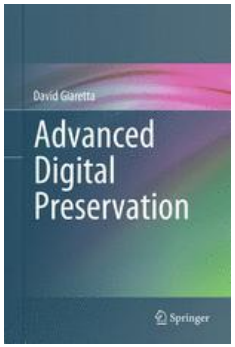


The DPHEP Collaboration & Project(s)

Services, Common Projects, Business Model(s)

Jamie.Shiers@cern.ch

PH/SFT Group Meeting

December 2013

DPHEP

International Collaboration for Data Preservation and Long Term Analysis in High Energy Physics

Brief Background (<2013)

- DPHEP started as a **Study Group** in 2008/9
  - In the "grid world", this was between CCRC'08 / STEP'09 & the time of EGEE III / EGI_DS

- It delivered a Blueprint in May 2012, a summary of which was input to the ESPP in Krakow

  - *"as well as infrastructures for data analysis, **data preservation** and distributed data-intensive computing should be maintained and further developed."*

➢ **The main recommendations of the Blueprint – including the appointment of a full time project manager – are now being implemented**

- **This includes moving to a "Collaboration" (difficult)**

Entity	Description	Input and Positioning	Output
DPHEP Project Manager	Project management, administrative, technical, funding	Main operational coordinator, maintain contacts, organises meetings, **lead proposals for funding**	Reports to the steering committee

DPHEP – 1st Workshop

- *"The problem is substantial and past experience shows that **early preparation** is needed and **sufficient resources** should be allocated."*

- *"The "raison d'être" of data preservation should be clearly and convincingly formulated, including a **viable economic model**."*

➢ The Full Costs of Curation workshop will address these questions!

2020 Vision for LT DP in HEP

- ***Long-term – e.g. LC timescales**: disruptive change*
  - By 2020, all archived data – e.g. that described in Blueprint, including LHC data – easily findable, fully usable by designated communities with clear (Open) **access policies [example later]** and possibilities to annotate further
  - **Best practices**, **tools** and **services** well run-in, fully documented and sustainable; built in common with other disciplines, based on **standards**
  - ✓ **DPHEP portal**, through which data / tools accessed

➢ **Vision achievable, but we are far from this today**

ICFA Statement on LTDP

- *The International Committee for Future Accelerators (ICFA) supports the efforts of the Data Preservation in High Energy Physics (DPHEP) study group on long-term data preservation and welcomes its transition to an active international collaboration with a full-time project manager.* ***It encourages laboratories, institutes and experiments to review the draft DPHEP Collaboration Agreement with a view to joining by mid- to late-2013.***

- *ICFA notes the lack of effort available to pursue these activities in the short-term and the possible consequences on data preservation in the medium to long-term.* ***We further note the opportunities in this area for international collaboration with other disciplines and encourage the DPHEP Collaboration to vigorously pursue its activities****. In particular, the effort required to prepare project proposals must be prioritized, in addition to supporting on-going data preservation activities.*

- ***ICFA notes the important benefits of long-term data preservation to exploit the full scientific potential of the, often unique, datasets****. This potential includes not only future scientific publications but also educational outreach purposes, and the Open Access policies emerging from the funding agencies.*

- 15 March 2013

2013+

- During this year, we have built / strengthened links with other communities & projects
  - **This (IMHO) has helped us a great deal!**

✓ We have converged on a small set of services
  - Instantiated at multiple sites / collaborations

✓ And a similar number of (potential) joint projects

➢ **But big questions still remains: how to support long-term (multi-decade) data preservation**
  - M+P; budget lines (APT), resource review (RRB) etc.
  - Interaction with projects / collaborations such as APA(RSEN), 4C, etc.

LHCC – June 2013

- *"The LHCC has welcomed a presentation on Data Preservation in High Energy Physics"*
- *"Multi-disciplinary projects have started and are being planned in Europe and in the US at national and international levels"*
- *"A strategic vision on data preservation is prepared at CERN and the LHC experiments are encouraged to actively take part to this process"*

Data Harmonization Guidelines

- Common tacit points of agreement between LHC experiments:
  - ✔ **level 1 data:** All experiments already make data from papers and supporting information available through HEPDATA/Inspire, support open access journals etc..
  - ✔ **level 2 data:** All experiments already support limited access of samples in simple formats for outreach and teaching.
  - ! **level 3 data:** Full reconstruction outputs for analysis (AOD, DPD/ntuples) might be made available after an embargo period – but suggested durations range from 3 to 10 years, and there is a question of usefullness. The resource implications to make this useful are high.
  - ✔ **level 4 data:** General agreement RAW data is preserved for the experiment and future – open data access is not usually possible even to the collaboration members. (In ATLAS access to RAW data on tape is restricted).

- **Tools like Rivet, HEPDATA & Recast may make data (information) usefully available, bridging level 3 and level 1.**

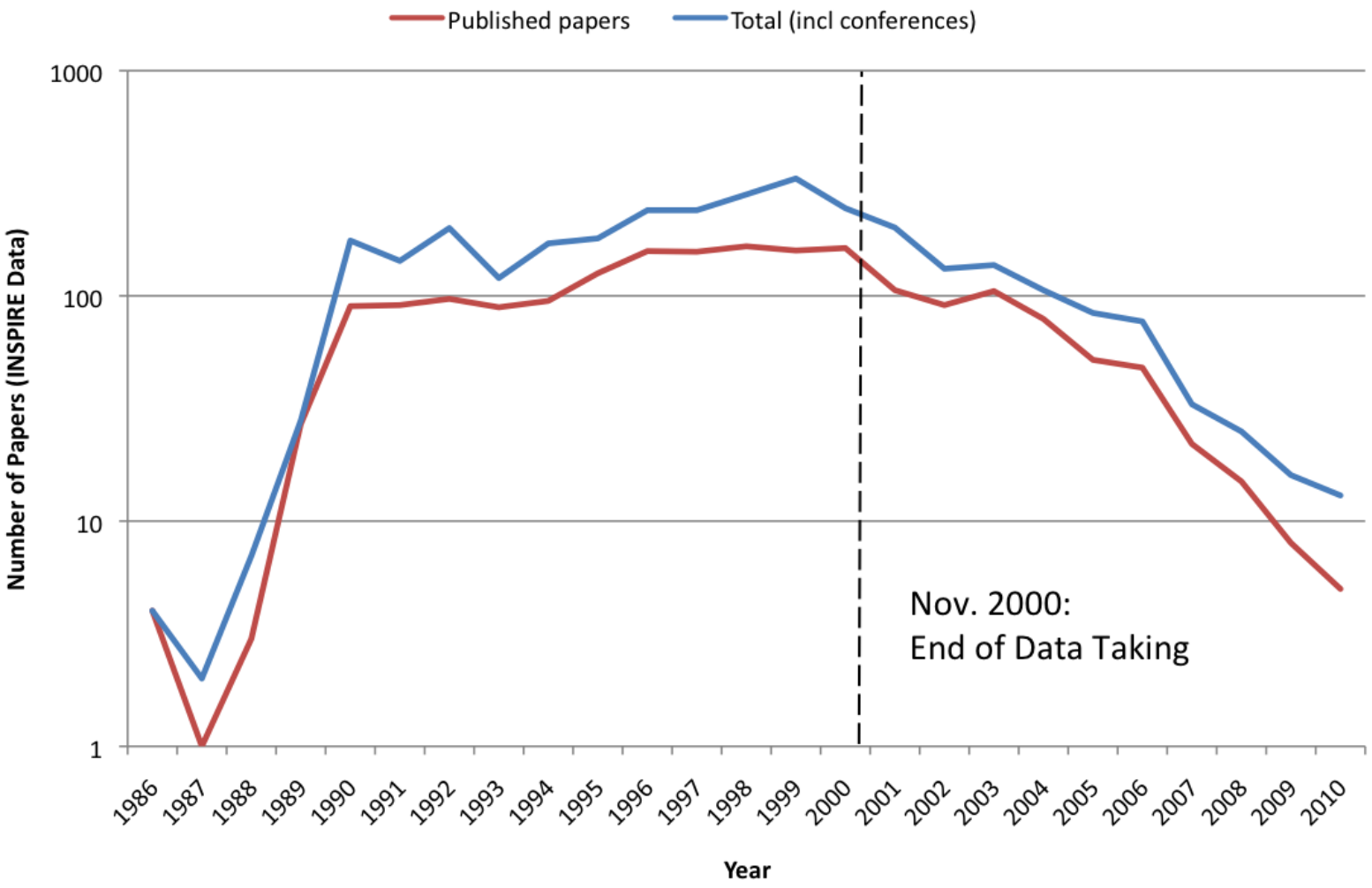
“Summary” of CHEP Workshop

- **Services**: sustainable bit-level preservation for multiple decades; INSPIRE, CDS, HEPData, …
- **Projects**: Rivet, Recast, “CERNLIB consortium”, DPHEP Portal, Validation Tools, Virtualisation Tools etc.
- **Business Plan**: based on clear Use + Business Cases and Costs -> explicit funding in MTPs

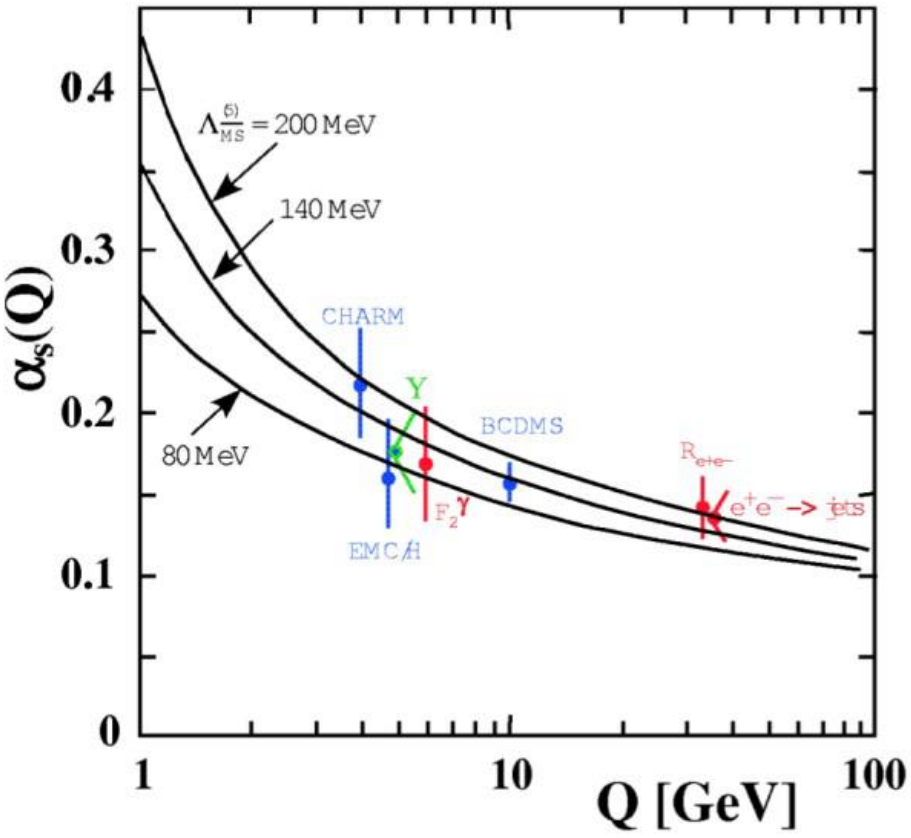
Use Cases

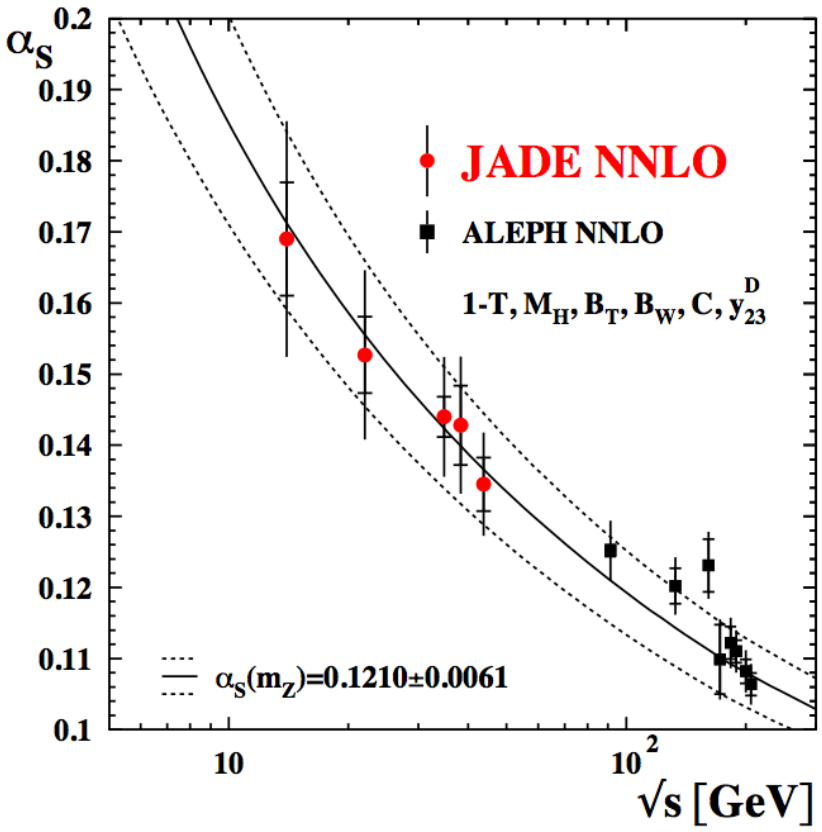
- Three Use Cases have been identified, based on the “Problem Statement(s)” in the DPHEP Blueprint
- They are simple enough for discussions with non-experts
- They may be over-simplified but IMHO this does not dramatically alter the bottom line

1 – Long Tail of Papers



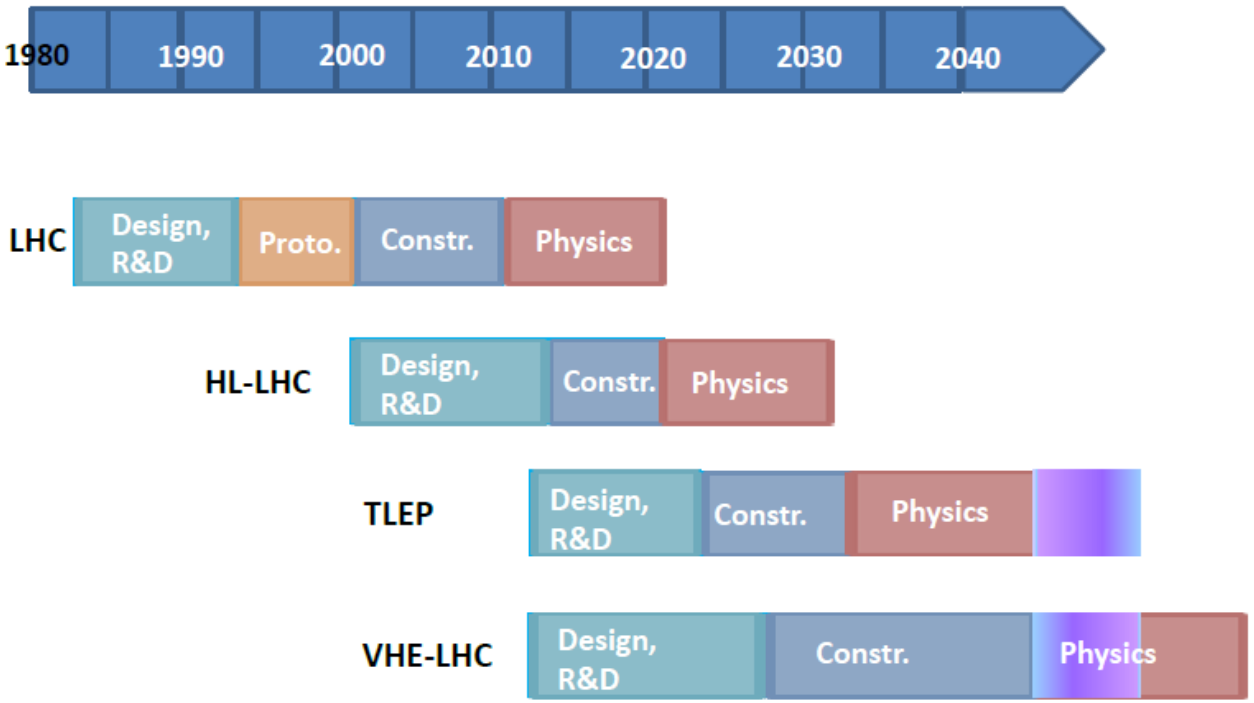
2 – New Theoretical Insights





3 – “Discovery” to “Precision”

possible long-term time line



4 – (whatever)

- There is a general feeling that “we” should preserve data “forever” “just in case”
- No clear business case
- An understanding of the costs can help clarify the strategy (e.g. “best effort” – bit preservation + ?)
- Preservation of data + software + knowledge beyond human lifetimes not obvious…
- (Cost benefit analysis)
  - **See PV2013 “South Atlantic Anomaly”**

Use Case Summary

1. Keep data usable for ~1 decade

2. Keep data usable for ~2 decades

3. Keep data usable for ~3 decades

- Re-visit after we have understood costs & cost models, plus potential “solutions”

COSTS AND COST MODELS

Costs – Introduction

- We do not know *exactly* what the costs will be in the future
- But, we can make estimates, based on our "knowledge" and experience
- In some areas these estimates will be relatively accurate
- In others, much less so
- "Acceptable" costs compared to what?
  - Cost of LHC? WLCG? A specific service, **such as DB**?

A DB Service

- Costs include:
  - Hardware;
  - Licenses & maintenance;
  - People.
- There is also value = business case

➢ **10 FTEs @EUR100K/year = EUR1M/year**

Costs of Curation Workshop

- Within DPHEP, and in collaboration with external projects (e.g. 4C), we are planning a "no stone left un-turned" workshop
- Look at the **many migrations** we have performed in the (recent) past – plus those foreseen
- ➢ **Estimate / calculate costs**
- Come up with scenarios for the future:
  - **10 year preservation = 3 media migrations + n build systems + p s/w repositories + q O/S versions + …**
  - **20 year preservation: more disruptive changes**
  - **30 year preservation: more still**
- ➢ **Manpower almost certainly the dominant cost**
- What can we do to optimize it?
  - Coordinate validation activities -> service
  - Streamline emulation activities -> tool-kit(s)
  - Best practices & support for migration activities -> support activity
- Can we do things in a way that costs less in the future – and make our data more "preservational"?

Summary

- Your input and experience is needed to make the workshop successful
  - Jan 13/14
- We will start to build agenda now – output will be a report with costs & cost models
- This should help guide our work – and IMHO is a pre-requisite for obtaining funding / resources

Conclusions

- Unless there are real surprises (IMHO not consistent with “experiment”), the real and necessary costs of curation are **affordable**

- **Affordable** means business case is valid / strong

➢ **Knowing the numbers can only help**