

# Data Preservation in High Energy Physics

and the DPHEP Collaboration.



David South (DESY)

HAP Workshop Topic 2  
The Non-Thermal Universe

Erlangen  
September 22, 2016



[hep-project-dpheap-portal.web.cern.ch](http://hep-project-dpheap-portal.web.cern.ch)  
[dpheap.org](http://dpheap.org)



# Outline

## > **Since 2008 : Formation of the DPHEP Study Group**

- Assessing the landscape, defining the problem
- The last generation of HEP experiments.. what lessons have we learned?

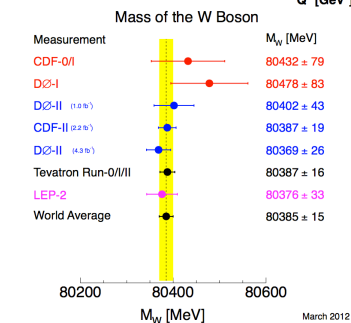
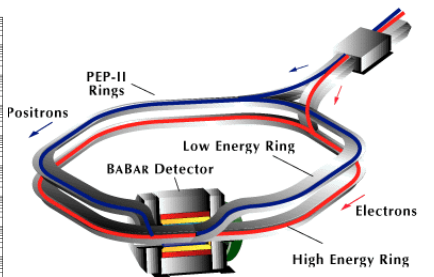
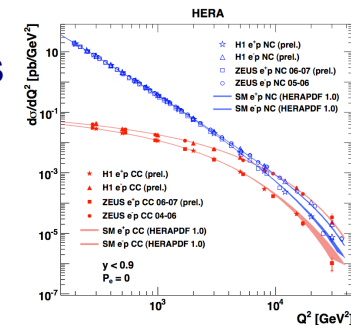
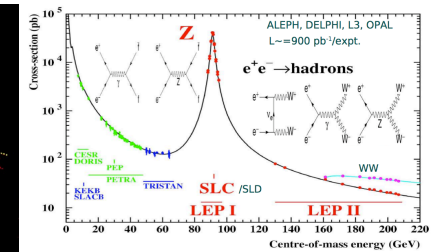
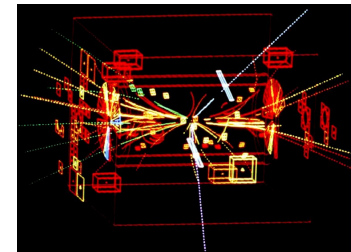
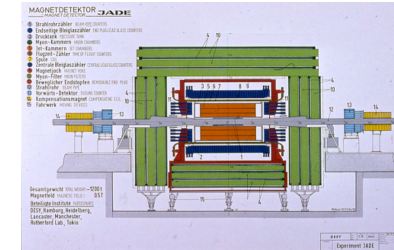
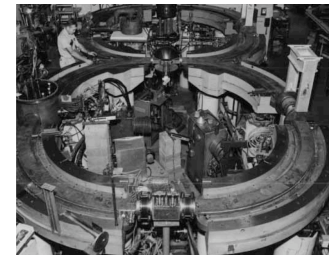
## > **Since 2014 : The DPHEP Collaboration and the LHC era**

- Moving forward, the mandate and the “2020 vision”
- Current activities of the LHC experiments

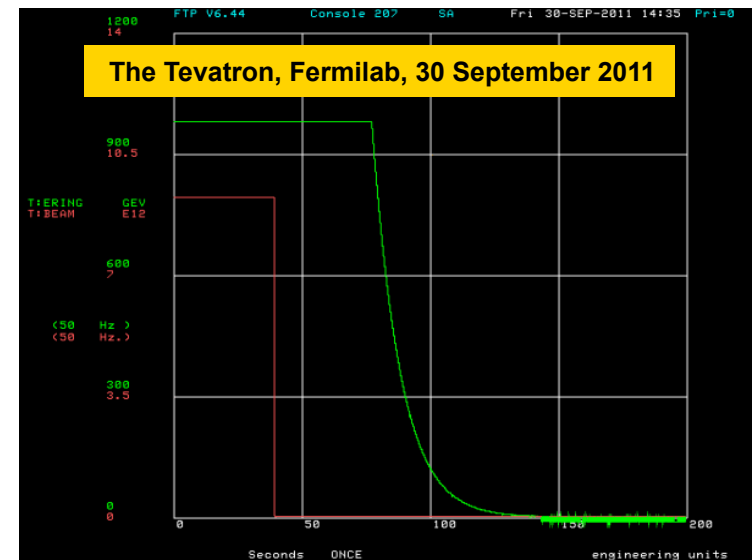
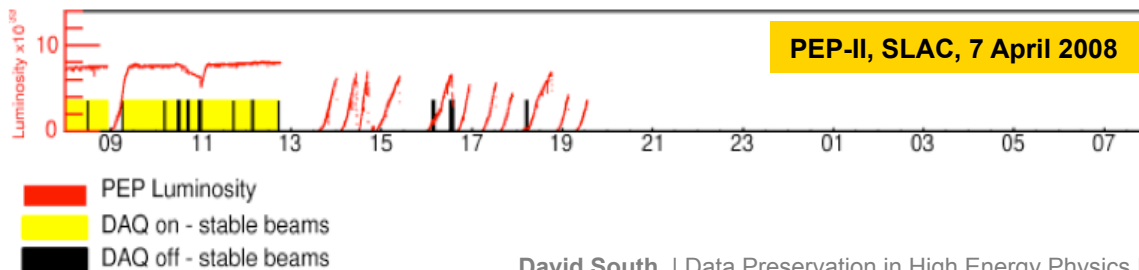
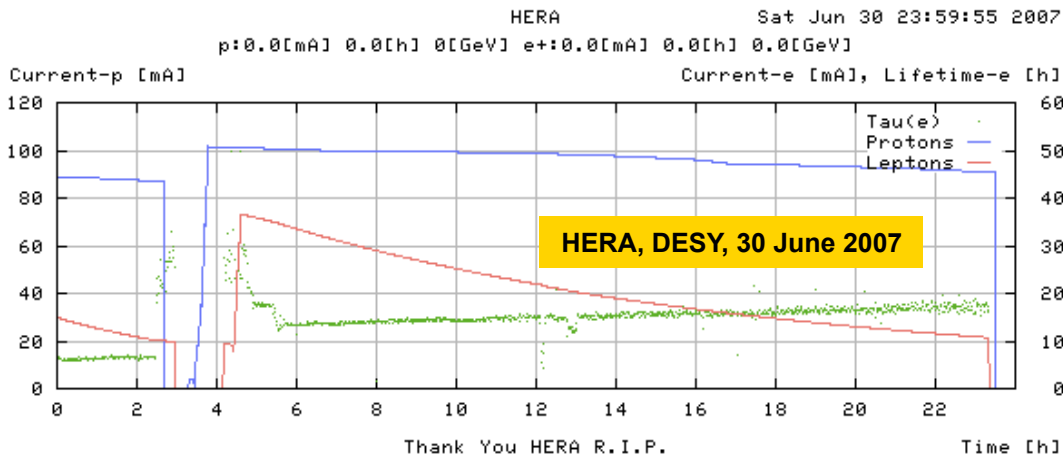
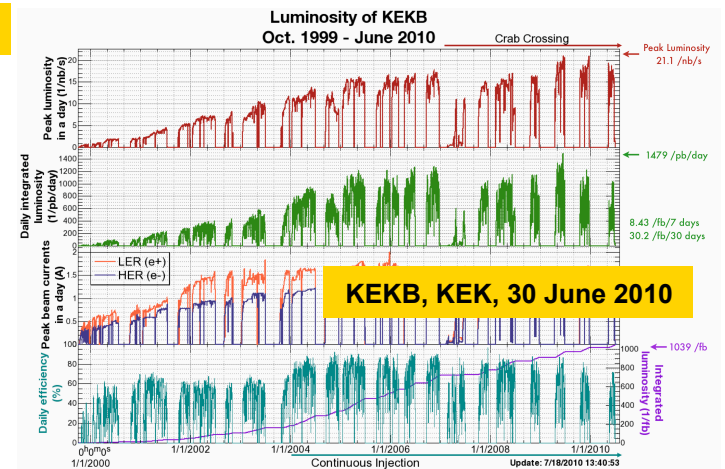
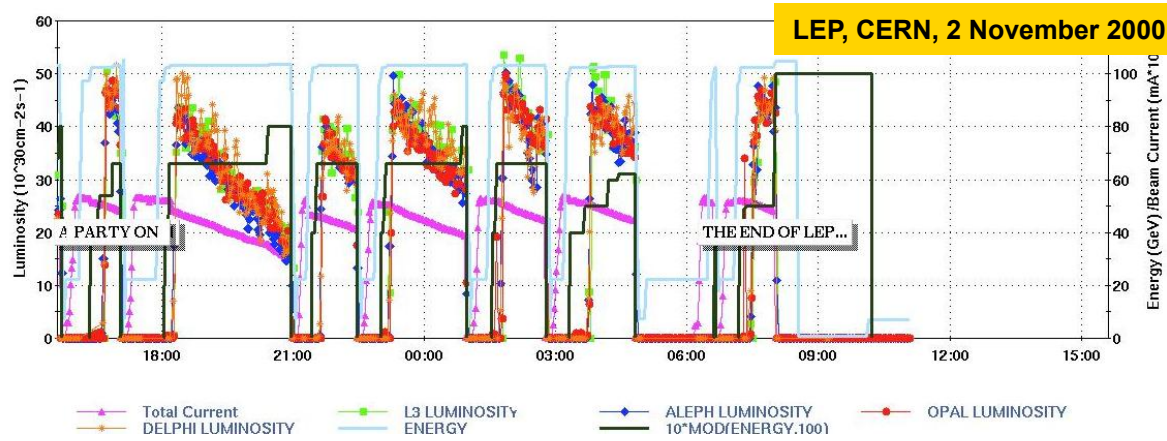


# Experimental particle physics in the collider era

- A wide variety of physics results from many, often very different experiments
- Energy frontier probed with increasingly complex accelerator installations
  - From single room colliders in the late 1950s to installations measured in 10s of kilometres
  - Results from newer experiments typically, but not always, supersede those of similar older ones
- Growth in size of the international collaborations, increase in the diversity of the data management
- We are now in the age of the LHC
  - Belle 2, HL-LHC, and other projects such as the ILC or the next e-p/A collider are to come



# The start of the 21<sup>st</sup> century: the end of several experiments



# What do you do when the collisions have stopped ?

- Finish the analyses! But then what do you do with the data?
  - Up until recently, there was no clear policy on this in the HEP community
  - In the main, older HEP experiments have simply lost the data
- Data preservation, including long term access, is generally not part of the planning, software design or budget of an experiment
  - Again, up until recently, HEP data preservation initiatives have been in the main not planned by the original collaborations, but rather the effort a few knowledgeable people
  - The now infamous example is the recovery of the unique JADE data taken at DESY 1979-86, which led to the discovery of the gluon. That data is still being analysed today

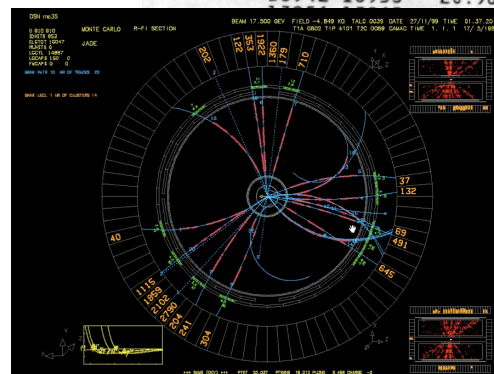
- The Exabyte Cartridge Collection (in the J.O. Office)



traveled to MPI Munich, was transferred to disk, and is now a (very very small) part of the ATLAS Data Storage.

- Programs were developed to read the FPACK-ed Data, and to convert each of the JADE BOS Banks into the original sequences of I\*4, I\*2, F and A data.

RUNS	BEAM	BARREL	LUMINOSITY
13856	13864	20.840	0.474029E+02 +- 0.779300E+01
13865	13872	20.855	0.538850E+02 +- 0.831464E+01
13873	13885	20.870	0.719484E+02 +- 0.961450E+01
13886	13895	20.885	0.694769E+02 +- 0.945461E+01
13896	13906	20.900	0.579792E+02 +- 0.864303E+01
13907	13919	20.915	0.516098E+02 +- 0.816022E+01
13920	13931	20.930	0.555588E+02 +- 0.847264E+01
13932	13941	20.945	0.465800E+02 +- 0.776333E+01
13942	13953	20.960	0.285056E+02 +- 0.607743E+01
			0.609841E+02 +- 0.889545E+01
			0.519744E+02 +- 0.821787E+01
			0.442404E+02 +- 0.758717E+01
			0.508176E+02 +- 0.813734E+01
			0.678519E+02 +- 0.940937E+01
			0.770938E+02 +- 0.100368E+02
			0.667339E+02 +- 0.934461E+01
			0.497930E+02 +- 0.807749E+01
			0.524870E+02 +- 0.829892E+01
			0.499324E+02 +- 0.810010E+01
			0.467388E+02 +- 0.722265E+01



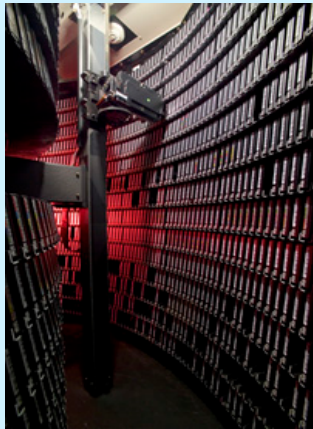
# What do you do when the collisions have stopped ?

- The conservation of tapes is not data preservation!  
Some quotes from computer centres in 2010:



- *“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!”*

# An important question: What is HEP data?

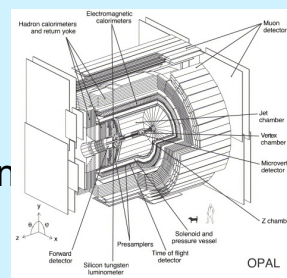


## Digital information

The data themselves, volume estimates for preservation data of the order of **a few to 10 PB** for pre-LHC experiments. Other digital sources such as databases to also be considered

## Software

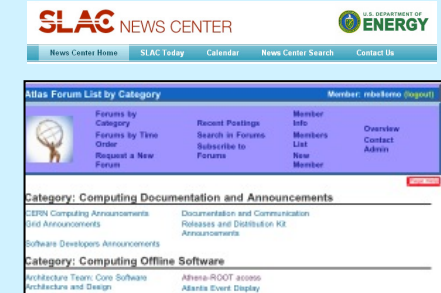
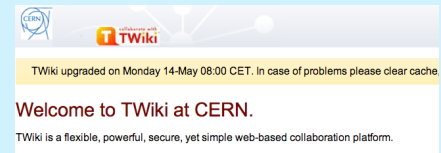
Simulation, reconstruction, analysis, user, in addition to any external dependencies



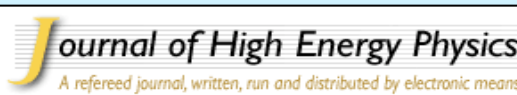
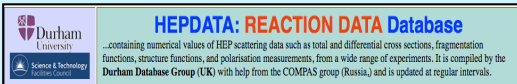
Software license for LHC++  
**CERNLIB Access**  
 • Access to the CERN Program Library is free of charge to all HEP users worldwide.  
 • Non-HEP academic and not-for-profit organizations: 1KSF/year

## Meta information

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

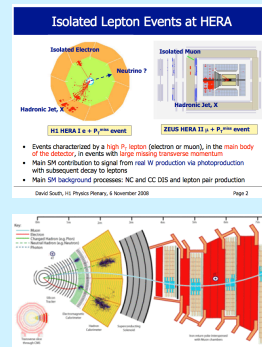
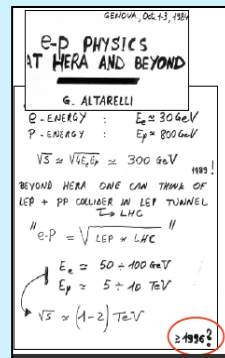


## Publications



## Documentation

Internal publications, notes, manuals, slides



## Expertise and people



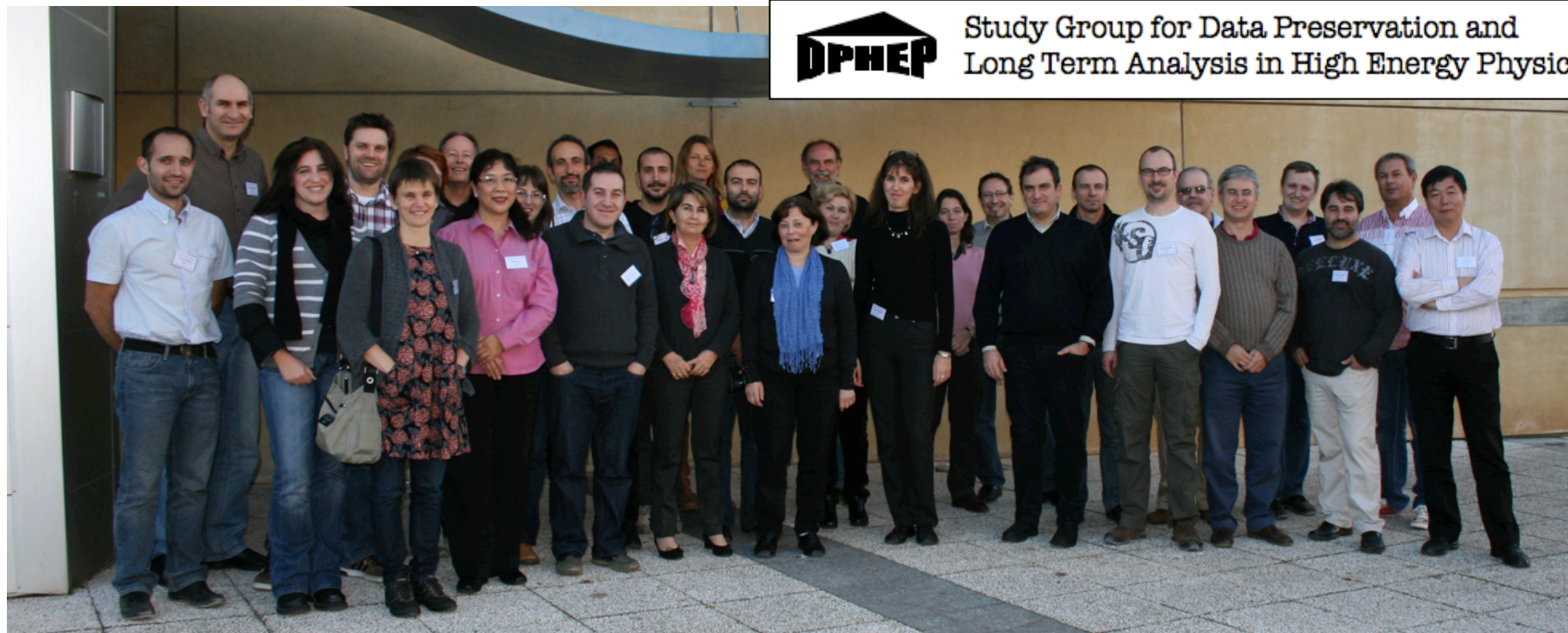
# Issues and difficulties concerning HEP data preservation

- > The experiments are generally interested in the here and now
  - Up until recently the issue of data preservation was not really considered at the LHC
- > Handling HEP data involves large scale traffic, storage and migration
  - The distribution of HEP data and evolving access methods may complicate the task
- > Who is responsible for the data? The experiments? The computing centres?
  - Problem of older, unreliable hardware: unreadable tapes after 2-3 years
- > The software is often very complex, multi-layered and distributed
  - Infrastructure, versioning, compatibility vary considerably over the lifetime of the experiment
- > Key resources, funding and expertise, decrease after data taking stops
- > And importantly: *Who says we want to do all this anyway ?*
  - Is the potential benefit really worth the cost and effort? And how much does it cost?
  - Can the relevant physics cases be made?





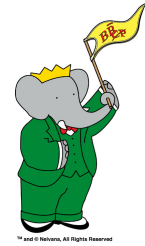
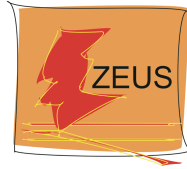
# DPHEP: An international Study Group on data preservation



- First contacts established 2008, endorsed as an ICFA panel 2009
  - Group has grown to over 100 contact persons
  - Initial make up of the group was driven by the coincidence of the end of data taking at several large colliders – SLAC, HERA, Tevatron
  - Has since grown to include many others including the LHC experiments from 2011



# DPHEP: An international Study Group on data preservation



Institute of High Energy Physics  
Chinese Academy of Sciences



Jefferson Lab



Science & Technology  
Facilities Council



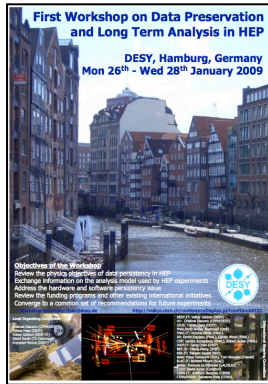
INSPIRE

**BROOKHAVEN**  
NATIONAL LABORATORY



# DPHEP: An international Study Group on data preservation

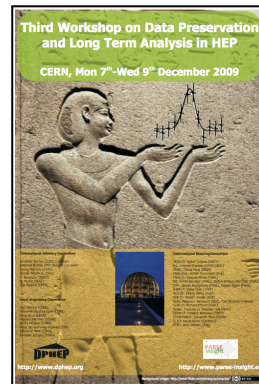
## ➤ Series of DPHEP workshops held 2009-2012



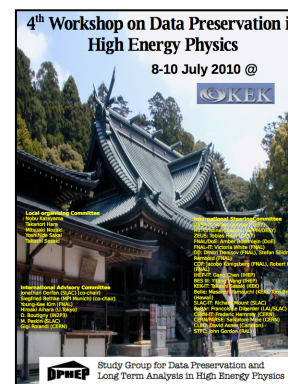
Jan 2009: DESY



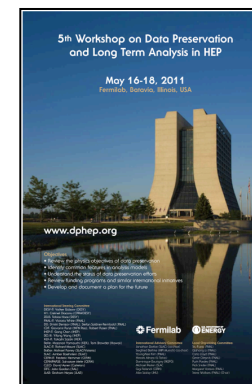
May 2009: SLAC



Dec 2009: CERN



Jul 2010: KEK



May 2011: Fermilab

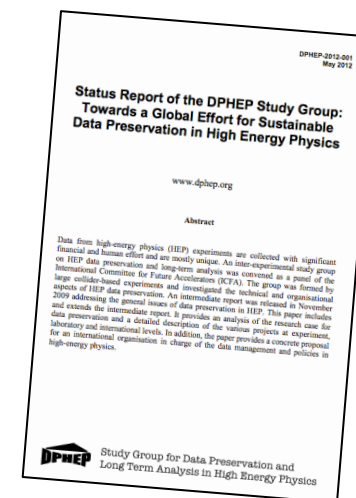


Nov 2012: CPPM

## ➤ Initial findings published in a short interim report, published December 2009

## ➤ Full report of the activities of the DPHEP Study Group, published May 2012

- 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

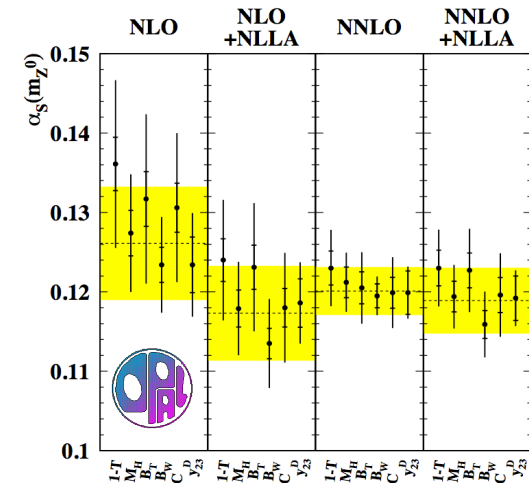
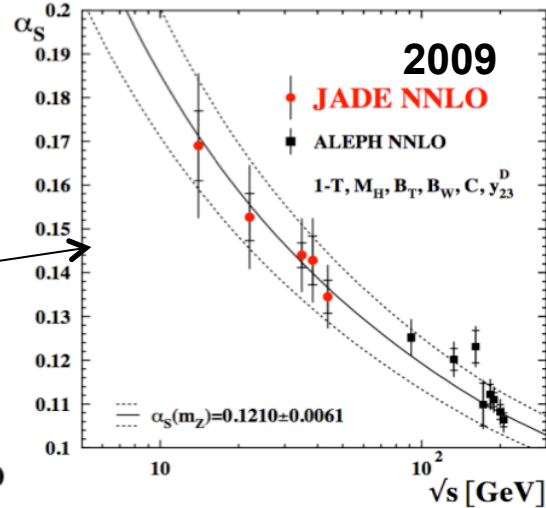
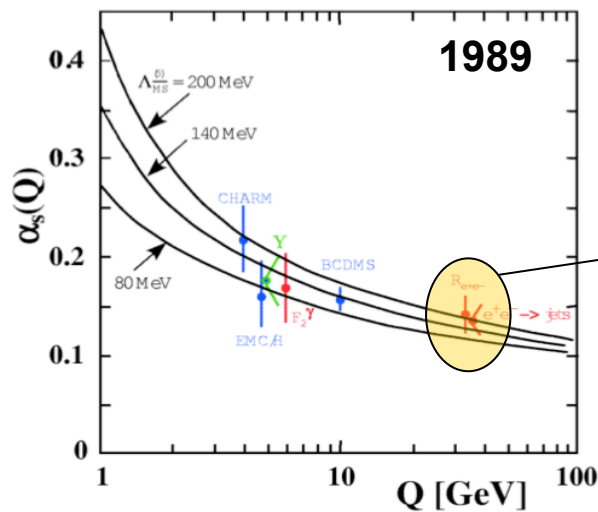


# Building the physics case: Reasons to preserve HEP data

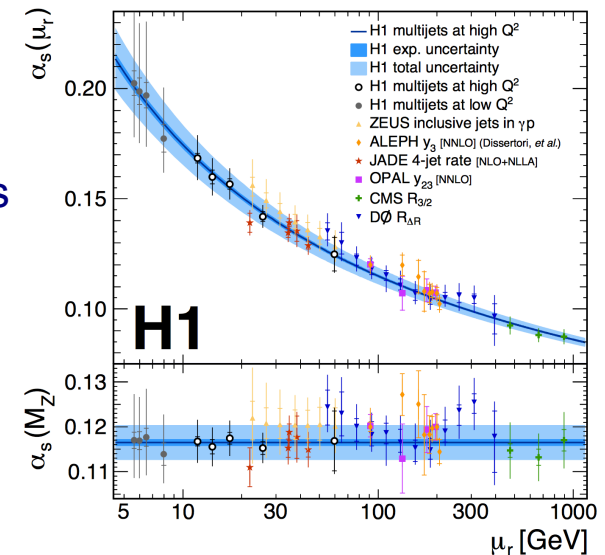
- > Long term completion and extension of an existing physics program
  - Up to 10% of papers are finalised in the “archival mode”
  - Gain in scientific output of the experiments
- > Cross-collaboration and combinations of physics results
  - During the active lifetime of similar experiments at one 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
  - Simplified formats: associated exercises to perform e.g. composite-particle reconstruction, finding signals in the background, ...



# Examples: Revisit old measurements or perform new ones




- Access to newly developed techniques, comparisons to new theoretical models
  - History may be repeated with the HERA  $\alpha_s$  measurements
- Unique data sets are available in terms of initial state particles and energy
  - If no LHeC or alternative, HERA  $e^+p$  data are all we have
  - Tevatron  $p\bar{p}$  are also unique:  $A_{FB}$ , high-x jets, ...
  - Fixed target experiments, others, ...



# DPHEP data preservation levels

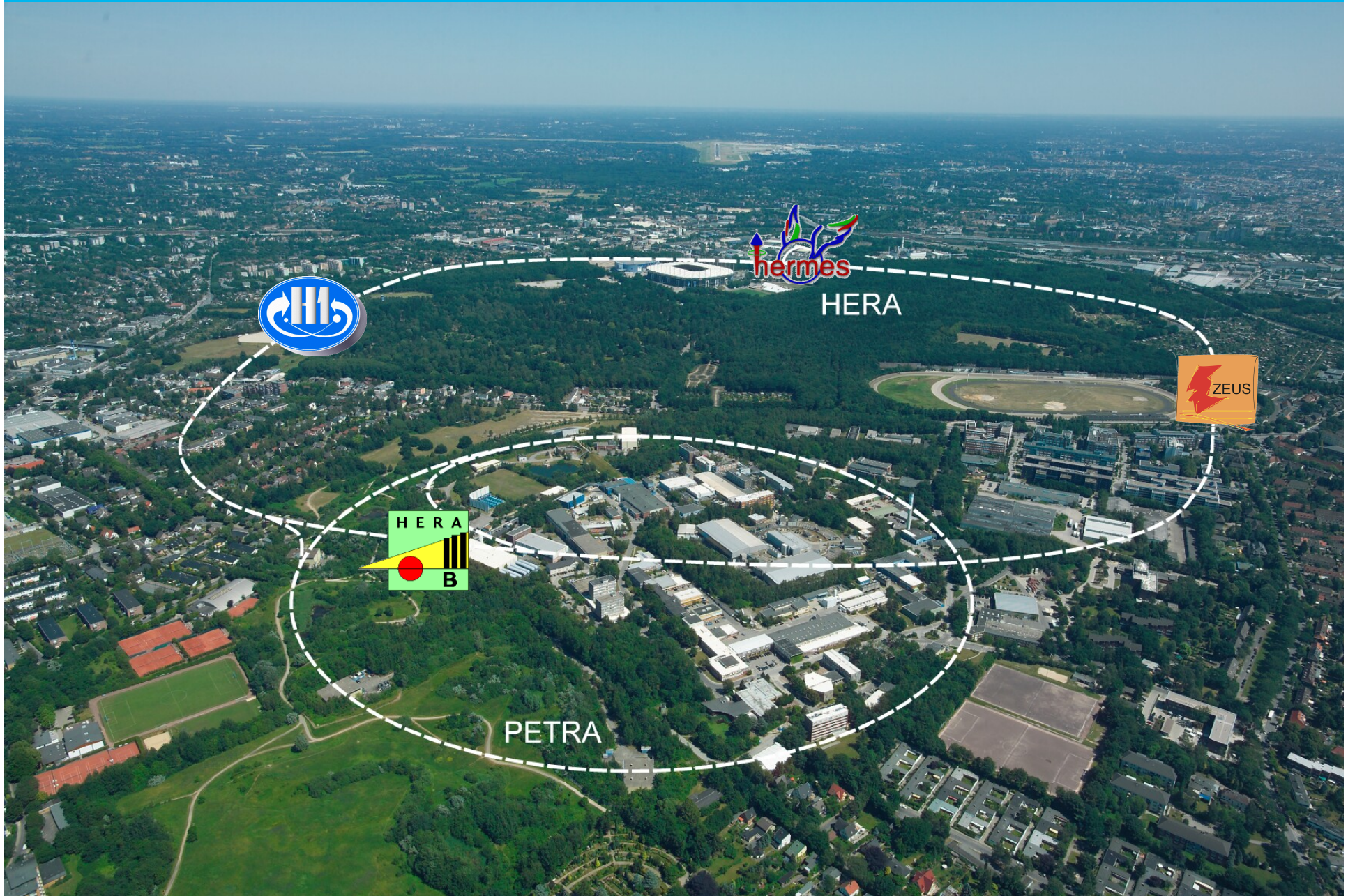
- > So called “levels of data preservation” defined by the DPHEP Study Group now common language in high energy physics

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

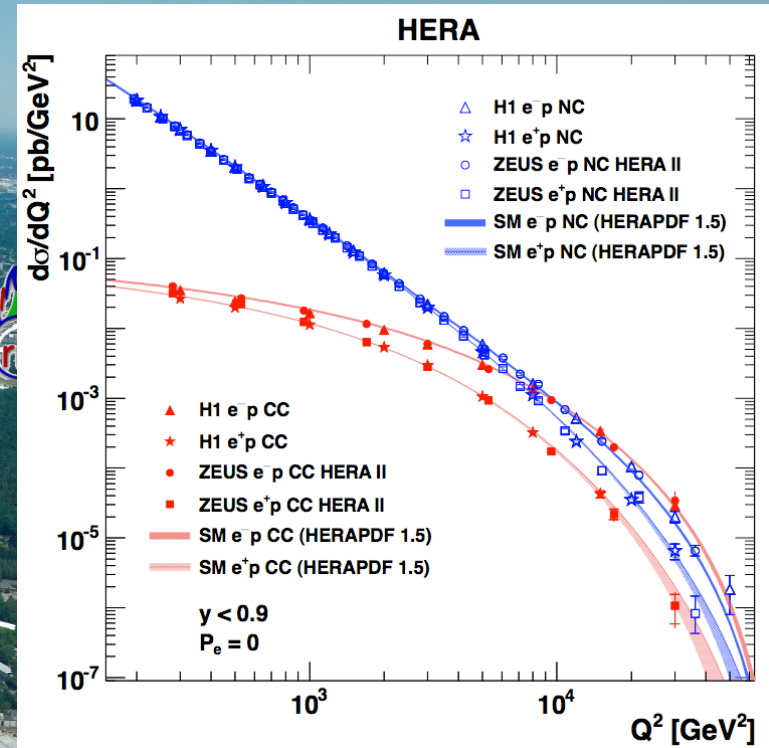
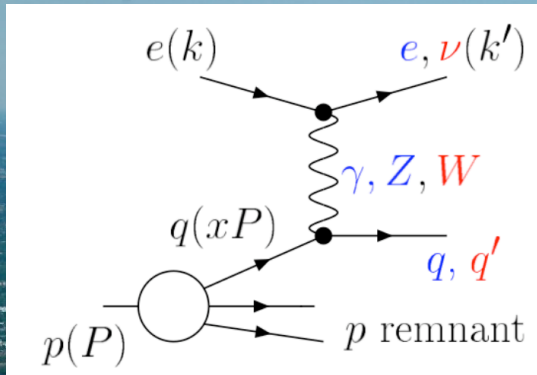
- > Originally idea was more of a progression, almost like an inclusive level structure, but now seen as complementary initiatives
- > Three levels representing three areas:
  - **Documentation, Outreach and Technical Preservation Projects**
- > Now as an example a few highlights from the data preservation efforts carried out at DESY



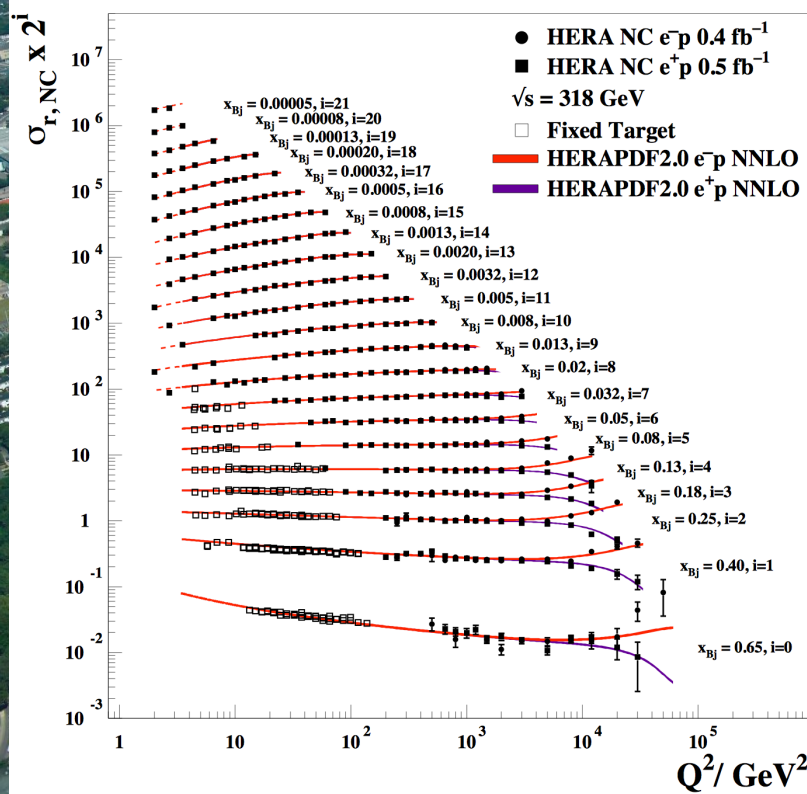
# 1992-2007: Hadron-Electron Ring Accelerator (HERA) @ DESY



# 1992-2007: Hadron-Electron Ring Accelerator (HERA) @ DESY



## H1 and ZEUS

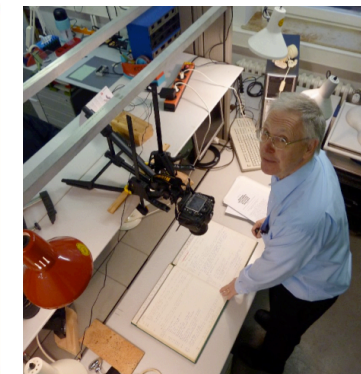
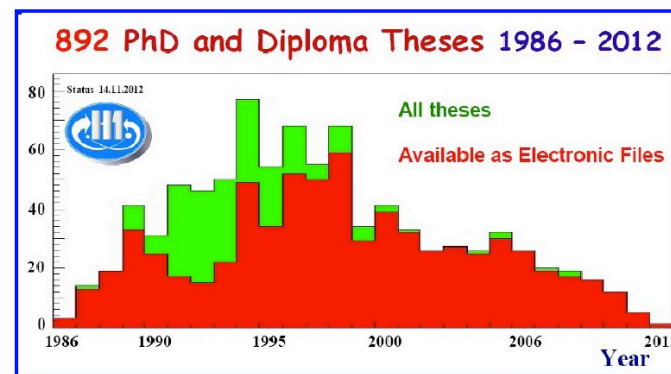
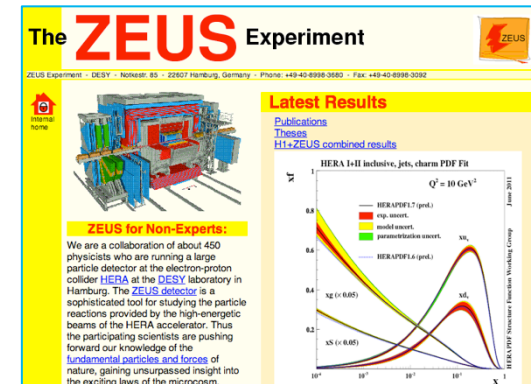
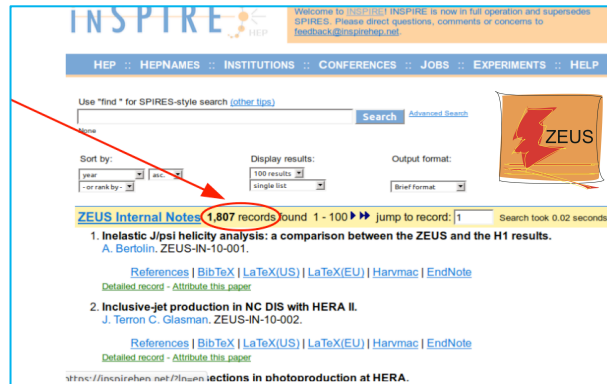


- The world's only electron-proton collider, collisions at H1 and ZEUS 1994-2007
- Precise picture of the proton, crucial measurements for hadron colliders
- Many other unique physics measurements



# DP @ DESY: Documentation

- Successful collaboration between INSPIRE, the experiments and the DESY Library
- Digital documentation such as web-pages revised, reduced and streamlined for future use
- Lots of effort done sorting the vast amount of non-digital documentation
  - Many new (re-)discoveries along the way
- Work possible only by key people with the *right expertise and necessary experience*



# DP @ DESY: Data for preservation and archival storage

- > Deciding which data (and MC) are needed for the long term depends on the preservation model assumed: Remember “level 4” goes back to the raw data
- > Final production version of HERA data for preservation only completed in 2013, a full 6 years after data taking stopped!
- > Estimates for the final HERA DPHEP dataset volume (including MC samples):
  - Two tape copies and an “always online” (disk-based) component
  - Data which should be archived, but not online all the time re-packed into larger files
  - Costs not prohibitive on data volume basis
- > Dedicated system too costly in both hardware and support required
  - All collaborations use dCache for mass storage and this system will continue at DESY-IT for the LHC, photon-physics and others. Natural solution for DPHEP dataset
  - Changes “transparent” for user, but relies on IT support

Expt	Online (TB)	Total (TB)
H1	250	500
ZEUS	250	250
HERMES	100	300
<b>Total</b>	<b>600</b>	<b>1050</b>

Different strategies visible



# DP @ DESY: Software preservation & validation: sp-system

arXiv:1310.7814

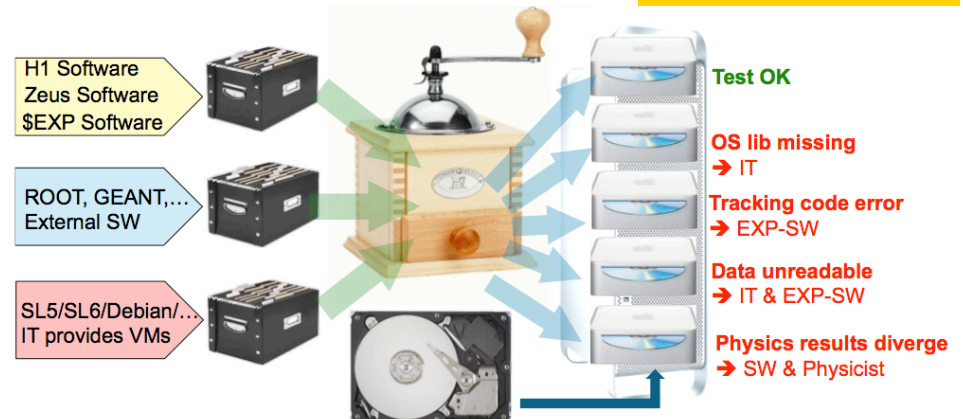
- > Fairly early on, H1 and HERMES decided to try to **migrate their software for as long as possible** rather than **freezing the current data and environment**

- > *Briefly:* The idea of the sp-system is to help perform migrations to newer software versions and environments, where transitions are performed often and validated by a comprehensive set of tests provided by the expts

- Idea is not to run analysis within the system itself
- The output of such a system is a recipe for deployment on (future) external resource(s)
- Future analysis resources maybe local batch farm, grid, cloud, whatever

- > Ambitious project, which showed what may be possible with enough effort

- Due to available resources and changes in personnel, the implementation was slower than expected and system could not be fully deployed
- Still facilitated the transition from SL5 (32 bit) to SL6 (64 bit)



# DP @ DESY: Common Ntuples (ZEUS)

## > Motto: keep it simple!

- Flat (simple) ROOT-based ntuple (same format as PAW ntuple converted with h2root) containing high level objects (electrons, muons, jets, energy flow objects, ...) as well as low level objects (tracks, CAL cells, ...)

## > Well tested!

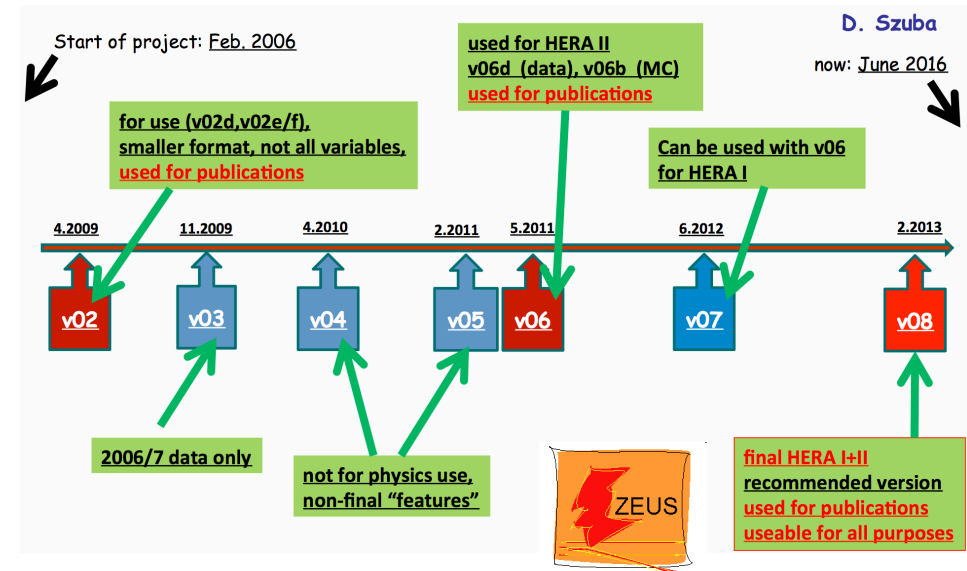
- Almost all recent ZEUS papers based on Common Ntuples

## > Easy to use

- Several recent ZEUS papers based on results produced by Master students from remote institutes, using resources at DESY
- PhD students can produce a ZEUS paper within only a fraction of their PhD time (e.g. ~6 months - 1 year)

## > Parallel, fall back solution for H1 and HERMES when software migration restricted or no longer possible

Slide credit: A. Geiser (ZEUS/DESY)



# Lessons learned from DP@DESY applicable to LHC and beyond

- The **physics output** tail seen by LEP also became true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**



# Lessons learned from DP@DESY applicable to LHC and beyond

- The **physics output** tail seen by LEP also became true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**
- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1 HERA-1 took 3 years to get 3 months of work)



# Lessons learned from DP@DESY applicable to LHC and beyond

- The **physics output** tail seen by LEP also became true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**
- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1 HERA-1 took 3 years to get 3 months of work)
- Best to **avoid dedicated material solutions**: use what's currently available. And who knows what this may look like in the future, so keep as generic and flexible as possible!



# Lessons learned from DP@DESY applicable to LHC and beyond

- The **physics output** tail seen by LEP also became true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**
- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1 HERA-1 took 3 years to get 3 months of work)
- Best to **avoid dedicated material solutions**: use what's currently available. And who knows what this may look like in the future, so keep as generic and flexible as possible!
- There is a **great reduction** in person power (and available expert knowledge) as well as funding for an experiment as soon data taking stops. Budgets become much tighter towards the end, competing with other projects that are just beginning





# Lessons learned from DP@DESY applicable to LHC and beyond

- The **physics output** tail seen by LEP also became true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**
- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1 HERA-1 took 3 years to get 3 months of work)
- Best to **avoid dedicated material solutions**: use what's currently available. And who knows what this may look like in the future, so keep as generic and flexible as possible!
- There is a **great reduction** in person power (and available expert knowledge) as well as funding for an experiment as soon data taking stops. Budgets become much tighter towards the end, competing with other projects that are just beginning
  - **Don't start too late**, projects should be well in place before data taking ends



# Lessons learned from DP@DESY applicable to LHC and beyond

- The **physics output** tail seen by LEP also became true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**
- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1 HERA-1 took 3 years to get 3 months of work)
- Best to **avoid dedicated material solutions**: use what's currently available. And who knows what this may look like in the future, so keep as generic and flexible as possible!
- There is a **great reduction** in person power (and available expert knowledge) as well as funding for an experiment as soon data taking stops. Budgets become much tighter towards the end, competing with other projects that are just beginning
  - **Don't start too late**, projects should be well in place before data taking ends
  - **Don't underestimate the required person-power**: for funding or practical reasons



# Lessons learned from DP@DESY applicable to LHC and beyond

- The **physics output** tail seen by LEP also became true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**
- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1 HERA-1 took 3 years to get 3 months of work)
- Best to **avoid dedicated material solutions**: use what's currently available. And who knows what this may look like in the future, so keep as generic and flexible as possible!
- There is a **great reduction** in person power (and available expert knowledge) as well as funding for an experiment as soon data taking stops. Budgets become much tighter towards the end, competing with other projects that are just beginning
  - **Don't start too late**, projects should be well in place before data taking ends
  - **Don't underestimate the required person-power**: for funding or practical reasons
  - **Dedicated manpower** is needed, people working on this part time or in spare time is not enough: such initiatives cannot “run for free”



# Lessons learned from DP@DESY applicable to LHC and beyond

- The **physics output** tail seen by LEP also became true for the experiments at HERA, where there is much physics output in the years after data taking stopped
- In addition, the final data for preservation is not ready immediately after data taking
- Data volume, when the final data are available, **may not be such a decisive issue**
- Getting all of the data for preservation **to the same level** is quite some work but absolutely necessary: OS, software version (in house and external), calibrations, methodologies..
- This should be started **as soon as possible** (H1 HERA-1 took 3 years to get 3 months of work)
- Best to **avoid dedicated material solutions**: use what's currently available. And who knows what this may look like in the future, so keep as generic and flexible as possible!
- There is a **great reduction** in person power (and available expert knowledge) as well as funding for an experiment as soon data taking stops. Budgets become much tighter towards the end, competing with other projects that are just beginning
  - **Don't start too late**, projects should be well in place before data taking ends
  - **Don't underestimate the required person-power**: for funding or practical reasons
  - **Dedicated manpower** is needed, people working on this part time or in spare time is not enough: such initiatives cannot “run for free”
  - Losing the best people for the best roles is almost inevitable and finding support for unfinished things is extremely difficult. Difficult to capture the best candidates without providing a **long term perspective**



# Moving on to the present: The Large Hadron Collider @ CERN

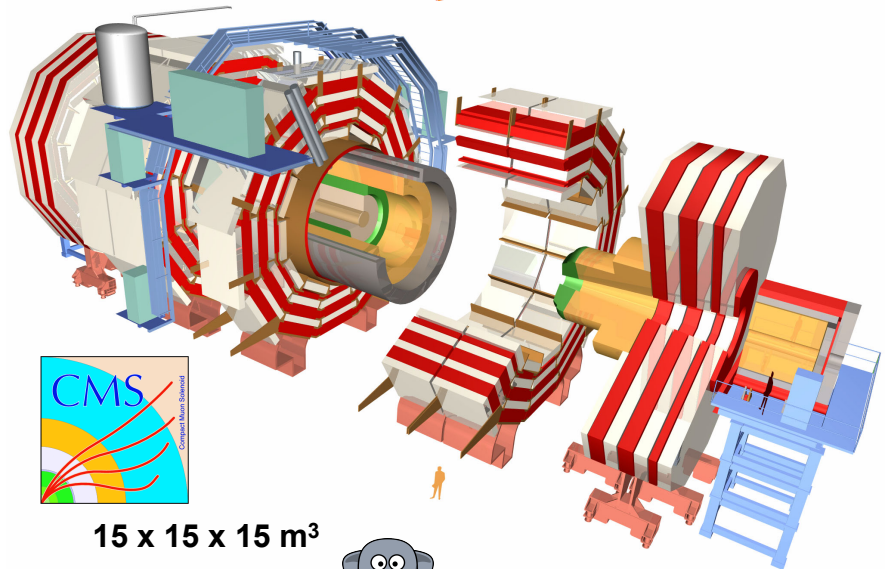
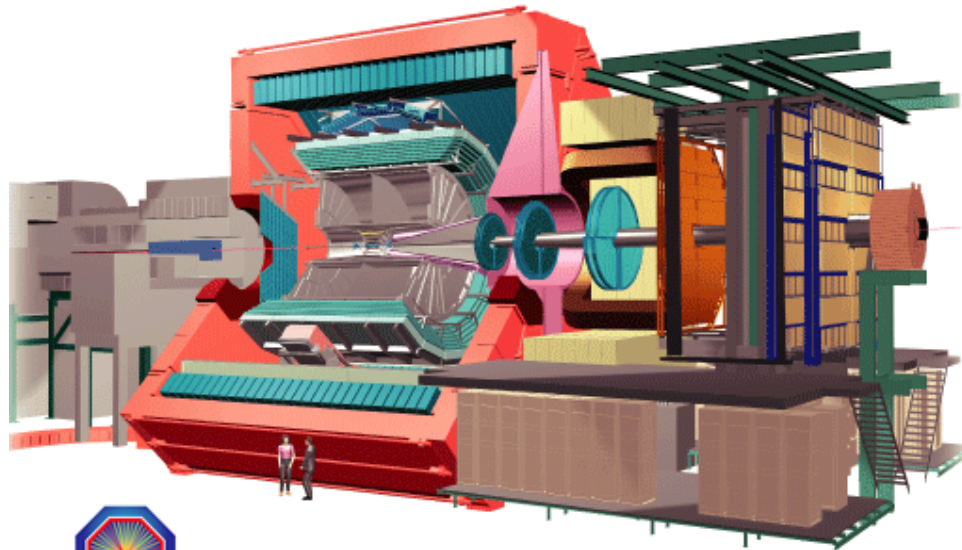
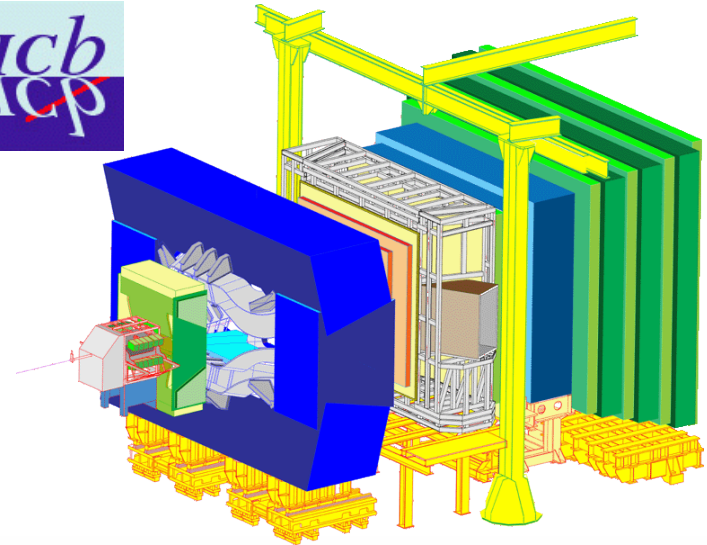
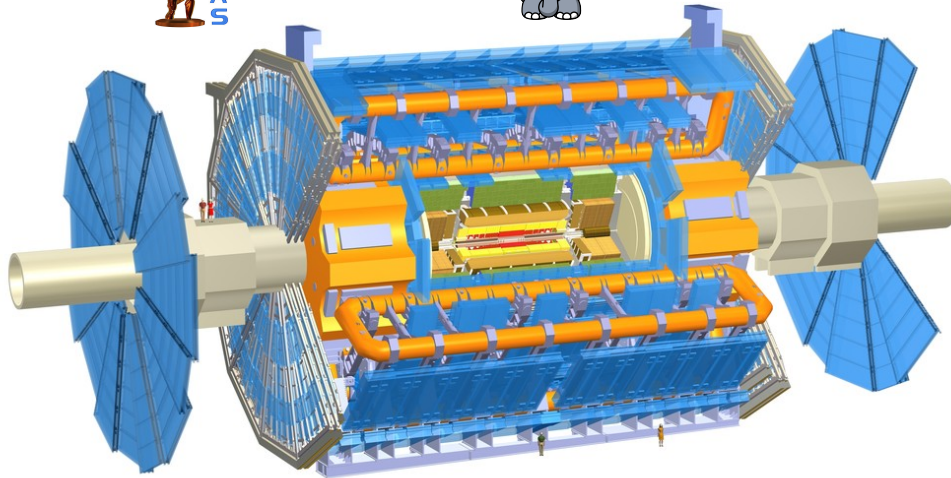


# Four experiments with some large detectors..



22 x 22 x 45 m<sup>3</sup>

7,000t = 1400 x



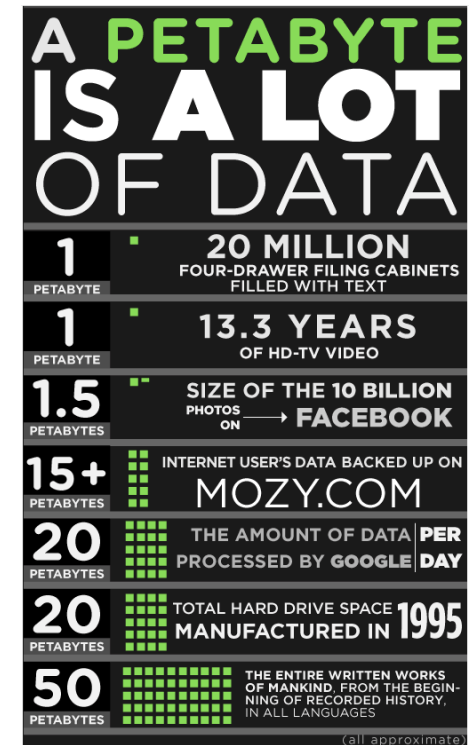
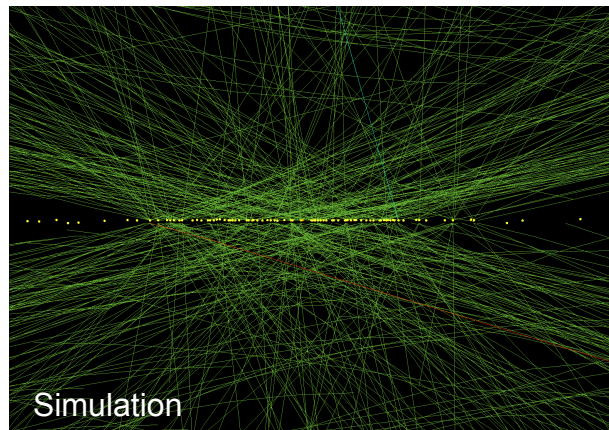
15 x 15 x 15 m<sup>3</sup>

12,500t = 2500 x



## ..which will record A LOT of data

- > Reminder: **100TB** per LEP experiment, **1-10PB** for experiments at the HERA collider at DESY, the TeVatron at Fermilab or BaBar experiment at SLAC
- > The LHC experiments are already in the several hundred PB range (**x00PB**)
- > This will increase to **10EB** or more including the High Luminosity upgrade of the LHC (HL-LHC)
- > It's also worth noting that we throw away the vast majority of the events at the very first opportunity, which allows us to write out those we want to keep
  - **ATLAS high level trigger** writes out at up to 1 kHz



Source: <https://visual.ly/how-much-petabyte>

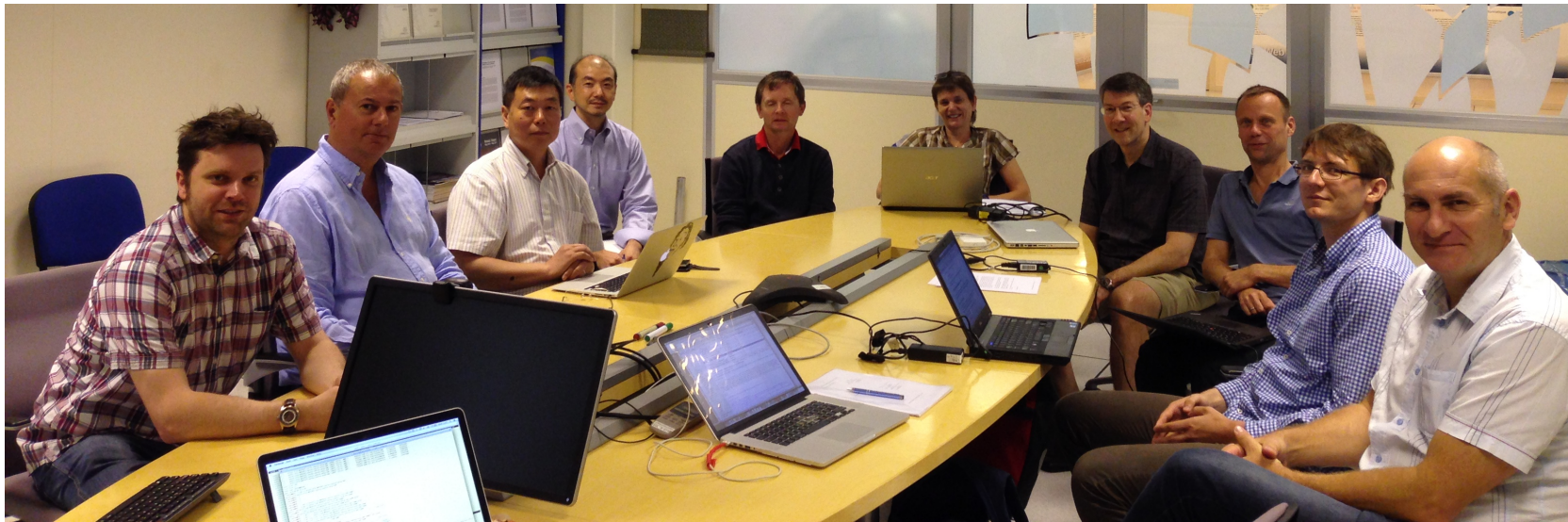
LHC has bunch crossings every 25ns, i.e. a rate of 40 MHz

HL-LHC pile-up of 78, would mean 3 billion events every second!



# DPHEP has made the transition to a Collaboration

- > Following on from the activities of the Study Group, the **DPHEP Collaboration Agreement** was signed in 2014 by the initial partners: CERN, DESY, HIP (Finland), IHEP (China), IN2P3, IPNS (Japan), MPP
  - Additional partners from the Study Group intending to join: BNL, CSC (Finland), FNAL, IPP (Canada), INFN, SLAC, STFC,...



- > **First Collaboration Meeting at CERN, and Collaboration Board, June 2015**
  - Just to note: At this point, this all may like seem an obvious need.. but a few years ago such cooperation between experiments, labs, groups was simply not there



# The road forward for DPHEP

arXiv:1512.02019

- > Attempting to *learn from those lessons* of the pre-LHC experiment, a "**2020 vision**" is established in a new publication, declaring:
  - All archived data described in the 2012 DPHEP publication, as well as LHC data, should be easily **findable** and fully usable by the **designated communities** with clear (open) access policies and possibilities to annotate further;
  - Best practices, tools and services should be well run-in, **fully documented** and **sustainable**; built in common with other disciplines, based on standards;
  - There should be a DPHEP **portal**, or **portals**, through which data / tools accessed;
  - Clear **targets & metrics** to measure the above should be agreed between **funding agencies, and the experiments**
- > For the LHC experiments, this is being formalised as **Data Management Plans**, "DMPs", following advice from funding agencies, where:
  - The DMP should describes how data generated through the course of the proposed research will be **shared and preserved** or explains why data sharing and/or preservation are not possible or scientifically appropriate
  - The DMP should also describe how data sharing and preservation will enable **validation of results**, or how results could be validated if data are not shared or preserved



## Result: The LHC experiments are active in data preservation!

- > All LHC experiments are now very active in this field, and taking ATLAS as an example, the collaboration has produced over the last few years:
  - A policy document outlining the **general principles of data preservation for ATLAS**: the data themselves, data formats and reproducibility of physics results has been prepared. *This in particular I think is a **major achievement** of DPHEP*
  - A note outlining the requirements for preserving ATLAS data **for use by ATLAS**
  - A policy document on **data access** rules, based on the DPHEP preservation levels
  - A note outlining **datasets for outreach purposes and open access**
  - An ATLAS mandate for **analysis preservation**

where much of this work is in collaboration with other experiments / CERN-IT

- > For the last few minutes I want to talk a little about the two currently most active areas of the LHC experiments, both in collaboration with CERN-IT:
  - **Open access to LHC data**
  - **Analysis preservation**



## CERN press office

[Français](#) [English](#)

[Media visits](#)

[Press releases](#)

[For journalists](#)

[Contact us](#)

# CERN makes public first data of LHC experiments

20 Nov 2014

Geneva, 20 November 2014. CERN<sup>1</sup> launched today its Open Data Portal where data from real collision events, produced by the LHC experiments will for the first time be made openly available to all. It is expected that these data will be of high value for the research community, and also be used for education purposes.

*"Launching the CERN Open Data Portal is an important step for our Organization. Data from the LHC programme are among the most precious assets of the LHC experiments, that today we start sharing openly with the world. We hope these open data will support and inspire the global research community, including students and citizen scientists,"* said CERN Director General Rolf Heuer.

The principle of openness is enshrined in CERN's founding Convention, and all LHC publications have been published Open Access, free for all to read and re-use. Widening the scope, the LHC collaborations recently approved Open Data policies and will release collision data over the coming years.

The first high-level and analysable collision data openly released come from the CMS experiment and were originally collected in 2010 during the first LHC run. This data set is now publicly available on the CERN Open Data Portal. Open source software to read and analyse the data is also available, together with the

### CONTACT PRESS OFFICE

[press.office@cern.ch](mailto:press.office@cern.ch)

+41 (0)22 767 34 32

+41 (0)22 767 21 41

+41 (0)22 767 41 01

### CERN PEOPLE: SIGN IN FOR MORE RESOURCES

Sign in to see more resources for [CERN people](#)

### FOLLOW CERN

[CERN Twitter feed](#)

[CERN TV](#)

[Quantum diaries blog](#)

[CERN Press office Twitter feed](#)

### RESOURCES

[Images](#)



## CERN press office

Français English

- Media visits
- Press releases
- For journalists
- Contact us

## CERN makes public first data of LHC experiments

### CONTACT PRESS OFFICE

[press.office@cern.ch](mailto:press.office@cern.ch)  
+41 (0)22 767 34 32  
+41 (0)22 767 21 41

20 Nov 2014

Geneva, 20 November 2014. CERN data produced by the LHC experiments will be of high value for the rest of the world.

"Launching the CERN Open Data Portal are among the most precious assets we have. We hope these open data will support scientists," said CERN Director General Fabrice Gianfranceschi.

The principle of openness is ensured by published Open Access, free for all approved Open Data policies and with the support of the Open Access Policy Group.

The first high-level and analysable data were originally collected in 2010 during the LHC Run 1. The Open Data Portal is an Open Access Data Portal. Open source software

opendata CERN

ABOUT SEARCH EDUCATION RESEARCH

### Education

Visualise events, check reconstructed data, run tools or build your own!

Start learning

### Research

Get the genuine working environments, virtual machines and datasets to start your research

Start analysing

Home > Data-Policies

## Data-Policies

This collection contains data policies.

### ATLAS Data Access Policy

This document contains the policy document regarding the access to ATLAS data by non-ATLAS members which was endorsed by the ATLAS Collaboration Board in June 2014.

Collection Data-Policies DOI 10.7483/OPENDATA.ATLAS.T9YR.Y7MZ

### ALICE data preservation strategy

This document contains the ALICE data preservation strategy and policy.

Collection Data-Policies DOI 10.7483/OPENDATA.ALICE.54NE.X2EA

### CMS data preservation, re-use and open access policy

This document describes the CMS collaboration's policy on long-term data preservation, re-use and open access. The policy has been approved by the CMS Collaboration Board in March 2012.

Collection Data-Policies DOI 10.7483/OPENDATA.CMS.UDBF.JKR9

### LHCb External Data Access Policy

This document contains the LHCb Data Access Policy. This was adopted at the Collaboration Board meeting on 27th Feb 2013.

Collection Data-Policies DOI 10.7483/OPENDATA.LHCb.HKJW.TWSZ Author Clarke, Peter

CERN

Media visit

CERN

exper

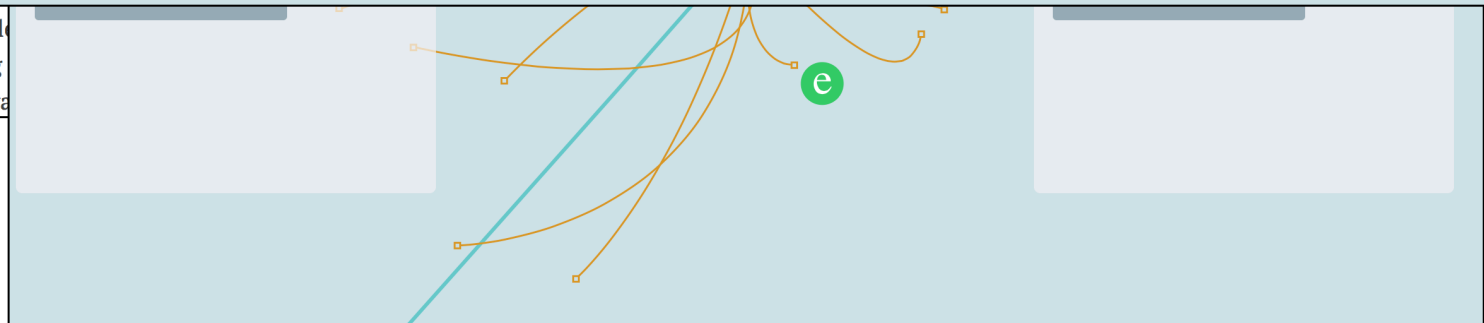
20 Nov 2014

Geneva, 20  
produced b  
data will be

"Launching  
are among  
hope these  
scientists,"

The princip  
published C  
approved C

The first high-level and analysabl  
originally collected in 2010 during  
Data Portal. Open source softwa



## Education



The CMS (Compact Muon Solenoid) experiment is one of two large general-purpose detectors built on the Large Hadron Collider (LHC). Its goal is to investigate a wide range of physics such as the characteristics of the Higgs boson, extra dimensions or dark matter.

[Explore CMS >](#)



ALICE

ALICE (A Large Ion Collider Experiment) is a heavy-ion detector designed to study the physics of strongly interacting matter at extreme energy densities, where a phase of matter called quark-gluon plasma forms. More than 1000 scientists are part of the collaboration.

[Explore ALICE >](#)



The ATLAS (A Toroidal LHC ApparatuS) experiment is a general purpose detector exploring topics like the properties of the Higgs-like particle, extra dimensions of space, unification of fundamental forces, and evidence for dark matter candidates in the Universe.

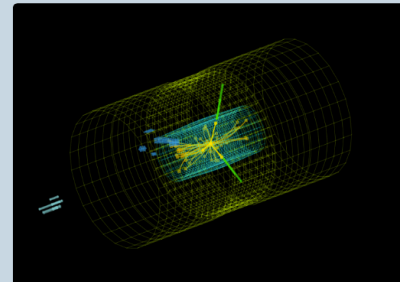
[Explore ATLAS >](#)



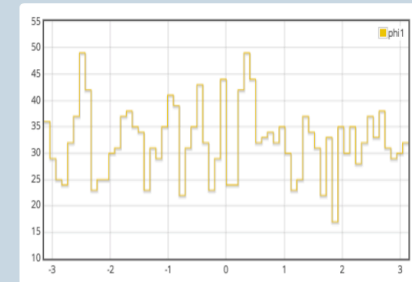
The LHCb (Large Hadron Collider beauty) experiment aims to record the decay of particles containing b and anti-b quarks, known as B mesons. The detector is designed to gather information about the identity, trajectory, momentum and energy of each particle.

[Explore LHCb >](#)

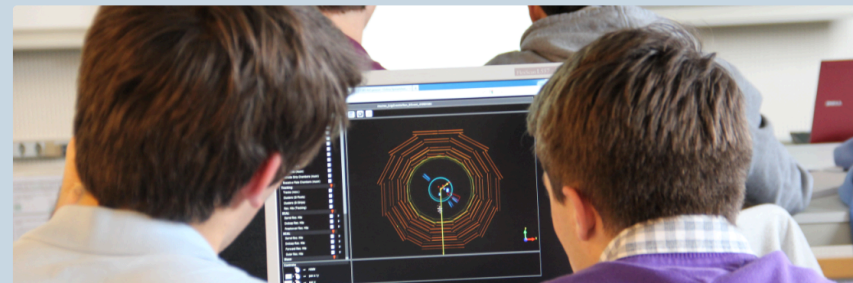
For education purposes, the complex primary data need to be processed into a format (examples below) that is good for simple applications. Get in touch if you wish to build your own applications similar to those shown here



[Visualise events >](#)



[Visualise histograms >](#)

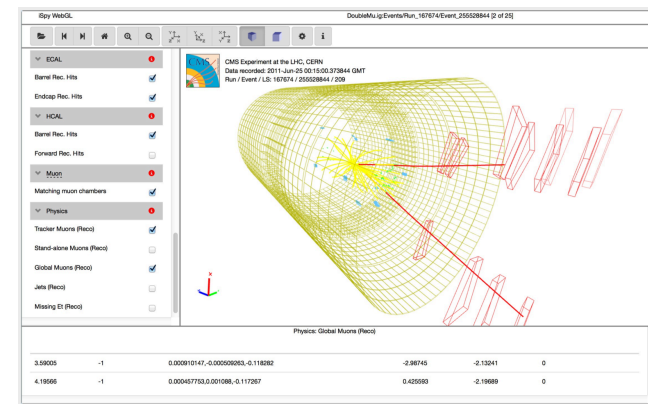


[Learning Resources >](#)

# CMS Open Data

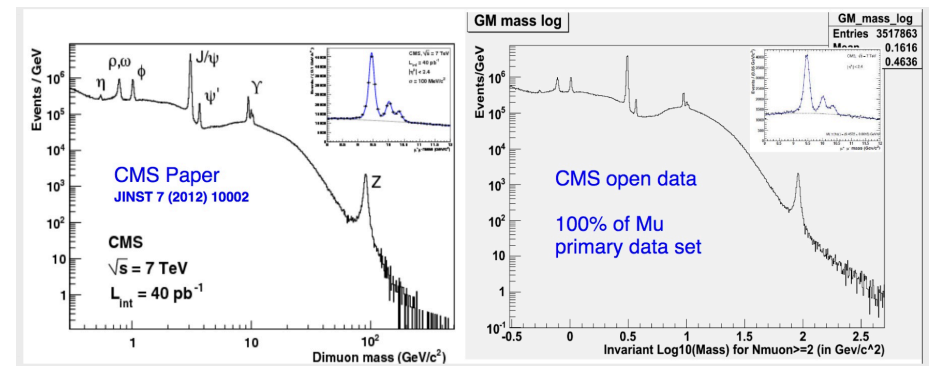
- Hosted directly in the CERN portal, CMS has actually released up to “level 3” data, the same AODs used in analysis
  - Nov 2014: half of 2010 pp collision data at  $\sqrt{s} = 7$  TeV released, 27 TB in size
  - April 2016: half of 2011 pp collision data at  $\sqrt{s} = 7$  TeV released, 100 TB in size, together with 200 TB of Monte Carlo samples

- Datasets maybe visualised using an interactive event display



- Examples provided with detailed instructions, e.g. produce the di-muon spectrum from a CMS 2010 dataset <http://opendata.cern.ch/record/560>

- Results from open data comparable to those in the CMS publication

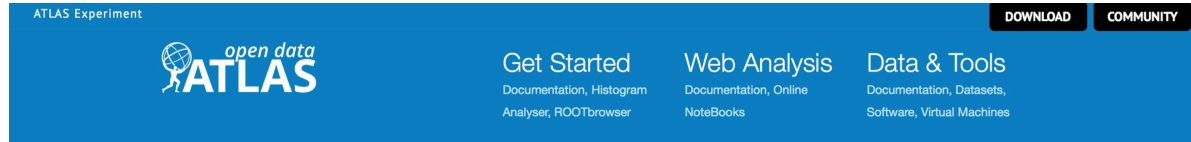


# ATLAS Open Data

<http://atlasopendata.web.cern.ch>  
(currently linked from the CERN Open Data Portal)

> Initial focus: undergraduate and postgraduate students (but eventually to expand target audience)

- Activities from visualizations, to web analysis, to more complex analysis
- Also provides a “**Open Data Community**” for users to interact within and share their experiences

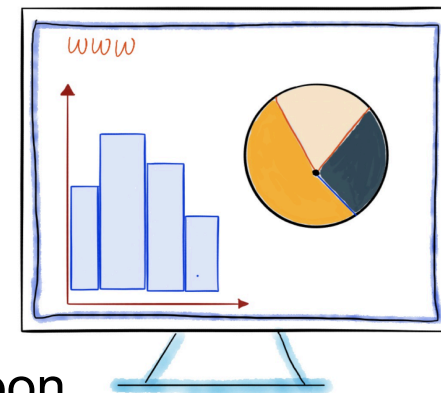


## Level 1: Get Started

Physicists at the ATLAS Experiment visualise collision data with histograms. They are used in every publication, from simple analyses to headline-making discoveries. In this section, you will learn how the data is visualised.

### Explore:

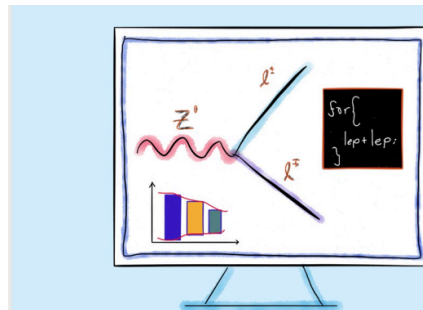
- **Documentation:** a step-by-step guide to using Histogram Analyser and ROOTbrowser
- **Histogram Analyser:** a web based tool for fast, cut-based analysis of data. Visualise data using online histograms
- **ROOTbrowser:** a web based tool for displaying and analysing data. Visualise data online
- **Live events:** see live events from the ATLAS experiment



> Fraction of ATLAS 2012

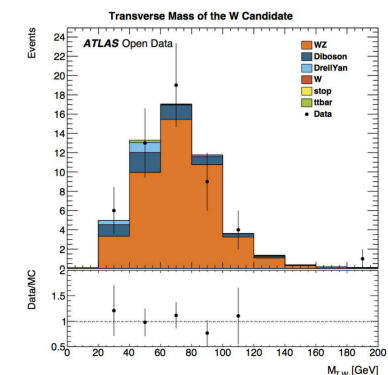
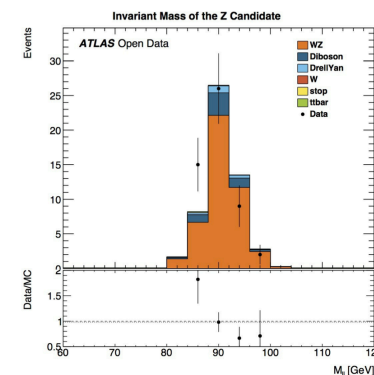
$\sqrt{s} = 8$  TeV data released, to be followed by 13 TeV soon

> Example: Understanding Z bosons



## The Z Analysis ROOTbook

Many analyses selecting leptons suffer from Z + jets as a contributing background due to its large production cross section. It is therefore vital to check the correct modelling of this process by the Monte-Carlo simulated data. It is important to measure well known Standard Model particles, to confirm that we understand properly the detector and software. We are then ready to search for new physics.



Excellent ICHEP 2016 talk on LHC Open data:

<http://indico.cern.ch/event/432527/contributions/2205592/attachments/1321257/1981477/mccauley-opendata-ichep2016.pdf>





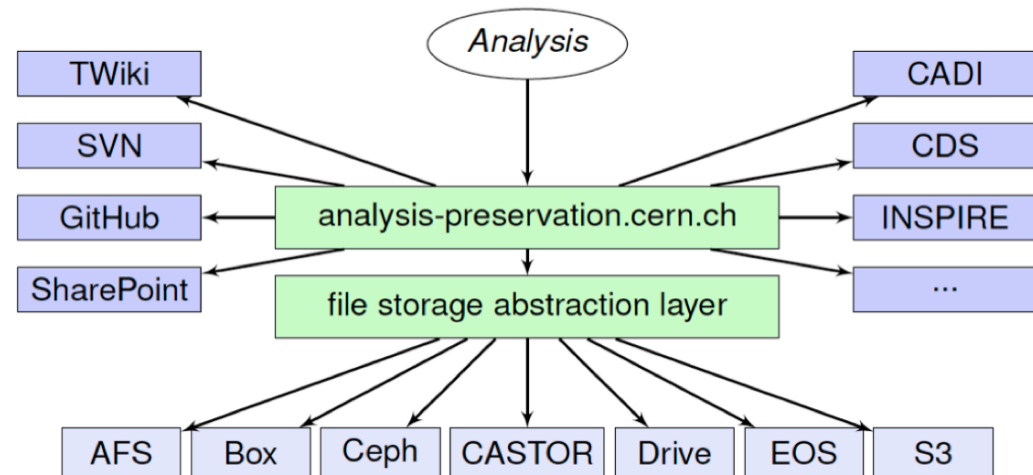
# Analysis preservation

- > A relatively new concept and initiative
- > It is clearly desirable to be able to “*preserve an analysis*” for the future, to fully *encapsulate* **what was done at the time** into an easy to understand and deploy package **for the host collaboration**
  - What is *not* primarily understood as Analysis Preservation is the fully flexible, level 3 or 4 data preservation programme, or the release of the data/software for use by non-collaboration members
- > There are many identified reasons to do this, including:
  - To address concerns about published analyses, internal or otherwise
  - To assist in knowledge transfer if a person leaves the collaboration and has to hand over the know-how to other members To compile a comprehensive set of metadata concerning a presentation or publication which is to be submitted for internal review
  - To allow an existing analysis to be reinterpreted for a new model
  - To allow an existing analysis to be repeated, which may be desirable due to improved precision, combination with new data, interaction with theory



# Analysis preservation: The CAP

- > There are two complementary strategies to analysis preservation
  - To capture in as much detail as possible a description of the analysis, to provide the possibility to recreate the analysis, and thus reproduce and/or re-use the analysis
  - Encapsulating the actual code and workflow that was used for the analysis organised so that it can be re-run exactly as before, to faithfully reproduce the analysis and provide opportunities for reinterpretation
- > Here again CERN-IT is providing central infrastructure to all LHC experiments, via the CERN Analysis Preservation (CAP) portal
  - The work of the experiments is to identify all of the many resources holding information about an analysis
  - The work of CERN-IT revolves around being able to talk to those many resources and to design an interface applicable to the workflow of the experiments



# Analysis preservation: A sneak preview of the CAP

**Basic Information** ▼

**Analysis Number**

**AOD Processing** ▼

Provide AOD Processing information

**Primary Datasets**

**Monte Carlo Datasets**

**Selection Triggers**

**Physics Information** ▼

Provide information about datasets, triggers, physics objects, etc

Item #1 + - ^ v

**Additional Information**

**Number of Events**

**pt\_hat**

**Collision Energy**

**Collision Species**

- None
- PbPb
- pPb
- PP



# Analysis preservation: A sneak preview of the CAP

### Basic Information

Analysis Number

### AOD Processing

Provide AOD Processing information

Primary Datasets

Monte Carlo Datasets

Selection Triggers

### Physics Information

Provide information about datasets, triggers, physics objects, etc

Item #1	
Additional Information	<input type="text"/>
Number of Events	<input type="text"/>
$p_{t\_hat}$	<input type="text"/>
Collision Energy	<input type="text"/>
Collision Species	<input type="text"/>
	<input type="radio"/> None
	<input type="radio"/> PbPb
	<input type="radio"/> pPb
	<input type="radio"/> PP

### Documentations

Provide documentation and other things

Item #1	
Comment	<input type="text"/>
Keyword	<input type="text"/>
Reference ID	<input type="text"/>
URL	<input type="text"/>

### Internal Discussions

Add Internal Discussions

Item #1	
URL	<input type="text"/>

### Presentations

Add Presentations

Item #1	
URL	<input type="text"/>

### Publications

Add publications

Item #1	
Editorial Board	<input type="button" value="+ Add New Item"/>
Full Title	<input type="text"/>
Reference Code	<input type="text"/>
Short Title	<input type="text"/>
Journal Title	<input type="text"/>
Journal Year	<input type="text"/>
Journal Volume	<input type="text"/>
Journal Issue	<input type="text"/>
Journal Page	<input type="text"/>

# Analysis preservation: A sneak preview of the CAP

**CERN Analysis Preservation** Create new analysis sunje@cern.ch

ATLAS | Analyses | Analysis 1

## ATLAS SUSY EQ 2L (e/mu)

Searches for direct production of charginos, neutralinos, and sleptons in final states with leptons and missing transverse momentum in pp collisions at  $\sqrt{s} = 8\text{TeV}$  with the ATLAS detector.  
Started Thursday 20th March 2014

**Overview** | Publications | Files | Workflow | Measurements | Contributors | RECAST

1 Publication

Searches for direct production of charginos, neutralinos, and sleptons in final states with leptons and missing transverse momentum in *Eur.Phys.J. C76 (2016) 451, 2016*  
DOI 10.1140/epjc/s10052-016-4286-3

23 Files


SLHA	3.24MB
Figure 2A	3.24MB
Figure 2B	3.24MB

[View More](#)

Links

- INSPIRE
- HEPData

Workflow



```
graph TD; RawData[Raw Data] --> Reco[Reconstruction]; Reco --> Selection[Selection]; Selection --> Simulation[Simulation]; Simulation --> Histograms[Histograms]; Histograms --> Fitting[Fitting]; Fitting --> Results[Results];
```

1 Measurement

### SUSY EQ 2L (e/mu)

Searches for direct production of charginos, neutralinos, and sleptons in final states with leptons and missing transverse momentum in pp collisions at  $\sqrt{s} = 8\text{TeV}$  with the ATLAS detector.

# Summary

- > The DPHEP Collaboration is now well established in high energy physics
- > The **early experiences** of the pre-LHC experiments in the DPHEP Study Group phase were crucial in shaping the recommendations for present and future experiments
- > There is now much activity at the LHC in a diverse range of areas concerning **data preservation**, **analysis preservation** and providing **open access** to the data themselves
- > **The main message is that it is never early to consider data preservation: early planning is likely to result in cost savings that may be significant**
  - Furthermore, resources (and budget) beyond the data-taking lifetime of the projects must be foreseen from the beginning
- > Align with the overall strategy and even implementation of other data preservation activities at your institute / laboratory or globally
- > Adopt mainstream and supported technologies wherever possible
- > Understand the target communities for your data preservation activities, the use cases and the expect benefits and outcomes
- > Try to understand the costs – in particular those that are specific to your collaboration (and not “external” – e.g. host laboratory bit preservation services)

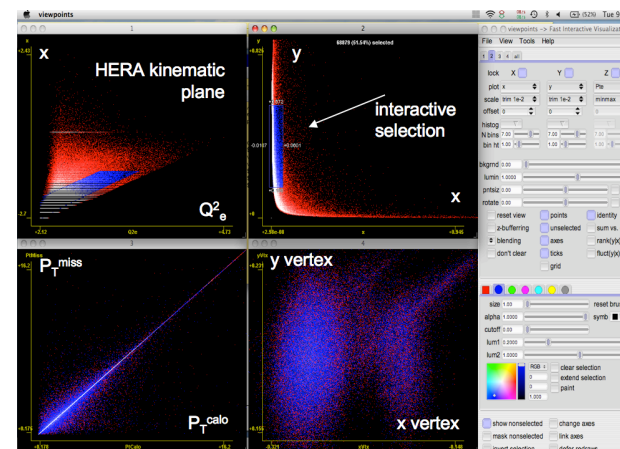
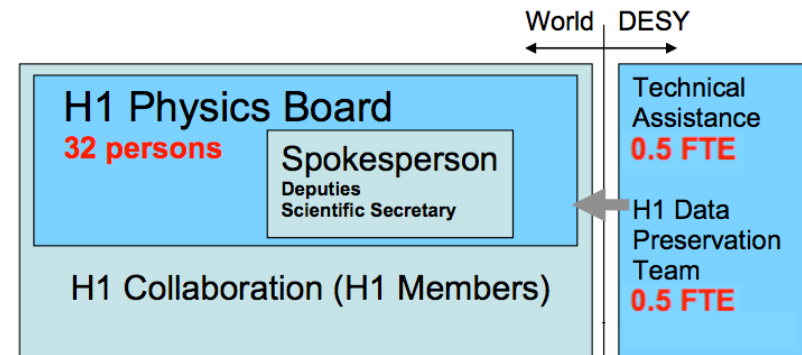


# EXTRAS



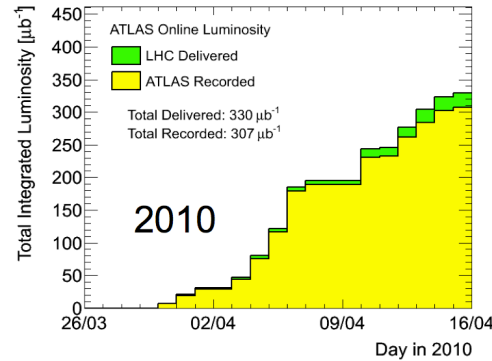
# DP @ DESY: Governance, open access and outreach

- H1 collaboration moved to a new management model in July 2012
  - Formation of *H1 Physics Board*, to replace Collaboration Board (institute based)
  - Future author list policies also set down in new constitution approved by collaboration
- ZEUS and HERMES management teams retain same model as before, but similarly to H1 the collaborating institute layer is now removed
  - Remaining physics ZEUS working groups consolidated to a single physics group
- Open access still to be considered and/or defined by the HERA experiments
- Outreach is a great idea, but was not possible without dedicated resources
  - Already dropped in 2011 table shown earlier
  - Ideas existed, but nothing concrete came of it



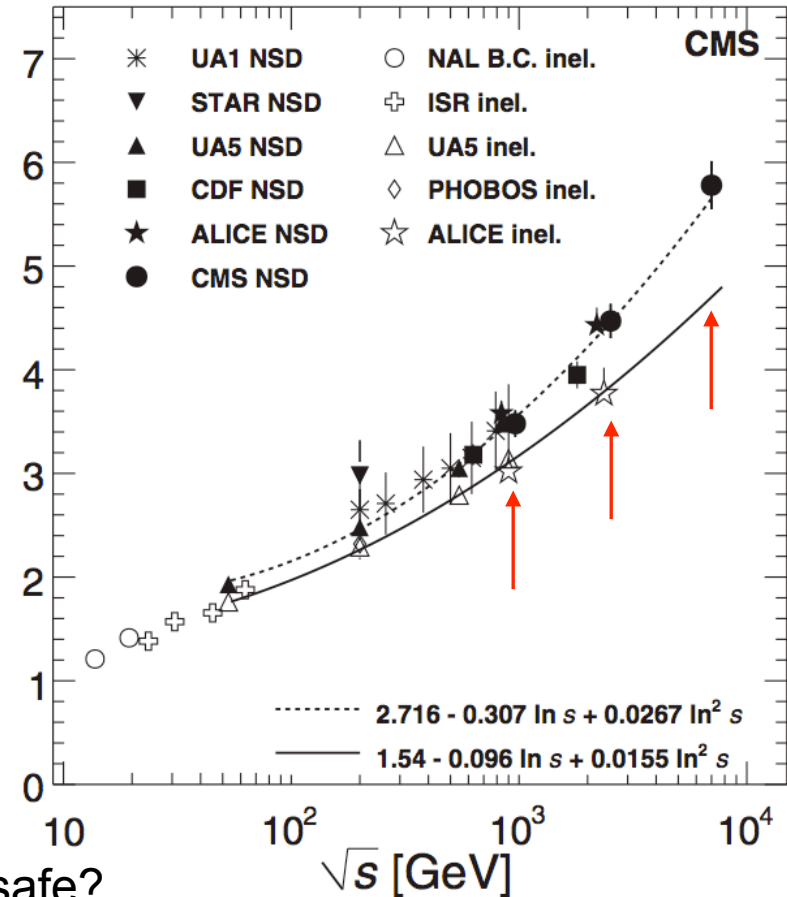


# What about LHC 900 GeV and 2.32 TeV data? 7 TeV data?



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

$$\frac{dN_{ch}}{dn} \Big|_{\eta \approx 0}$$



- > Early LHC measurements made using data at a unique centre of masses
- > Is the 35 pb<sup>-1</sup> of 2010 low pile up 7 TeV data safe?
- > What happens to Run 1 data when the 14 TeV collisions come? Hopefully not something like what happened at the TeVatron..



# Transition scenario and resources at the experimental level

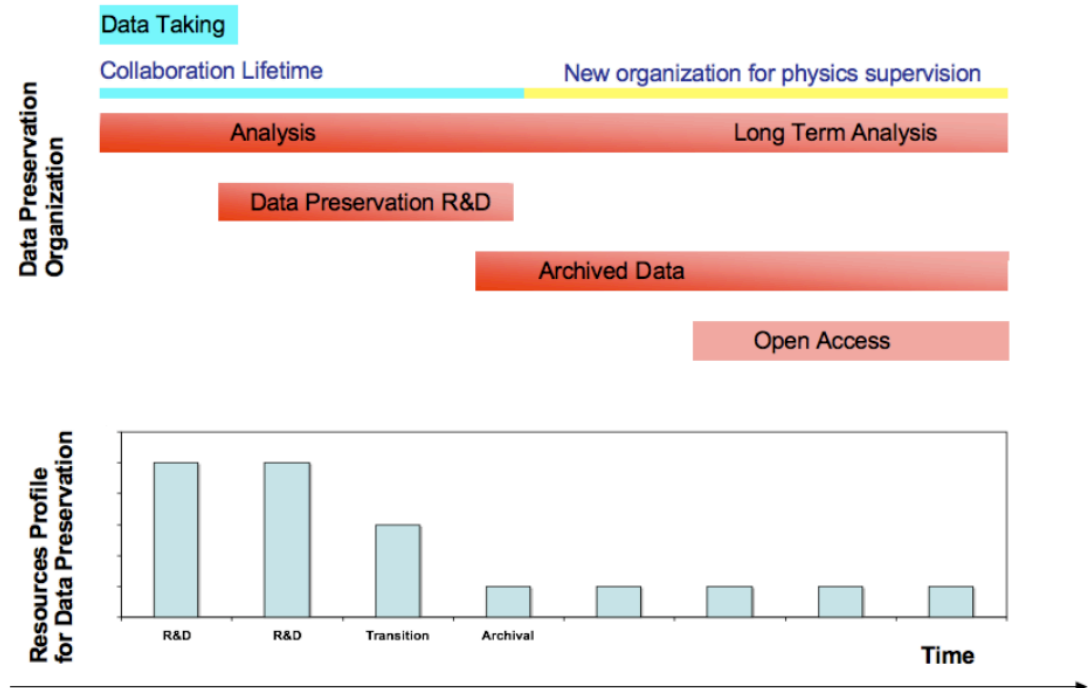
➤ Planning the transition to a long term analysis model

➤ 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!

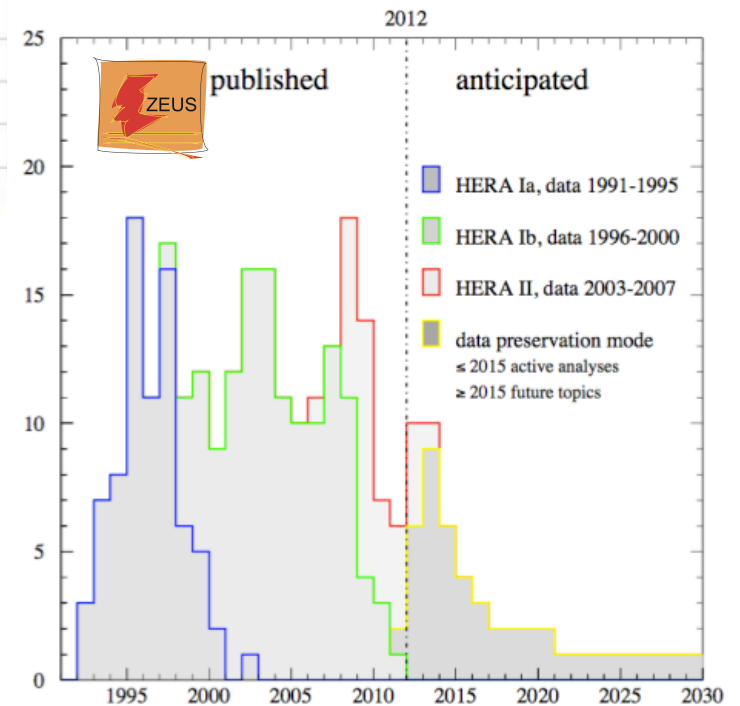
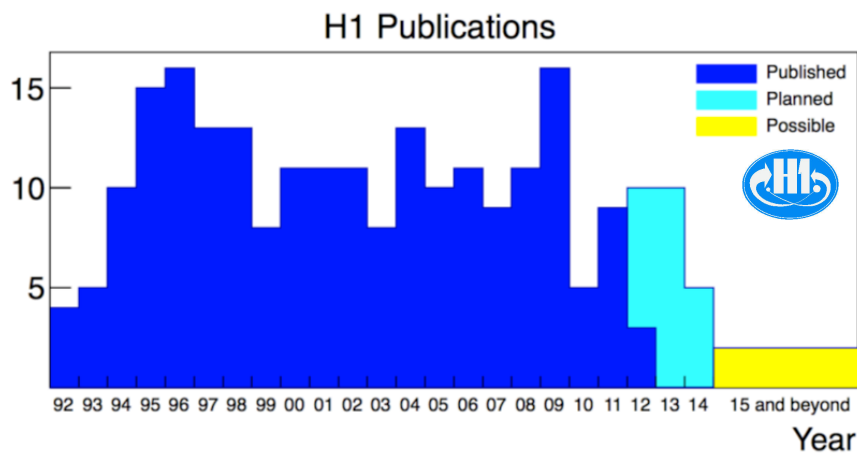
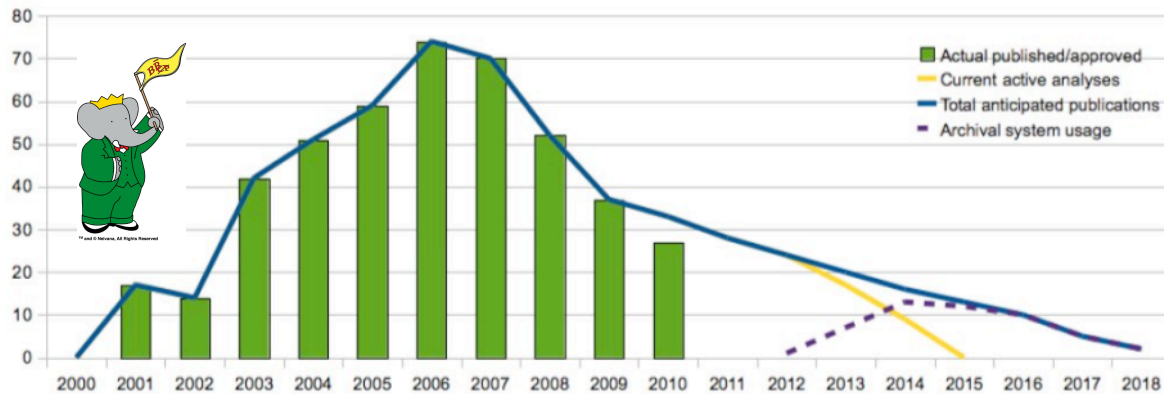


Cost estimates represent typically **much less than 1%** of the original investment

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



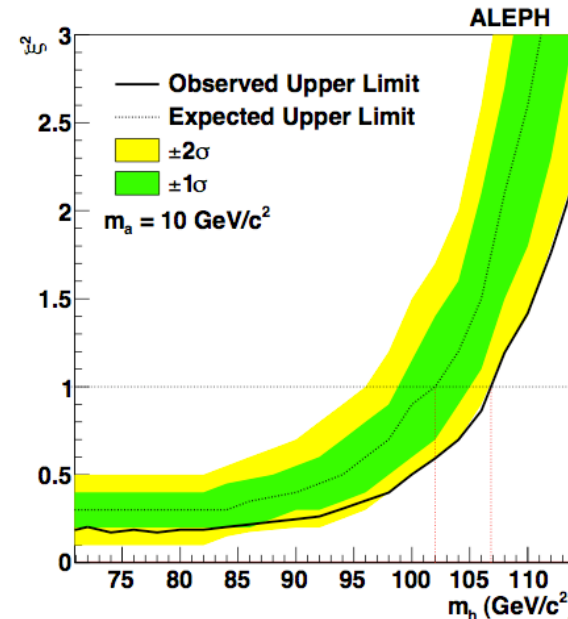
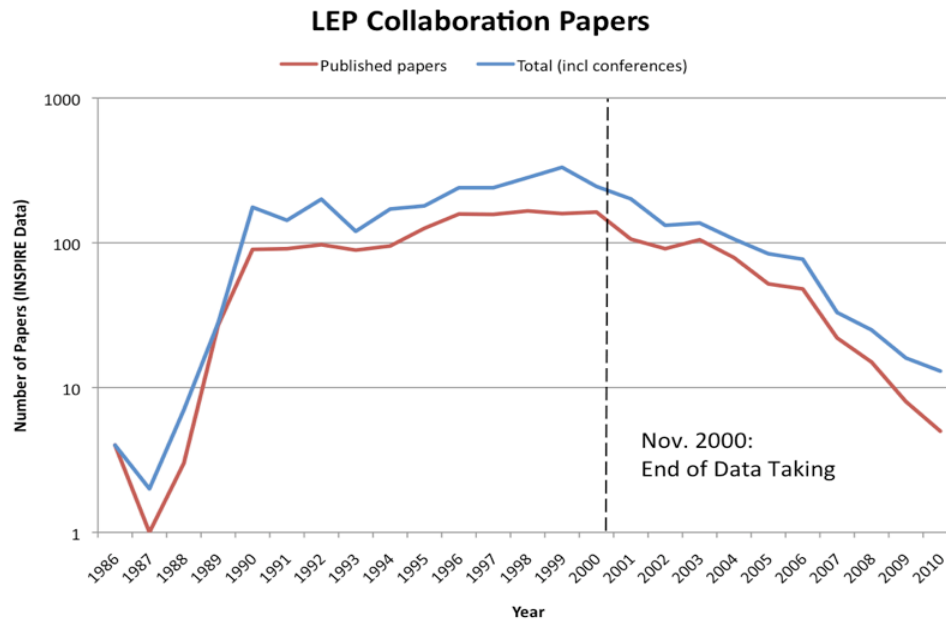
# Long term completion of the physics programme



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



# Long term completion of the physics programme



- 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



# Cross-collaboration combinations of physics results

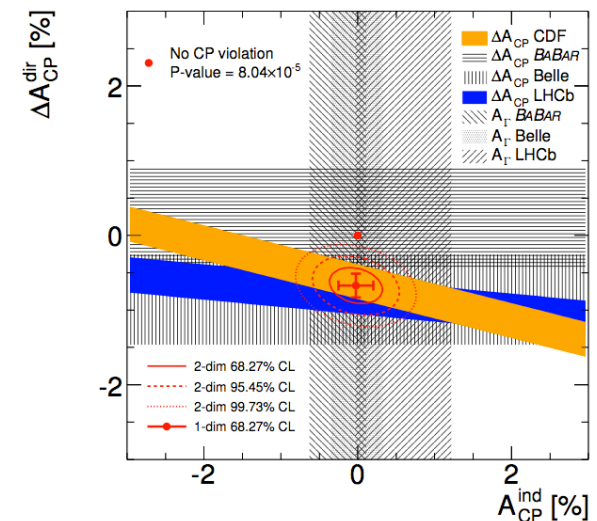
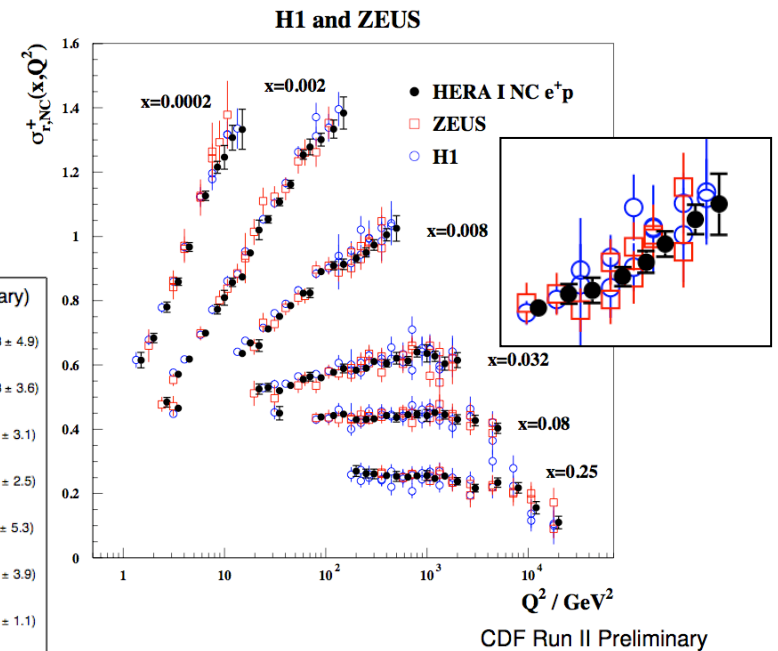
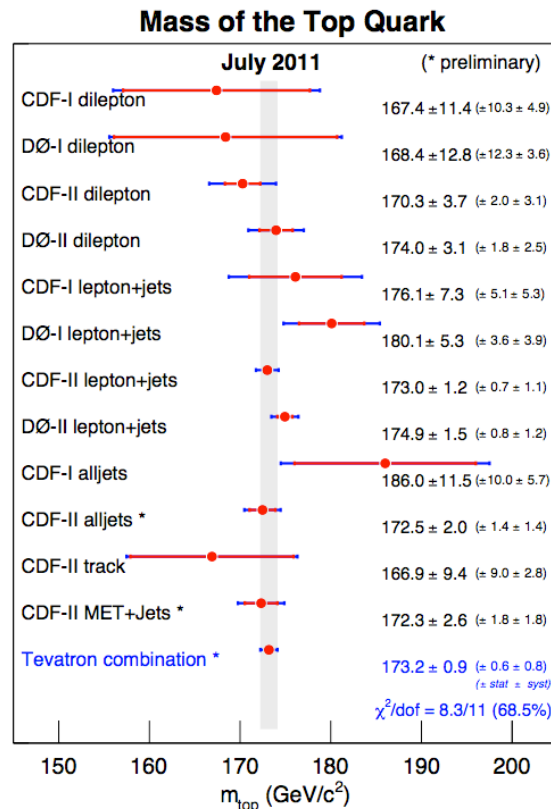
> Combination of data from multiple experiments to produce new scientific results

- Improved precision and increased sensitivity

> Comparison of experimental results

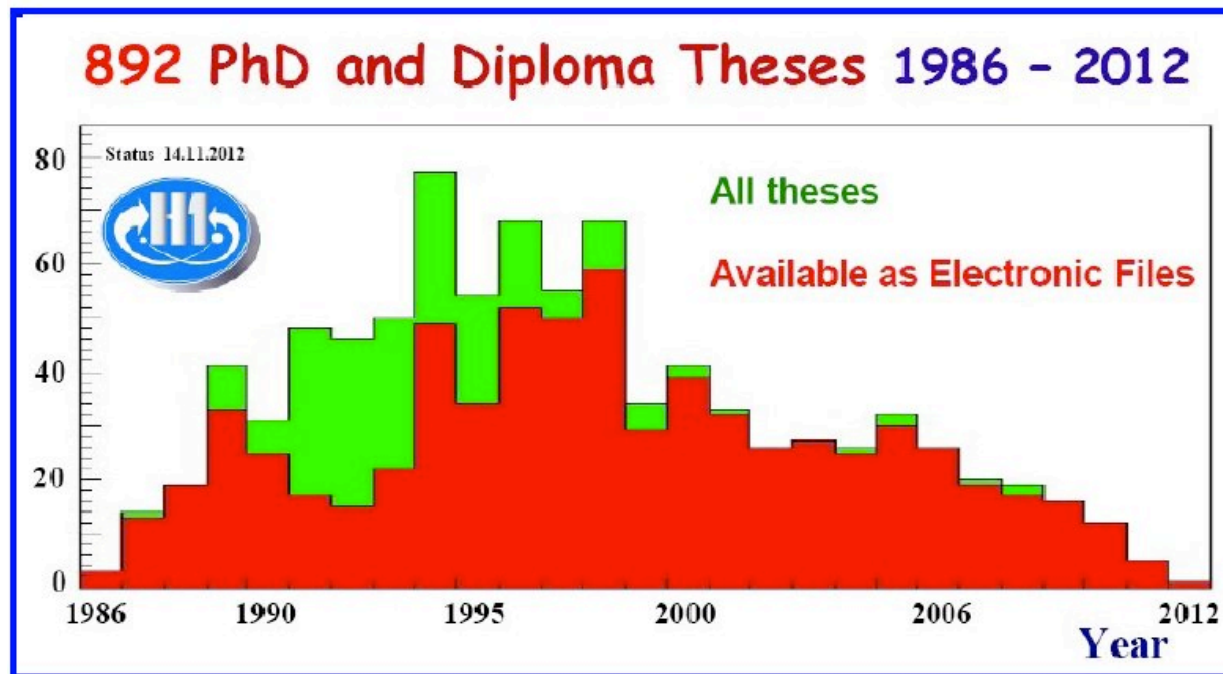
- Complimentary information from different physics
- Verification of experimental observations

> Both objectives facilitated by data preservation



# H1 Theses

- > Since October 2010, **106** H1 theses discovered not previously known to the collaboration; **18** since this summer, latest ones only last week
- > Scanning and linking these to the official H1 pages is given high priority



- > **Currently, of the 892 known H1 theses 197 are not available in electronic form: ~ 22% not available to the H1 community!**



# 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 screenshot shows the INSPIRE website interface. At the top left is the INSPIRE logo with 'HEP' underneath. To the right, a message states: 'Welcome to INSPIRE! INSPIRE is out of beta and ready to replace SP please email us at [feedback@inspirehep.net](mailto:feedback@inspirehep.net)'. Below this is a navigation bar with links for HEP, INST, HELP, SPIRES, and HEPNAMES. The main content area is titled 'ZEUS Internal Notes'. It contains a search instruction: 'Use "find" for SPIRES-style search ([other tips](#))'. There is a search input field with a 'Search' button and links for 'Easy Search' and 'Advanced Search'. A link 'find in ZEUS-IN-10004' is also present. A message reads: 'This collection is restricted. If you are authorized to access it, please click on the Search button.' At the bottom, there is a footer with links for 'HEP', 'Search', and 'Help', and text: 'Powered by [Invenio](#) v1.0.0-rc0+', 'Problems/Questions to [feedback@inspirehep.net](mailto:feedback@inspirehep.net)', and 'Last updated: 19 Oct 2011, 03:15'.



# 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 screenshot shows the INSPIRE login interface. At the top, there is a navigation bar with the INSPIRE logo and a welcome message: "Welcome to INSPIRE! INSPIRE is out of beta and ready to replace SPIRES. Please email us at [feedback@inspirehep.net](mailto:feedback@inspirehep.net)". Below this is a secondary navigation bar with links for HEP, INST, HELP, SPIRES, and HEPNAMES. The main content area is titled "Login" and includes a message: "This collection is restricted. If you think you have right to access it, please authenticate yourself." There are input fields for "Username:" (containing "zeus") and "Password:". A checkbox labeled "Remember login on this computer." is present, along with a "login" button and a link for "(Lost your password?)". A note at the bottom states: "Note: You can use your nickname or your email address to login." The footer contains the text: "HEP: Search: Help Powered by Invenio v1.0.0-rc0+ Problems/Questions to [feedback@inspirehep.net](mailto:feedback@inspirehep.net)".





# 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 screenshot displays the INSPIRE website interface. At the top, there is a navigation bar with the INSPIRE logo and a welcome message: "Welcome to INSPIRE! INSPIRE is out of beta and ready to replace SP...". Below this, a search bar is visible. The main content area shows a search result for "ZEUS Internal Notes" with "10 records found". The results are listed as follows:

- 1. Inclusive-jet production in NC DIS with HERA II.**  
J. Terron C. Glasman, ZEUS-IN-09-004.  
[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)  
[Detailed record](#) - [Similar records](#)
- 2. Three-subjet distributions in neutral current deep inelastic scattering.**  
E. Ron C. Glasman, J. Terron. ZEUS-IN-09-003.  
[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)  
[Detailed record](#) - [Similar records](#)
- 3. 2009 Guide to Funnel: The ZEUS Monte Carlo Production Facility.**  
A. Parenti. ZEUS-IN-09-002.  
[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)  
[Detailed record](#) - [Similar records](#)
- 4. Automated calculation of radiative correction to electron-proton charged current DIS at HERA.**  
I. Marfin. ZEUS-IN-09-001.  
[References](#) | [BibTeX](#) | [LaTeX\(US\)](#) | [LaTeX\(EU\)](#) | [Harvmac](#) | [EndNote](#)  
[Detailed record](#) - [Similar records](#)



# 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 screenshot displays the INSPIRE website interface. At the top, there is a navigation bar with 'HEP', 'INST', 'HELP', 'SPIRES', and 'HEPNAMES'. Below this, a list of 'ZEUS Internal Notes' is shown, including:

- Inclusive-jet production in** J. Terron C. Glasman, ZEUS-IN-09-004. [References](#) | [BibTeX](#) | [Detailed record](#) - [Similar records](#)
- Three-subjet distributions** E. Ron C. Glasman, J. Terron C. Glasman, ZEUS-IN-09-003. [References](#) | [BibTeX](#) | [Detailed record](#) - [Similar records](#)
- 2009 Guide to Funnel: The** A. Parenti, ZEUS-IN-09-002. [References](#) | [BibTeX](#) | [Detailed record](#) - [Similar records](#)
- Automated calculation of** I. Marfin, ZEUS-IN-09-001. [References](#) | [BibTeX](#) | [Detailed record](#) - [Similar records](#)

The detailed view of the first note, 'Inclusive-jet production in NC DIS with HERA II - C. Glasman, J. Terron . ZEUS-IN-09-004', is shown. It includes a 'Files' tab with a document icon and the text: 'ZEUS-09-004 version 1 [ZEUS-09-004.ps.gz](#) [130.74 KB] 21 Sep 2011, 18:13'. The footer of the page contains the text: 'HEP : Search : Help Powered by Invenio v1.0.0-rc0+ Problems/Questions to [feedback@inspirehep.net](mailto:feedback@inspirehep.net)'.

- 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



# INSPIRE: Paper histories



Welcome to INSPIRE  $\beta$ . Please go to SPIRES if you are here by mistake.  
Please send feedback on INSPIRE to [feedback@inspire-hep.net](mailto:feedback@inspire-hep.net)

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

[Home](#) > Events with Isolated Leptons and Missing Transverse Momentum and Measurement of W Production at HERA

Information | References (52) | Citations (8) | **H1 internal**

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

H1 Collaboration (F.D. Aaron (Bucharest, IFIN-HH & Bucharest U.) *et al.*) [Show all 256 authors.](#)  
2009

**Eur.Phys.J. C64 (2009) 251-271**  
e-Print: [arXiv:0901.0488 \[hep-ex\]](https://arxiv.org/abs/0901.0488)

**Abstract:** Events with high energy isolated electrons, muons or tau leptons and missing transverse momentum are studied using the full  $e^+p$  data sample collected by the H1 experiment at HERA, corresponding to an integrated luminosity of  $474 \text{ pb}^{-1}$ . Within the Standard Model, events with isolated leptons and missing transverse momentum mainly originate from the production of single W bosons. The total single W boson production cross section is measured as  $1.14 \pm 0.25 \text{ (stat.)} \pm 0.14 \text{ (sys.) pb}$ , in agreement with the Standard Model expectation. The data are also used to establish limits on the  $WW\gamma$  gauge couplings and for a measurement of the W boson polarisation.

**Keyword(s):** INSPIRE: [W: production](#) | [transverse momentum: missing-energy](#) | [DESY HERA Stor](#) | [H1](#)

Record created 2009-01-05, last modified 2010-04-11 [Similar records](#)

[Abstract](#) and [Postscript](#) and [PDE](#) from arXiv.org  
[Journal Server](#)  
[Reaction Data \(Durham\)](#)

Export  
[BibTeX](#), [EndNote](#), [LaTeX\(US\)](#), [LaTeX\(EU\)](#), [NLM](#), [DC](#)

- > 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](mailto:feedback@inspire-hep.net)

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

[Home](#) > [Events with Isolated Leptons and Missing Transverse M](#)

[Home](#) >> [Search Results](#)

Information | **References (52)** | Citation

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

## 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 Monod \(Electroweak\) | 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](#)

Abs data with prod also

Key

Record created 2009-01-05, last mod

[Abstract and Postscript](#)  
[Journal S](#)  
[Reaction Data](#)



# HERA data for preservation



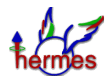
Final data reprocessing to mDST completed in 2009

- Basic preserved data format: ROOT based “Common Ntuples”
- Ultimately RAW, MDST data and MC removed from robots, keep only cNuptles
- Final production of data/MC cNuptles started, to be completed early 2013



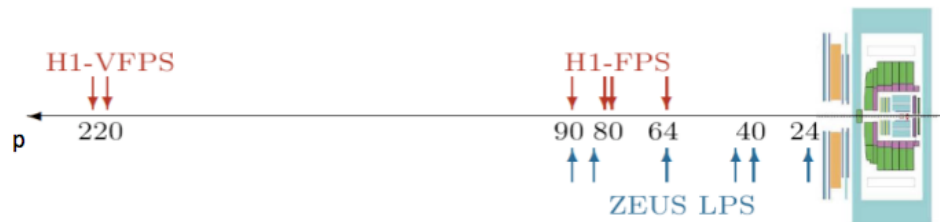
Final reprocessing (DST-7) of HERA II data in 2009, HERA I done in 2012

- Final version of *common analysis software environment + files*, H1OO also done
- Preserve RAW data, as well as DST-7 and H1OO 4.0 versions
- Large MC production of up  $2 \cdot 10^9$  events / year, preserved MC sets to be decided



Final data and MC production completed in 2012

- Main format for analysis is the mDST, this is the one to be preserved
- Importantly for HERMES, all data/MC productions now moved to dCache



Dialogue with DESY machine group concerning their HERA data



# HERA data for preservation



Final data reprocessing to mDST completed in 2009

- Basic preserved data format: ROOT based “Common”
- Ultimately RAW, MDST data and MC removed from ... nUptles
- Final production of data/MC cNuptles started ... 2013



Final reprocessing (DST-7) of HERA ... done in 2012

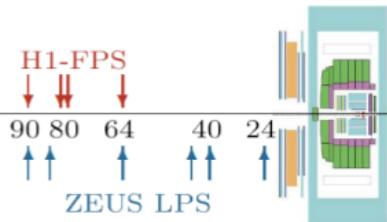
- Final version of common analysis ... + files, H100 also done
- Preserve RAW data, as ... 4.0 versions
- Large MC production ... year, preserved MC sets to be decided



Final data and ... ed in 2012

- Main for ... mDST, this is the one to be preserved
- In ... all data/MC productions now moved to dCache

Total for HERA experiments: ~ 1 PB  
Data preservation is not about the data!



Dialogue with DESY machine group concerning their HERA data



# Isn't it obvious, virtualisation will solve everything?

## My first and very naïve ansatz

- > OK, why don't we just put everything in a virtual machine?
  - Data archival is done elsewhere, just need "to plug that into the VM"
  - Your VM contains everything you need to develop and run code and analysis
- > The problem would then be reduced to maintain virtual images, and maintain their ability to run. In the Cloud era, seems like a trivial task
- > Problems: Everything in IT is a moving target:
  - Will your network always be the same?
  - Will your access protocol always be the same?
  - Are you sure you do not need new software (e.g. MC generators) that require a new OS?
  - Are you sure your i386/SL4 VM will produce the same results when emulated on a quantum computer in NN years?
  - What about service you need, like CondDB,...
- > Naïve virtualization will not work... but still, virtualization can help



# Freezing vs rolling

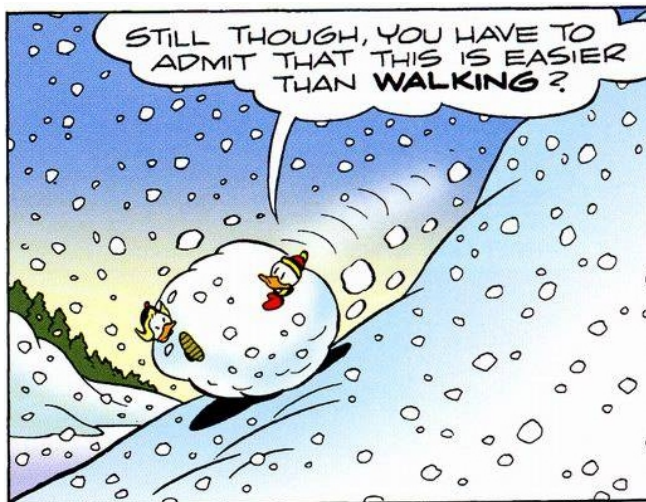


## > Pro Freezing

- One-time effort, very small maintenance outside of analysis phase
- Also allows software w/o code (but might fail with DRM / licensing issues)

## > Cons Freezing

- Rely on certain standards and protocols that may evolve
- Potential performance problems



## > Pro Test-driven migration

- Usability and correctness of code is guaranteed at every moment
- Data accessibility and integrity can be checked as well
- Fast reaction to standard/protocol changes
- General code quality can improve, as designed for portability and migration

## > Cons Test-driven migration

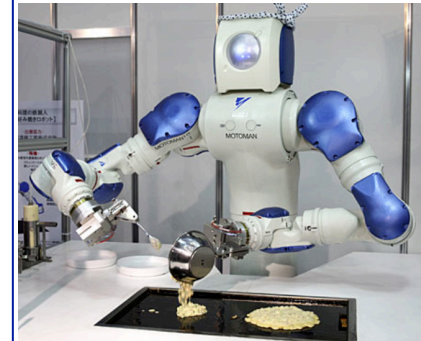
- Needs long-time intervention, more man-power and resources needed
- Some knowledge of the frameworks must be passed to maintainers



# Pizza Preservation



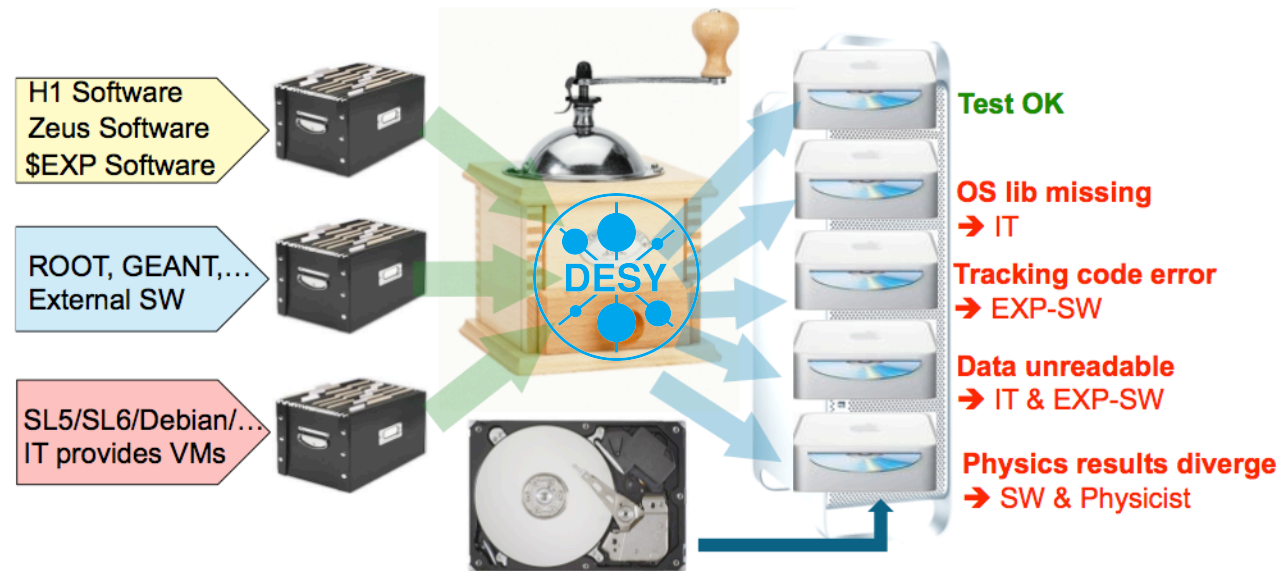
- > Couple of days
  - Fridge
- > Couple of month
  - Deep freezer
- > Couple of years???
  - Preserve the recipe
  - Practice it often: You will not forget the recipe and you can detect variations in external dependencies



- > Whilst freezing the software and environment is easy to do, long term use and correctness of the results not guaranteed
  - Naïve assumption virtualisation solves everything breaks down at the first security hole
- > Freezing software is *OK* if the timeline and scope are reduced, but if changes are needed this is more difficult the longer software is frozen
- > Better to cook the same recipe again and again (and maybe even allow it to be improved), validating the output *automatically*
  - Virtualisation can help!



# The Software Preservation System @ DESY



- > Automated validation system to facilitate future software and OS transitions
  - Uses virtualisation techniques to repeatedly run well defined tests
  - Perform checks of different and evolving environments (OS, s/ware configuration)
  - Stand alone system: No hidden dependencies or /afs access etc: rigorous testing
  - Automatically check these results against predefined, default values
  - Notify when test results differ from these values
  - Separate responsibilities of IT and the experiments



# The Software Preservation System @ DESY



- > Automated validation system to facilitate future software and OS transitions
  - **Uses virtualisation techniques to repeatedly run well defined tests**
  - **Perform checks of different and evolving environments (OS, s/ware configuration)**
  - **Stand alone system: No hidden dependencies or /afs access etc: rigorous testing**
  - **Automatically check these results against predefined, default values**
  - **Notify when test results differ from these values**
  - **Separate responsibilities of IT and the experiments**

# The Software Preservation System @ DESY



- > Automated validation system to facilitate future software and OS transitions
  - Uses virtualisation techniques to repeatedly run well defined tests
  - Perform checks of different and evolving environments (OS, s/ware configuration)
  - Stand alone system: No hidden dependencies or /afs access etc: rigorous testing
  - Automatically check these results against predefined, default values
  - Notify when test results differ from these values
  - Separate responsibilities of IT and the experiments

# The Software Preservation System @ DESY



- > Automated validation system to facilitate future software and OS transitions
  - Uses virtualisation techniques to repeatedly run well defined tests
  - Perform checks of different and evolving environments (OS, s/ware configuration)
  - Stand alone system: No hidden dependencies or /afs access etc: rigorous testing
  - Automatically check these results against predefined, default values
  - Notify when test results differ from these values
  - Separate responsibilities of IT and the experiments

# The Software Preservation System @ DESY



- > Automated validation system to facilitate future software and OS transitions
  - Uses virtualisation techniques to repeatedly run well defined tests
  - Perform checks of different and evolving environments (OS, s/ware configuration)
  - Stand alone system: No hidden dependencies or /afs access etc: rigorous testing
  - Automatically check these results against predefined, default values
  - Notify when test results differ from these values
  - Separate responsibilities of IT and the experiments

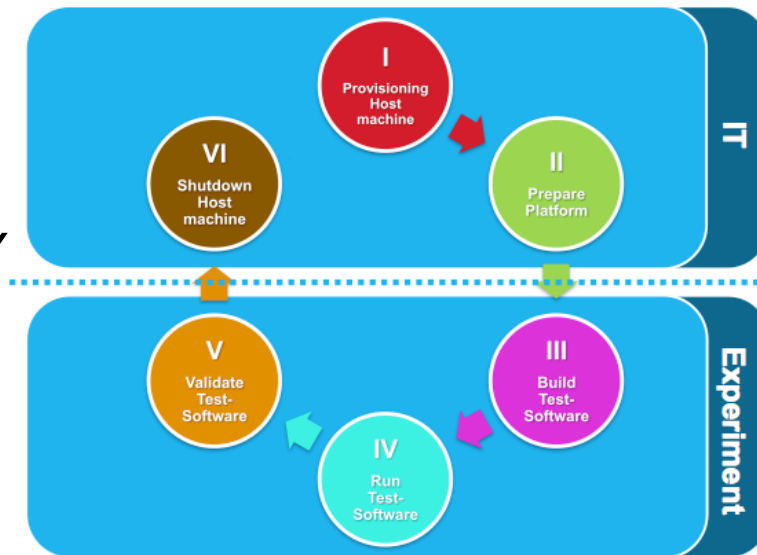
# First test runs in pilot project at CHEP 2010

	SL4	SL5	Fedora 13	
ROOT V5.26	-no F77 compiler gfortran found -libX11 MUST be installed	Estimated ROOTMARKS: 1534.29	Estimated ROOTMARKS: 1512.76	Compilation
H1Data analysis	Processed 47243 events with J/Psi candidates Histogram written to jpsi_mods.root	Processed 47243 events with J/Psi candidates Histogram written to jpsi_mods.root	Processed 47243 events with J/Psi candidates Histogram written to jpsi_mods.root	Run pre- compiled tgz using compat libs
ZEUS MC prod	> ls -lh ZEUSMC.HFSZ627.E89 54.GRAPE.Z01 4.2 MByte	> ls -lh ZEUSMC.HFSZ627.E89 54.GRAPE.Z01 4.2 MByte	> ls -lh ZEUSMC.HFSZ627.E89 54.GRAPE.Z01 4.2 MByte	Run pre- compiled tgz using compat libs
HERA-B	Compilation OK  DB connect fails	Compilation OK  DB connect fails	Compilation failed – needs code change	Compilation



# The sp-system: Towards the full implementation

- > Pilot project in 2010
  - Single configuration, simple tests
- > Full implementation now installed at DESY
- > Common baseline of SLD5 / 32-bit achieved in 2011 by all experiments
  - Sound starting point for validation
- > Following OS configurations now available in sp-system:
  - sl5.6/64(gcc4.4), sl5.7/32(gcc4.4), sl5.7/64(gcc4.1), sl5.7/32(gcc4.1), sl6.2/64(gcc4.4)
- > In addition, to multiple ROOT versions
  - 5.26.00d, 5.28.00c, 5.30.05, 5.32.00, 5.34.01
- > 64-bit systems a major step toward migrations to future OS and hardware
  - SL6 will only be supported in 64 bit variant at DESY
  - NFS4.1 technology, to be used in dCache, native only in SL6.2/64 or higher

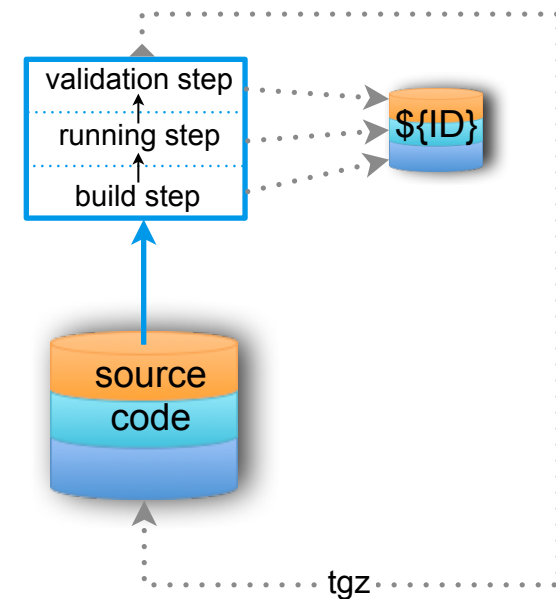




# Running jobs in the sp-system

## > 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



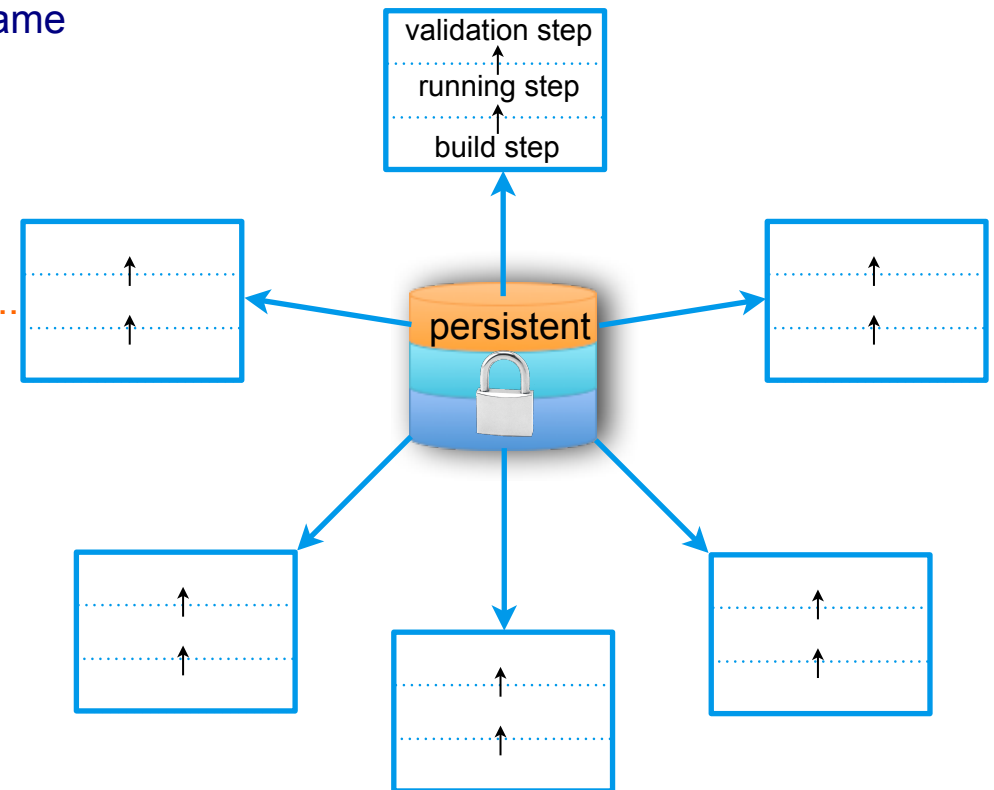
# Running jobs in the sp-system

## > 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, ...



# 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

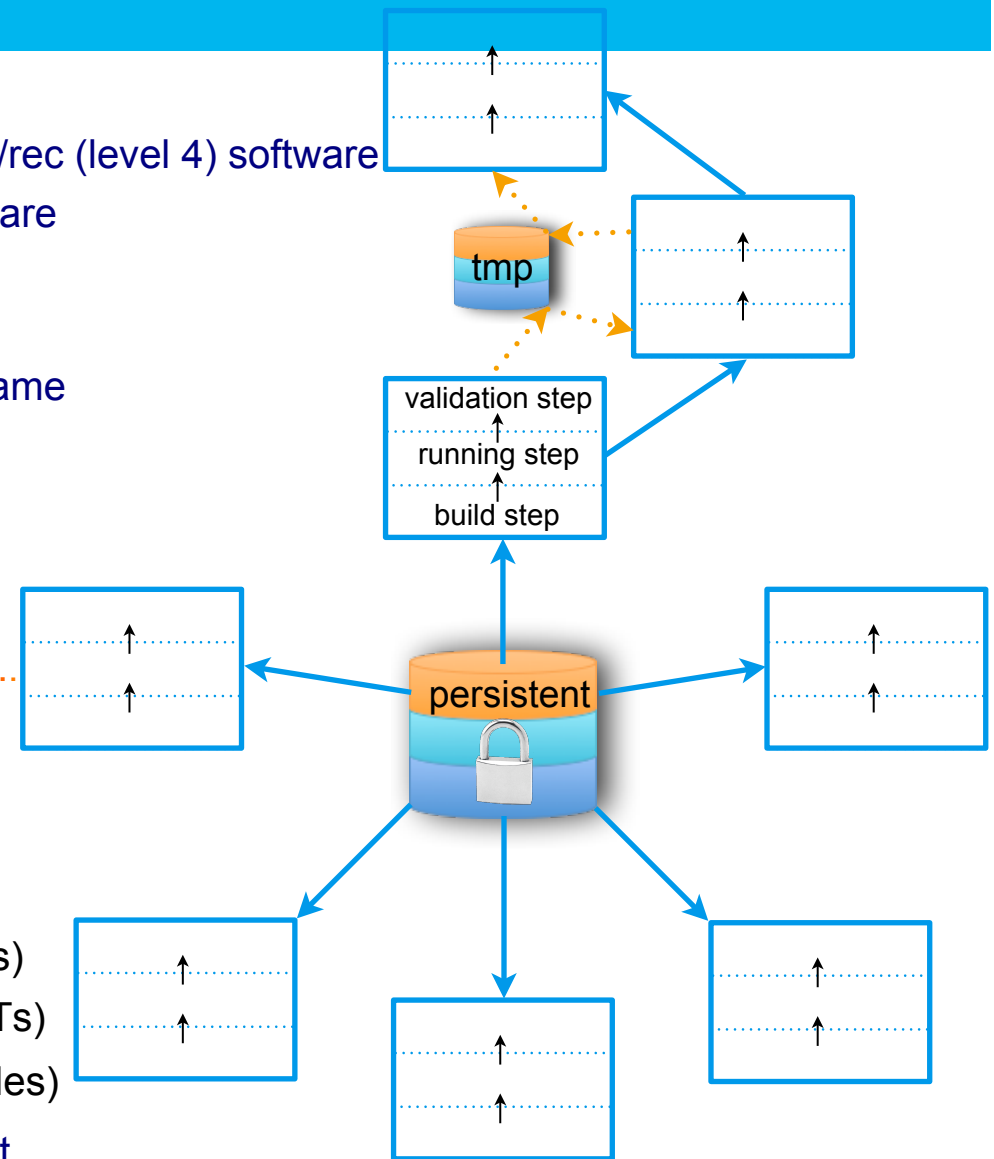
## > Run parallel tests

- 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



# 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

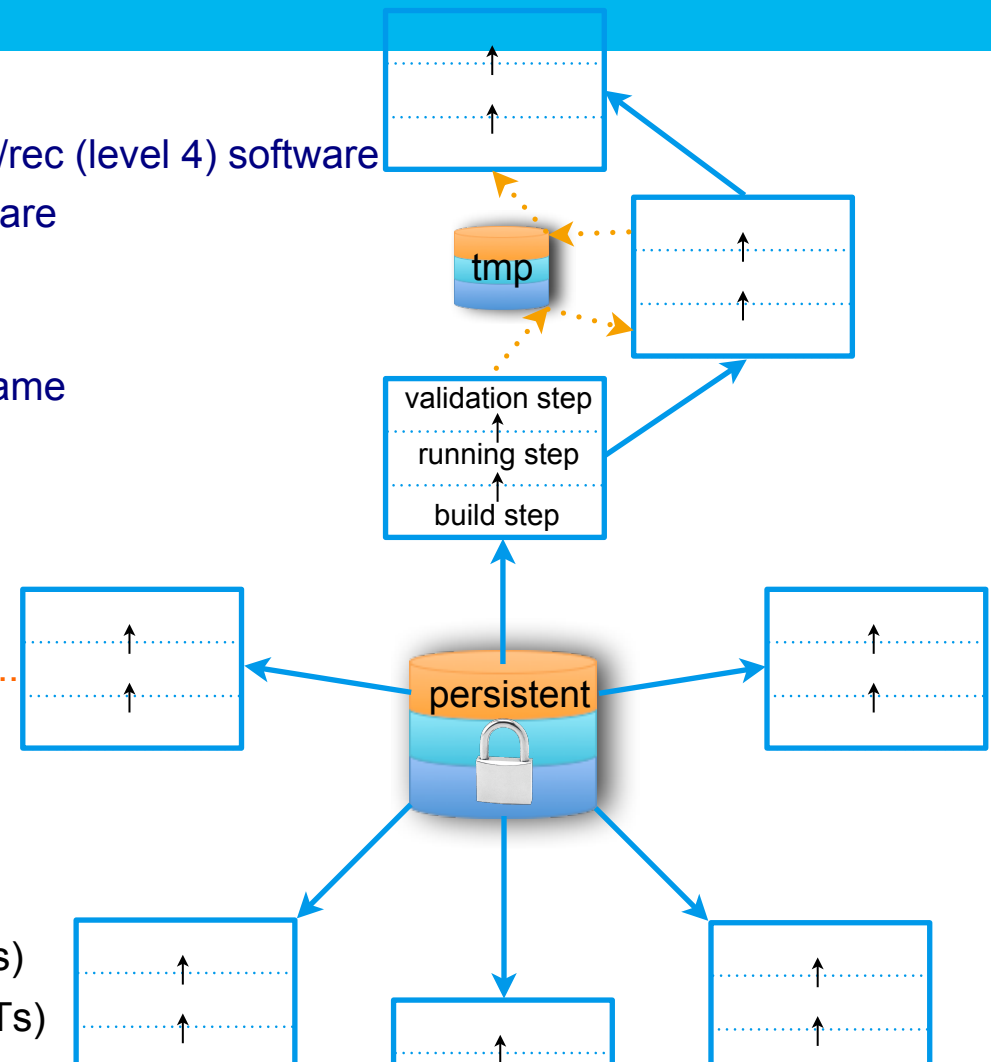
## > Run parallel tests

- 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 → ROC)
- Tests use output of previous test as

## > Results remain accessible or can



→ It is essential to have robust definition of complete set of experimental tests  
 The nature and number dependent on desired preservation level

# First sketch of H1 tests

```

*****
*****
++h1 executables
*****
*****
antlr
batch_kinit
carli
chk_tree
dlg
fpack
fpist
fpmerge
fpsubset
h1ftemu
h1geanonly
h1ieeefp.o / h1ieeefp.cpp
h1rec
h1sim
h1simcheck
h1simrec
hostr
M4shis
M4his
M4m
M4s
look
ltab
ndbint
nqs2pbs
pbs_tclsh
pbs_wish
pbsdsh
pbsnodes
printjob
printtracking
qaller
qdel
qdisable
qenable
qhold
qmgr
qmove
qmsg
qorder
qreun
qris
qrun
qselect
qsig
qstart
qstat
qstop
qsub
qterm
refresh
refresh_init
tracejob
xpbs
xpbsmon
    
```

```

*****
*****
++h1 libraries
*****
*****
#cemlib-gcc44
libLHAPDF.so
libriadne412.a
libbases.a
libbos.a
libcascade2.a
libdatman.a
libdiffm.a
libfpack.a
libfpack.so
libgksdummy.a
libh1bstruc.a
libh1eclass.a
libh1ftemu.a
libh1geang.a
libh1geanh.a
libh1geant.a
libh1i4.a
libh1look.a
libh1mcutit.a
libh1ndb.a
libh1phan.a
libh1qt.a
libh1rec.a
libh1sim.a
libh1trig.a
libh1util.a
libheracles*.a
libheracles*.so
libhztool.a
libjset74.a
liblook.a
libpythia62.a
libpythia64.a
librappap31.a
libshift.a
    
```

55

20  
??

```

*****
*****
++h1oo packages
*****
*****
H1Analysis
H1AnalysisExample
H1Arrays
H1Banks
H1Benchmarks
H1Binning
H1Bos2oop
H1CalcPointers
H1CalcWeights
H1Calculator
H1CalibTrigger
H1CaloTrigger
H1Clusters
H1Cuts
H1ElecCalibration
H1Examples
H1Filter
H1Finder
H1Geom
H1HadronicCalibration
H1Hat
H1HatFilter
H1HfsFinder
H1JetFinder
H1Macros
H1Mods
H1MuonFinder
H1NonepBgFinder
H1OOBanks
H1Ods
H1PartEmFinder
H1PhysUtils
H1Pointers
H1QCDFunc
H1Red
H1SVFit
H1Selection
H1Skeleton
H1SoftLeptonId
H1Steering
H1SubDetInfo
H1Tools
H1Tracks
H1TrkFinder
H1UserCim
H1UserDstar
H1UserFit
H1UserLifetime
H1Wrappers
oo_tools
#share
    
```

36

"only" UseTiming(v2)

x2  
H1User

x3

x2

x2

x2

+ Mayfield  
lots

51

```

*****
*****
++h1oo binaries
*****
*****
AnalysisExample
AnalysisExampleExtraction
AnalysisExamplePlots
H1Bos2oop
H1Makeptr
L12Root
MakeInputTable
TestQCDFunc
batchAnalysis
boosted Jets
checkcim
cintsteering
clusters_ods
copyMyEvents
create_eventlist
dbaccess
deleteJobs
dst2all
dst2ods
dstar_mods
empz_hat
h1red
h1root
jpsi_mods
kaonfind_ods
l1te_hat
lumicalc
mergeAnalysis
mynkicim
ods2modshat
oolist
oolumi
oomclumi
oomove
oosubset
read_dstartree
read_eventlist
read_ods
read_usertree
rerun_finder
rerun_rec
resubChains
snapshot
steermanager
test_binning
write_eventlist
    
```

x2

46

```

*****
*****
++h1oo libraries
*****
*****
libH1Analysis.so
libH1AnalysisExample.so
libH1Arrays.so
libH1Benchmarks.so
libH1Binning.so
libH1CalcPointers.so
libH1CalcWeights.so
libH1Calculator.so
libH1CaloTrigger.so
libH1Clusters.so
libH1Cuts.so
libH1ElecCalibration.so
libH1Filter.so
libH1Filter_odsonly.so
libH1Finder.so
libH1Finder.so
libH1Geom.so
libH1HadronicCalibration.so
libH1Hat.so
libH1HatFilter.so
libH1HfsFinder.so
libH1JetFinder.so
libH1MagFieldOO.so
libH1Mods.so
libH1MuonFinder.so
libH1NonepBgFinder.so
libH1OOBanks.so
libH1Ods.so
libH1PartEmFinder.so
libH1PhysUtils.so
libH1Pointers.so
libH1QCDFunc.so
libH1Red.so
libH1RedLook.so
libH1Red_bos.so
libH1SVFit.so
libH1Selection.so
libH1Skeleton.so
libH1SoftLeptonId.so
libH1SoftLeptonId_impl...so
libH1Steering.so
libH1SubDetInfo.so
libH1Tools.so
libH1Tracks.so
libH1TrkFinder.so
libH1UserCim.so
libH1UserDstar.so
libH1UserDstar_fill.so
libH1UserFit.so
libH1UserFit_Filter.so
libH1UserLifetime.so
libH1UserTiming.so
libH1UserTiming_fill.so
libH1Wrappers_bos.so
libH1Wrappers_fastjet.so
libH1Wrappers_geom.so
    
```

74

```

libH1Wrappers_Jumi.so
libH1Wrappers_ndb.so
libH1Wrappers_neurobayes.so
libSISconePluginOO.so
libUser.so
libbosutil.so
libcemlibOO.so
libfastjetOO.so
libfortran.so
libfortranpatchOO.so
libfortranshared.so
libfortranstat.a
libfpackOO.so
libh1ndbOO.so
libh1recOO.so
libmbddummy.so
libneurobayesOO.so
libsisconeOO.so
libutildummy.so
    
```

copy ~100...so compiler  
~60 core files

exec 25+12 H1mods + H1oo  
H1oo "tests" ~37

1. Suptec  
2. dst2all → 10, includes MC, DR?

2. fpack 3  
+ ndbint, includes only ~60

S>10 analysis, including: ↑  
+ all hat → for each  
+ all calc 15020

event deploy → h1oo core  
+ h1oo  
+ h1oo



# First sketch of H1 tests

## > Validate compilation of

- ~100 (shared) library objects
- ~60 executables
- MC generators not yet included
- most important:

- simrec - reconstruction / dst production
- dst2all - h1oo file production

## > Validate correct running of

- 37 x h1oo
- 4 x fpack, ndb
- ≥10 x h1simrec → dst2all → analysis

One test for every run period + MC / (DQHat)

- ... Let's say about 60 executables

## > Run and validate physics analyses

- (At least) one test for every run period
- Inclusive & all HAT/H1Calculator variables
- 5-10 'real' physics analyses

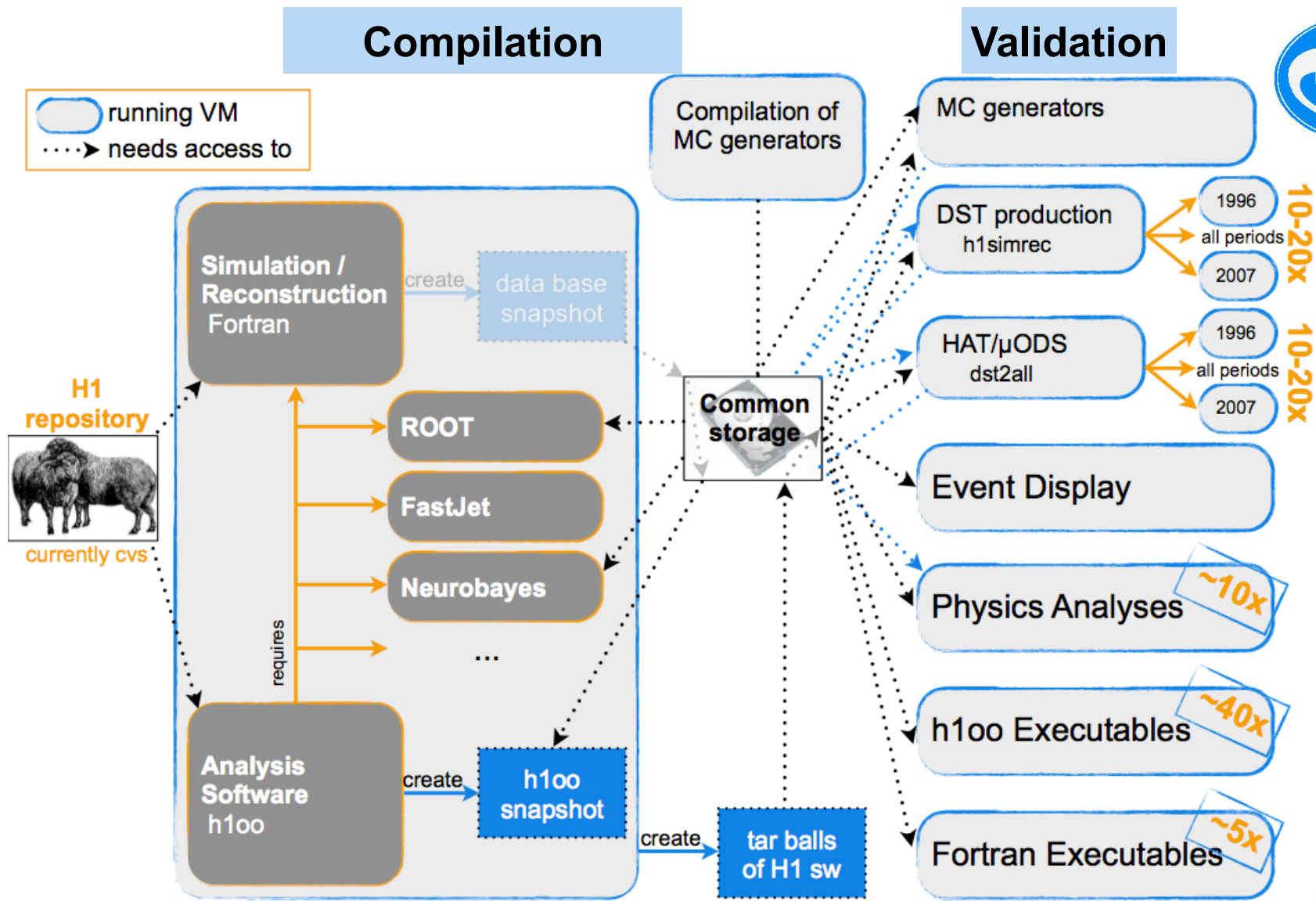
## > Check event display

Handwritten notes on a grid background, enclosed in an orange box. The notes are organized into sections:

- comp**: ~100 .so compile hrs.
- exec**: ~60 executables
- exec**: 25 + 12 H1Exmde & H1moro
- H1oo "tests"**: (37)
- + **simrec** → (10), including MC, DQ?
- + **dst2all**
- + **fpack** (3)
- + **ndbint**, including exit (~60)
- 5 > 10 analyses, including:**
  - + all hat → for each
  - + all calc (15 > 20)
- event display** → hndg evts
  - mzo
  - r.hly detem.
  - trks (mult etc)
  - dmp.
- Σ ≈ 250 tests**



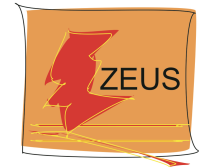
# Example structure of experimental tests: H1 (Level 4)



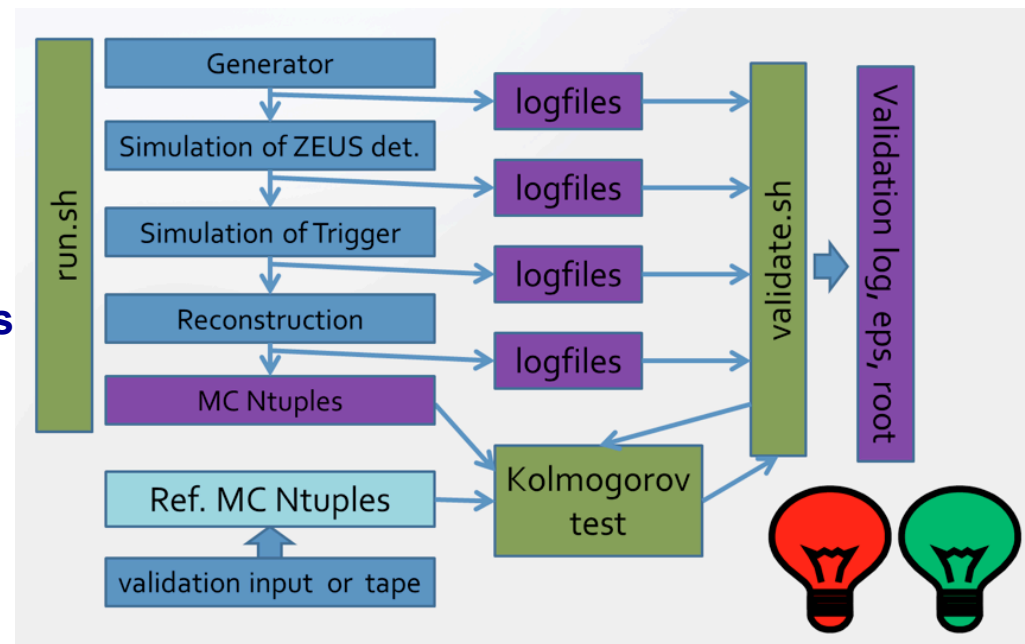
Including compilation of individual packages: about 250 tests planned by H1



# 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 pre-compiled executables will also be preserved as a standalone package
  - **Remaining ZEUS SL3 executables continue to work on the SL6/64 OS**
- > In addition, an interface for new generators is developed, which is also included in the validation system







# Putting it all together

Process	Operating System	SL5 32bit				SL5 64bit					SL6 64bit	
		External Dependencies	5.26	5.28	5.30	5.32	ROOT		Cernlib		Fastjet	Neuro-bayes
						2005	2006	2.3.3	2008 0312	3.3.0		
 Accessing cNtuples (Data/MC)												
Creating cNtuples (Data/MC)												
ZMCSP (simulate/reconstruct MC)						No dependence						
Validation												
Compilation of s/w												
Generating MC files												
Producing DST files (Data/MC)												
Producing h1oo files (Data/MC)												
Accessing h1oo files (Data/MC)												
Accessing ndb snapshot												
Validation												
Compilation of s/w												
Accessing uDST (precompiled s/w)												
Reconstruction (precompiled s/w)								No dependence				
Producing uDST (precompiled s/w)												
Validation												

Full chain, including compilation of all H1 software, from MC generation, through to validation of analysis level (e.g. high  $Q^2$  neutral current) histograms now in place within the sp-system



# Putting it all together



Process	Operating System	SL5 32bit				SL5 64bit					SL6 64bit
		External Dependencies	5.26	5.28	5.30	5.32	ROOT		Cernlib	Fastjet	Neuro-bayes
						2005	2006	2.3.3	2008 0312	3.3.0	
Accessing cNtuples (Data/MC)											
Creating cNtuples (Data/MC)											
ZMCSP (simulate/reconstruct MC)						No dependence					
Validation											
Compilation of s/w											
Generating MC files											
Producing DST files (Data/MC)											
Producing h1oo files (Data/MC)											
Accessing h1oo files (Data/MC)											
Accessing ndb snapshot											
Validation											
Compilation of s/w											
Accessing uDST (precompiled s/w)											
Reconstruction (precompiled s/w)								No dependence			
Producing uDST (precompiled s/w)											
Validation											

Here a much finer granularity needed for displaying the results !

# 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, ...

### H1 Validation Results

List of available validation runs: success error(s) work to be done not in list

- [H1\\_64bit\\_VT79\\_4.0.21](#)

Description of used software version:  
**H1\_64bit\_VT79\_4.0.21**

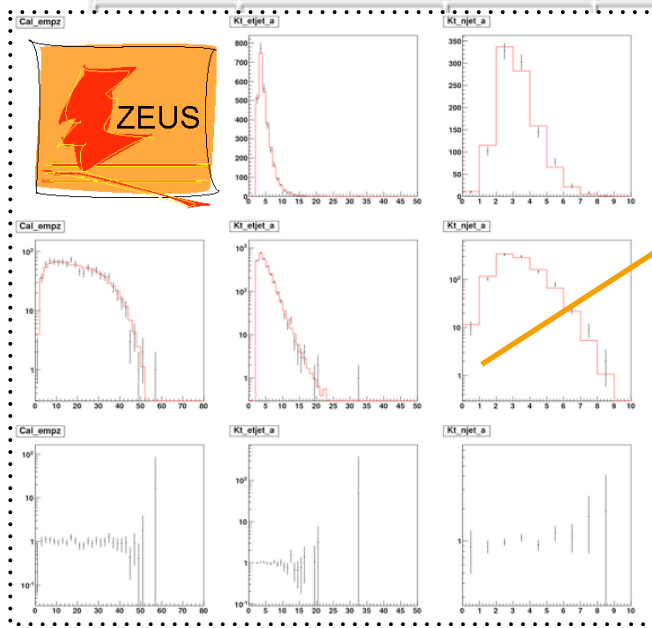
	12:23 17.01. 2012	17:38 30.01. 2012	08:06 04.02. 2012
cernlibs			
fastjet			
neurobayes			
h1unix			
h1icefp			
box			
...			

### Results of Tests

Tests run with software version:  
**H1\_64bit\_VT79\_4.0.21**  
 created on: 04.02.2012 (08:06)

	HERA I					HERA II				
no specific year	1996	1997	1998	1999	2000	2003	2004	2005	2006	2007
dst2all										
dumpHATvariables										
jpsi_mods										
ndbint										

test number	operating system	root version	staus	std output file	error file	plots	root file
58	s15.6_64	5.28.00c	OK	out	err	plots	root



Opening ZEUSMC.HFIX627.F15419.4B70.TEEST.Z01.root

You have chosen to open

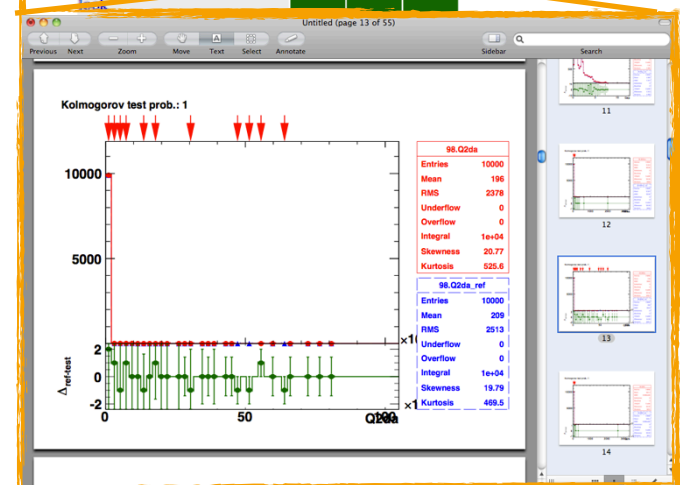
ZEUSMC.HFIX627.F15419.4B70.TEEST.Z01.root  
 which is a: root File (40.7 MB)  
 from: <http://www-zeus.desy.de>

What should Firefox do with this file?

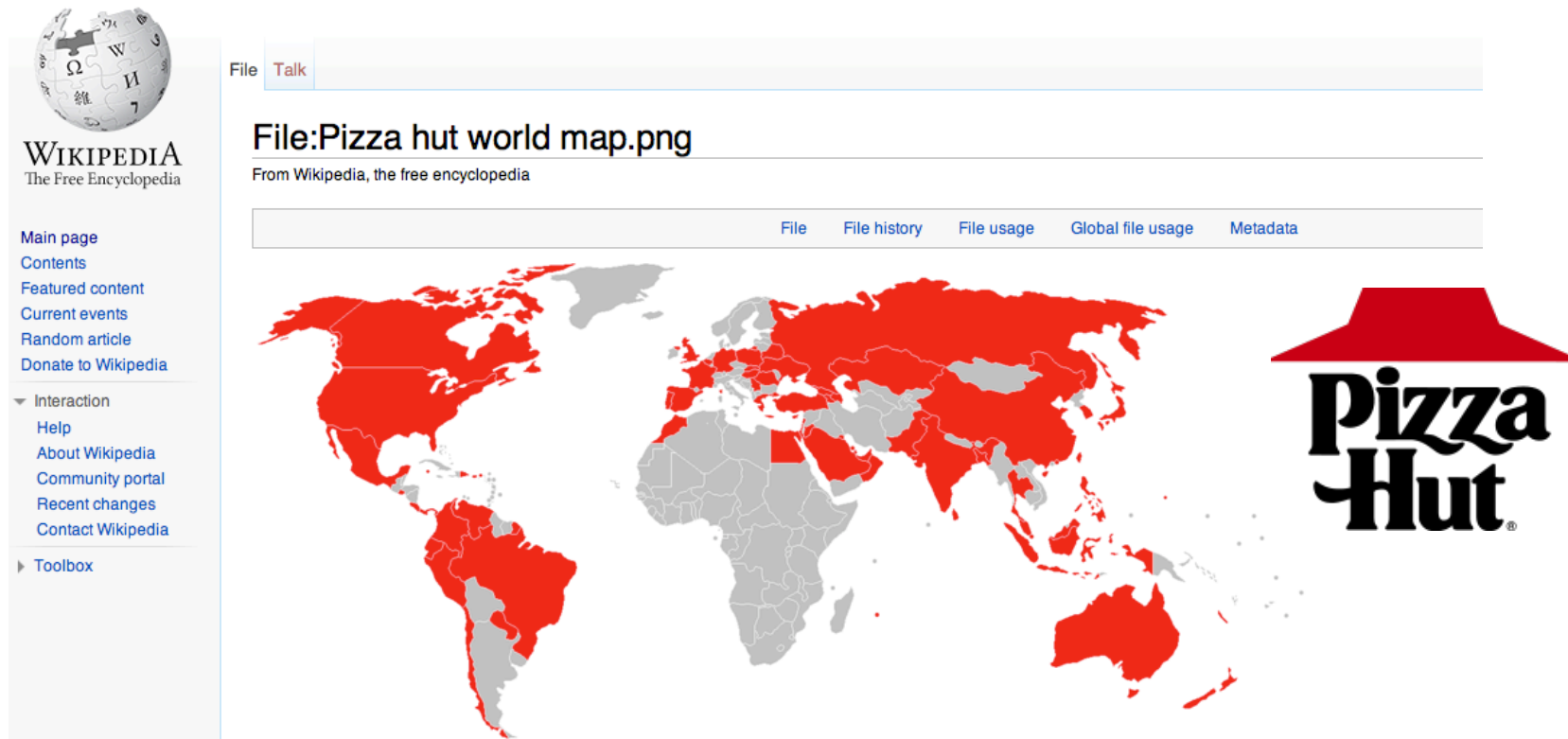
Open with

Save File

Do this automatically for files like this from now on.

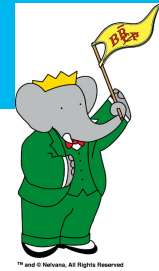


# Deployment

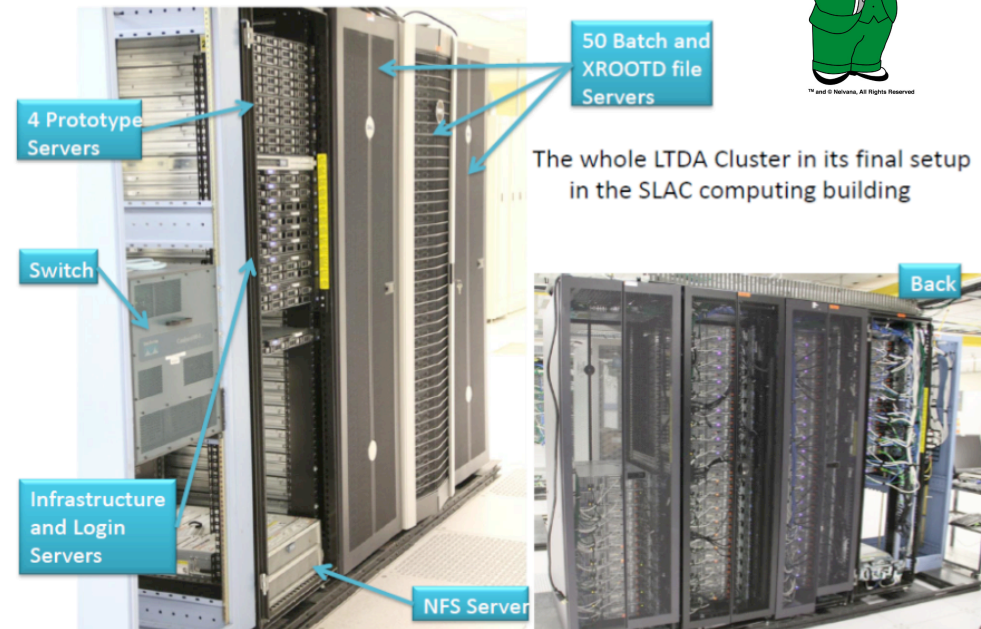


- The whole point of the `sp-system` is **not** to provide a future resource for the experiments, but rather to provide a recipe which can be deployed
  - At DESY, this means for example exploring alternative resources such as the local BIRD cluster, the National Analysis Facility (dedicated to LHC, unlikely) or the Grid

# The BaBar Long Term Data Access archival system



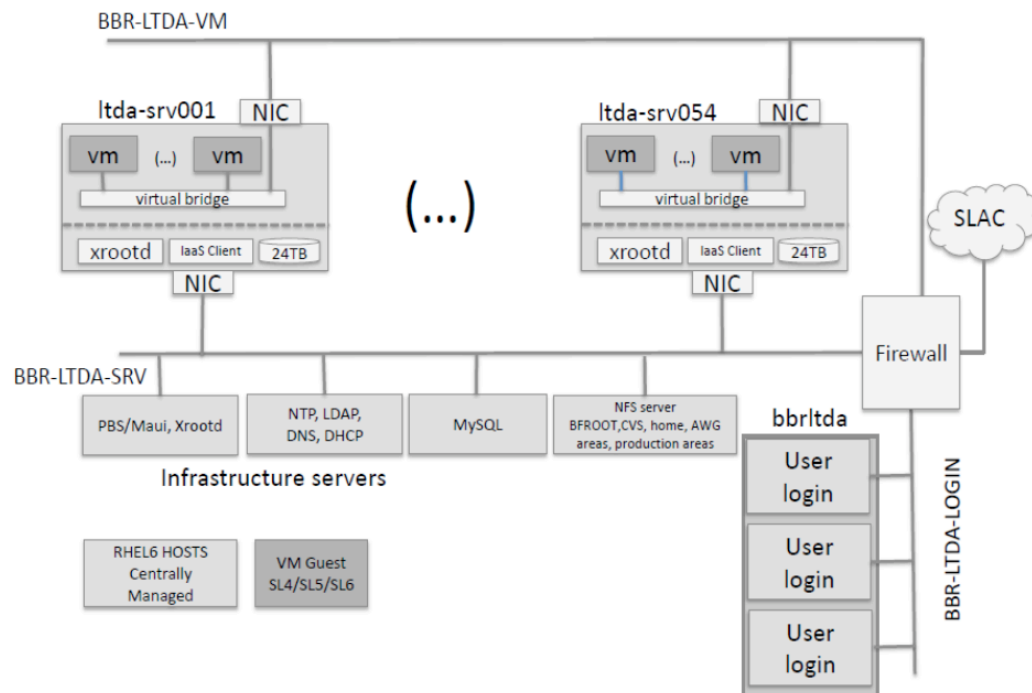
- > New BaBar system installed for analysis until at least 2018
- > Isolated from SLAC, and uses virtualisation techniques to preserve an existing, stable and validated platform
- > Complete data storage and user environment in one system



- > 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

# Summary of information from the (pre-LHC) experiments

	BaBar	H1	ZEUS	HERMES	Belle	BESIII	CDF	DØ
<b>End of data taking</b>	07.04.08	30.06.07	30.06.07	30.06.07	30.06.10	2017	30.09.11	30.09.11
<b>Type of data to be preserved</b>	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)
<b>Data Volume</b>	2 PB	0.5 PB	0.2 PB	0.5 PB	4 PB	6 PB	9 PB	8.5 PB
<b>Desired longevity of long term analysis</b>	Unlimited	At least 10 years	At least 20 years	5-10 years	5 years	15 years	Unlimited	10 years
<b>Current operating system</b>	SL/RHEL3 SL/RHEL 5	SL5	SL5	SL3 SL5	SL5/RHEL5	SL5	SL5 SL6	SL5
<b>Languages</b>	C++ Java Python	C C++ Fortran Python	C++	C C++ Fortran Python	C C++ Fortran	C++	C C++ Python	C++
<b>Simulation</b>	GEANT 4	GEANT 3	GEANT 3	GEANT 3	GEANT 3	GEANT 4	GEANT 3	GEANT 3
<b>External dependencies</b>	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





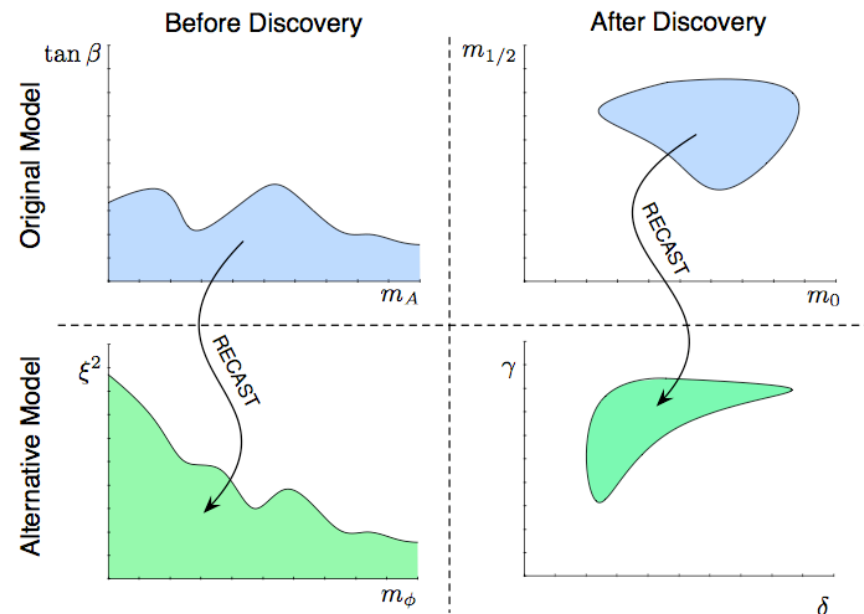
# A multi-preservation level tool: RECAST

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

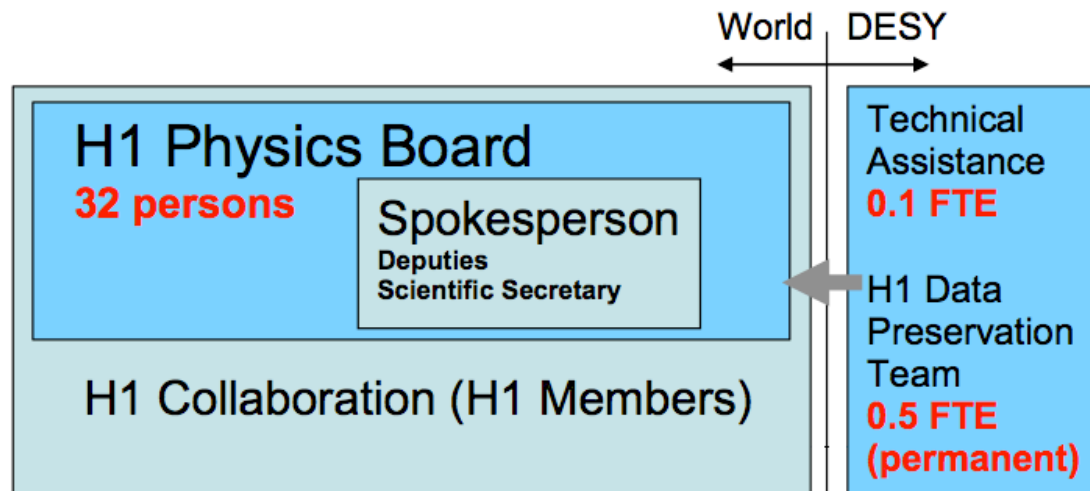
- Complete information from original analysis contained in the data
- Already performed on ALEPH data, LHC experiments investigating



- > 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



# Changing face of the HERA collaborations



- > H1 moved to a new collaboration management model in July 2012
  - Formation of *H1 Physics Board*, to replace Collaboration Board (institute based)
  - Future author list policies also set down in new constitution approved by collaboration
- > ZEUS (and HERMES) management teams retain same model as before, but similarly to H1 the collaborating institute layer is now removed
  - Remaining physics ZEUS working groups are now consolidated to a single physics group



# Identified use cases and action areas by DPHEP

- > **Bit preservation** as a basic “service” on which higher level components can build;
- > **Preserve data, software, and know-how** in the collaborations; Basis for reproducibility;
- > **Share data and associated software** with (wider) scientific community, such as theorists or physicists not part of the original collaboration;
- > **Open access** to reduced data sets to general public (LHC experiments)



# CERN Services for LTDP

1. State-of-the art "**bit preservation**", implementing practices that conform to the ISO 16363 standard
2. "**Software preservation**" - a key challenge in HEP where the software stacks are both large and complex (and dynamic)
3. Analysis **capture and preservation**, corresponding to a set of agreed Use Cases
4. Access to **data behind physics publications** - the [HEPData portal](#)
5. An **Open Data portal** for released subsets of the (currently) LHC data
6. A **DPHEP portal** that links also to data preservation efforts at other HEP institutes worldwide.

➤ **Each of these is a talk topic in its own right!**

