Data Preservation and Long Term Analysis in High Energy Physics

David South (TU Dortmund)

Study Group for Data Preservation and Long Term Analysis in High Energy Physics

The 44th Rencontres de Moriond on QCD and High Energy Interactions

La Thuile, Aosta Valley, Italy
March 14 - 21, 2009
High Energy Physics Data are Unique
High Energy Physics Data are Unique

A generation of HEP experiments are concluding their data taking and winding up their physics programmes.
High Energy Physics Data are Unique

The experimental data from these experiments still has much to tell us, from the existing analyses still to be completed..

A generation of HEP experiments are concluding their data taking and winding up their physics programmes.
High Energy Physics Data are Unique

A generation of HEP experiments are concluding their data taking and winding up their physics programmes.

The experimental data from these experiments still has much to tell us, from the existing analyses still to be completed.

..but they may also contain things we do not yet know, which may only come to light at a later date via LHC data or a new theory.
Why Should We Preserve HEP Data?

• We may want to re-do previous measurements
  – Increased precision, reduced systematics
  – New and improved theoretical calculations / MC models
  – Newly developed analysis techniques

• We may want to perform new measurements
  – At energies and processes where no other data are available (or will become available in the future)

• Investigate if new phenomena found today
  – Go back and check in the old data
The Importance of Data Preservation

In your opinion, how important is the issue of data preservation?
(top/blue: theorists, bottom/green: experimentalists)

- Irrelevant: 0.4% (0.9%),
- Moderately important: 8.7% (3.3%),
- Important: 25.6% (15.2%),
- Very important: 40.5% (41.7%),
- Crucial: 24.3% (39.5%)

- Most of surveyed HEP physicists are in favour of data preservation.
The Challenge of Preserving HEP Data

• HEP has little or no tradition or clear current model of long term conservation of data in a meaningful and useful way

• The preservation of and supported long term access to the data is generally not part of the planning, software design or budget of a HEP experiment
  – The main assumption has probably been that the data will always be superseded by the next experiment: but this is not always the case!
  – Another (sometimes wrong) assumption is that the physics potential is exhausted at the end of the program

• There is also little tradition of useful open access of HEP data beyond the walls of the original collaboration
  – This is clearly a difficult prospect, with many issues like control, correctness and reputation of the experiment, not to mention a lack of portability and the state of the documentation
Re-analysis of JADE Data from PETRA

Precision measurements of $\alpha_S$ and tests of asymptotic freedom

In NLO QCD: $\alpha_S(35 \text{ GeV}) = 0.14 \pm 0.02$
No indication of a running $\alpha_S$ signature

In re-summed NNLO QCD: $\alpha_S(M_Z) = 0.1172 \pm 0.0051$
Significant evidence of running $\alpha_S$ and asym. freedom
An Inter-Experimental Study Group

• Start a common enterprise between experiments and associated computing centres
  – First contacts made and meetings held autumn 2008

• International Steering Committee
  – Made up of the Spokespersons of the HEP experiments and the Directors of the associated computing centres

• International Advisory Committee
  – Chaired by J. Dorfan and S. Bethke

• First Workshop at DESY in January 2009
  – Initial discussions, to share ideas and to see the current picture

• Follow up workshop at SLAC in May 2009
  – To achieve some concrete goals and converge on a set of recommendations to ICFA in form of a blueprint for data preservation
Broad Representation of the HEP Community

- All large HEP centres represented at the workshop, common enterprise between the experiments and associated computing centres

- Additional contributions from older experiments (LEP, PETRA) in dedicated session on past experiences

- Also representatives from funding agencies, (US/DoE, UK/STFC, EU/FP7) and a guest speaker from astrophysics
The First DPLTA Workshop at DESY

- Workshop held at DESY with ~50 participants and lots of useful discussion
- Two days of pre-prepared talks, third morning of dedicated discussions
- Conference webpage: [http://indico.cern.ch/conferenceDisplay.py?confId=42722](http://indico.cern.ch/conferenceDisplay.py?confId=42722)
## DPLTA Workshop Agenda

### Reports from Experiments: Data Analysis and Computing Models
Conveners: Homer Neal; Robert Roser (Fermilab)
(Seminar 4 (EVO): 09:00 - 12:20)

### Computing Centres and Technologies
Conveners: Volker Guelzow (Unknown); Frederic Hemmer (CERN)
(Seminar 4 (EVO): 14:00 - 17:00)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(Seminar 4 (EVO): 17:00 - 18:00)</td>
<td>(Seminar 3a: 17:00 - 18:00)</td>
<td>(Seminar 5: 17:00 - 18:00)</td>
</tr>
</tbody>
</table>

### Reports from Past Experiences of Data Preservation
Conveners: Tobias Haas; Takashi Sasaki
(Seminar 4 (EVO): 09:30 - 12:45)

### Open Access and Long Term Collaborative Governance
Conveners: Richard Mount; Siegfried Bethke (Max-Planck-Institut fur Physik)
(Seminar 4 (EVO): 14:00 - 17:05)

### Workshop Discussions: Options for Long Term Data Analysis (Summary Discussions)
Conveners: Cristinel Diaconu (Faculte des Sciences de Luminy)
(Seminar 4 (EVO): 09:30 - 12:40)

- 2.5 days, 24 talks, 3 parallel sessions, 8 organised discussions

---

David South, Data Preservation in HEP, Moriond QCD 2009
• Familiar descriptions of data analysis chain, from reconstruction to analysis level
  - RAW → POT → DST → ntuple
Some Information from the Experiments

• **Data Format and Volume**
  – Differences appear early on: US experiments use skims at early stage: different ntuples for different physics working groups
  – Event sizes vary from a few to 100 kB/event; total size of expected data and MC to conserve 0.5 to 10 PB

• **Software**
  – Reconstruction C++ / C / Fortran; Simulation GEANT 3 (Fortran) or 4 (C++)
  – ROOT and C++ almost universal as analysis level software (Belle: Fortran)

• **Reprocessing**
  – Done by many experiments but (for example) CDF plans not to

• **Simulation**
  – MC production generally done on GRID, analysis on local farm (≈10^3 CPUs)
The Challenge of Handling HEP Data

Total: 15 PB
The Challenge of Handling HEP Data

• Massive data traffic, storage and migration is within the scope of all HEP computing centres

• The conservation of tapes is not equivalent to data preservation
  – Older tapes are often not accessible after 2-3 years

• The distribution of data complicates the task

• There is a grey area between the experiments and the computing centres concerning the long term preserved data
  – Missing Hardware: unreadable tapes
  – Software for accessing the data is usually under the control of the experiments
Data Migration

- Continuous migration of archived data works if:
  - The next generation media costs half as much
  - All media is robot managed (no shelved tapes!)
  - Migration is possible on a short timescale

- Future mass storage media: SSD (Flash..)
  - More reliable than disk or tape
  - Requires less power than disk (x10)
  - Currently more expensive (x20), but coming down
  - Density increasing with respect to disks (smaller footprint)

- Consistent data migration policy still to be defined
  - Data loss and risk analysis
  - Geographical placement of the data sets
  - Uniform framework for scientific data management

Total cost of data migration
= double current costs:
1 + 1/2 + 1/4 + 1/8 .. = 2
Past Experiences of Data Preservation: PETRA

- Recent re-analysis of JADE data is such a case (S. Bethke, J. Olsson et al.)
  - Conversion of old data format done in 2005 and 2008 by Jan
  - Successful revitalisation and validation of complete JADE reconstruction and simulation software and event display
  - Involved conversion, translation and some rewriting of original code

- There were several interesting JADE anecdotes along the way, including:
  - A hand-typed recovery of a luminosity / calibration file from a (green) paper copy found in Jan’s DESY office
  - An old version of BOSLib79 found at the Tokyo computing centre
  - Original JADE MC 9-track tapes found in a Heidelberg University cupboard

- Only through careful documentation will we avoid such things
Past Experiences of Data Preservation: LEP

- There is still a significant number of new LEP publications, approaching 10 years after the end of taking and even after the official end of collaborations

  Plot by Travis Brooks (SPIRES/SLAC)

- However, no coherent approach was attempted, no project to preserve the full data analysis capabilities
  - Several analyses still alive (ALEPH laptop model, Higgs group high level data)
  - But in general the preserved data are lacking in standardisation and have limited, model dependent usage

- There is an imminent danger that the LEP data will be absorbed in the "digital black hole" in a few years, if nothing is done
Past Experiences of Data Preservation

- It is likely that most older HEP experiments have in fact simply lost the data

- For the few known preserved data examples, in general the exercise has not been a planned initiative by the collaboration, but a push by knowledgeable people

- The task in hand is to provide a coherent set of guidelines for future experiments to ensure the longevity of our data
Plans for a Systematic Approach to Data Preservation in HEP

- Physics Cases for Data Preservation in HEP
  - Survey of possible benefits from data preservation
  - Including business models
  - Including links with other research fields

- Preservation Models
  - Input from ee, ep, pp experiments
  - Priorities, costs and benefits, links to technology

- Collaborations, Governance, and Data Access Policies
  - Including contacts with general initiatives

- Technologies and Facilities
  - Survey and assessment of existing infrastructures in HEP and their adaptability to data preservation demand
  - Reflection on the impact of new technologies on the data preservation methods
Physics Case for Data Preservation

- The ep collisions recorded at HERA are a unique data set, unlikely to be superseded in the near future (LHeC?)

- The pp collision data from Tevatron will also provide a contingency for LHC data, as well as a lower energy point

- As already shown, ee data may still provide future interest, e.g. for comparison to Super-B, in particular if data sets can be combined (BaBar + Belle in trials now)
Possible Future Use Cases

What types of use cases can be imagined?
What’s the justification: why will we need to access the data again?

- *Real* open access: essentially available for everyone and anyone

- A new analysis to be done by experts who know the analysis 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 a 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: need to go back to RAW data
## Models for Preservation

<table>
<thead>
<tr>
<th>Minimum Level of Preservation</th>
<th>The basic level to conserve</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>0</strong> RAW data</td>
<td>Essentially frozen, but reconstruction software still compiles, so changes are possible…</td>
</tr>
<tr>
<td><strong>1</strong> Reconstruction</td>
<td>A new simulation: can it use old reconstruction (issue of F vs C++)?</td>
</tr>
<tr>
<td>Simulation</td>
<td>DST level expects no further development, this is the final version</td>
</tr>
<tr>
<td>Database considerations?</td>
<td>Rolling model, fluid preservation from here up: gives regular verification of full chain</td>
</tr>
<tr>
<td>Commercial software?</td>
<td>Fixed ntuple, “all” analysis level info</td>
</tr>
<tr>
<td><strong>2</strong> DST</td>
<td>Common format ntuple (repository?)</td>
</tr>
<tr>
<td><strong>3</strong> Ntuple / analysis level data (and MC?) <em>production</em></td>
<td>Not enough for full analysis(?), but rather for open access / outreach</td>
</tr>
<tr>
<td><strong>4</strong> Existing ntuple / analysis level</td>
<td></td>
</tr>
<tr>
<td><strong>5</strong> Combined analysis with a (for example) H1+ZEUS “ep ntuple”</td>
<td></td>
</tr>
<tr>
<td><strong>6</strong> Outreach : very simple format</td>
<td></td>
</tr>
</tbody>
</table>

Use of Virtualisation / Emulation techniques?
Data Access and Supervision in Long Term

- Data preservation costs money and it needs organisation
  - Laboratories need to take into account appropriate budget planning

- HEP data preservation makes sense if the appropriate supervision structures are defined

- Long term evolution of the HEP collaborations need to be initiated during the active lifetime

- Data access and publication procedures for a hibernated collaboration should be defined

- Data stewardship in computing centres must be defined taking into account the preservation models and included in the economical considerations of the DPHEP projects
News from SPIRES

SURVEY OF OVER 2000 PHYSICISTS
Which HEP information system do you use the most?

- SPIRES soon to transform into INSPIRE, whereby they will offer custodianship of more than scientific articles

- The addition of metadata and perhaps even the data themselves provides an opportunity of defining standards and practices
Summary

- HEP data is a long term investment and contains a true potential for physics results beyond the collaborations lifetime

- A study group has been formed to reflect on data preservation and long term analysis in HEP: [http://dphep.org](http://dphep.org)

- The aim of this initiative is to provide a written document to ICFA containing guidelines on this subject

Next workshop is at SLAC, May 26-28 2009.