Long Term Data and Knowledge Preservation

Jamie.Shiers@cern.ch

For the DPHEP[1] Collaboration

1 The Problem

The data from the world's particle accelerators and colliders is both **costly** and **time consuming** to produce.

It contains a **wealth** of scientific potential, plus high value for educational **outreach**.

Given that much of the data is **unique**, it is essential to preserve not only the data but also the **full capability** to reproduce past analyses and perform new ones.

There are numerous cases where data from a past experiment was re-analyzed: we must retain the ability in the future.

2 The Approach

Whilst retaining a holistic view, the problem is broken down into a number of key areas. Each is addressed using state-of-the-art techniques, that include:

1. **Digital library** tools (Invenio[2]) & services (CDS[3], INSPIRE[4], ZENODO[5])
2. **Sustainable software**, coupled with advanced **virtualization** techniques[6] and **validation** frameworks[7]
3. Proven bit preservation at the 100PB scale, together with a **sustainable** funding model with an outlook to 2040/50

3 Results

Several tens of PB of data – from the BaBar experiment at SLAC, the CDF and D0 experiments at Fermilab, as well as the H1, HERMES and ZEUS experiments at DESY are preserved.

Data from the LEP and LHC experiments at CERN as well as that from many others, is also addressed with a strong focus on sustainable solutions.

These preservation activities build on the tools and services listed above.

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

4 Partners

DPHEP involves all major laboratories and experiments worldwide, plus a number of key funding agencies.

Working through the **RDA**, the **APA** and others, the intent is to share experience, services and tools as widely as possible.

Bit preservation experience and services at the the **100PB** scale were presented recently at the **RDA Europe** event.

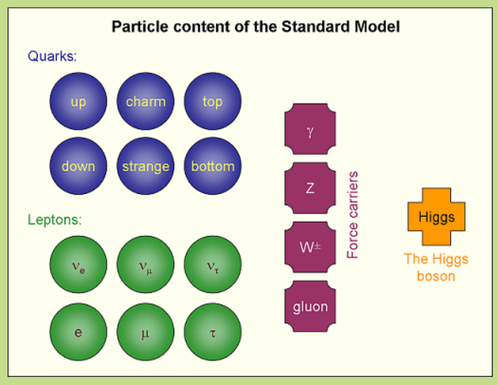
Detailed curation costs were shown at the **4C** workshop held after **iPRES**.

5 Conclusions

DPHEP can bring valuable experience, tools and services to help tackle the **Long Term Data Preservation** issue for **all** disciplines.

Peer-to-peer **collaboration** with other communities – and targetted **funding** – are the keys to building a sustainable long-term solution.

Significant experience with designing software for **long-term sustainability** as well as migrations across many generations of computing infrastructures are also expected to be important.



References

1. http://www.dphep.org/
2. http://invenio-software.org/
3. http://cds.cern.ch/?ln=en
4. http://inspirehep.net/
5. http://www.zenodo.org/
6. DOI: 10.1088/1742-6596/396/3/032064
7. DOI: 10.1088/1742-6596/396/6/062011