Discussions between ep Experiments

David South (H1, TU Dortmund)

Eduard Avetisyan (HERMES), Cristinel Diaconu (H1),
Larry Felawka (HERMES), Sergey Levonian (H1), Bogdan Lobodzinski (H1),
Jan Olsson (H1), Dmitri Ozerov (H1, DESY-IT), Gunar Schell (HERMES),
Janusz Szuba (ZEUS), Krzysztof Wrona (ZEUS)

First Workshop on Data Preservation and Long Term Analysis

26 - 28 January 2009

Physics with HERA: A Unique Collider

The collisions recorded from HERA are a unique data set

A rich physics program, from DIS to unique searches, measurement of the longitudinal structure function F_L and investigations into the spin structure of the proton

The data should of course be preserved, especially as they are unlikely to be superseded in the near future

Lively discussion of "Use Cases", some "Models for Preservation" and the idea of a "Common Repository"

Possible Future Use Cases

What types of Use Cases can be imagined - why would we need to access the data again?

- Essentially available for everyone and anyone: real open access: “anything”

- New analysis to be done by experts who know the (analysis level) software
    - Re-do existing analysis but in new phase space
    - Re-do existing analysis but with more data (from other experiment?)

- A new theory comes out: need the new simulation - how, and how difficult?

- But new theory / observation means new reconstruction is desirable, ie the new idea is currently killed by a harsh cut: back to RAW

Models for Preservation

Level	
0	RAW data
1	Reconstruction Simulation Database considerations..
2	DST
3	Ntuple / analysis level data (and MC?) *production*
4	Existing ntuple / analysis level
5	Combined analysis with an H1+ZEUS ntuple
6	Outreach : very simple format

- The basic level to conserve

- Essentially frozen, but reconstruction software still compiles, so changes are possible…
- New simulation, can use old reconstruction?

- DST level expects no further development, (but see above…)

- Rolling model proposed for by H1, fluid preservation from here: gives regular verification of full chain

- Fix the ntuple now, more like ZEUS

- See next slide

- Not enough for full analysis (?), but rather for open access

H1 vs ZEUS: A Common Repository?

A comparison of some H1 and ZEUS numbers:

	H1	ZEUS
RAW (kB/event)	75	125
POT (kB/event)	200	-
(m)DST (kB/event)	18	75
MC (m)DST (kB/event)	40	200
μODS (kB/event)	3	-
HAT (kB/event)	0.4	-
Common ntuple (kB/event)	-	10
Number of data events	1 billion	0.51 billion
Total data to conserve (TB)	100TB (raw + HERA I+II DST + μODS +HAT)	30TB (HERA II mDST)
Total MC (TB)	100-200	400
Estimated storage needed (TB)	200-300	430

Same format H1+ZEUS data is an idea, *could do same time as "outreach" format*

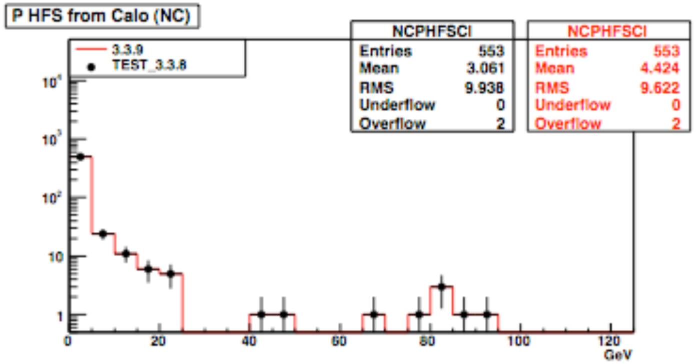
- As example: full (searches) analysis in place in one experiment: take 500 pb^{-1} data from the other experiment and produce improved HERA limit
- But different experimental set ups, resolutions: how to factorise out the detectors?

Certification, Validation Models

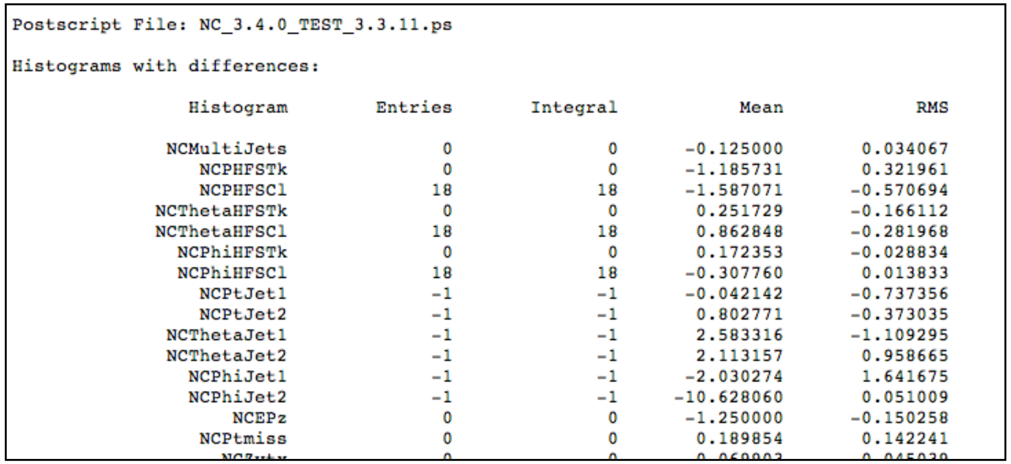
- H1 would like to recompile analysis software, recreate analysis ntuple regularly, say every 3 months

- Use benchmark analyses to check and compare results

- Both the above already exist for example in “H1Validation” package



3.3.8 vs 3.3.9



```
Postscript File: NC_3.4.0_TEST_3.3.11.ps

Histograms with differences:

         Histogram       Entries      Integral          Mean           RMS

       NCMultiJets             0             0     -0.125000      0.034067
          NCPHFSTk             0             0     -1.185731      0.321961
          NCPHFSCl            18            18     -1.587071     -0.570694
      NCThetaHFSTk             0             0      0.251729     -0.166112
      NCThetaHFSCl            18            18      0.862848     -0.281968
        NCPhiHFSTk             0             0      0.172353     -0.028834
        NCPhiHFSCl            18            18     -0.307760      0.013833
          NCPtJet1            -1            -1     -0.042142     -0.737356
          NCPtJet2            -1            -1      0.802771     -0.373035
       NCThetaJet1            -1            -1      2.583316     -1.109295
       NCThetaJet2            -1            -1      2.113157      0.958665
         NCPhiJet1            -1            -1     -2.030274      1.641675
         NCPhiJet2            -1            -1    -10.628060      0.051009
            NCEPz              0             0     -1.250000     -0.150258
          NCPtmiss             0             0      0.189854      0.142241
```

3.3.11 vs 3.4.0

Risk Analysis

- Database: no longer needed? Snap shot file of (probably) no longer changing items could remove dependence on commercial software (Oracle)

- What about shared libraries (CERNLIB)

- Un-maintained open-source software..

- Changes in operating system may be non-trivial (DL5 to SL4 experience)

- Documentation losses..

Infrastructure, What We Need

- Define clearly the data and MC sets that represent the legacy

- DST/RAW: Storage and the reading back of old files (who will fund this?)

- Hardware: A few (powerful) machines (like latest h1wgs) + up to 1 FTE

- External software specifications (commitment of ROOT, ready to collaborate)

- Need guaranteed GRID resources in current model for data+MC mass production

- Virtualization (or emulation? : “Black-boxing”) only as last resort.. Rather have rolling preservation model..

- Begin to set out program now, plan to meet again soon