Use & Business Cases; Cost Models (DPHEP)

Introduction

This document lists a number of key Use Cases regarding Data Preservation for Long-Term Analysis / Re-use in the High Energy Physics domain. It refers to material published in the DPHEP Blueprint – the output of the Study Group for Data Preservation and Long-Term Analysis in High Energy Physics (HEP).
For each of the Use Cases listed, we give the corresponding Business Case. We then conclude with comments about the associated costs.
The purpose of this document is both internal to DPHEP (now a world-wide collaboration), as well as for input to external communities, such as the APARSEN and 4C projects and the Research Data Alliance Preservation e-Infrastructure Interest Group.
The Use Cases cited are high-level but indicate the primary motivators for attempting long-term data preservation for the HEP community.

#	Use Case	Business Case
1	Continued Ability to Perform Analysis 10-15 years after end of data taking.	Experience shows that a significant number of publications and conference presentations are made in the 5-10 years following the end of data taking. However, this period also sees a significant drop in – or end of – funding to directly support the experiment(s) in question. The business case for continued funding is to ensure the maximum scientific potential of the experiment(s) / facility. Lack of funding may result in a "loss" of some 20% of the potential output.
2	Ability to Re-analyse past data in the light of new theoretical models / in-sights.	Improved and / or new theoretical models can have a major impact on the interpretation of data. Re-analysis of past data using such new insights has, in the past, led to significantly improved results. As opposed to Use Case 1, the time scale involved is much less clear: there is no guarantee that new models will appear in 10, 50 or even 100 years. On the other hand, if the data – including the full capability to re-analyse them – are not preserved,

		then the "outcome" is assured.
3	Ability to Compare Results from a Future Facility with those of a Current / Past one, including full re-analysis if the new results justify it.	Machines (accelerators, colliders) used in HEP often follow a "discovery machine" followed by "precision machine" pattern. There is strong scientific motivation to retain the data from the "discovery era" to the "precision era" to perform comparisons and, if necessary, new analyses of the old data. The duration of the preservation period is typically known, although potential delays in funding / construction / commissioning of the new facilities needs to be taken into account. A "successor" machine to the LHC maybe operational in the 2030s, so the curation period is a factor longer than Use Case 1.

The above non-exhaustive Use Cases suggest that for core scientific reasons, HEP data should be preserved for one to a few decades. The motivations for preserving the data are exactly those that led to its being acquired, given a strong suggestion that the same funding agencies could or should be targeted to ensure this preservation.
Other Use Cases, such as the preservation for an indefinite period for unknown future re-use, are much less clear.
It is also important to note that at least some host laboratories may change their core business over such a period and so alternative partners for providing some of the key infrastructure, such as one of the data repositories, may be needed.
The primary focus of the DPHEP Collaboration is on scientific re-use, although the Cost Analysis described below is equally applicable to other Use Cases.

Costs and Cost Models

The above Use and Business Cases motivate the preservation of HEP data for a few decades. HEP data – currently around the 100PB mark – is characterized not only by its volume but also by the significant amount of software, meta-data, documentation and "knowledge" that is required to process it.
The precise costs of preserving the data and this knowledge until the middle of this century are clearly unknown. However, past experience and current and projected costs can give us a good handle on what to expect.
To arrive at reasonable Cost Models for each of the above Use Cases, a comprehensive workshop is foreseen for early 2014. This will cover all known and expected services / areas in the full offline computing environment for long-term data preservation and detail costs and likelihood of occurrence.
Thus, for a scenario where data is preserved for one decade one can estimate (say) three media migrations, one change of software repository, no changes in

the “digital library” infrastructure and so forth. For longer periods, more disruptive changes, such as change of data format / storage interface(s), computing infrastructure etc should most likely be factored in.
These *costs* can then be turned into a *cost model*, whose predictions would be compared with reality, and tuned as necessary, over the running period of the LHC and its successors (up to 2030 / 40).
A likely outcome of this cost analysis is the identification of key areas for optimization(s). For example, whereas HEP experiments share common e-Infrastructures and storage services, they differ – sometimes significantly – in Computing and Analysis models. Such differences may be justified historically but are an impediment to long-term preservation and re-use *within* the HEP community, let alone in the wider context.