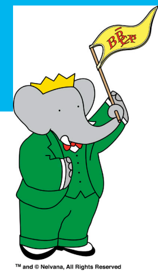
Technical Projects: DPHEP preservation levels 3 and 4

- This is really the main focus of the data preservation effort
  - Level 3: Access to analysis level data, MC and the analysis level software
  - Level 4: Access to reconstruction and simulation software, retain the full capability
- Deciding on level 3 or 4 depends on the scope of your project
  - What do you want to be able to do in *N* years time?
  - Only level 4 gives full flexibility, but this also means not relying on frozen executables and binaries but rather retaining the ability to recompile: more work

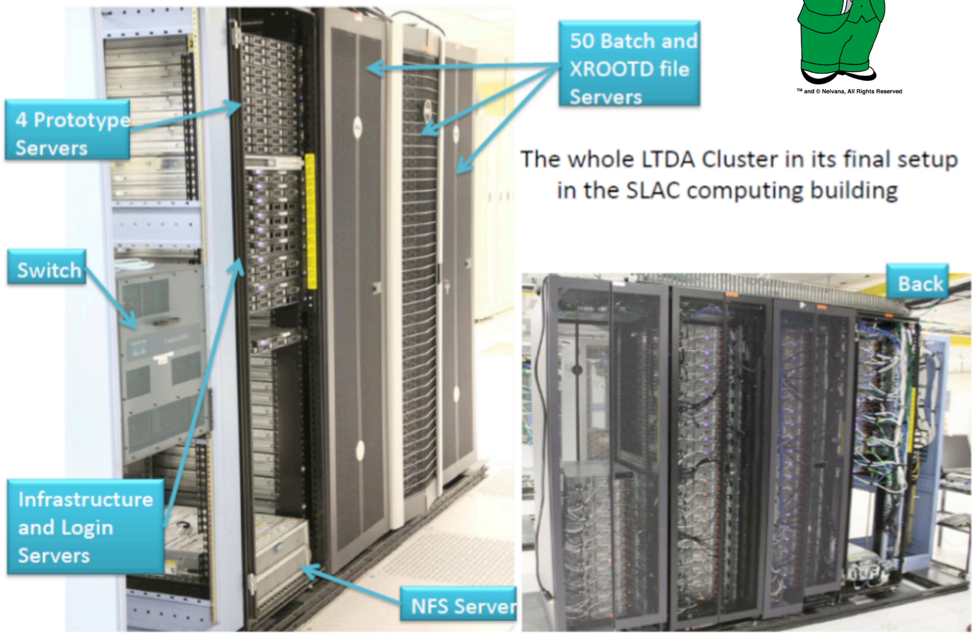
The majority of DPHEP experiments aim for DPHEP level 4 preservation

- Remember: it's not about the data, but about still being able analyse it
  - Either keep your current environment alive as long as possible
  - Or adapt and validate your code to future changes as they happen
- Two complimentary approaches taken at SLAC and DESY
  - Both employing virtualisation techniques, but in rather different ways

The BaBar Long Term Data Access archival system

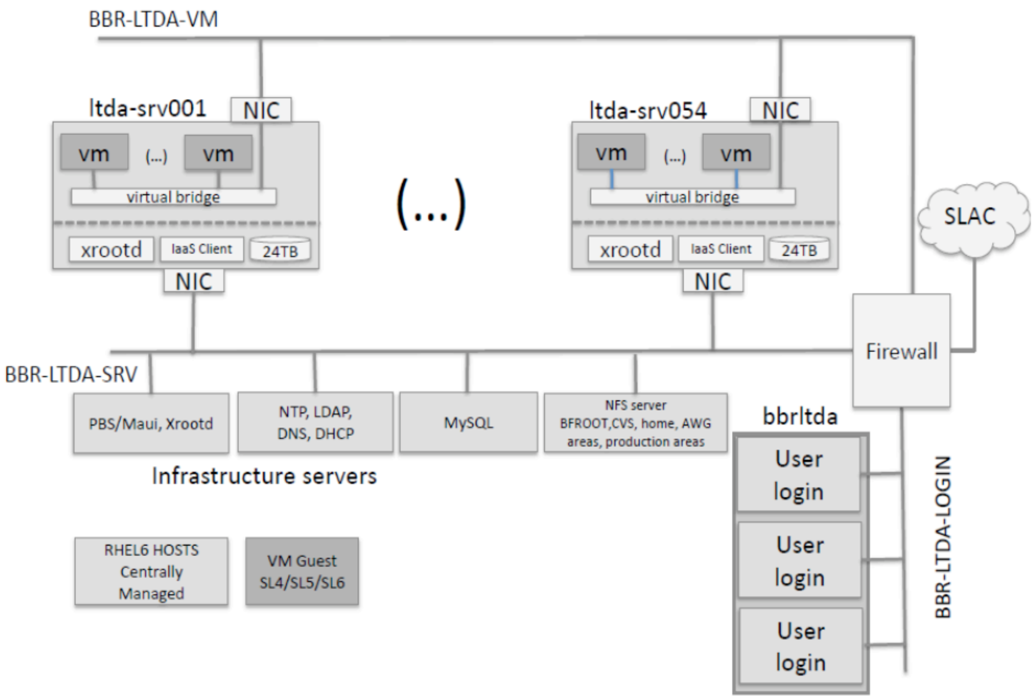


- New BaBar system installed for analysis until at least 2018
- Isolated from SLAC, and uses virtualisation techniques to preserve an existing, stable and validated platform
- Complete data storage and user environment in one system



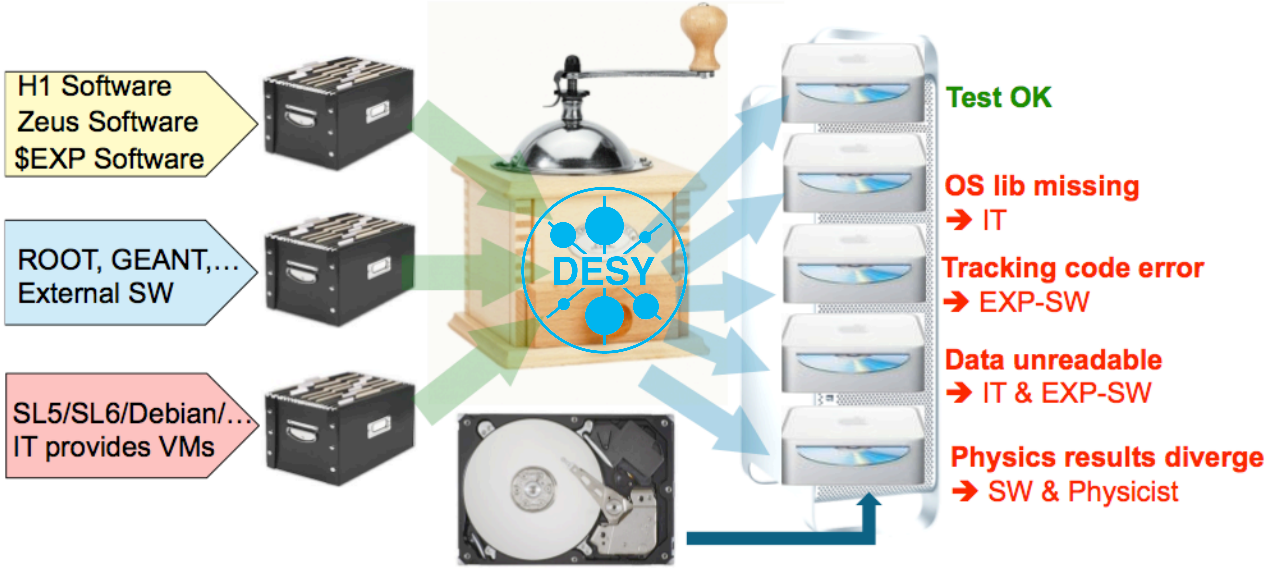
- Required large scale investment: 54 R510 machines, primarily for data storage, as well as 18 other dedicated servers
  - Resources taken into account in experiment's funding model during analysis phase!
- From the user's perspective, very similar to existing BaBar infrastructure

The BaBar Long Term Data Access archival system



- Crucial part of design is to allow frozen, older platforms to run in a secure computing environment
- *Naïve* virtualisation strategy, not enough
  - Cannot support an OS *forever*
  - Security of system under threat using old versions
- Achieved by clear network separation via firewalls of part storing the data (more modern OS) and part running analysis (the desired older OS)
- Other BaBar infrastructure not included in VMs is taken from common NFS
- More than 20 analyses now using the LTDA system as well as simulation

The `sp-system` at DESY



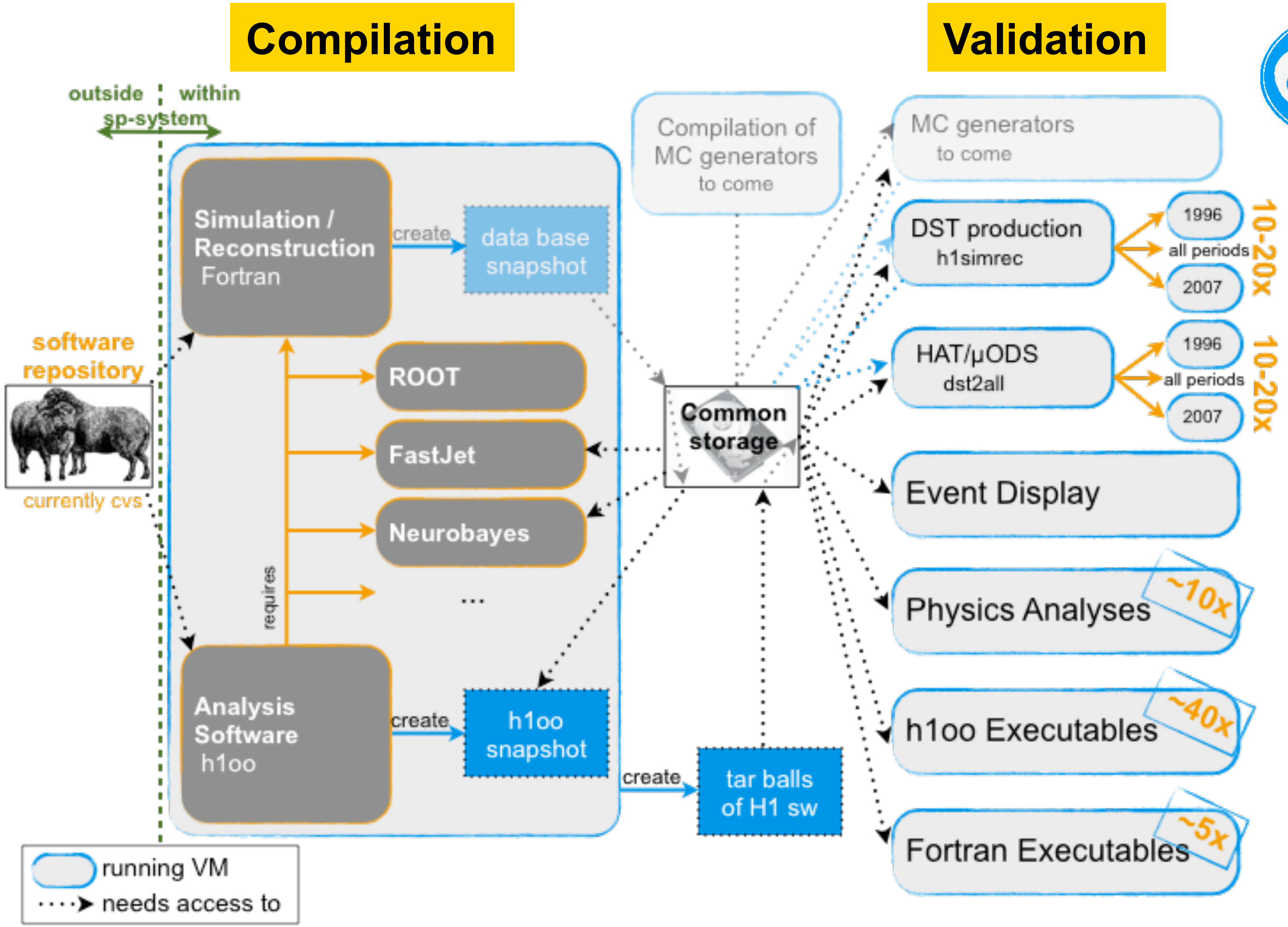
> Automated validation system to facilitate future software and OS transitions

- Utilisation of virtual machines offers flexibility: OS and software configuration is chosen by experiment controlled parameter file
- Successfully validated recipe to be deployed on future resource, e.g. Grid or IT cluster
- Pilot project at CHEP 2010, full implementation now installed at DESY

> Essential to have a robust definition of a complete set of experimental tests

- Nature and number dependent on desired preservation level

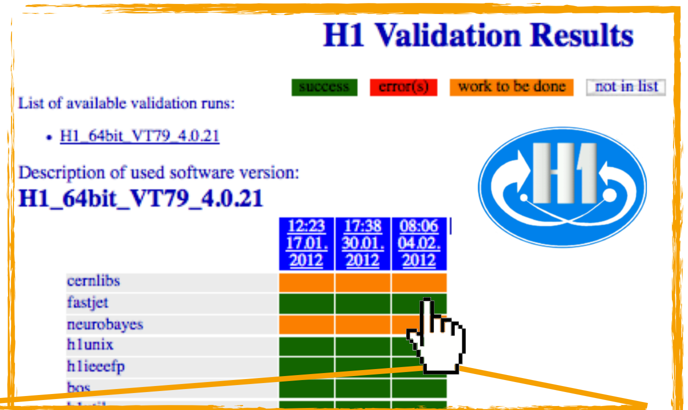
Example structure of the experimental tests: H1 (Level 4)

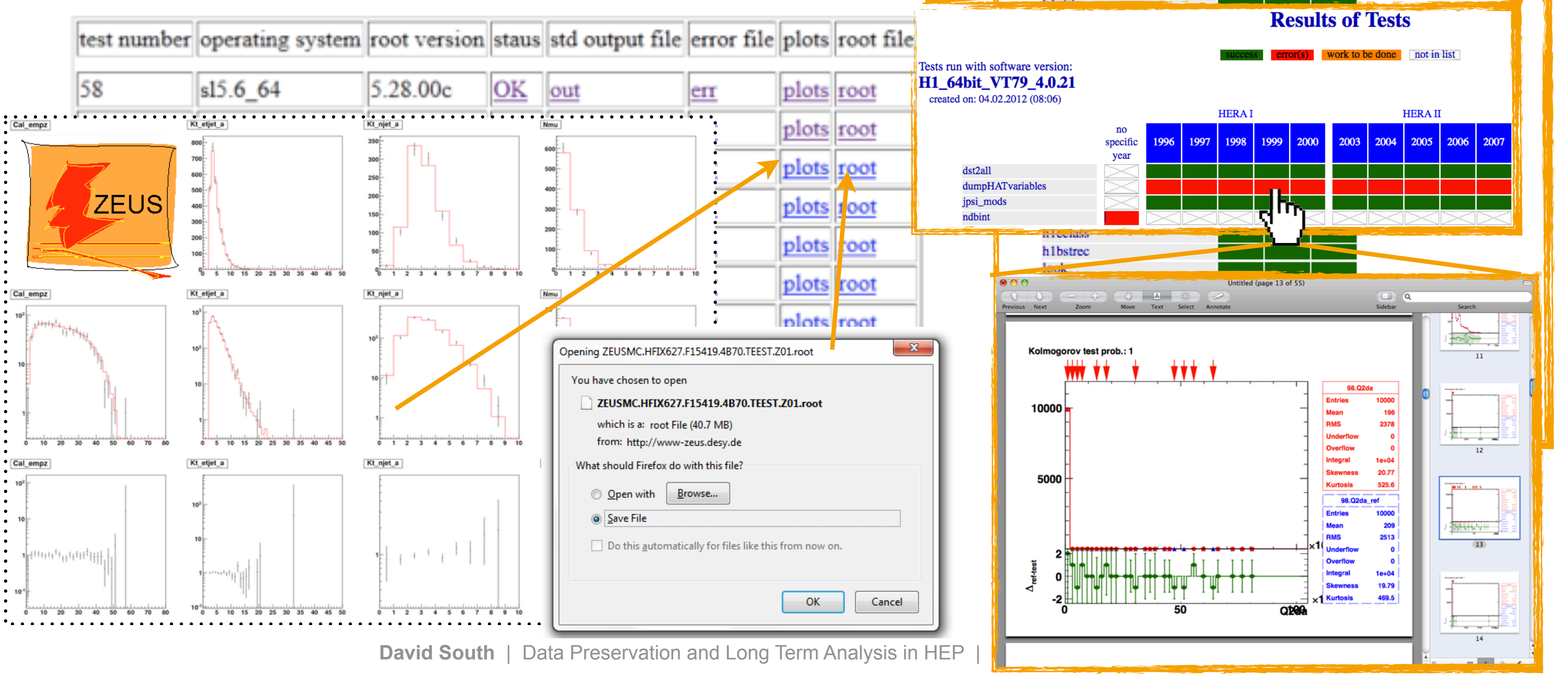


Including compilation of individual packages: about 250 tests planned by H1

Digesting the validation results

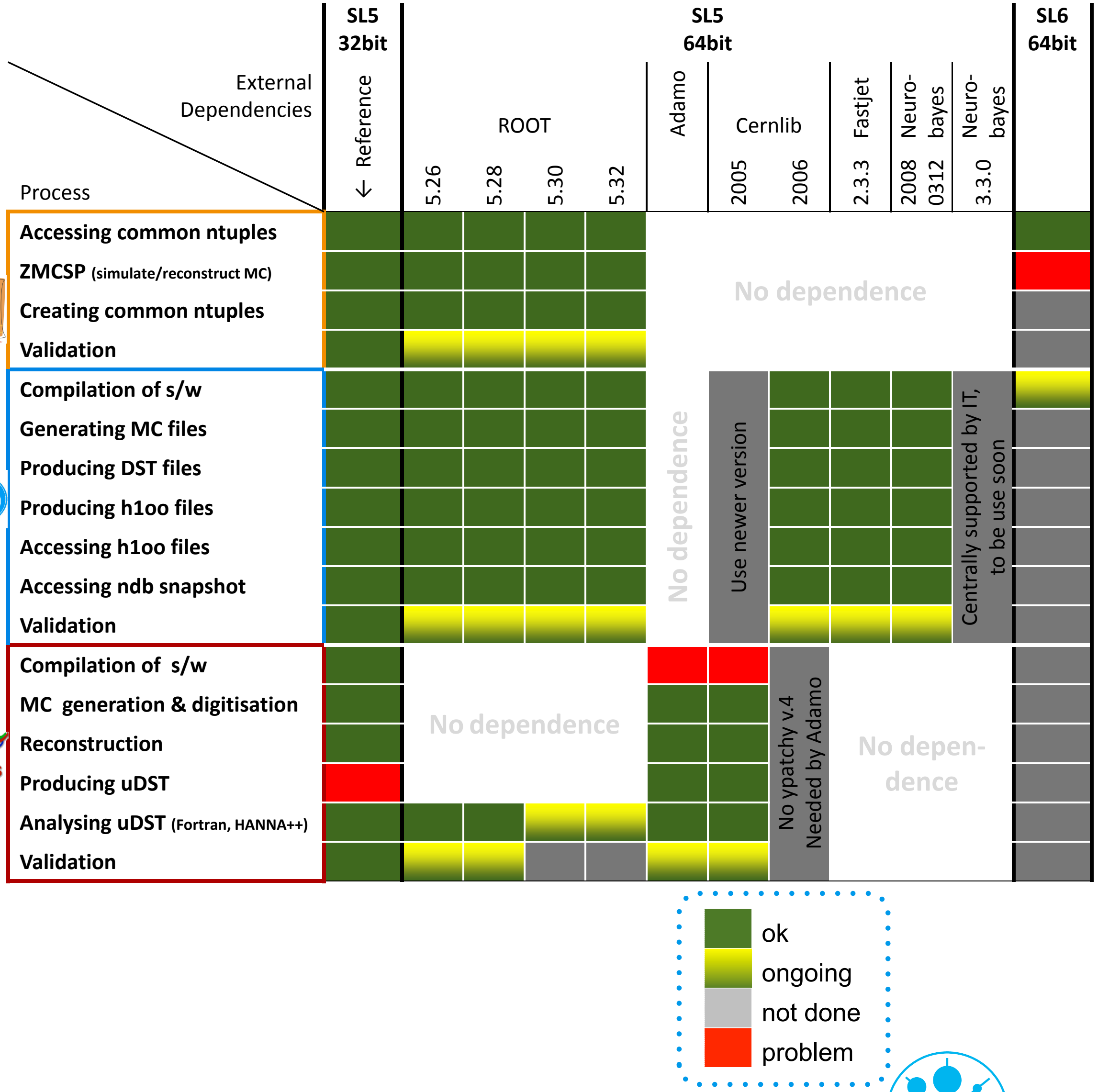
- Display the results of the validation in a comprehensible way: web based interface
- The test determines the nature of the results
  - Could be simple yes/no, plots, ROOT files, text-files with keywords or length, ...





Current status of the HERA experiments software

- Common baseline of SLD5 / 32-bit achieved in 2011 by all experiments
  - Validation of 64-bit systems is a major step towards migrations to future OS
  - The system has already been useful in detecting problem visible only in newer software
- Note that this system does not concern data integrity
  - The investigation into data archival options is underway

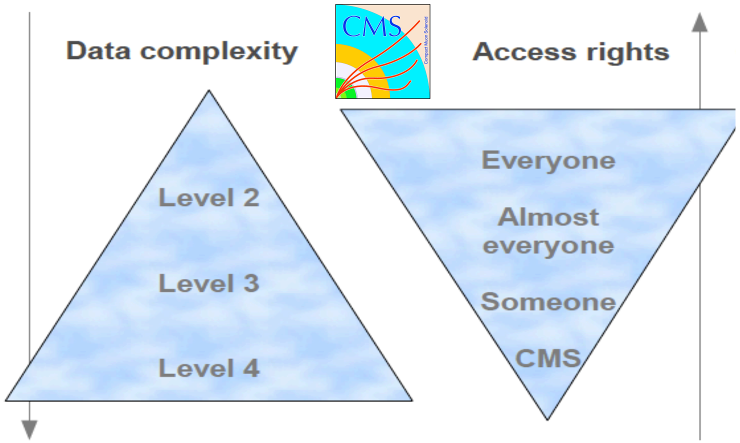


Summary of information from the (pre-LHC) experiments

	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 Preservation at the LHC

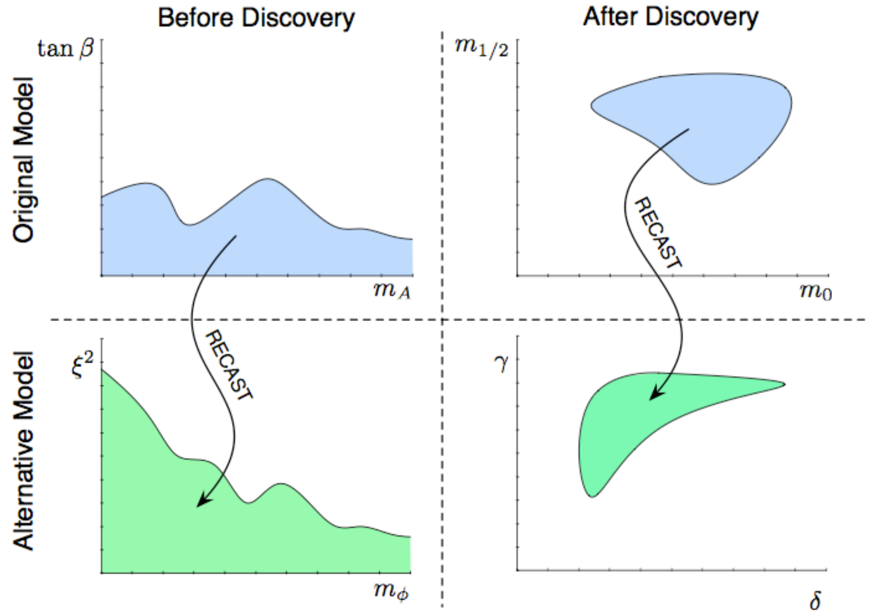
- Reflection just started in ATLAS, ALICE, CMS, LHCb
  - Common understanding that starting earlier will consolidate the long term future
  - Strong wish to develop a common policy at CERN and within DPHEP
  - Specific cases already identified: Lower energy data, trigger configurations, pile up.
- In terms of documentation, LHC experiments are in good shape
  - The electronic era: Twikis, accompanying notes, plans for extended use of INSPIRE
- Outreach projects and open access explored
- The distributed data model eases the worry of data loss
  - Although as previously stated: no successful preservation without associated long-term access
  - No concrete plans yet, but level 4 seen as the ultimate objective



A multi-preservation level tool: RECAST

arXiv:1010.2506

- Framework developed to extend impact of existing analyses
- Complementary approach of analysis archival, encapsulating the full event selection, data, backgrounds, systematics
- Idea is to ***recast*** existing physics search results to constrain alternate model scenarios
  - Complete information from original analysis contained in the data
  - Already performed on ALEPH data, LHC experiments investigating



- RECAST does not fit directly into the DPHEP preservation levels
  - Levels 3 and 4 are in the back-end, containing the complete archived analyses
  - However, only the selection in the publication is preserved, it could also be described as additional information, more like level 1

DESY

Summing up: What has been achieved so far?

- The DPHEP Study Group represents the first large scale effort to address data preservation in the field of high energy physics

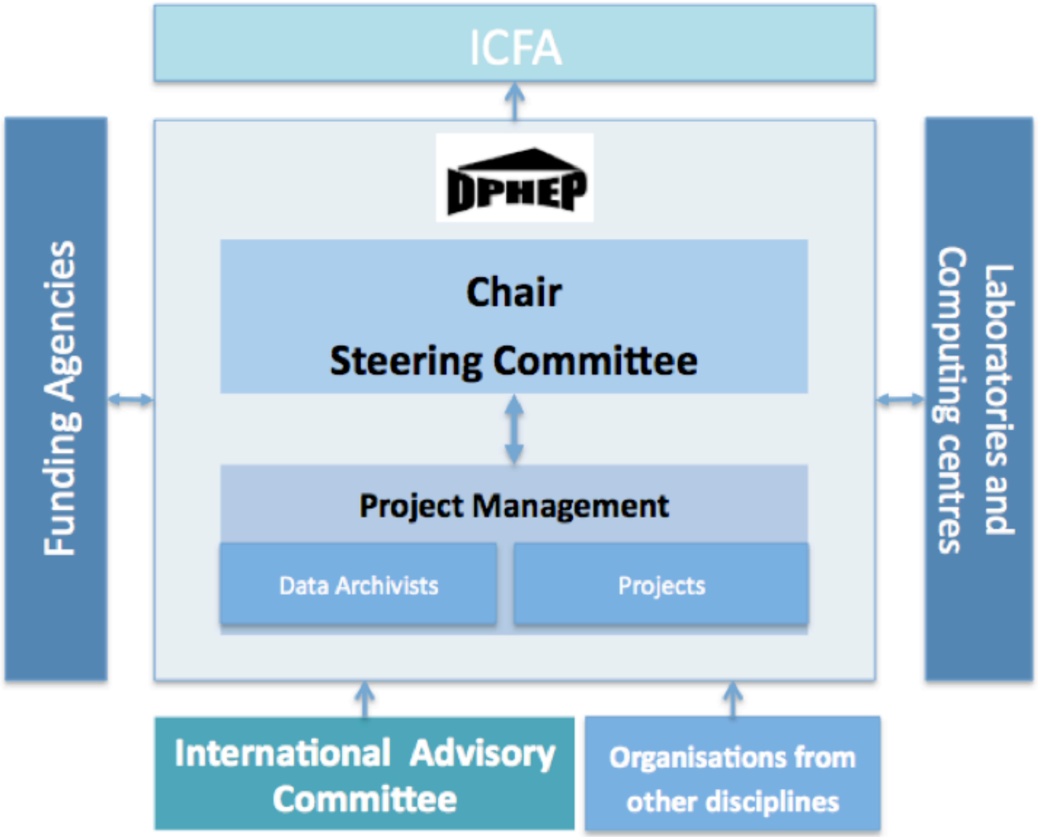
- The initial make up of the group was driven by the coincidence of the end of data taking at several large colliders, but had grown to include others including the LHC experiments

- The activity of the group over the last three years has led to an increased understanding of the relevant issues, enabling problems to be addressed, recommendations to be formulated and multi-experiment projects to begin

- To gain the most benefit from the work done so far, a transition from the current Study Group structure to a new, full time DPHEP Organisation

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



Conclusion and outlook

- The DPHEP Study Group has established itself in the HEP community and has reached a milestone in the publication of the latest report, which contains a comprehensive appraisal of data preservation in HEP
- The group will continue to investigate and take action in areas of coordination, preservation standards and technologies, as well as expanding the experimental reach and inter-disciplinary cooperation

- In order to do this a transition of the Study Group to the more structured DPHEP Organisation should occur
- It is foreseen that funding must come from different sources, in particular for common DPHEP enterprises or positions

- Take a look at the new DPHEP publication for more details

arXiv:1205.4667

DPHEP @ CHEP 2012

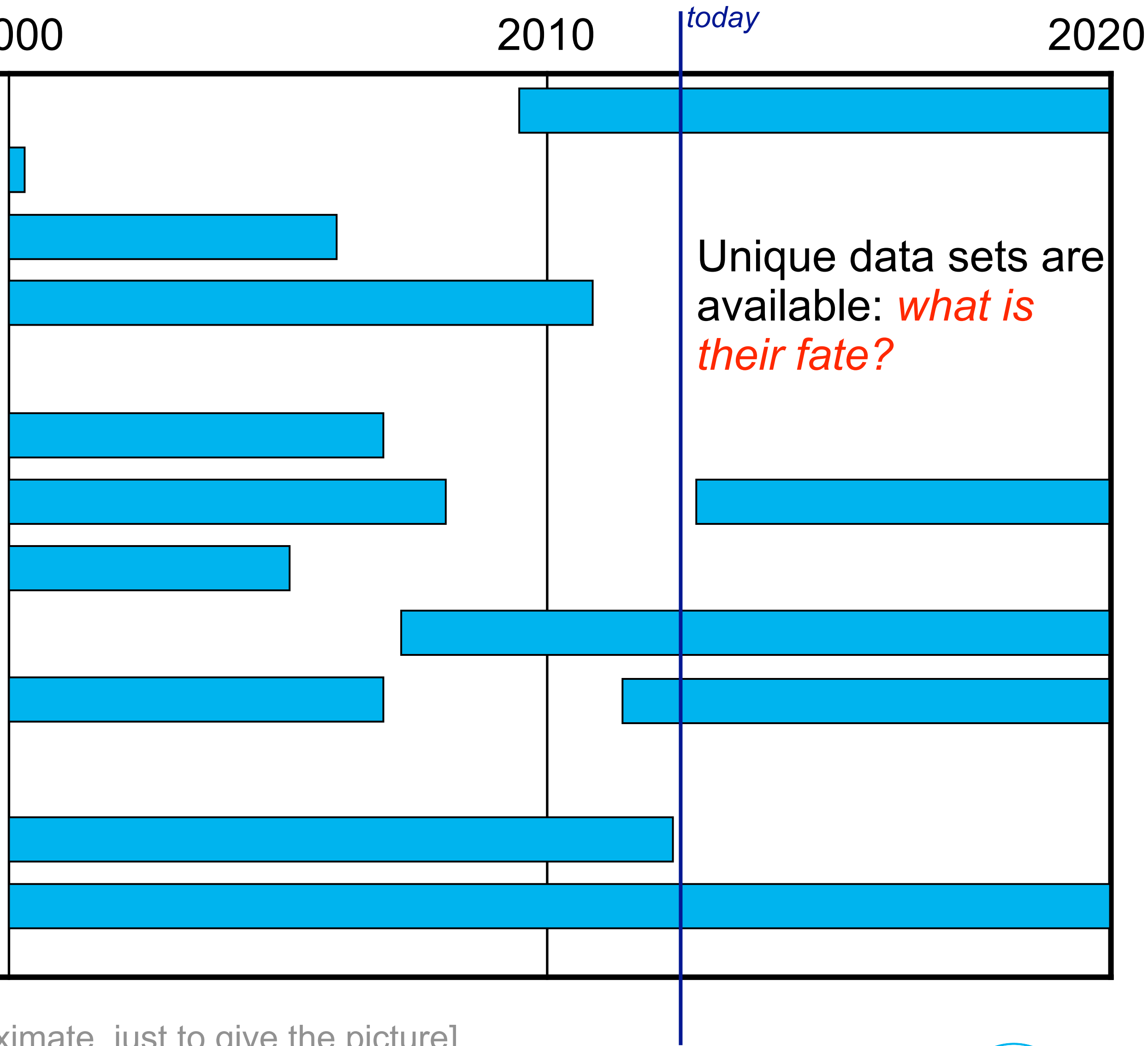
- “The workflow of LHC papers” (including INSPIRE), M. Ludmila, *Mon 2:45 PM*
- “Preparing expts' software for long term analysis and data preservation”, Y. Kemp, *Tues 2:20 PM*
- Data preservation posters by BaBar, H1, HERMES, ZEUS, as well as RECAST, *Thurs PM*

- **DPHEP session at CHEP,** Room 808, Thursday from 1:30 PM
  - Featuring reports from the experiments including a dedicated LHC session
  - https://indico.cern.ch/conferenceDisplay.py?confId=171962

EXTRAS

HEP experimental programmes ± 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]

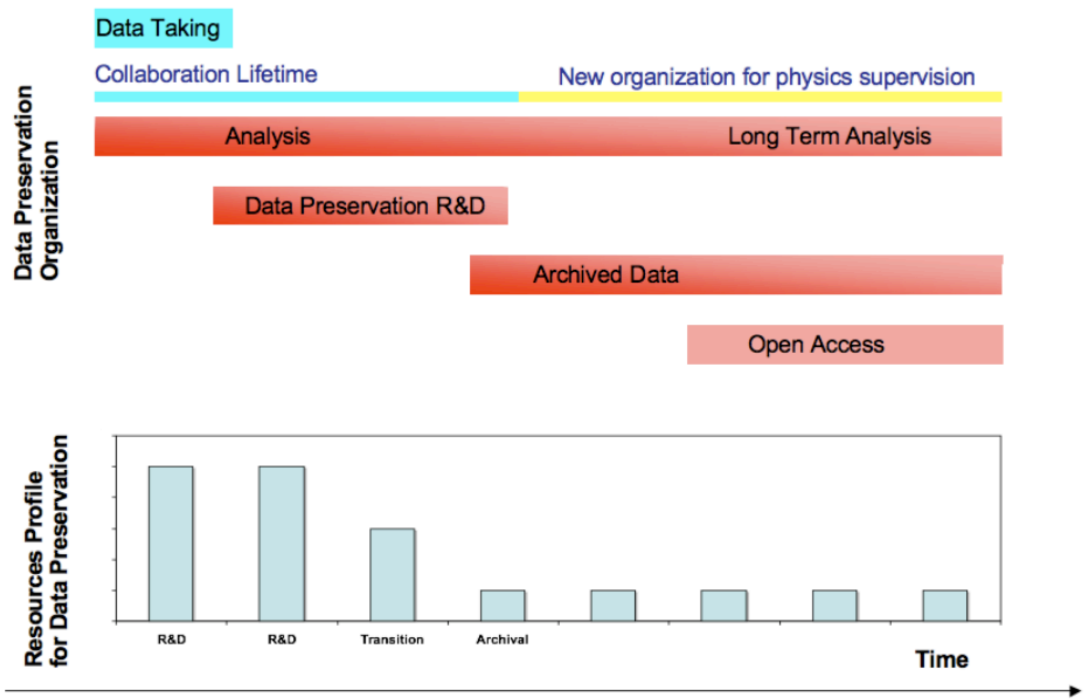
Transition scenario and resources at the experimental level

- Planning the transition to a long term analysis model

- R&D phase needed to develop the projects for the transition

- Long term custodianship of the physics data

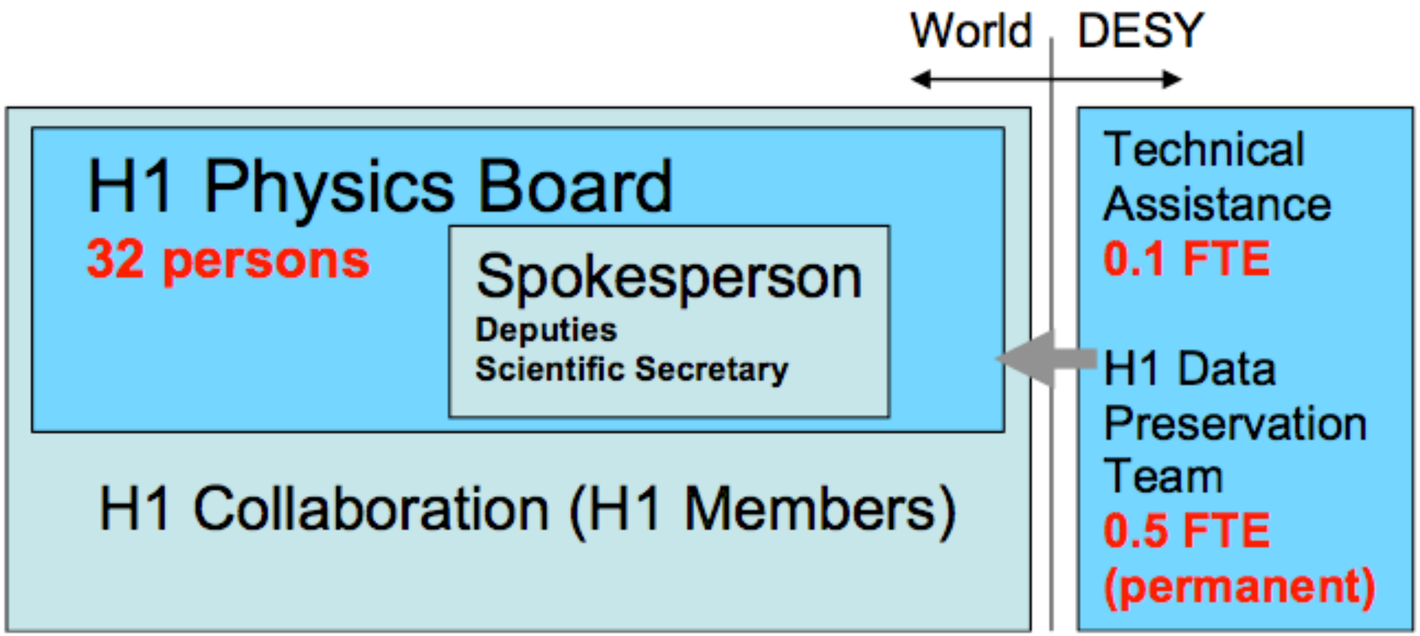
- Resources / experiment
  - Typically a surge of 2-3 FTEs for 2-3 years, followed by steady 0.5-1.0 FTE per experiment/lab
  - This should be compared to 300-500 FTEs for many years per experiment!



Cost estimates represent typically much less than 1% of the original investment

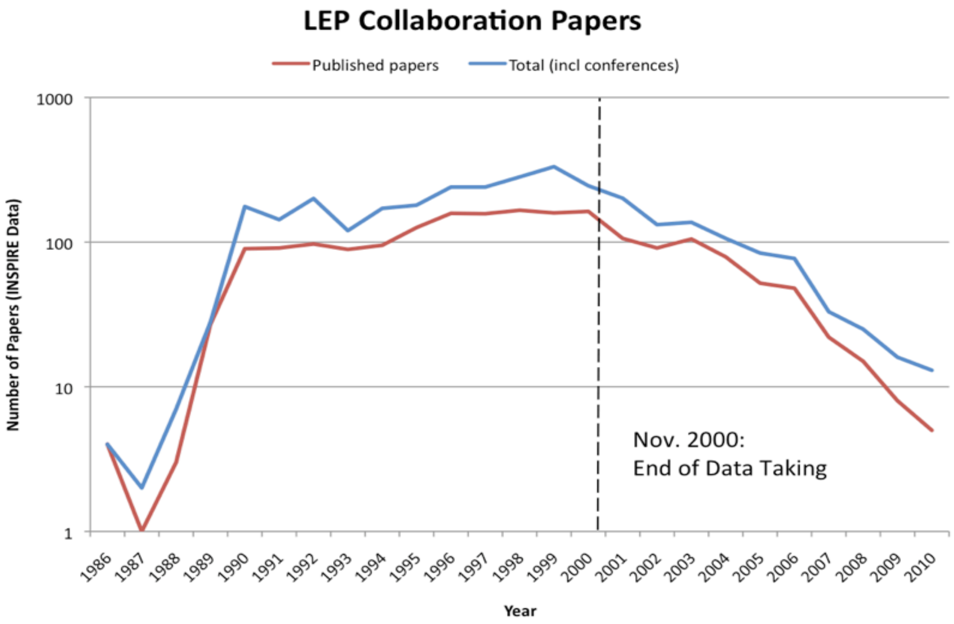
Scientific return: O(10%) in number of publications

Collaboration transitions

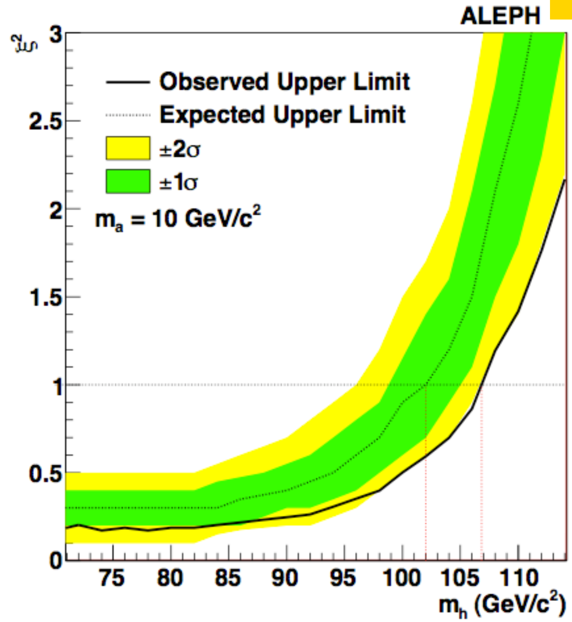


- Future structure collaborations should also be considered by experiments
  - Experimental organisation risks being left in an undefined state
  - Transition should also be planned in advance of the projected end date
  - Of particular note are authorship issues
  - Important when considering the future use of data and open access

Long term completion of the physics programme

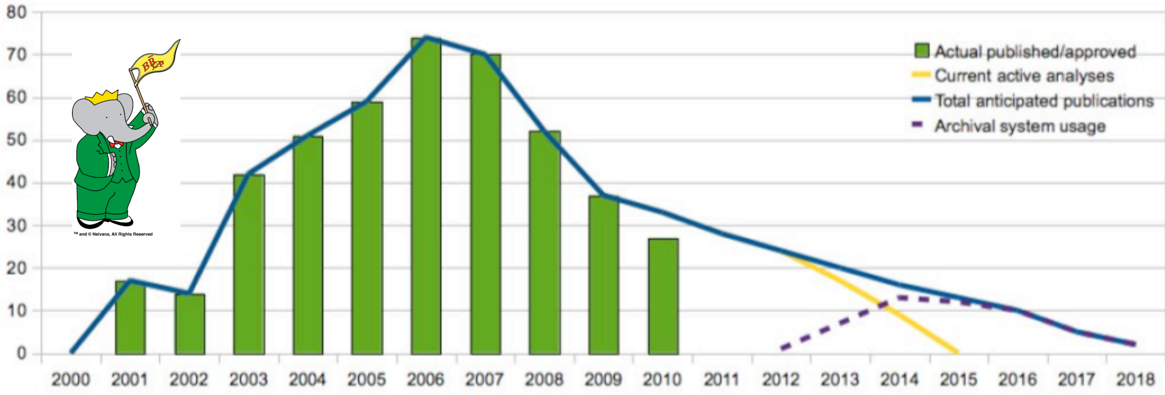


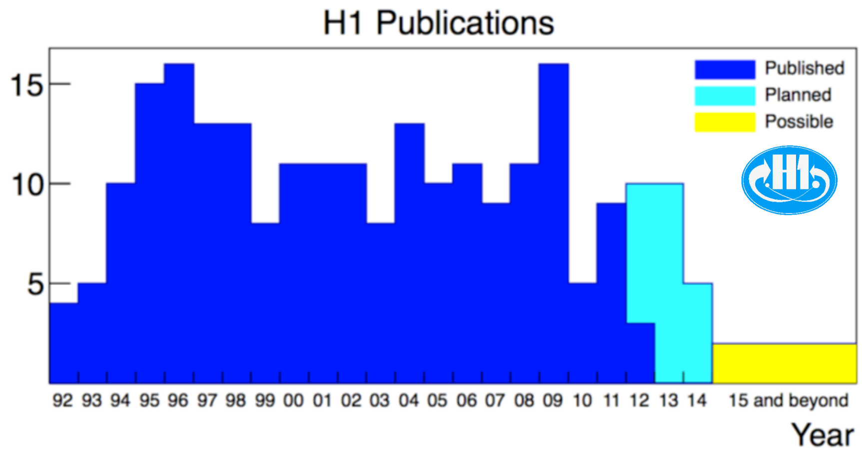
arXiv:1003.0705

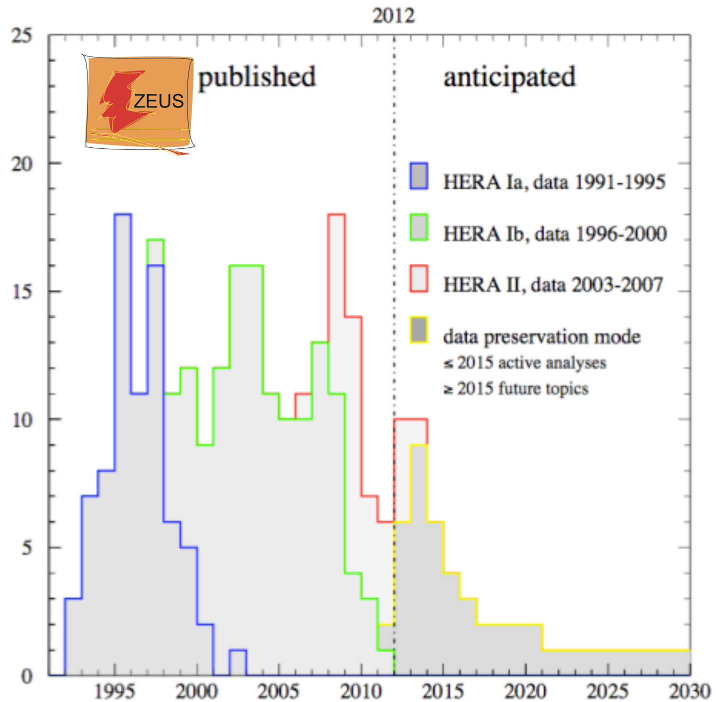


- The publication tail of LEP is long, with new papers still appearing
- Well over 300 papers produced since the end of collisions in 2000
- Recent analysis of LEP data gave unique limits on a novel Higgs model
- Similar, if not longer publication tails predicted by the BaBar, H1 and ZEUS experiments, after taking into consideration the plans for data preservation

Long term completion of the physics programme



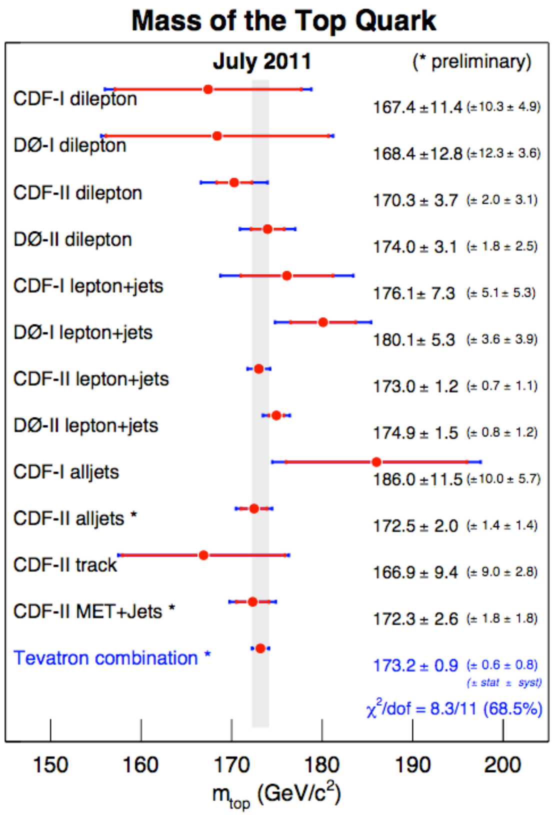


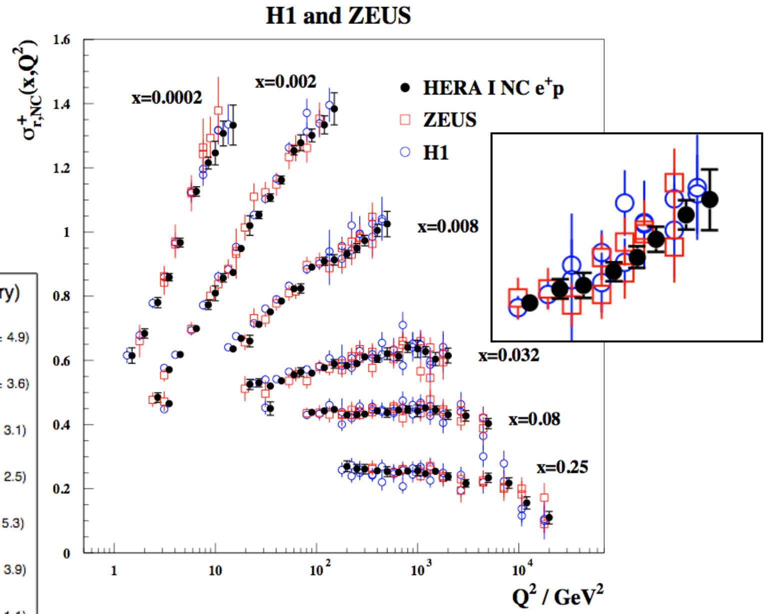


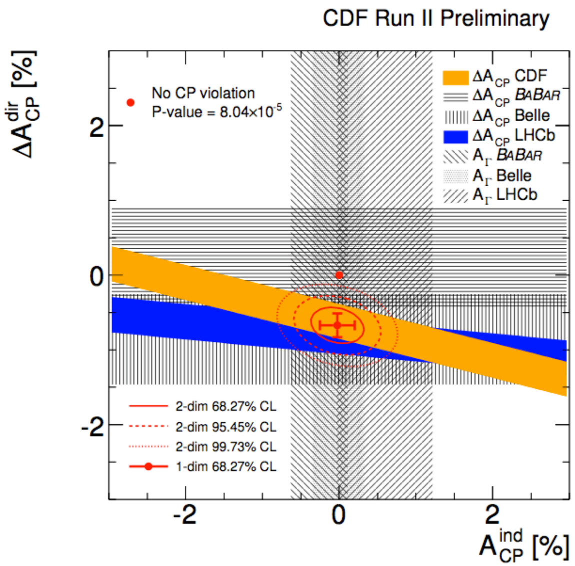
> Similar publication tails predicted by the BaBar, H1 and ZEUS experiments, taking into consideration the plans for data preservation

Cross-collaboration combinations of physics results

- Combination of data from multiple experiments to produce new scientific results
  - Improved precision and increased sensitivity
- Comparison of experimental results
  - Complimentary information from different physics
  - Verification of experimental observations

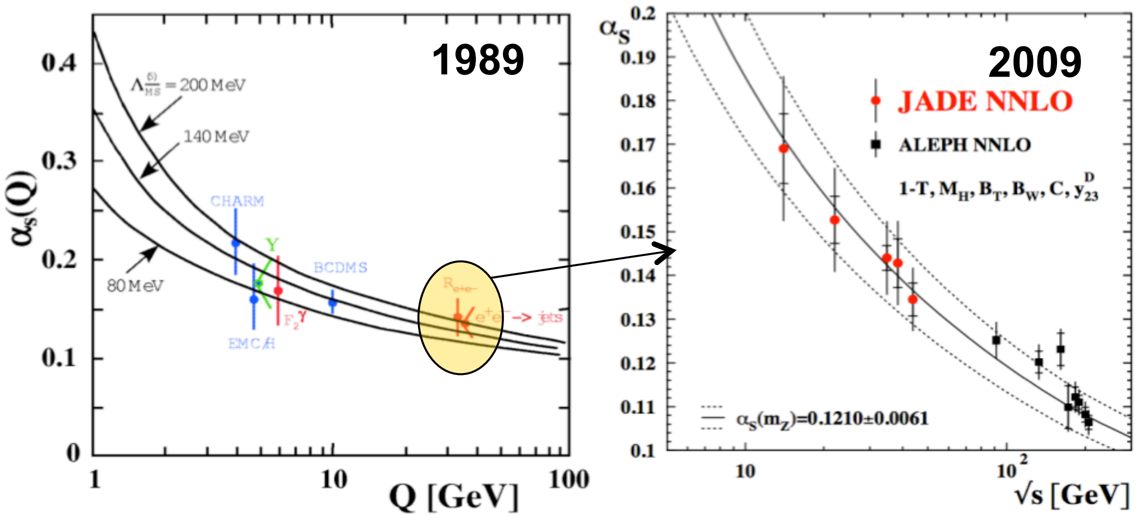


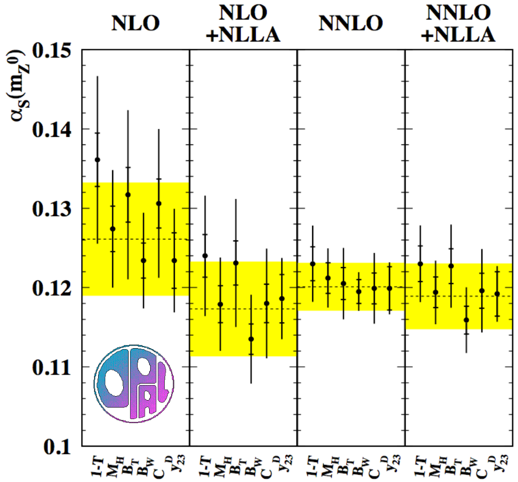




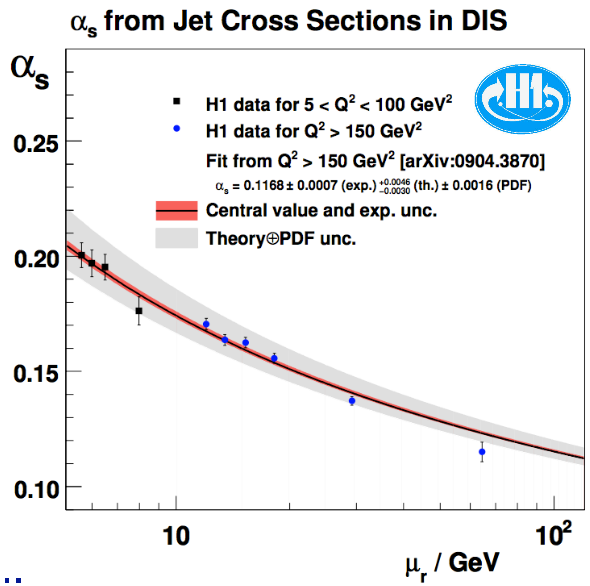
- Both objectives facilitated by data preservation

Revisit old measurements or perform new ones

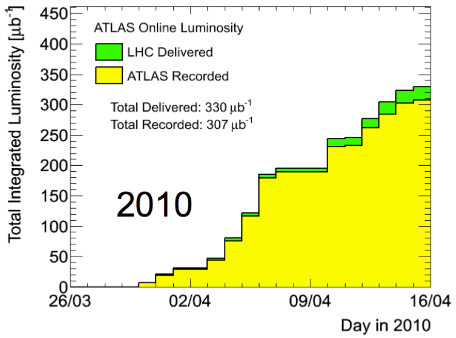




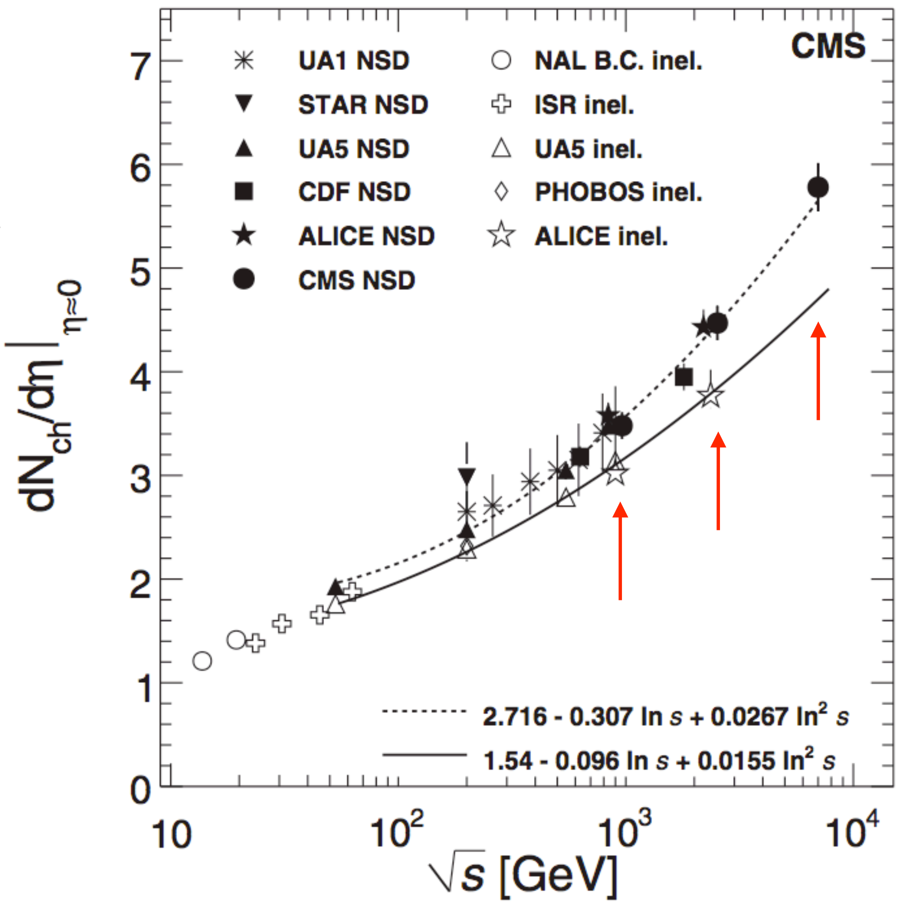
- **>** Access to newly developed techniques, comparisons to new theoretical models
  - History to be repeated with the HERA α_s measurements
- **>** Unique data sets are available in terms of initial state particles and energy
  - HERA $e^{\pm}p$, Tevatron $p\bar{p}$, fixed target experiments..
  - Early LHC data: 900 and 2.36 TeV, 2010 low pile-up 7 TeV...



What about LHC 900 GeV and 2.32 TeV data? And 7 TeV data?

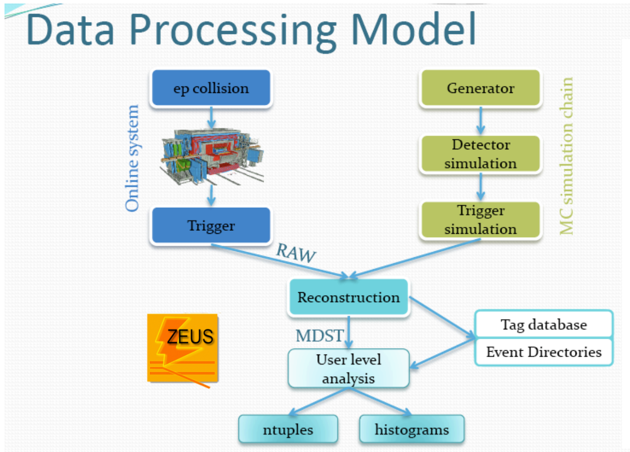


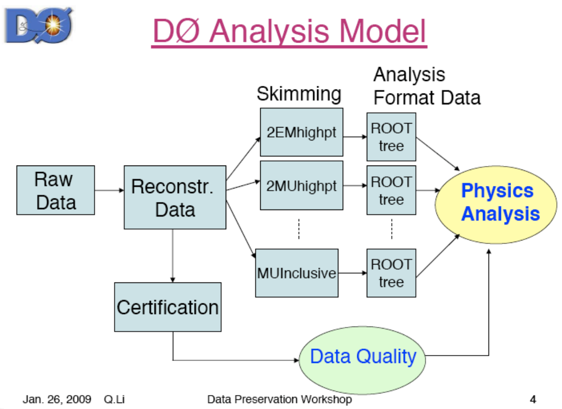
Centre-of-mass Energy	0.9 TeV	2.36 TeV
Selection	Number of Events	
BPTX Coincidence + one BSC Signal	72 637	18 074
One Pixel Track	51 308	13 029
HF Coincidence	40 781	10 948
Beam Halo Rejection	40 741	10 939
Beam Background Rejection	40 647	10 905
Valid Event Vertex	40 320	10 837

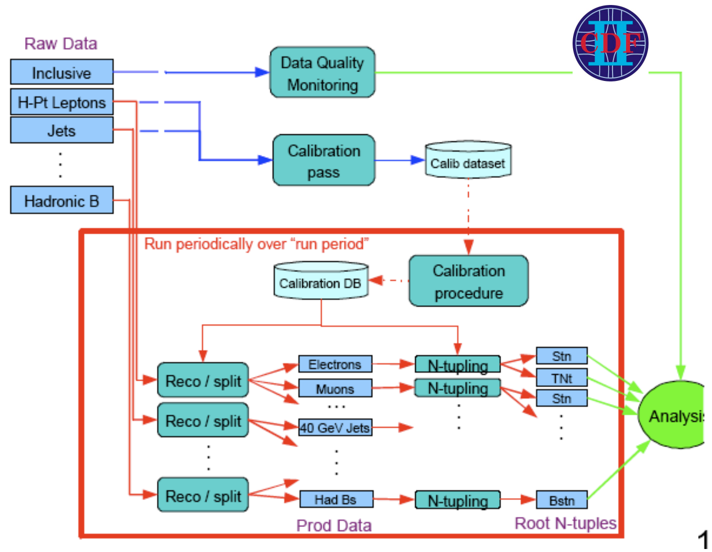


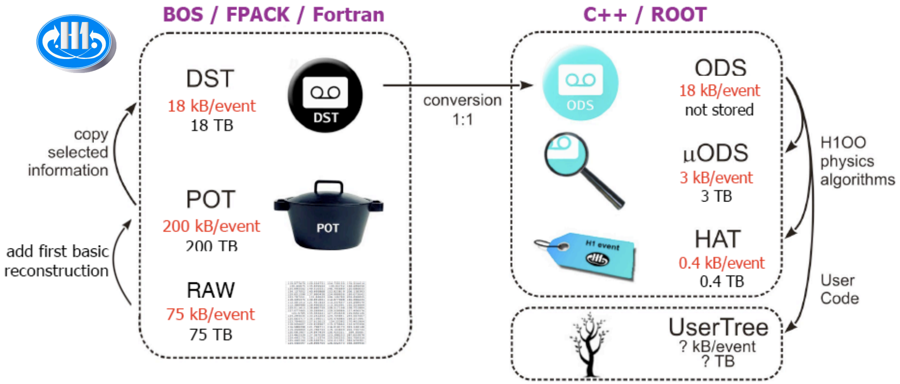
- Early LHC measurements made using data at a unique centre of masses
- 2010 low pile up 7 TeV data also at risk
- What happens when 14 TeV comes?

Data analysis models in HEP

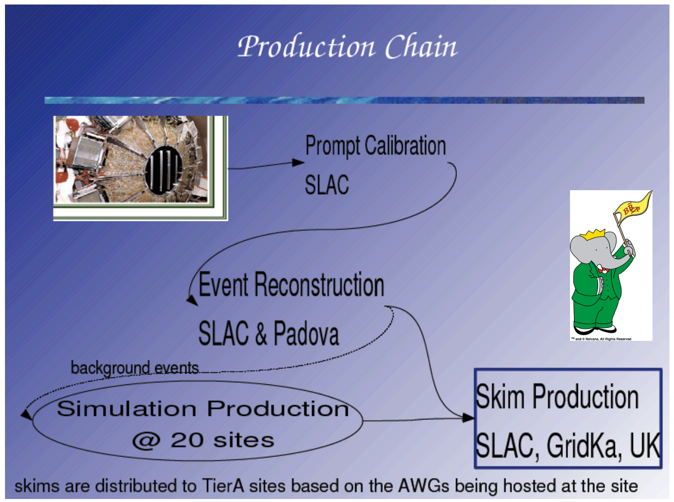




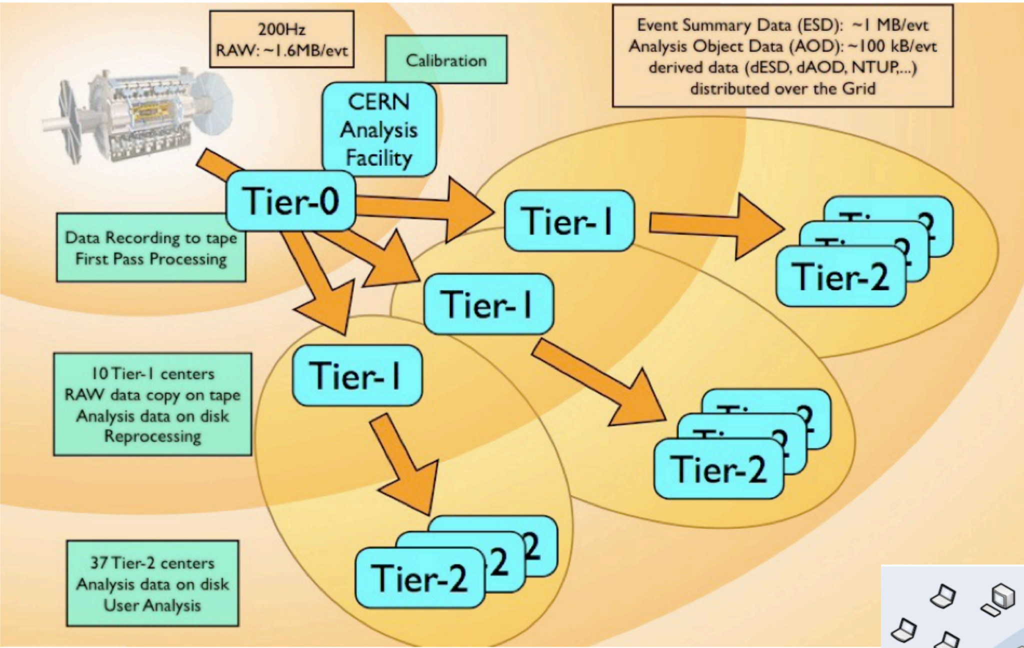


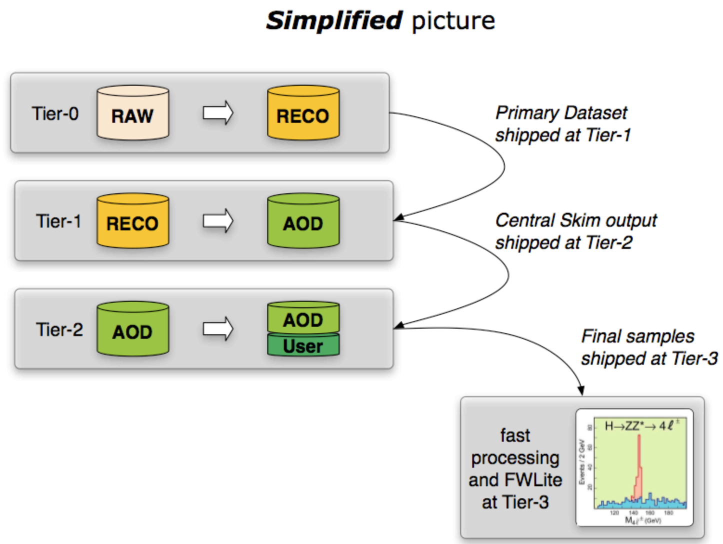


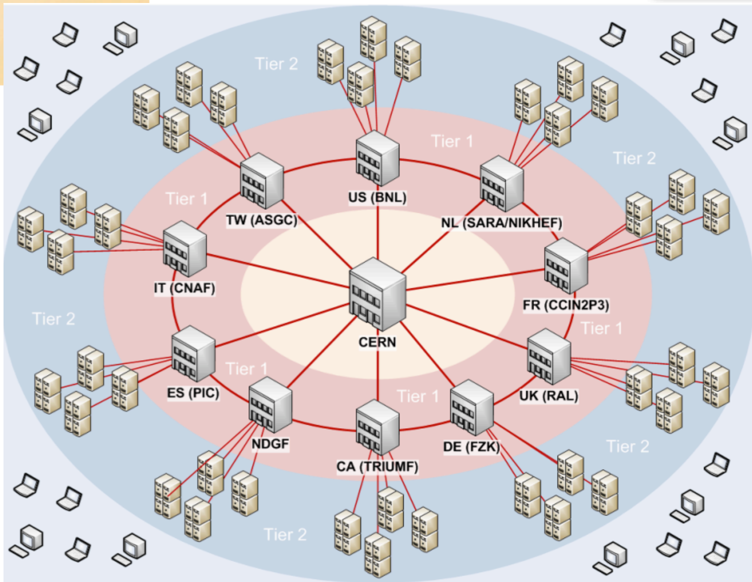
- Complicated, at first glance different
- Familiar descriptions of data analysis chain, from reconstruction to analysis level
  - RAW (→ POT)→ DST → ntuple → analysis



Data analysis models in HEP in the LHC era

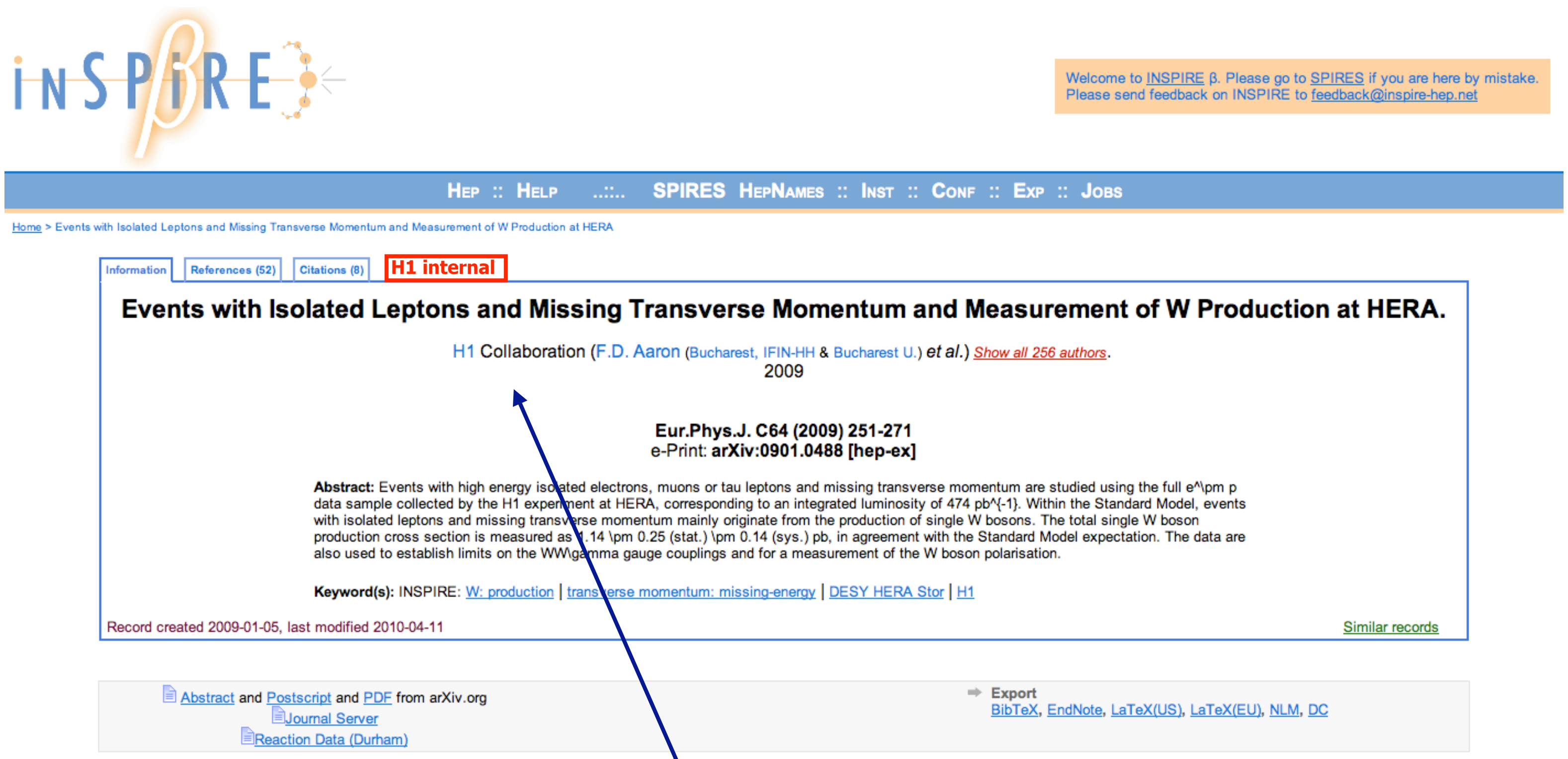






- More skims - *yes*
- More distribution - *certainly*
- More complexity - *perhaps..*
- Data placement is key, but analysis-wise it's still very similar to what we had before

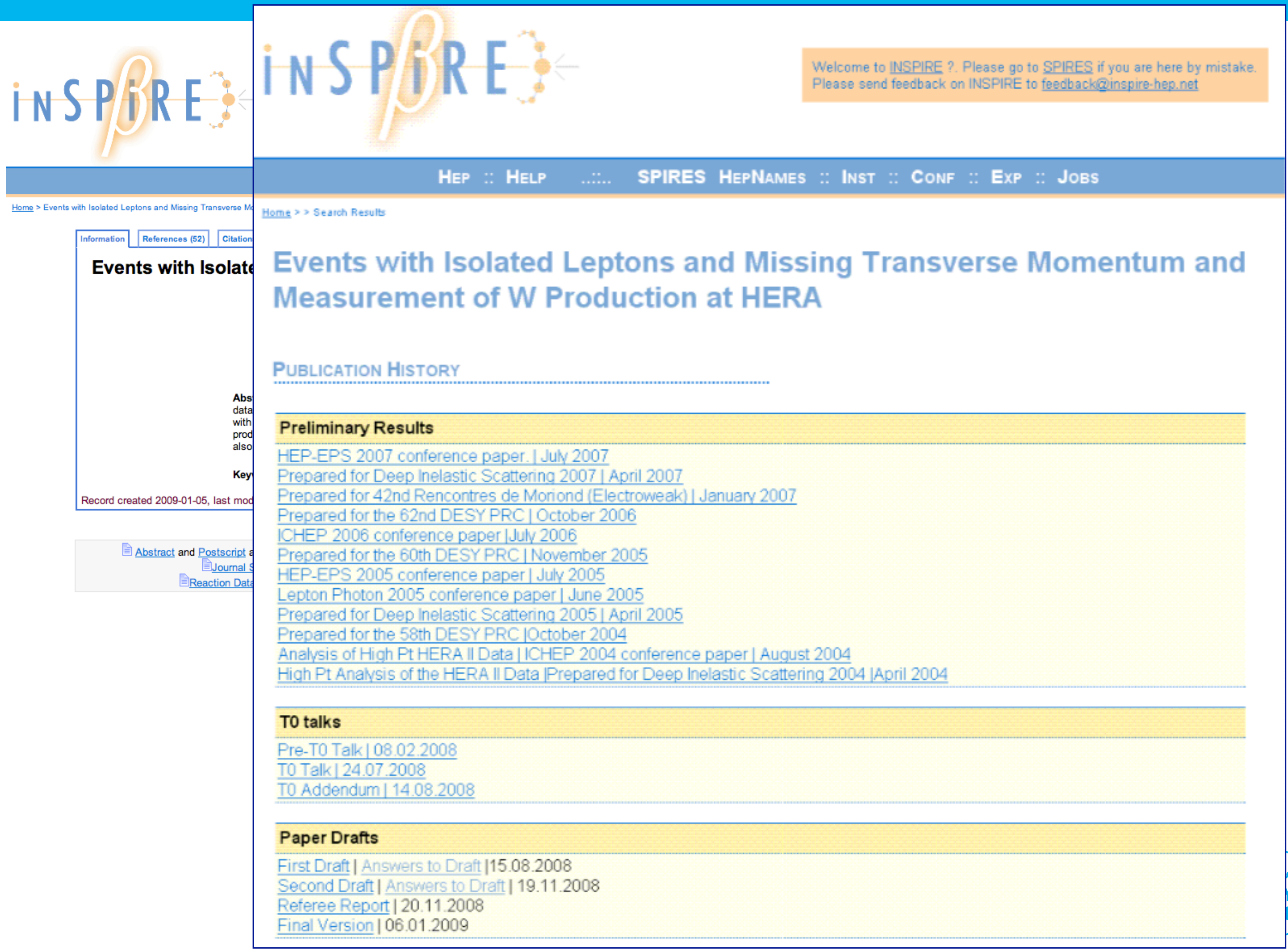
INSPIRE: Paper histories



> Envisage an additional link for H1 members only

> Provides additional information such as preliminary results, earlier draft versions and documentation from the publication procedure

INSPIRE: Paper histories



Welcome to INSPIRE ?. Please go to SPIRES if you are here by mistake.
Please send feedback on INSPIRE to feedback@inspire-hep.net

HEP :: HELP ..::.. SPIRES HEPNAMES :: INST :: CONF :: EXP :: JOBS

Home > > Search Results

Events with Isolated Leptons and Missing Transverse Momentum and Measurement of W Production at HERA

PUBLICATION HISTORY

Preliminary Results

HEP-EPS 2007 conference paper. | July 2007
Prepared for Deep Inelastic Scattering 2007 | April 2007
Prepared for 42nd Rencontres de Moriond (Electroweak) | January 2007
Prepared for the 62nd DESY PRC | October 2006
ICHEP 2006 conference paper |July 2006
Prepared for the 60th DESY PRC | November 2005
HEP-EPS 2005 conference paper | July 2005
Lepton Photon 2005 conference paper | June 2005
Prepared for Deep Inelastic Scattering 2005 | April 2005
Prepared for the 58th DESY PRC |October 2004
Analysis of High Pt HERA II Data | ICHEP 2004 conference paper | August 2004
High Pt Analysis of the HERA II Data |Prepared for Deep Inelastic Scattering 2004 |April 2004

T0 talks

Pre-T0 Talk | 08.02.2008
T0 Talk | 24.07.2008
T0 Addendum | 14.08.2008

Paper Drafts

First Draft | Answers to Draft |15.08.2008
Second Draft | Answers to Draft | 19.11.2008
Referee Report | 20.11.2008
Final Version | 06.01.2009

For completeness, the HERA data summary

> Final ZEUS data reprocessing to mDST completed in 2009

- Basic preserved data format: ROOT based "Common Ntuples" (CN)
- Ultimately RAW, MDST data and MC removed from robots, keep only CN
- Reduces total amount to be preserved for ZEUS from the current 1 PB to ~ **200 TB**

> Final H1 reprocessing of HERA II data 2009, HERA I repro almost there

- Common analysis software H1OO started in 2000, uses ROOT based data format, used by all H1
- In addition, a monthly MC production of up to 1/4 billion events
- H1 to preserve RAW data, as well as one DST version and one analysis level version
- Estimate total amount to be preserved for H1 to be ~ **200-500 TB**

> Main format for HERMES analyses is the mDST

- New production planned before final freeze
- Last years of data taking with recoil detector, still need improved calibrations
- MC productions on Grid for on-going analyses
- Total amount to preserve on tapes ~ **20-500 TB**

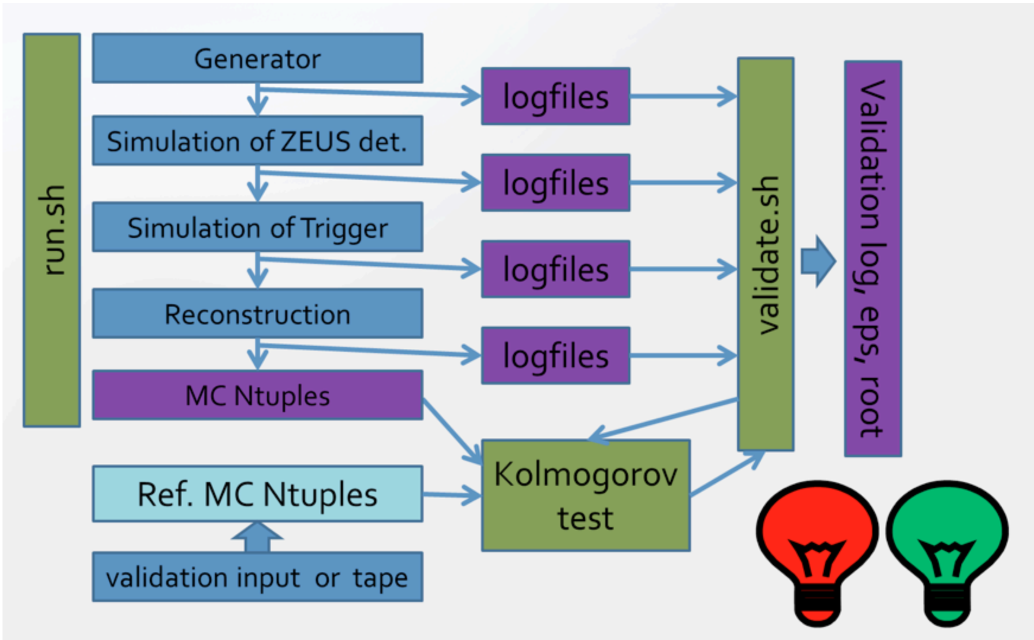
> Preservation of HERA-B data under investigation within DESY-IT

- Total amount of data currently ~ **250 TB**, decreases once preservation model established

Example structure of experiment tests: ZEUS (Level 3 + MC chain)

- ZEUS strategy: use ROOT based analysis level Common Ntuples as data format for preservation – DPHEP level 3
- Only external dependence is ROOT
  - Validation of new ROOT versions included as analysis level tests in the `sp-system`

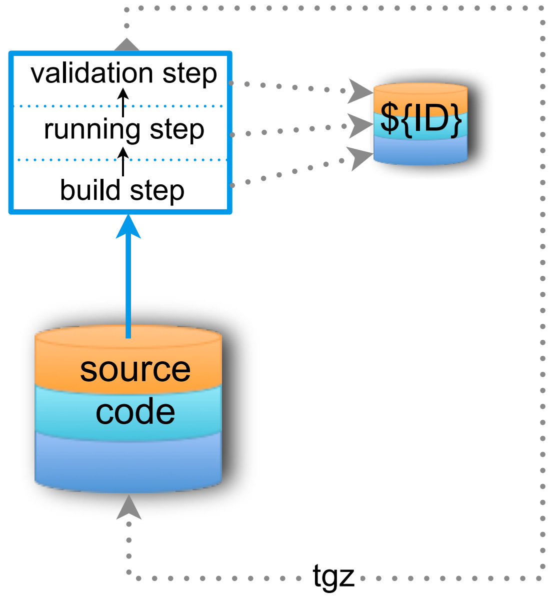
- However, the MC production chain executables will also be preserved as a standalone package
- In addition, an interface for new generators is developed, which is also included in the validation system



Running jobs in the `sp-system`

- Initial step
  - Compilation of analysis (level 3) and sim/rec (level 4) software
  - ***Or:*** use tar-balls with pre-compiled software
  - Provide access to software

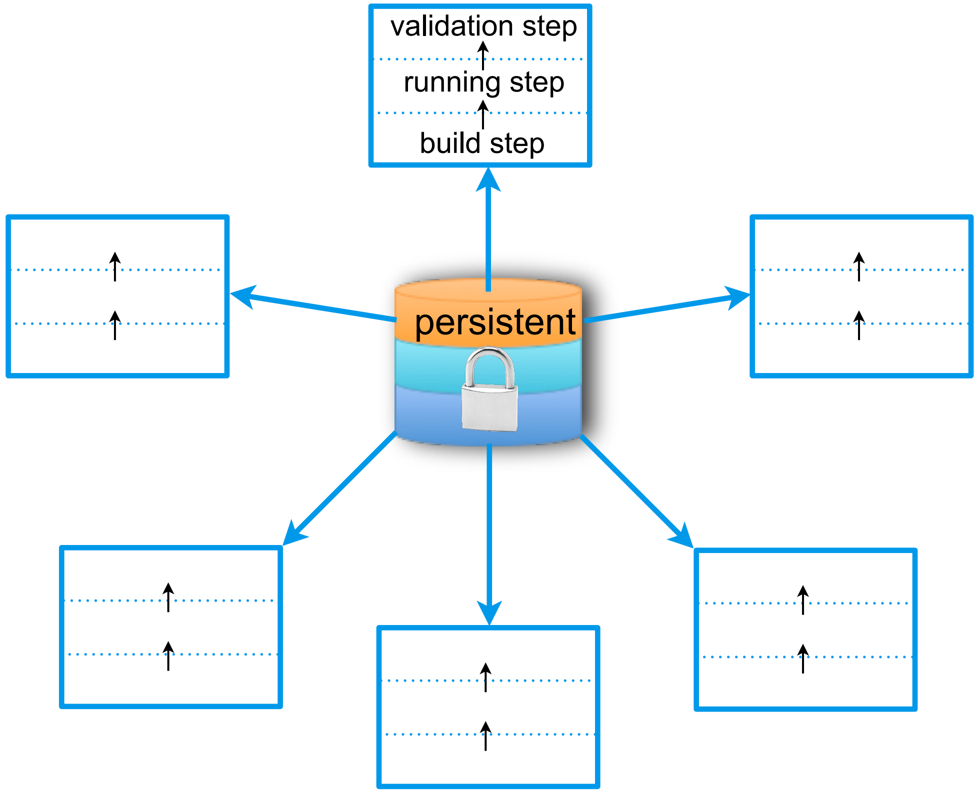
    Copy tar-balls to persistent storage
  - All output kept in directory with unique name



Running jobs in the `sp-system`

- Initial step
  - Compilation of analysis (level 3) and sim/rec (level 4) software
  - ***Or:*** use tar-balls with pre-compiled software
  - Provide access to software
    - Copy tar-balls to persistent storage
  - All output kept in directory with unique name

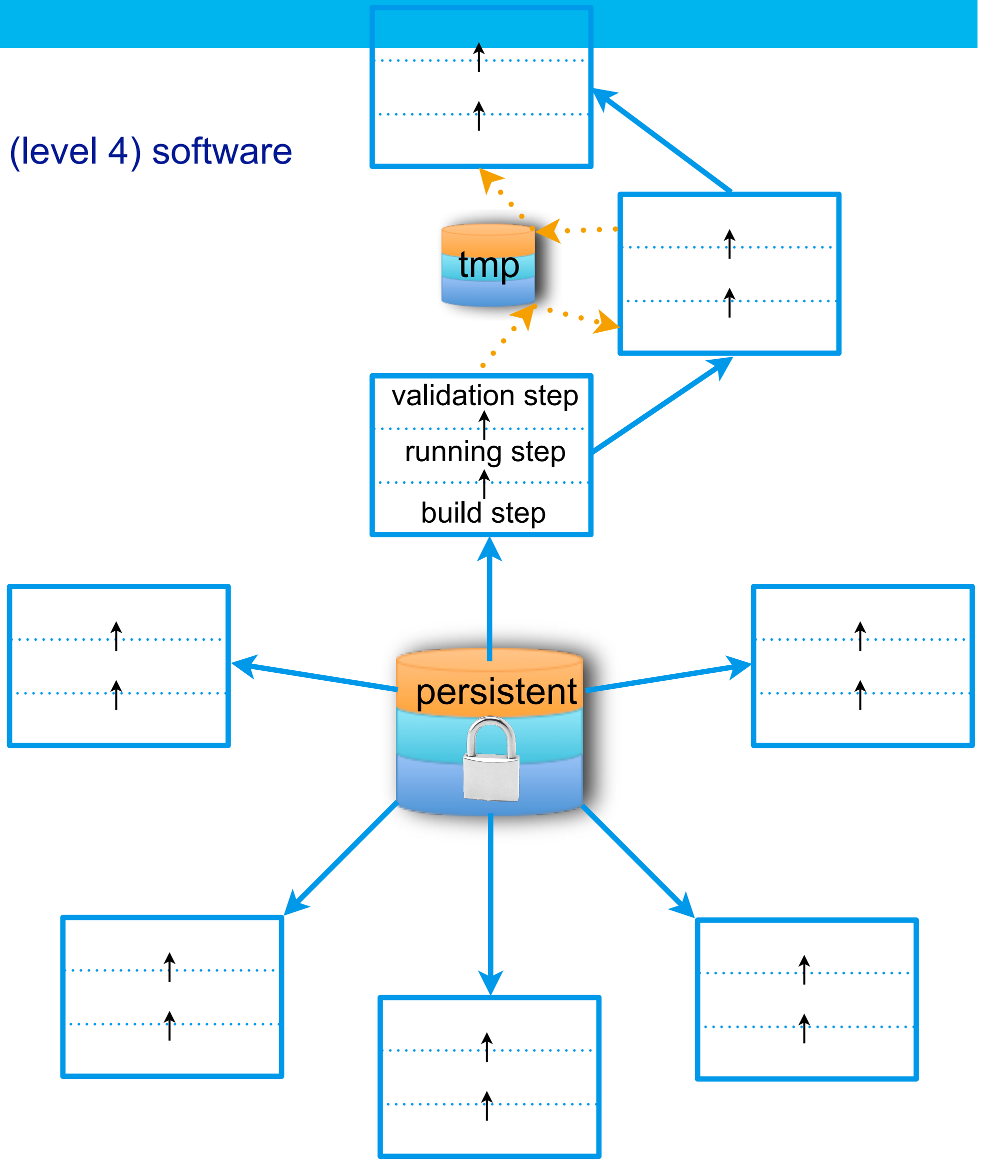
- Run parallel tests
  - Set up software environment
  - Validate binaries with persistent input
    - e.g. event display, database access, ...



DESY

Running jobs in the `sp-system`

- Initial step
  - Compilation of analysis (level 3) and sim/rec (level 4) software
  - ***Or:*** use tar-balls with pre-compiled software
  - Provide access to software
    - Copy tar-balls to persistent storage
  - All output kept in directory with unique name
- Run parallel tests
  - Set up software environment
  - Validate binaries with persistent input
    - e.g. event display, database access, ...
- Run sequential tests
  - Set up software environment
  - Validate file production
    1. MC generation (produce gen files)
    2. Reconstruction (gen. files → DSTs)
    3. Analysis level (DSTs → ROOT files)
  - Tests use output of previous test as input



- Results remain accessible or can be reproduced with identical results

Securing the resources

- The new DPHEP organisation will develop at least three levels:
  - Experiment / collaboration level projects
  - Multi-experiment level initiatives
  - Global DPHEP level projects or positions
- It is foreseen that funding must come from different sources, in particular for common DPHEP enterprises or positions
- The experiment and laboratory level projects are highest priority (1-2 FTE per site), followed by the appointment of the DPHEP Project Manager, which is a full time position
- Many potential multi-experiment projects also exist, including those shown today, which depend on additional funding, typically 0.5-1 FTE

DPHEP person power requirements

	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 program. 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.

LEP Paper Tables

	2001	2002	2003	2004	2005	2006	2007	2008	2009	Total	2004-2009
ALEPH	46	42	24	34	12	9	4	4	2	**607**	65
DELPHI	64	30	31	58	21	19	7	7	2	**678**	114
L3	51	40	23	52	16	11	5	2	0	**578**	86
OPAL	61	38	32	55	9	11	4	3	2	**675**	84
All	**222**	**150**	**110**	**199**	**58**	**50**	**20**	**16**	**6**	**2538**	**349**

Table 1: Statistics of peer-reviewed publications of the LEP collaborations.

Papers 2004-2009	ALEPH	DELPHI	L3	OPAL	All
Electroweak	17	26	22	24	89
QCD	19	25	19	22	85
Higgs Searches	6	14	8	9	37
SUSY Searches	4	7	5	9	25
Exotica Searches	5	12	10	7	34
Flavour Physics	6	15	4	5	30
Exclusive Channels	3	8	8	2	21
Cosmo-LEP	3	3	6	0	12
Other	2	4	4	6	16
Total	**65**	**114**	**86**	**84**	**349**

Table 2: Distribution of physics topics in LEP publications in the years 2004-2009.