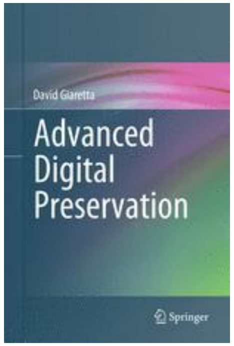

David Giaretta

Advanced Digital Preservation

Springer

Data Preservation in HEP

Use Cases, Business Cases, Costs & Cost Models

Jamie.Shiers@cern.ch

Grid Deployment Board

DPHEP

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

DPHEP-fest

- Today: DPHEP@GDB
- Monday Oct 14: DPHEP@CHEP
  - Update on progress since CHEP 2013
- Wednesday Oct 16: DPHEP WS @ CHEP
  - DPHEP “Common Projects”;
  - Moving from a “Problem Statement” (Blueprint) to Services, Solutions and Projects

DPHEP Implementation Board

- Equivalent to GDB / MB for DPHEP
  - Indico: https://indico.cern.ch/categoryDisplay.py?categId=4458
  - Twitter: https://twitter.com/search?q=%23DPHEP
  - Mail archives: https://groups.cern.ch/group/DPHEP-IB/default.aspx

DP in the Wider Context

- ***Many*** projects / disciplines active
- At least some “mature” in many aspects
  - **We can profit a lot by collaboration (bi-directional)**
- International / inter-disciplinary coordination:
  - Alliance for Permanent Access (APA) **[executive board candidate]**
  - RDA Preservation e-Infrastructure Interest Group **[vice-chair]**
- Several relevant conferences / workshops:
  - APA
  - iDCC
  - iPRES
  - PV
  - (RDA)

High Level Strategy wrt Others

- Make “them” **aware** of us
  - “Them” = other projects, funding agencies, …
- Clarify what we can **offer**
  - e.g. “bit preservation” at 100PB -> 1EB scale

➢ **This seems to be working**

The remainder of this talk will concentrate on:

- **Use Cases;**
- **Associated Business Cases;**
- **Costs & Cost Models.**

- **Why is this relevant for the GDB?**
  - Because there are messages and implications for the funders
  - As they may well be service and other implications (“best practices”)
  - Because members of the GDB can provide input to the elaboration of the costs & cost models

- **Once we have these we can prepare a “roadmap” for handling the key Use Cases**

- **An analysis of the costs is essential for future work…**

DPHEP – 1st Workshop

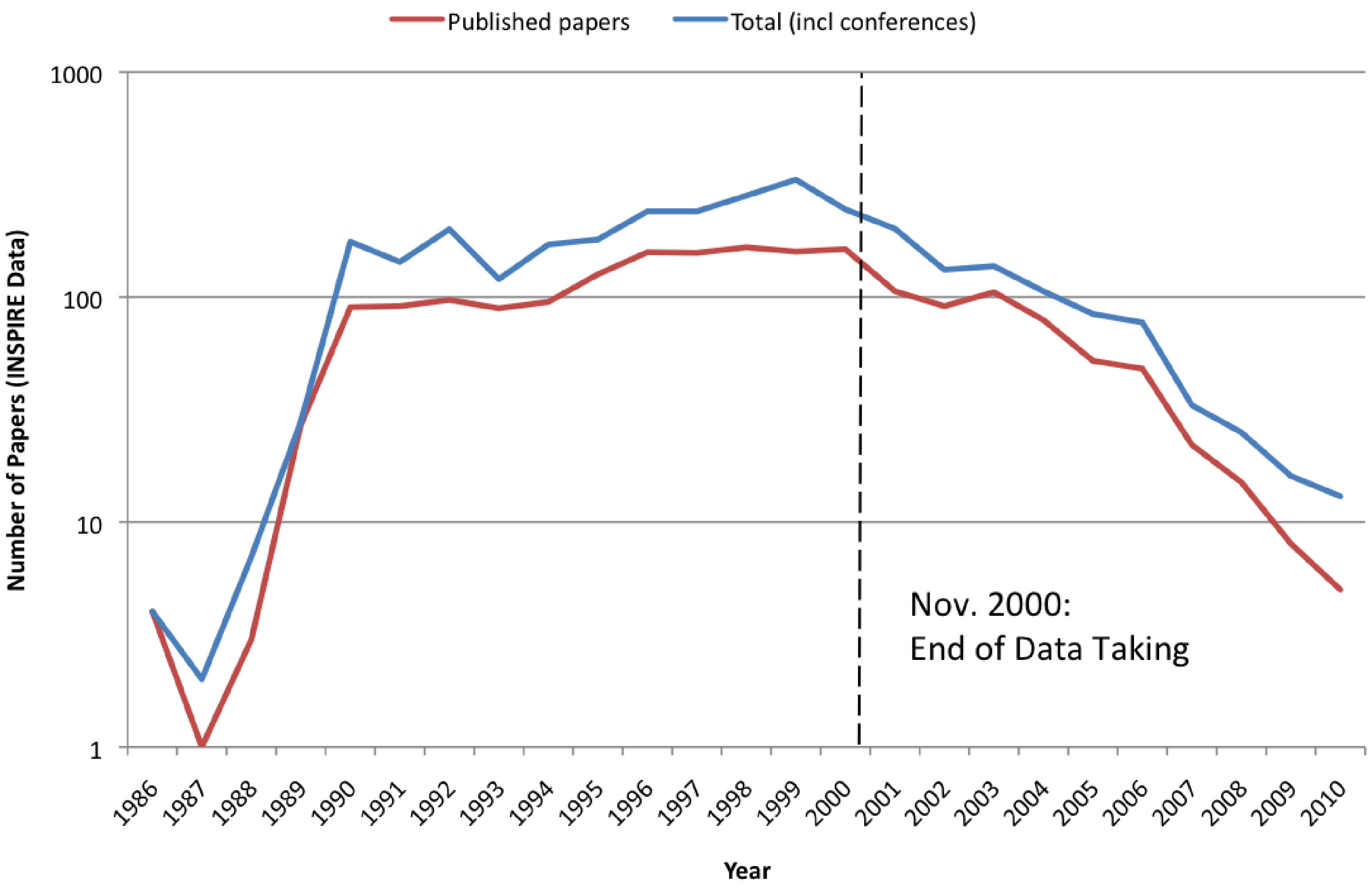
- *“The problem is substantial and past experience shows that early preparation is needed and sufficient resources should be allocated.”*

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

Use Cases

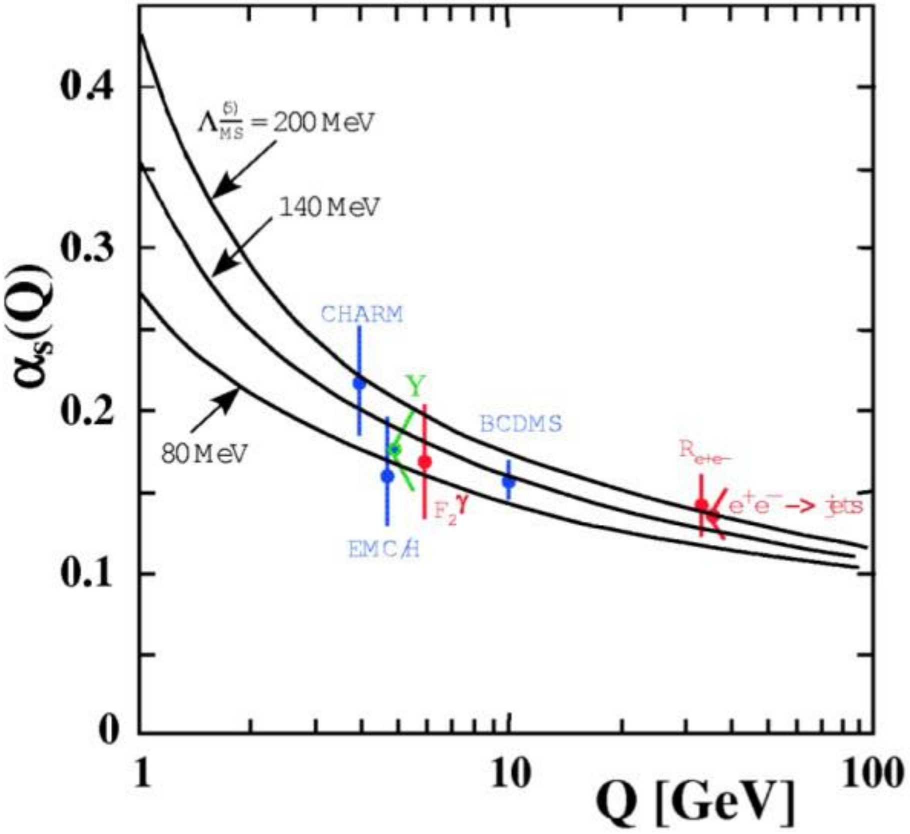
- Three Use Cases have been identified, based on the “Problem Statement(s)” in the DPHEP Blueprint
- They are simple enough for discussions with non-experts
- They may be over-simplified but IMHO this does not dramatically alter the bottom line

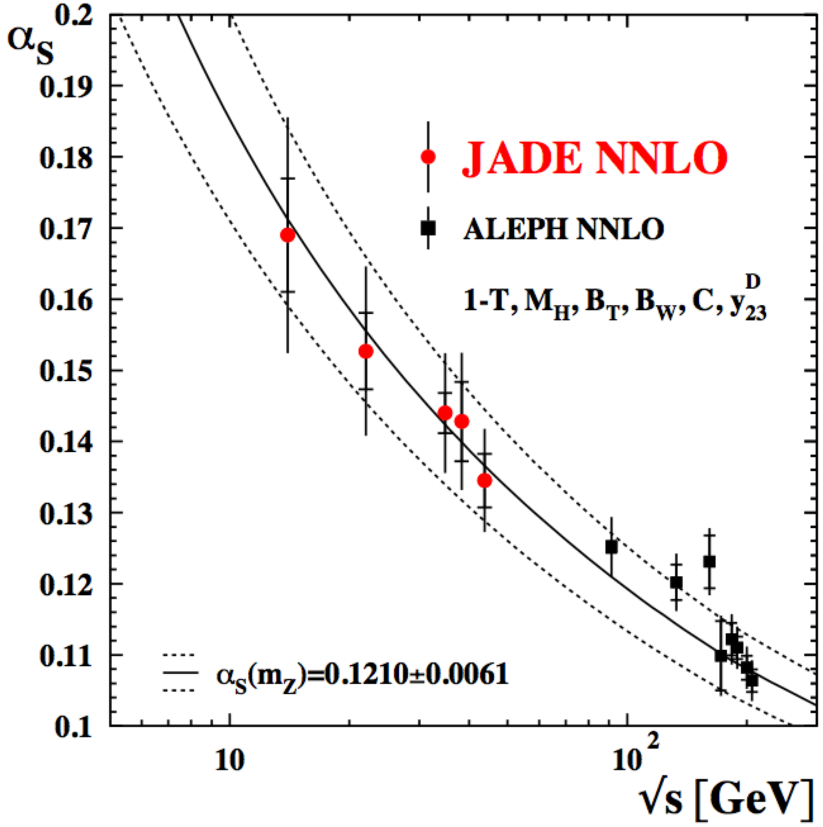
1 – Long Tail of Papers



Nov. 2000:
End of Data Taking

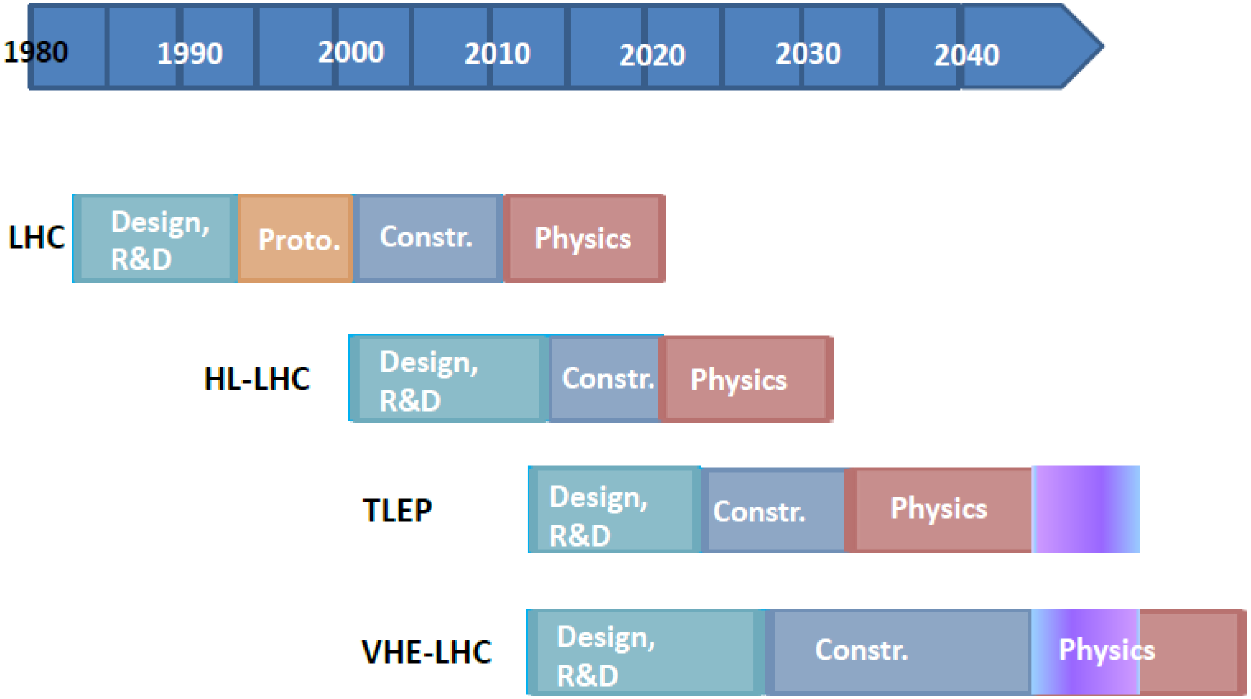
2 – New Theoretical Insights





3 – “Discovery” to “Precision”

possible long-term time line



1980 1990 2000 2010 2020 2030 2040

LHC: Design, R&D | Proto. | Constr. | Physics

HL-LHC: Design, R&D | Constr. | Physics

TLEP: Design, R&D | Constr. | Physics

VHE-LHC: Design, R&D | Constr. | Physics

Zimmermann

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)

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 (or Jan 27/28)

- 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**

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