technische universität
dortmund

H1 Data Preservation Project Status

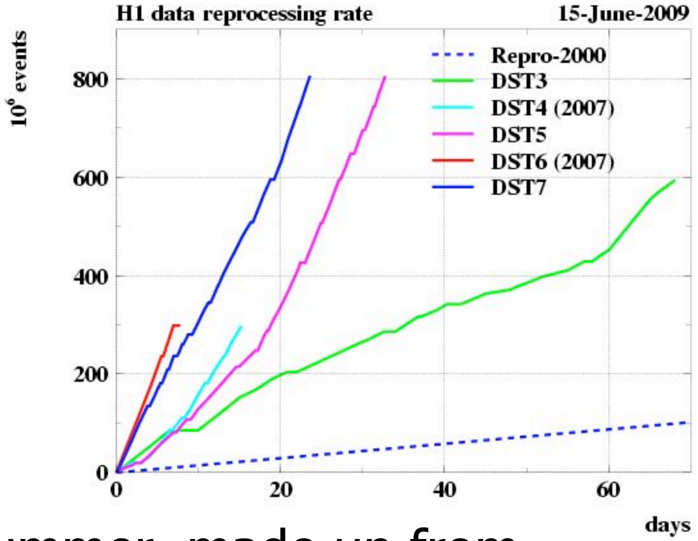
David South (TU Dortmund)

Third Workshop on Data Preservation and Long Term Analysis

7 - 9 December 2009

Since the 2nd Workshop at SLAC in May

- We completed the DST7 in the June 2009
  - Major achievement, scheduled to be the final reprocessing of the H1 HERA II data
  - Several further iterations over the summer for processing of MC events
  - HERA I data to follow start of 2010



- H1 data preservation task force set up in summer, made up from existing manpower*

- Project coordination and referees assigned from H1 and DESY-IT

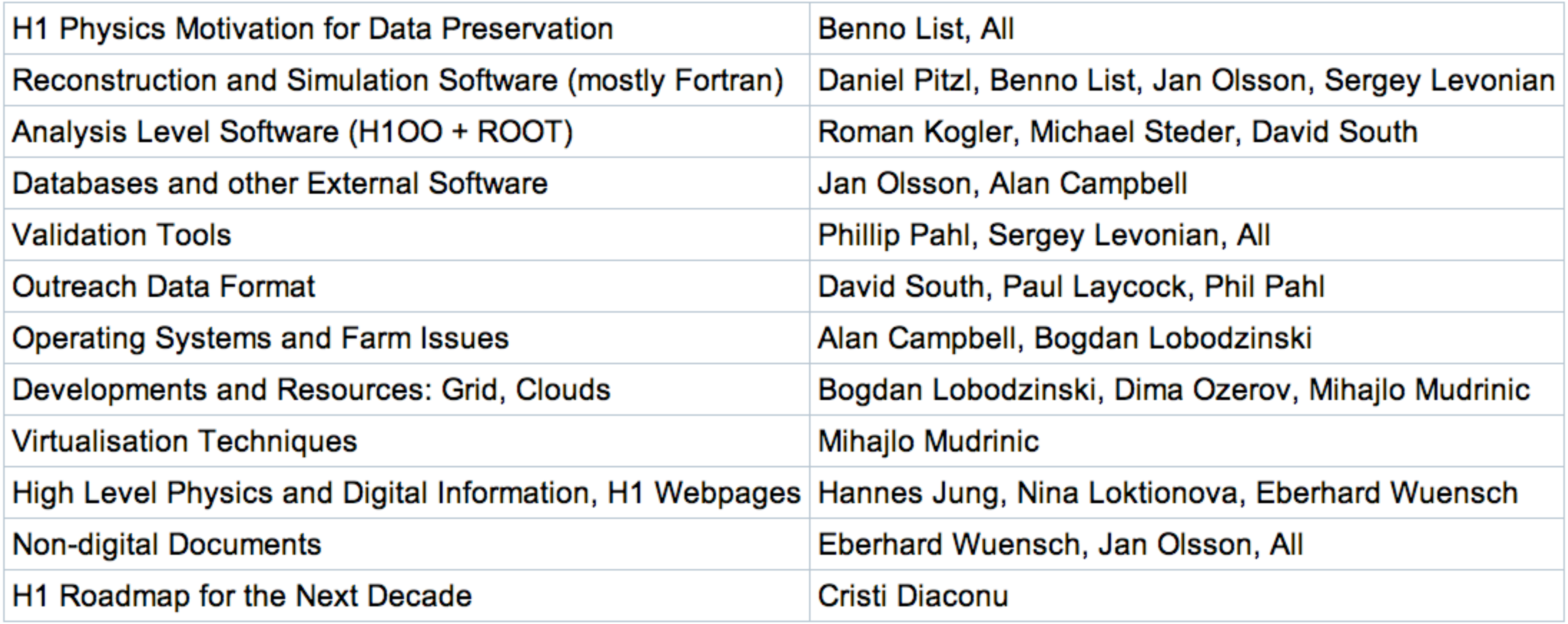
- Regular meetings to discuss different components of the projects: first meeting 2nd September 2009

- Key action areas identified: survey of all areas in such a project, using the DPHEP recommendations as a guide

**which is not all secured in the long term*

H1 Data Preservation Action Areas

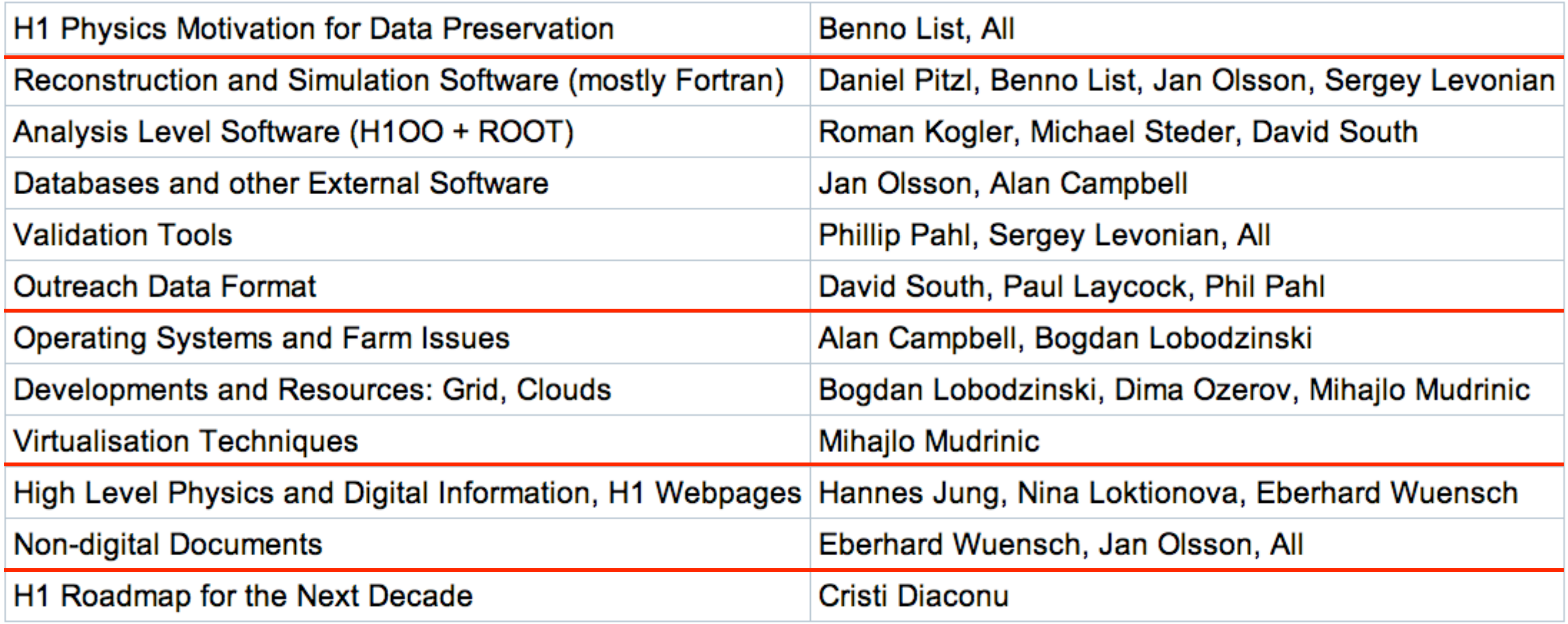
- This survey will form the basis of a document which, after internal refereeing within the collaboration, will be the proposal for a H1 data preservation project, including financial requirements



H1 Physics Motivation for Data Preservation	Benno List, All
Reconstruction and Simulation Software (mostly Fortran)	Daniel Pitzl, Benno List, Jan Olsson, Sergey Levonian
Analysis Level Software (H1OO + ROOT)	Roman Kogler, Michael Steder, David South
Databases and other External Software	Jan Olsson, Alan Campbell
Validation Tools	Phillip Pahl, Sergey Levonian, All
Outreach Data Format	David South, Paul Laycock, Phil Pahl
Operating Systems and Farm Issues	Alan Campbell, Bogdan Lobodzinski
Developments and Resources: Grid, Clouds	Bogdan Lobodzinski, Dima Ozerov, Mihajlo Mudrinic
Virtualisation Techniques	Mihajlo Mudrinic
High Level Physics and Digital Information, H1 Webpages	Hannes Jung, Nina Loktionova, Eberhard Wuensch
Non-digital Documents	Eberhard Wuensch, Jan Olsson, All
H1 Roadmap for the Next Decade	Cristi Diaconu

Contents

- Today I will try to summarise the activities in each of these different areas, so the list forms the contents of this talk



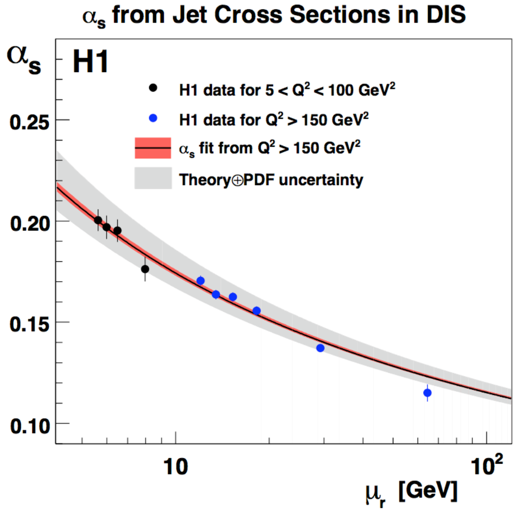
H1 Physics Motivation for Data Preservation	Benno List, All
Reconstruction and Simulation Software (mostly Fortran)	Daniel Pitzl, Benno List, Jan Olsson, Sergey Levonian
Analysis Level Software (H1OO + ROOT)	Roman Kogler, Michael Steder, David South
Databases and other External Software	Jan Olsson, Alan Campbell
Validation Tools	Phillip Pahl, Sergey Levonian, All
Outreach Data Format	David South, Paul Laycock, Phil Pahl
Operating Systems and Farm Issues	Alan Campbell, Bogdan Lobodzinski
Developments and Resources: Grid, Clouds	Bogdan Lobodzinski, Dima Ozerov, Mihajlo Mudrinic
Virtualisation Techniques	Mihajlo Mudrinic
High Level Physics and Digital Information, H1 Webpages	Hannes Jung, Nina Loktionova, Eberhard Wuensch
Non-digital Documents	Eberhard Wuensch, Jan Olsson, All
H1 Roadmap for the Next Decade	Cristi Diaconu

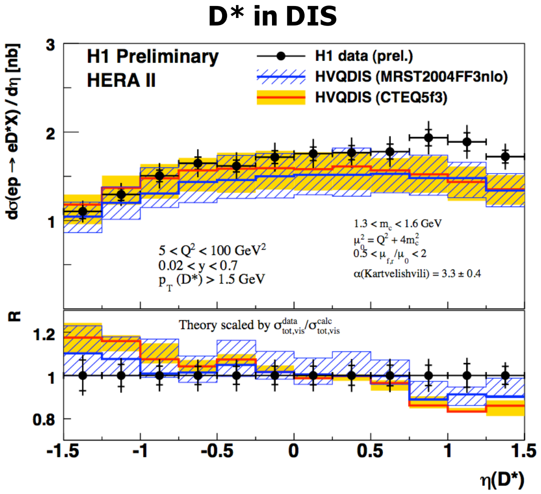
The Physics Case for Data Preservation

- The HERA data are a unique data set, unlikely to be superseded soon (LHeC?): But this is not enough to justify preservation project
  - We can’t say “if only we had the time / manpower”…
  - We are now at the end of HERA lifetime, the physics programme has already been set out, with no initial budget for preservation

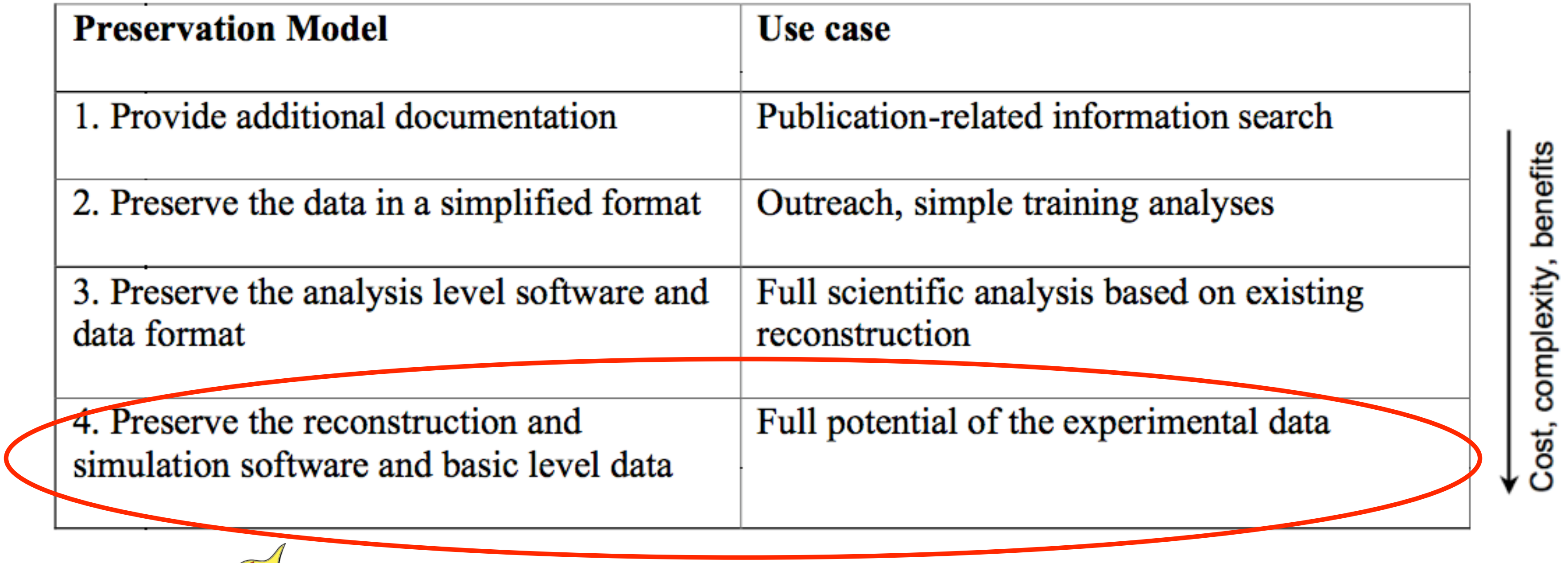
- Identify existing analyses which could be improved via
  - A better model, smaller theory uncertainties, more orders..
  - An improved analysis technique (event shapes, angular correlations..)

- H1 analyses where model uncertainty dominates:
  - Jet cross sections and α_S
  - Isolated photons in DIS
  - D* cross sections
  - Di-jets in diffraction
  - 3,4 jet production at low x



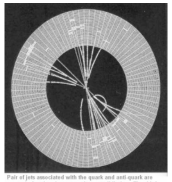


Data Preservation Models Identified by DPHEP

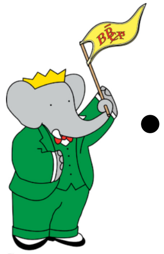


Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

JADE



Pair of jets associated with the quark and anti-quark are created after the electron-positron annihilation. The third jet comes from a gluon radiated in the process.



- Only when the full flexibility is retained does the full potential of the H1 data remain
  - A level 4 type programme was required by JADE re-analyses
- H1 aims for level 4 (full preservation) *and* at the same time tries a level 2 scheme (outreach), collaborating via DPHEP
  - And this goal will require new *manpower* to achieve it…

Preserving the Data Themselves

- Data formats to be preserved
  - RAW data of good and medium runs: 75 TB (copy to disk now complete for 96-07)
  - At least one full set of DSTs, total for HERA I+II: 18 TB
  - A version of analysis level format: μODS and HAT as well (< 3 TB)
  - In addition to calibration and cosmic runs, total data about 100 TB
  - Amount of MC to be decided, but will be of the same order

- Conservatively (x2) estimate total amount to preserve at 500 TB

- Do not expect to be limited by CPU or disk space
  - Preserved data/MC should be copied on to new media by IT at regular intervals, say every 2 years
  - Expect cost of migration to be x2 current costs: 1 + 1/2 + 1/4 + 1/8 + .. = 2
  - In terms of hardware to perform analysis, a few large working group servers should be enough (more on hardware later)

Reconstruction and Simulation Software

- Mostly written in Fortran (some C, some C++)
- Basic data format FPACK/BOS designed as machine independent

- No further major development after DST 7
  - But should still be possible…

- Some parts already frozen since a good few years

- First preservation steps undertaken
  - Movement of all code into CVS (some older code in CMZ)
  - Updating of documentation of bank description
  - Test migration to SLD5 has also allowed some clean up and bug finding

Analysis Level Software: H1OO

- Written entirely in C++ language
- Coherent framework for file production and analysis

- Model heavily reliant on ROOT framework: I/O, TTree..

- μODS: Particle finders
  - Pointers back to ODS (DST) information (original tracks, clusters, cells)
  - Most classes inherit from TObject
  - Much use of inheritance and class structure
- HAT: Contains around 200 selected basic event variables
  - Stored as flat ntuple format, mainly for fast event selections

- Next development series: focus on clean up of current data formats

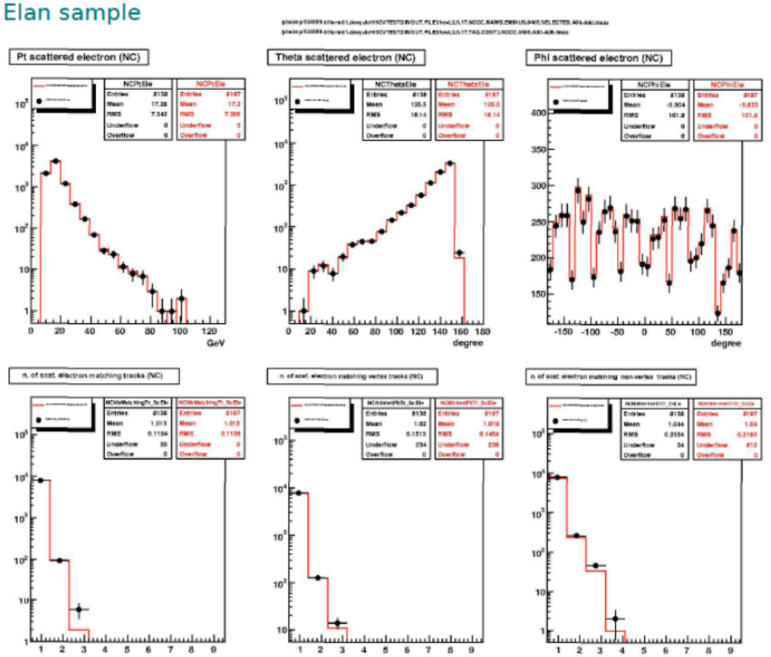
Databases and other External Software

- As well as ROOT, several other external software dependencies exist in the H1 Software
  - Try to isolate and phase out if possible, are there alternatives?
  - If they will remain, how much will they cost? Who maintains?
- Oracle
  - NDB database of run conditions etc: Data and MC Production (uses FPACK copy of NDB banks)
  - Slow Control (detector HV) database
  - Registration of MC production
  - Within the H1 webpage: members, institutes database
- CERNLIB: Analysis level executables
- GKS: Old event display (LOOK graphics)
- FastJet++: H1OO jet finder
- Neurobayes: H1OO cluster separation neural net algorithm

H1 Validation Tools

- If we want to have anything like a dynamic preservation model, then we will need validation tools
  - Such a scheme already exists to validate file content of analysis level software between releases
  - We are expanding this to include full analysis selections
  - Validation tools now being developed for the Fortran (simulation/reconstruction) part of the H1 software

NC Elan sample



- The eventual aim is to have a unified validation suite, which can compare different DSTs, H1OO releases and even analysis running under different operating systems

A Rolling Preservation Model

- In the longer term, for the analysis level we plan a rolling model of preservation, *with a timescale of say 3 months interval*
  - Regular recompilation of H1OO software
  - Full data production of μODS/HAT files, probably MC too

    <u>From current times:</u>

    Read copy 13.5 Tb of HERA II DST7 format data to Grid working nodes

    900 Grid jobs each running on average 20 hours

    Produce 1.3 Tb of HERA II μODS/HAT format data

    In ideal conditions: 1 day to produce data, 1 day to download from Grid

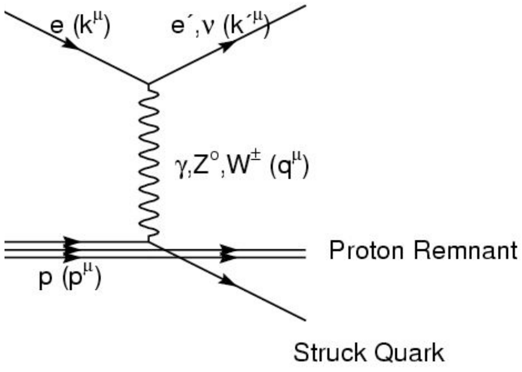
- Defining a strategy for a rolling preservation model
  - Always use newest versions or freeze external software?
  - Continue using the database / have a snapshot of it?
  - Would aim to incorporate ROOT updates
  - Extreme version: adopt change in OS, *requires guaranteed manpower*

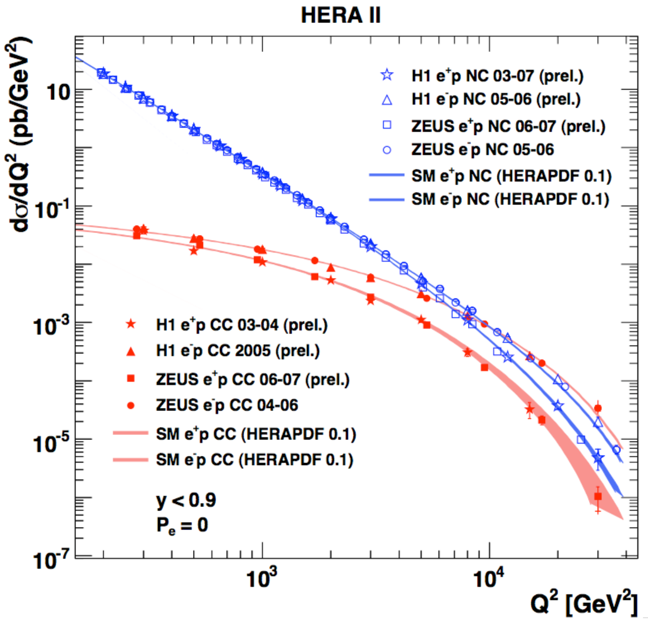
An H1 Data Format for Outreach

- Producing H1 data in a format suitable for outreach purposes is an attractive proposal
  - To run in parallel with the main preservation effort
- The physics content of such a format is essentially defined by the outreach plans
  - What can the user learn by studying ep collision data?

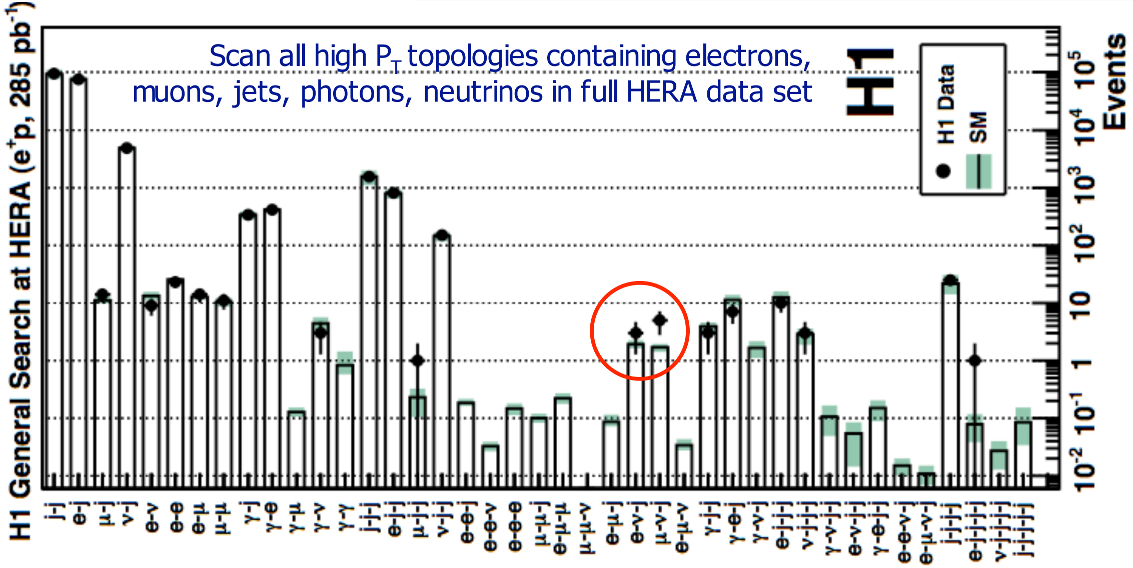
- This then starts to define the variables, quantities and even the outreach projects themselves

Outreach Potential of HERA Data



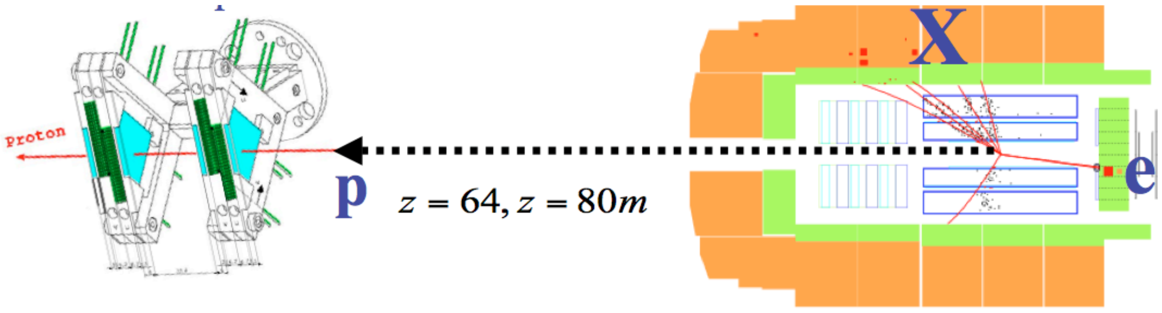


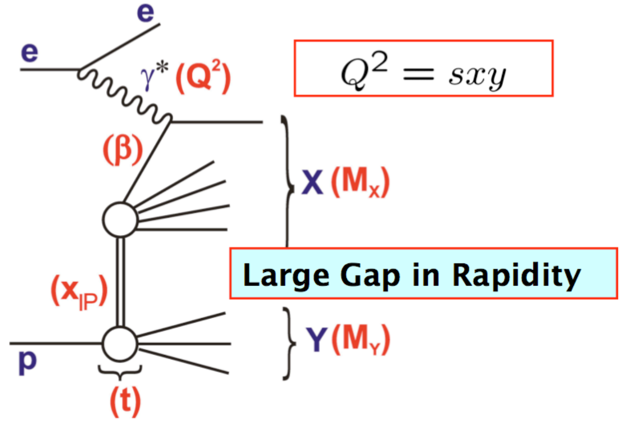
1) Basics of deep-inelastic ep scattering, and understanding the differences between NC and CC events, electroweak unification


Scan all high P_T topologies containing electrons, muons, jets, photons, neutrinos in full HERA data set

2) Looking at high P_T event topologies, and in regions where the SM expectation is low: look for deviations in the data

Outreach Potential of HERA Data

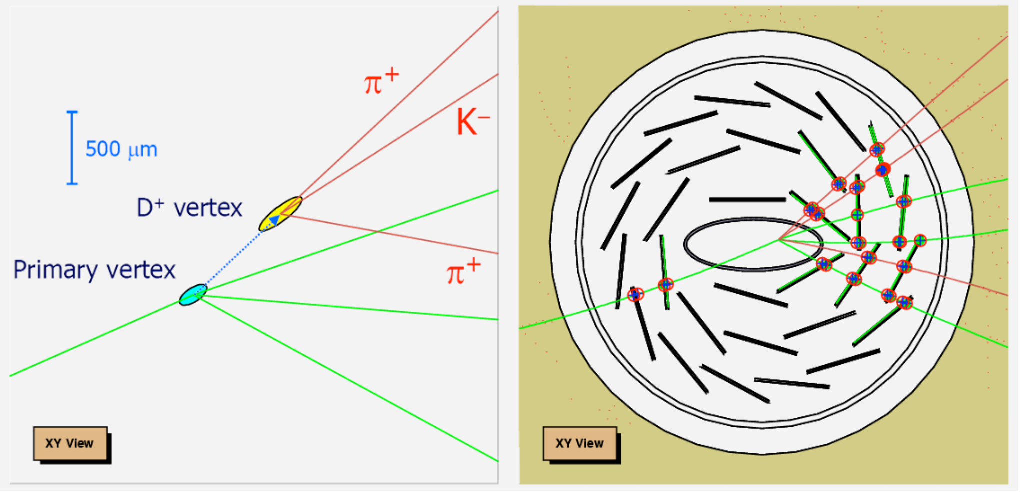


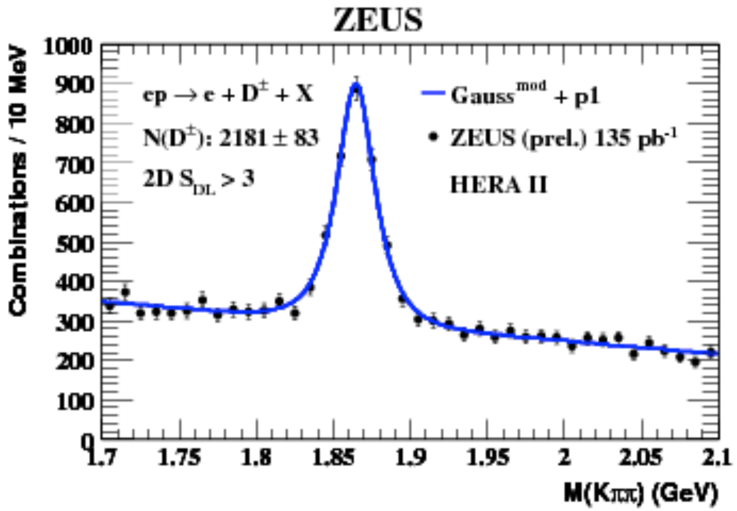


$Q^2 = sxy$

Large Gap in Rapidity

3) Diffractive events and their topologies





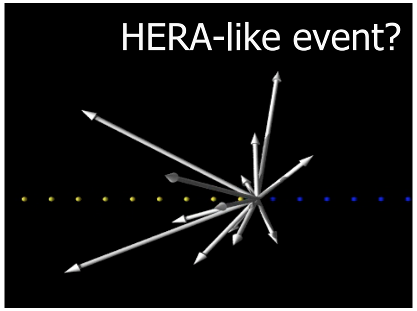
4) Fraction of total DIS cross section from heavy flavours (charm and beauty): particle spectroscopy, inclusive and maybe even lifetime methods (ambitious!)

Outreach Format: Technical Issues

- An outreach format seems reachable from the current software, and would come somewhere in content between the existing HAT and μODS formats

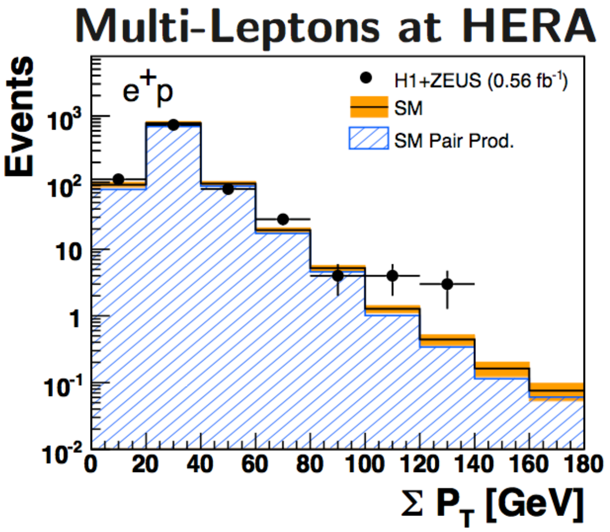
- What about the actual data format?
  - Should consist of simple data types: floats, ints, and arrays..
  - Independent of H1OO, but based on ROOT types (TClonesArrays etc)
  - A single format to cover all outreach projects would be preferable
  - If one wants to include comparison to MC, a universal event weighting scheme which takes into account all efficiencies from triggers, vertex finding and so on, may be prohibitively complex
  - If we only deal with data, then the situation is much simpler

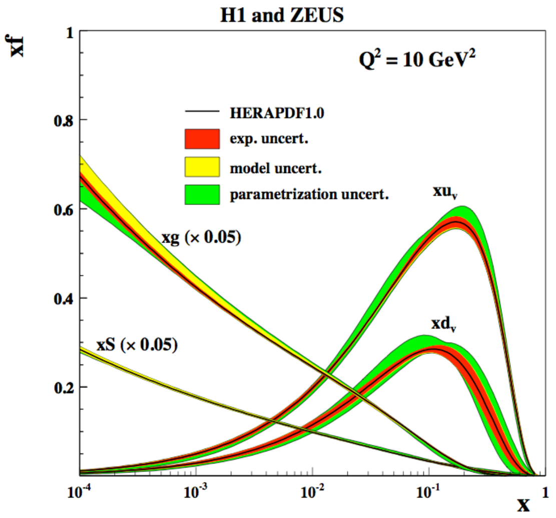
- Would be nice to have something that can interface to Matt Bellis' work in terms of user applications
  - Will certainly be followed up

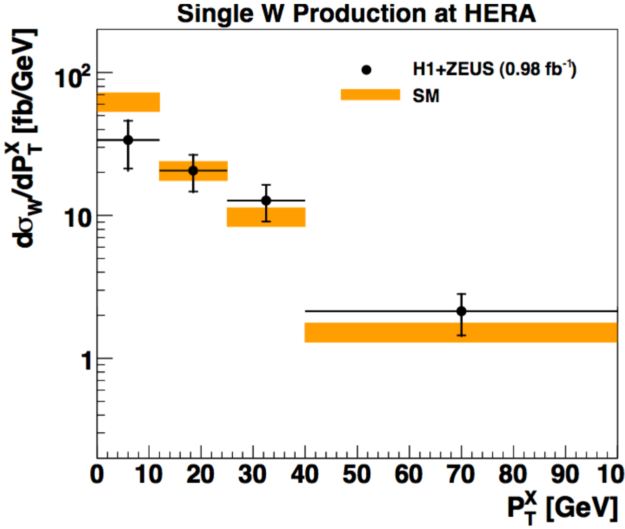


Outreach Format: HERA Format?

- Such a format would be a candidate for combining $e^{\pm}p$ data from the H1 and ZEUS (and even HERMES?) experiments
- 2009 saw the first combined H1+ZEUS publications:







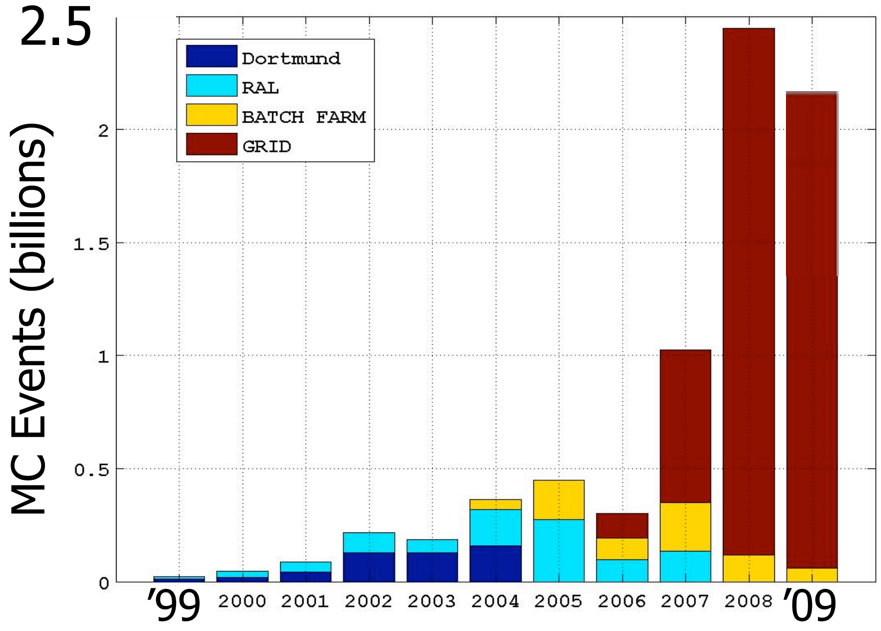
- Some ideas came out of the first HERA data preservation meeting
  - Different strategies in some areas: learn from shared experiences
  - Joint HERA financial proposal would give better chance of support?

Operating Systems and Resources

- Operating System and computing environment at DESY:
  - IBM to UNIX conversion done ‘96, since then a few Linux conversions
  - SLD5 transition happening at DESY now
  - Already had a lot of success, (almost) with full compilation of code, and revealed some missing parts (GKS…)
  - Define SLD5 for H1 with IT in coming months, full transition in 2010

- Mass Storage: how future proof are these systems?
  - Main storage is HERA dCache (/acs), using DESY-IT tape-robot system (duplicate system), with disk pools used for most commonly accessed files
  - Resilient dCache system also commonly used (~ 130/2 TB), benefit of disk only system and duplication
  - Increase in capacity of working group servers (latest models contain 12TB of usable disk space) means increasing use of such systems for analysis level

Large Scale MC Production on the Grid



- Some recent numbers
  - 500M events in 30 days
  - 40M events in one day
  - More than 2B for second year running, expect same for 2010
- Current level of MC production unlikely to be sustained in future

- Automatic μODS/HAT production follows afterwards as separate job

- Assume the Grid does not significantly change before 2014
- Recent global transition of Grid sites to SLD5 incorporated

- Hardware resources in preservation phase (> 2013) to be evaluated
- Will a few large capacity, multi-core machines to be sufficient?

H1 Batch Farm

- Current capacity: about 800 cores across 170 machines, expected to decrease to 700 across 110 in 2010
- Integral part of analysis at H1, most users run parallel analysis jobs
  - Also contains resilient dCache storage
- Some MC Production
  - SIM/REC for requests < 1M events
  - Analysis level μODS /HAT files
- Expected to last until at least 2013, albeit with a reduced capacity

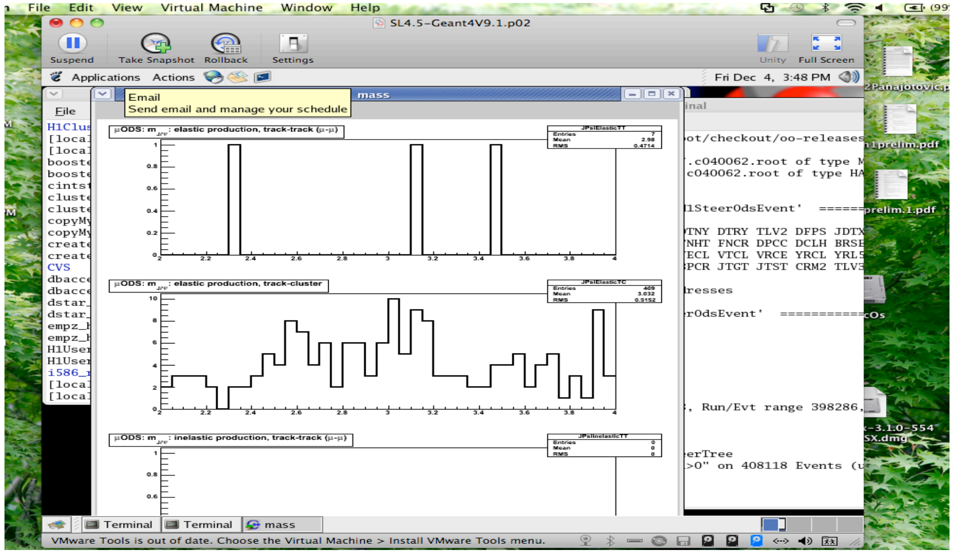
Virtualisation

- Currently no real use of virtualisation within H1
  - *And not planned to play a large role in current preservation model*

- Nevertheless, we try small feasibility project to investigate virtualisation solutions: may help with current SLD4 to SLD5 transition

- May also contribute to validation and tests of preservation model

- First tests underway with H1OO analysis software running on virtual machine (using VMware)



Email
Send email and manage your schedule

JPsiElasticTT	
Entries	7
Mean	2.98
RMS	0.4714

JPsiElasticTC	
Entries	409
Mean	3.032
RMS	0.5152

JPsiInelasticTT	
Entries	0
Mean	0
RMS	0

```
H1Clus
[local
[local
booste
booste
cintst
cluste
cluste
copyMy
copyMy
create
create
CVS
dbacce
dbacce
dstar_
dstar_
empz_h
empz_h
H1User
H1User
i586_r
[local
[local
```

```
ot/checkout/oo-releases

.c040062.root of type M
c040062.root of type HA

1SteerOdsEvent'  ======

TNY DTRY TLV2 DFPS JDTX
NHT FNCR DPCC DCLH BRSE
ECL VTCL VRCE YRCL YRL5
PCR JTGT JTST CRM2 TLV3

resses

rOdsEvent'  ===========




, Run/Evt range 398286,



erTree
>0" on 408118 Events (u
```

VMware Tools is out of date. Choose the Virtual Machine > Install VMware Tools menu.

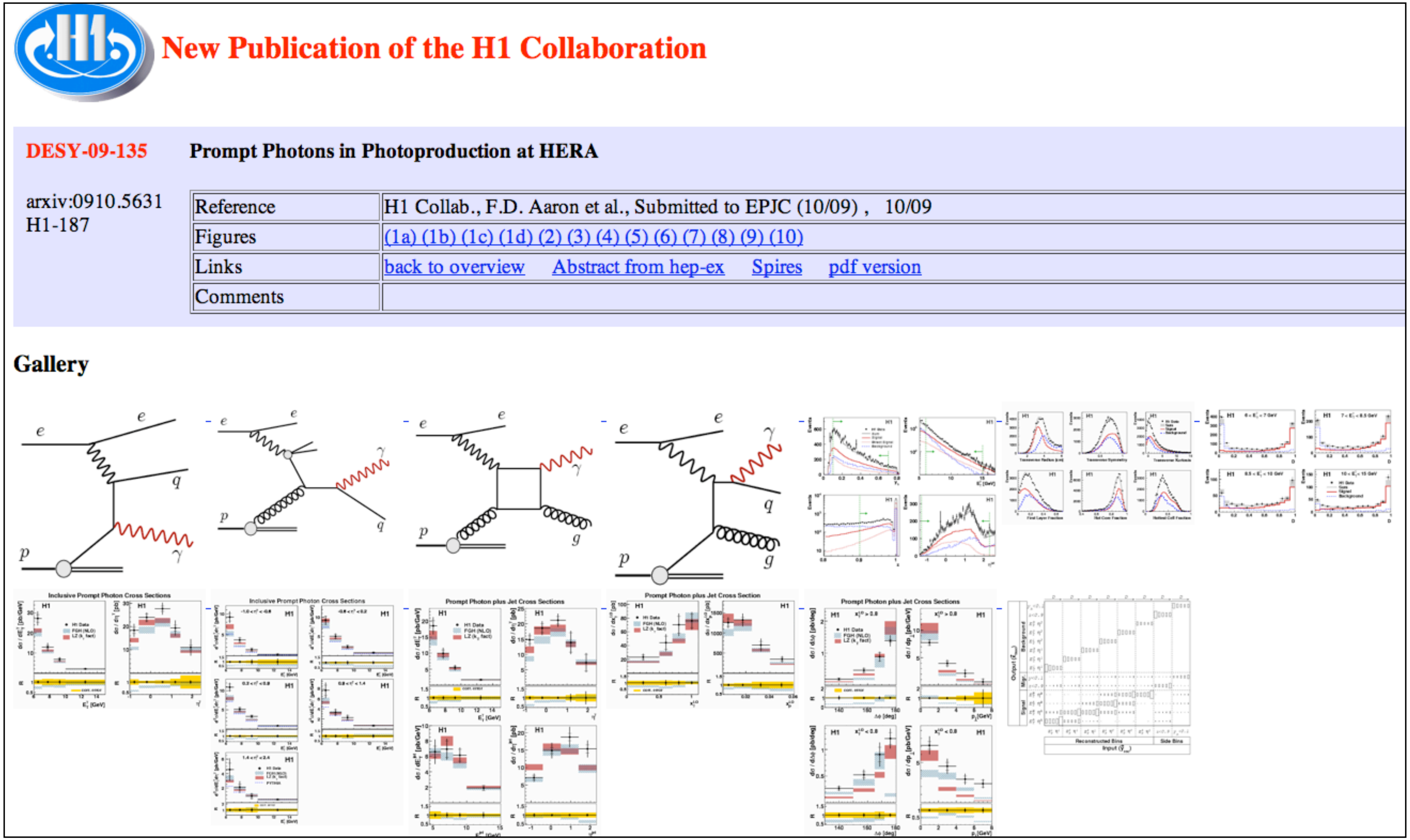
Digital Documentation: Web-based

- H1 Physics: Literature links
  - Published articles: review articles, expert articles, latest results
  - Preliminary results: preliminary experimental results, archive results
  - Talks at meetings, conferences, lectures, university courses
  - H1-Notes, unpublished articles and technical papers
  - Internal wiki pages extensively used by H1
  - Other electronic documentation: Software manuals and notes
  - *How much could INSPIRE / DESY-Library take over in all of this?*

Would be nice to have links to all the plots in all articles, preliminary results and talks, fully searchable…

- Encyclopaedia level of information
  - Explanation of the relevant keywords like: factorisation, coefficient-function, QCD, matrix elements, PDF, perturbation theory, structure functions, cross sections, deep inelastic scattering, factorisation scale, partonic cross section, off-shell, inclusive total cross section, semi-inclusive measurements, diffraction, systematic uncertainty, extrapolation…
  - Links to original literature for keywords, PDG articles, review articles
- Experimental data stored in computer readable form
  - Links to theory (Durham) and experimental databases, for easy comparison
  - Links to original data with detailed explanation (INSPIRE?)

Enhanced Presentation of H1 Results



New Publication of the H1 Collaboration

DESY-09-135 **Prompt Photons in Photoproduction at HERA**

arxiv:0910.5631
H1-187

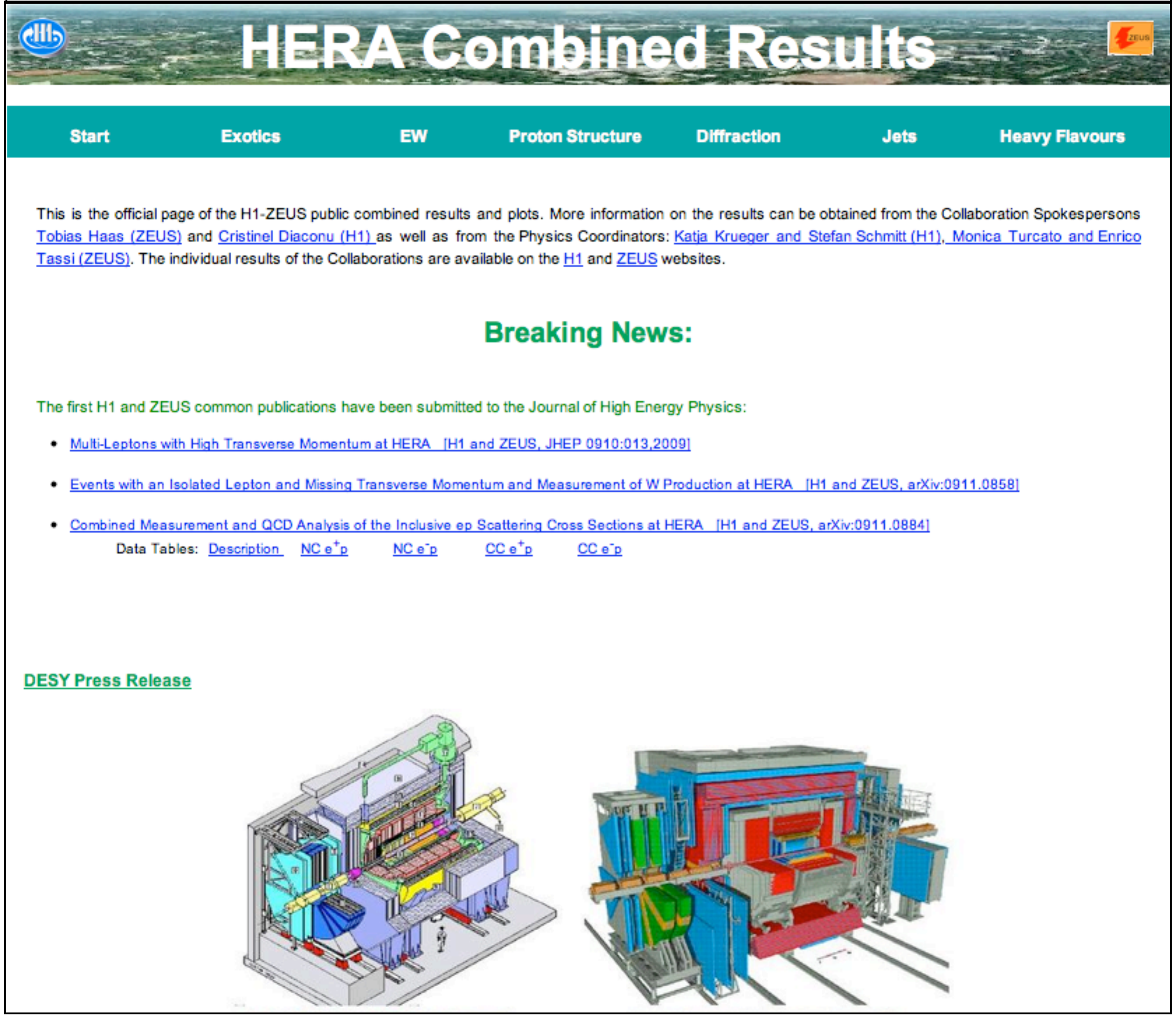
Reference	H1 Collab., F.D. Aaron et al., Submitted to EPJC (10/09) , 10/09
Figures	(1a) (1b) (1c) (1d) (2) (3) (4) (5) (6) (7) (8) (9) (10)
Links	back to overview Abstract from hep-ex Spires pdf version
Comments	

Gallery

e
e
q
p
γ
Inclusive Prompt Photon Cross Sections
Prompt Photon plus Jet Cross Sections
Prompt Photon plus Jet Cross Section
H1

One Last Plug for H1+ZEUS Combinations

http://www.desy.de/h1zeus/combined_results/index.php



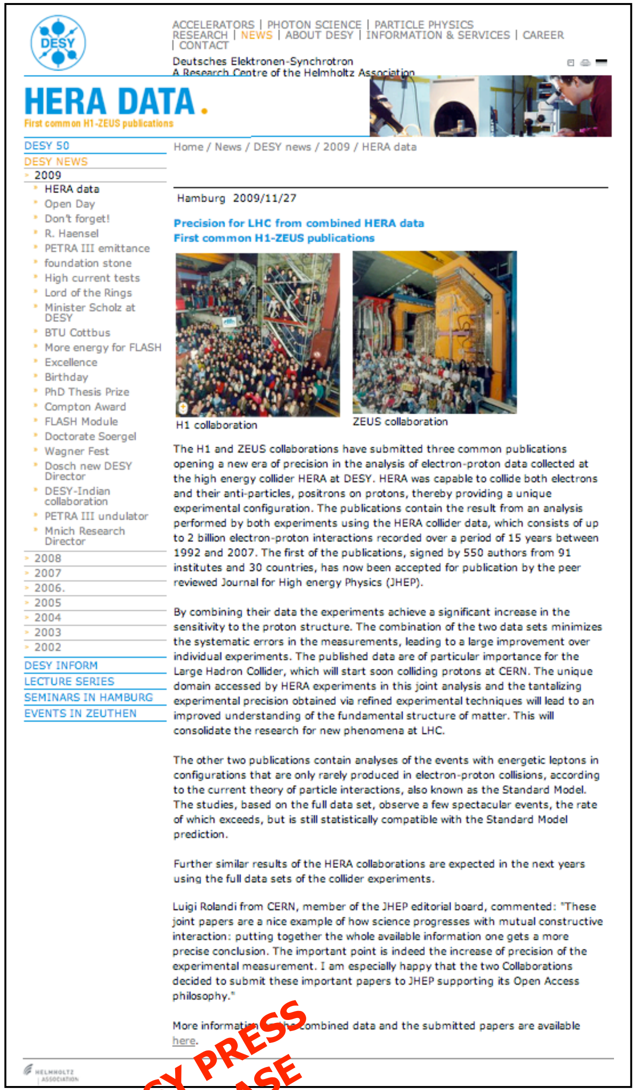
This is the official page of the H1-ZEUS public combined results and plots. More information on the results can be obtained from the Collaboration Spokespersons Tobias Haas (ZEUS) and Cristinel Diaconu (H1) as well as from the Physics Coordinators: Katja Krueger and Stefan Schmitt (H1), Monica Turcato and Enrico Tassi (ZEUS). The individual results of the Collaborations are available on the H1 and ZEUS websites.

Breaking News:

The first H1 and ZEUS common publications have been submitted to the Journal of High Energy Physics:

- Multi-Leptons with High Transverse Momentum at HERA [H1 and ZEUS, JHEP 0910:013,2009]
- Events with an Isolated Lepton and Missing Transverse Momentum and Measurement of W Production at HERA [H1 and ZEUS, arXiv:0911.0858]
- Combined Measurement and QCD Analysis of the Inclusive ep Scattering Cross Sections at HERA [H1 and ZEUS, arXiv:0911.0884]
  Data Tables: Description NC e^+p NC e^-p CC e^+p CC e^-p

DESY Press Release



Hamburg 2009/11/27

Precision for LHC from combined HERA data
First common H1-ZEUS publications

H1 collaboration

ZEUS collaboration

The H1 and ZEUS collaborations have submitted three common publications opening a new era of precision in the analysis of electron-proton data collected at the high energy collider HERA at DESY. HERA was capable to collide both electrons and their anti-particles, positrons on protons, thereby providing a unique experimental configuration. The publications contain the result from an analysis performed by both experiments using the HERA collider data, which consists of up to 2 billion electron-proton interactions recorded over a period of 15 years between 1992 and 2007. The first of the publications, signed by 550 authors from 91 institutes and 30 countries, has now been accepted for publication by the peer reviewed Journal for High energy Physics (JHEP).

By combining their data the experiments achieve a significant increase in the sensitivity to the proton structure. The combination of the two data sets minimizes the systematic errors in the measurements, leading to a large improvement over individual experiments. The published data are of particular importance for the Large Hadron Collider, which will start soon colliding protons at CERN. The unique domain accessed by HERA experiments in this joint analysis and the tantalizing experimental precision obtained via refined experimental techniques will lead to an improved understanding of the fundamental structure of matter. This will consolidate the research for new phenomena at LHC.

The other two publications contain analyses of the events with energetic leptons in configurations that are only rarely produced in electron-proton collisions, according to the current theory of particle interactions, also known as the Standard Model. The studies, based on the full data set, observe a few spectacular events, the rate of which exceeds, but is still statistically compatible with the Standard Model prediction.

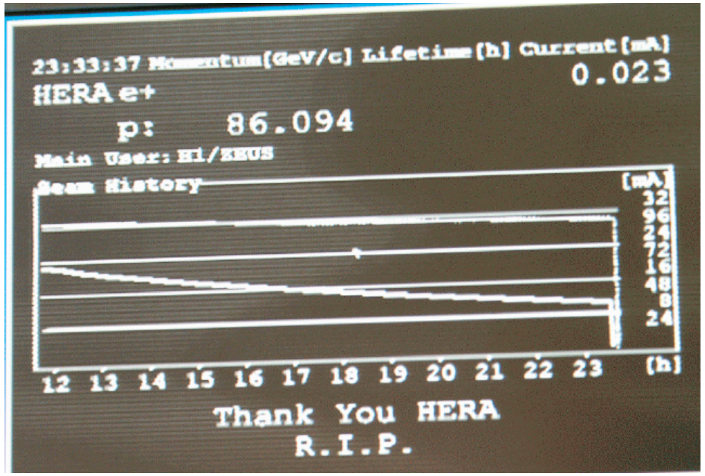
Further similar results of the HERA collaborations are expected in the next years using the full data sets of the collider experiments.

Luigi Rolandi from CERN, member of the JHEP editorial board, commented: "These joint papers are a nice example of how science progresses with mutual constructive interaction: putting together the whole available information one gets a more precise conclusion. The important point is indeed the increase of precision of the experimental measurement. I am especially happy that the two Collaborations decided to submit these important papers to JHEP supporting its Open Access philosophy."

More information on the combined data and the submitted papers are available here.

DESY PRESS RELEASE

Digital Documentation: Online Data



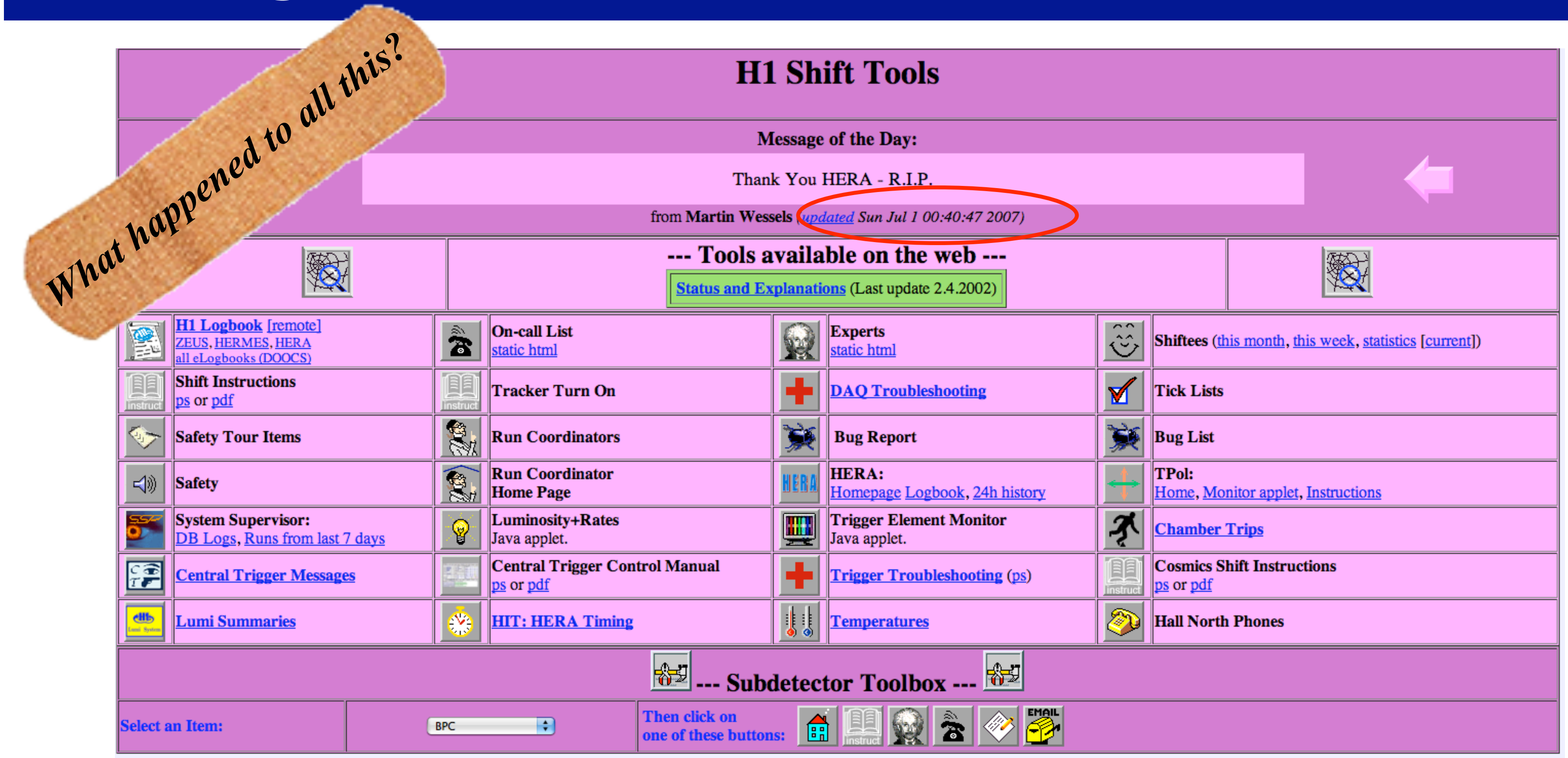
```
23:33:37 Momentum(GeV/c] Lifetime(h] Current(mA]
                                           0.023
HERA e+
    p:      86.094
Main User: H1/ZEUS
```

Thank You HERA
R.I.P.

- While we were marking the end of HERA running and data taking a collection of applications were running in the North Hall

- In fact I would guess (from memory) there were about 20 machines associated with different detector components ticking away..

Digital Documentation: Online Data



What happened to all this?

H1 Shift Tools

Message of the Day:

Thank You HERA - R.I.P.

from **Martin Wessels** (updated *Sun Jul 1 00:40:47 2007*)

--- Tools available on the web ---

Status and Explanations (Last update 2.4.2002)

H1 Logbook [remote] ZEUS, HERMES, HERA all eLogbooks (DOOCS)	On-call List static html	Experts static html	Shiftees (this month, this week, statistics [current])
Shift Instructions ps or pdf	Tracker Turn On	DAQ Troubleshooting	Tick Lists
Safety Tour Items	Run Coordinators	Bug Report	Bug List
Safety	Run Coordinator Home Page	HERA: Homepage Logbook, 24h history	TPol: Home, Monitor applet, Instructions
System Supervisor: DB Logs, Runs from last 7 days	Luminosity+Rates Java applet.	Trigger Element Monitor Java applet.	Chamber Trips
Central Trigger Messages	Central Trigger Control Manual ps or pdf	Trigger Troubleshooting (ps)	Cosmics Shift Instructions ps or pdf
Lumi Summaries	HIT: HERA Timing	Temperatures	Hall North Phones

--- Subdetector Toolbox ---

Select an Item: BPC

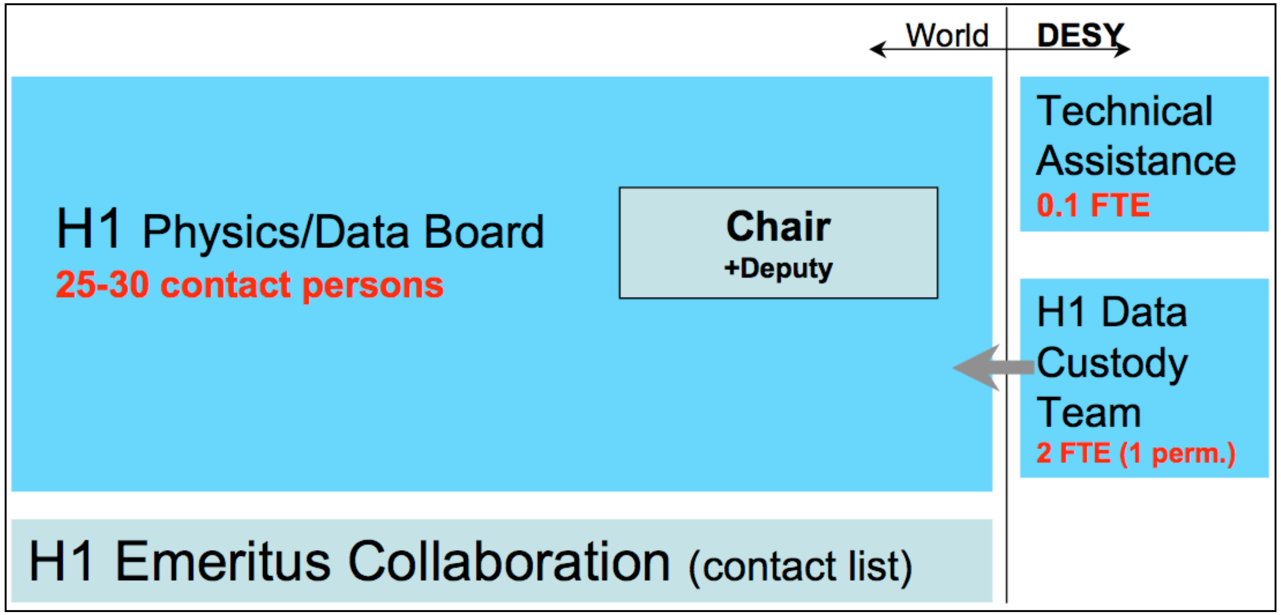
Then click on one of these buttons:

- Old online shift tools: may be particularly vulnerable to losses
  - Mostly not updated since July 2007
  - Electronic logbooks: H1, trigger, components, detailed run information
  - Calibration files on old hardware, was it all rescued / can it be rescued now?
  - Old applications like ZUBR (online L4 histograms: data quality) may be useful

Documentation: Non-Digital

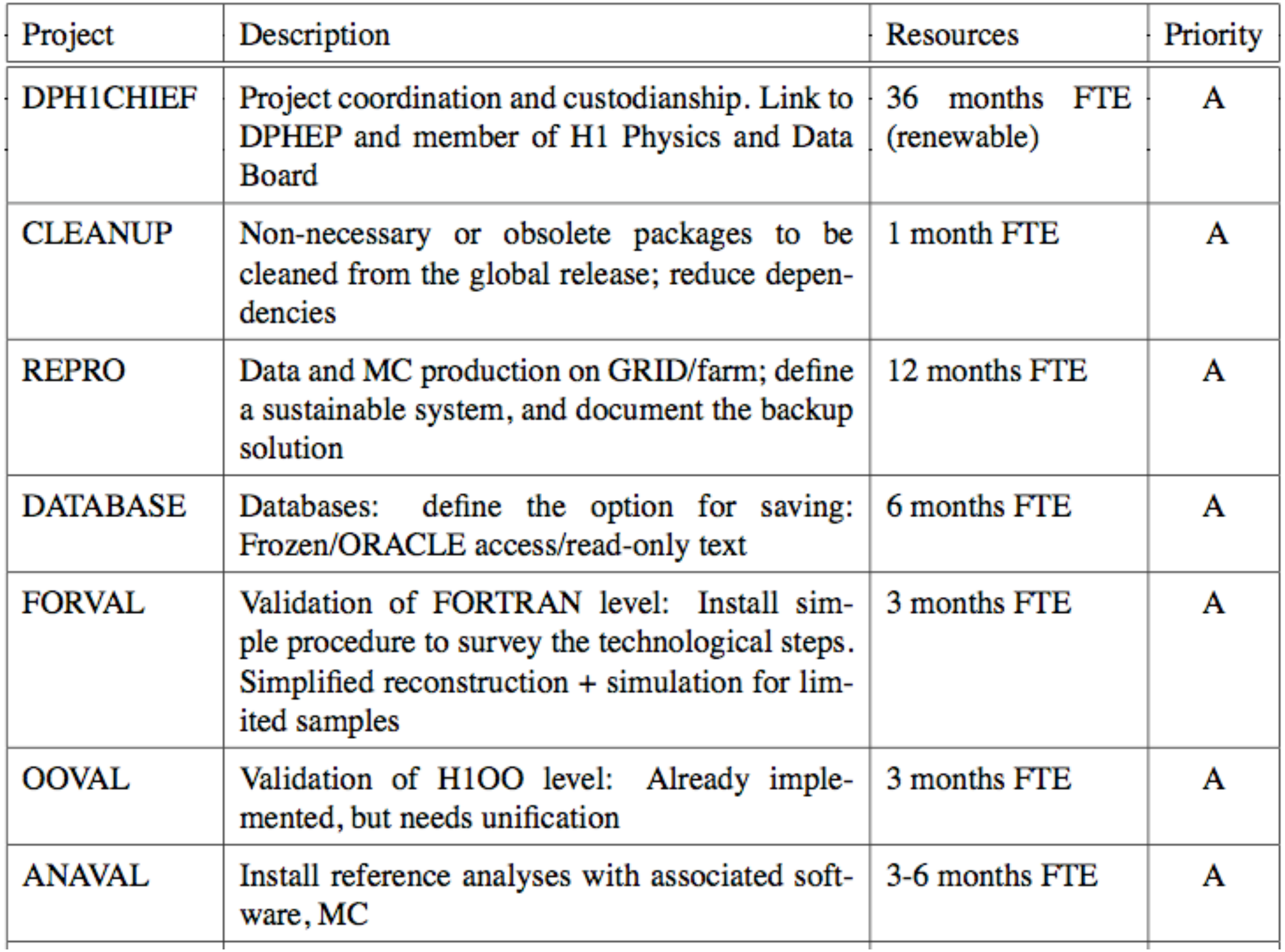
- Location: Where is everything now?
  - H1 physics and technical talks from pre-web days (pre-1995/6)
  - Detector schematics, blueprints..
  - North Hall *again*: artefacts: logbooks..?
- Is digitisation a viable solution?
  - Can we digitise paper documents?
  - Or pay someone to do it?
  - Which items, do we need it, how much?
- Future location: Where can we put everything?
  - The H1 documentation room is due to move (renovation at DESY)
  - Need to catalogue the exiting contents and consolidate documentation in one place from other areas at DESY
  - Can the (moved) library provide physical space?

Possible Model for Future Governance of H1



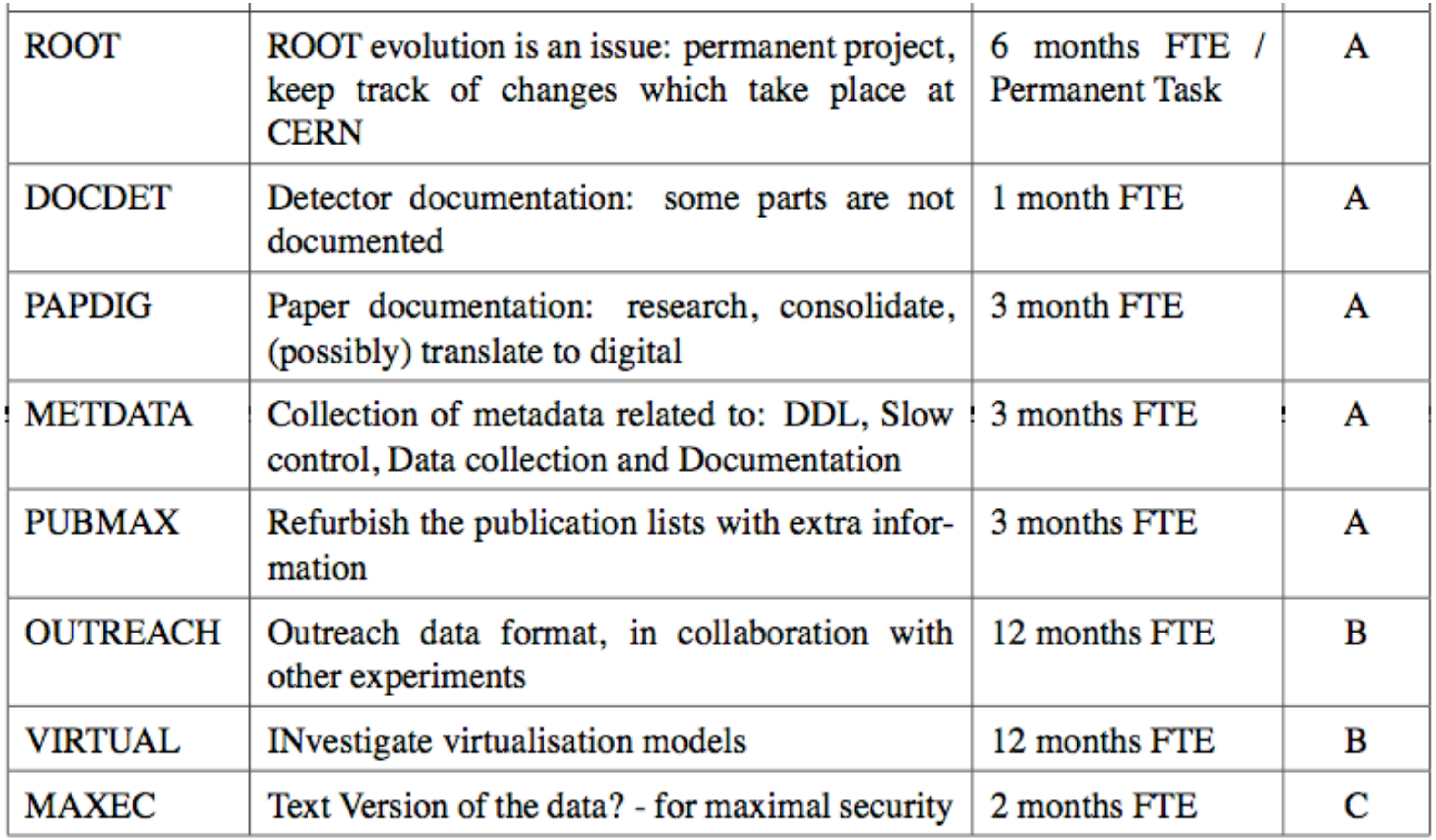
- Tasks of the H1 Physics/Data Board
  - General contact point for H1 Data/Physics beyond lifetime of the collaboration
  - Communicate to the host lab (DESY) and other experiments
  - Supervise H1 data: Data Custody Team report to H1PDB
  - Contact to DPHEP
  - Overview further publications using H1 Data
- Envisage one meeting/year (remote), reports, web-site, events

Start to define some H1 Projects



Project	Description	Resources	Priority
DPH1CHIEF	Project coordination and custodianship. Link to DPHEP and member of H1 Physics and Data Board	36 months FTE (renewable)	A
CLEANUP	Non-necessary or obsolete packages to be cleaned from the global release; reduce dependencies	1 month FTE	A
REPRO	Data and MC production on GRID/farm; define a sustainable system, and document the backup solution	12 months FTE	A
DATABASE	Databases: define the option for saving: Frozen/ORACLE access/read-only text	6 months FTE	A
FORVAL	Validation of FORTRAN level: Install simple procedure to survey the technological steps. Simplified reconstruction + simulation for limited samples	3 months FTE	A
OOVAL	Validation of H1OO level: Already implemented, but needs unification	3 months FTE	A
ANAVAL	Install reference analyses with associated software, MC	3-6 months FTE	A

Start to define some H1 Projects



ROOT	ROOT evolution is an issue: permanent project, keep track of changes which take place at CERN	6 months FTE / Permanent Task	A
DOCDET	Detector documentation: some parts are not documented	1 month FTE	A
PAPDIG	Paper documentation: research, consolidate, (possibly) translate to digital	3 month FTE	A
METDATA	Collection of metadata related to: DDL, Slow control, Data collection and Documentation	3 months FTE	A
PUBMAX	Refurbish the publication lists with extra information	3 months FTE	A
OUTREACH	Outreach data format, in collaboration with other experiments	12 months FTE	B
VIRTUAL	INvestigate virtualisation models	12 months FTE	B
MAXEC	Text Version of the data? - for maximal security	2 months FTE	C

Summary and Next Steps

- H1 will aim for a level 4 data preservation programme
  - The $e^{\pm}p$ collisions collected at HERA are a unique data set!
  - Physics motivation detailed, full flexibility desirable

- Task force set up, survey of the relevant preservation issues
  - Data and data formats
  - Technologies and Resources
  - Documentation

- An outreach format for H1 is a nice idea, project to run in parallel
  - Is attractive to physicists (fun) and financiers (global benefit)

- Isolate projects from the survey and start to attribute cost (FTEs)

- Formal written proposal for data preservation at H1 in progress
  - Joint HERA issues, should be considered carefully