Session 4: Data preservation lessons learnt and future prospects

LTDP in HEP: Status, lessons learnt and 2020 (2035 / 2050) outlook

Jamie Shiers
CERN & DPHEP

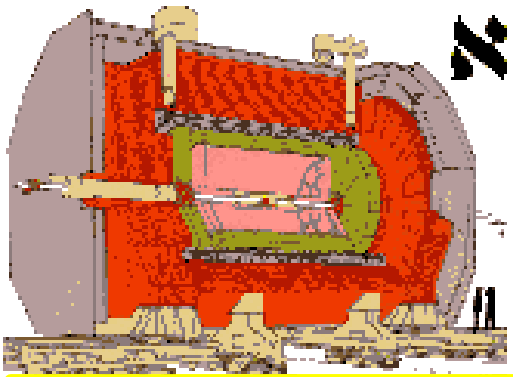
International Collaboration for Data Preservation and Long Term Analysis in High Energy Physics

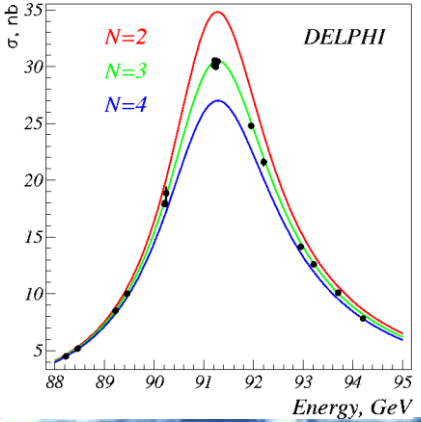
Outline

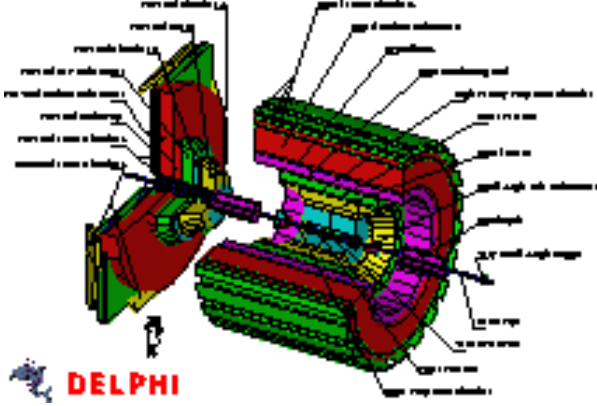
- Long-term
- Data Preservation
- Future Re-Use
- Lessons learnt & Outlook

LONG TERM


ifage




σ, nb
35
30
25
20
15
10
5
N=2
N=3
N=4
DELPHI
88
89
90
91
92
93
94
95
Energy, GeV

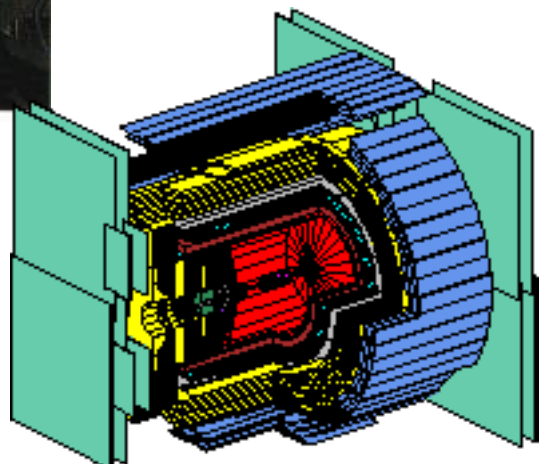

DELPHI

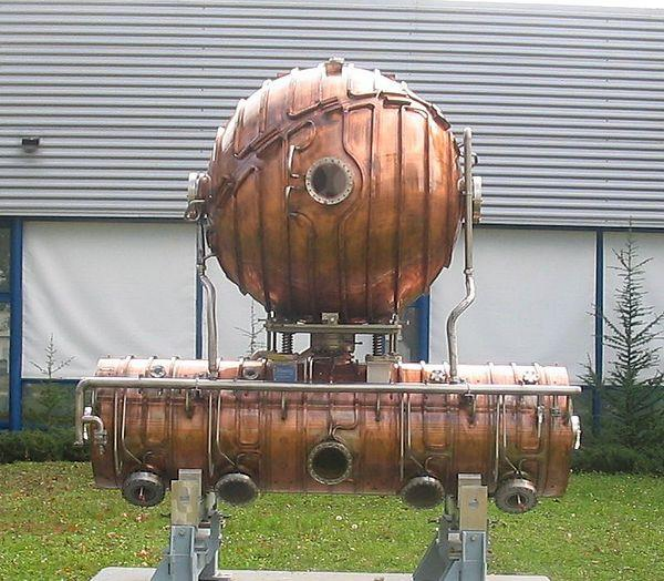

YEARS/ANS CERN
1954 2014


LEP
SPS
site
Prévessin
FRANCE
SUISSE


MONTH DAY YEAR PM HOUR MIN
OCT 26 1985 09 00
DESTINATION TIME
MONTH DAY YEAR PM HOUR MIN
OCT 21 2015 07 28
PRESENT TIME
MONTH DAY YEAR PM HOUR MIN
NOV 12 1955 06 38

L3 HOME PAGE





- LEP ran as a Z^0 factory;
- Then produced $W^{\pm}$ pairs;
- Energy scan up to 209 GeV

- **Total data: ~500TB (0.5PB)**
- This was “Big Data” at the time!

- **LEP experiments faced “constant change” – a first for HEP. Probably why data is still around!**

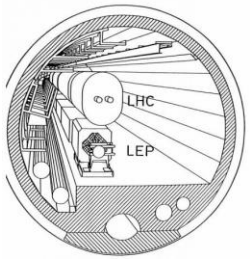
The Large Electron-Positron Collider

View

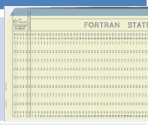
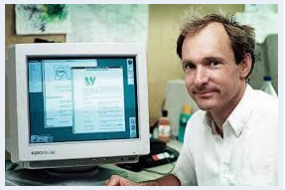
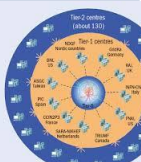
LEP – the largest electron-positron accelerator ever built – was dismantled in 2000. Its 27-kilometre tunnel now hosts the LHC

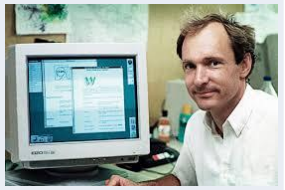
1981 1983 1988 1989 2000

LEP Events: approval, start / end of construction, start / end of data taking (~2 decades)

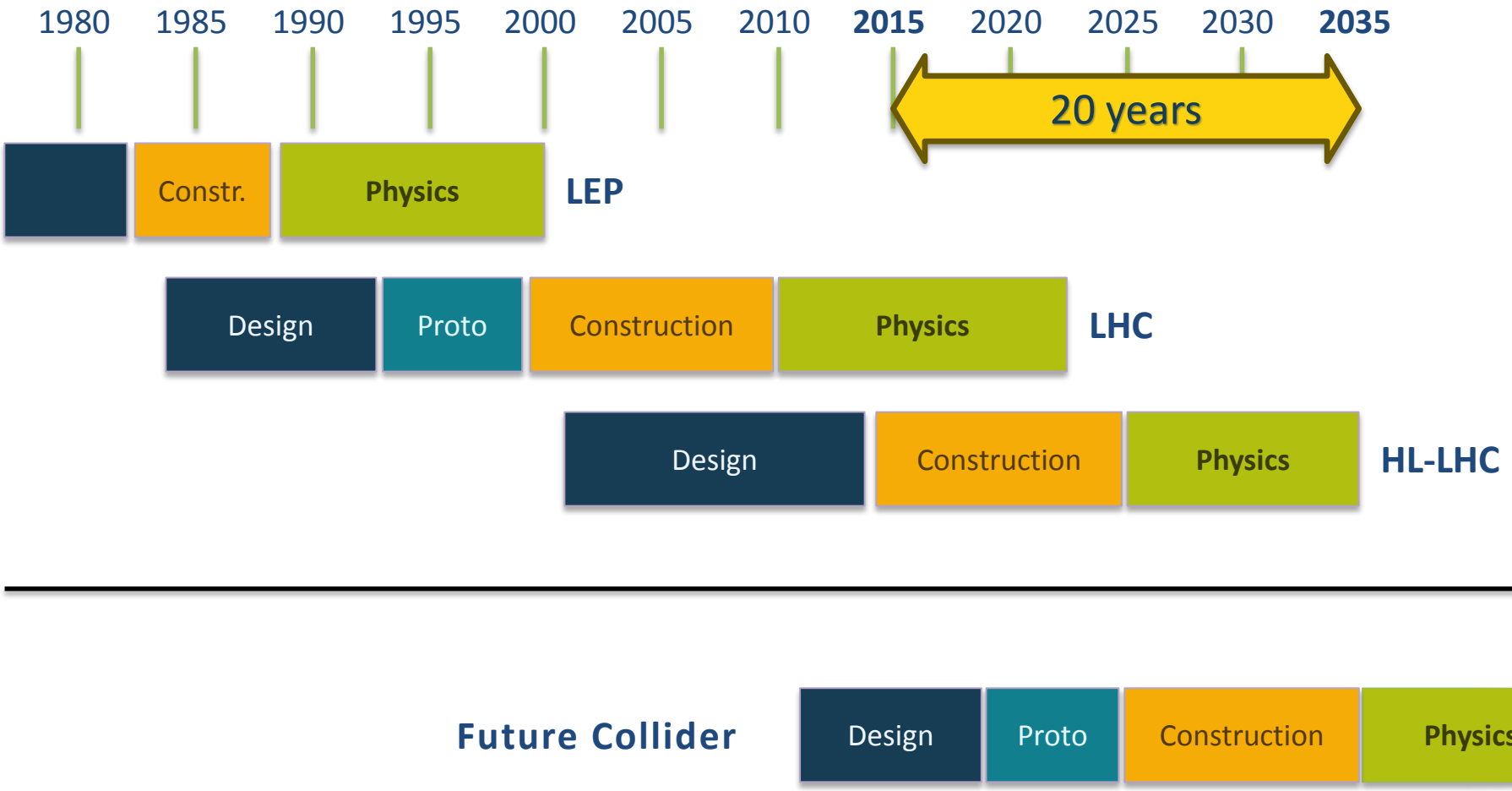


LEP Timeline

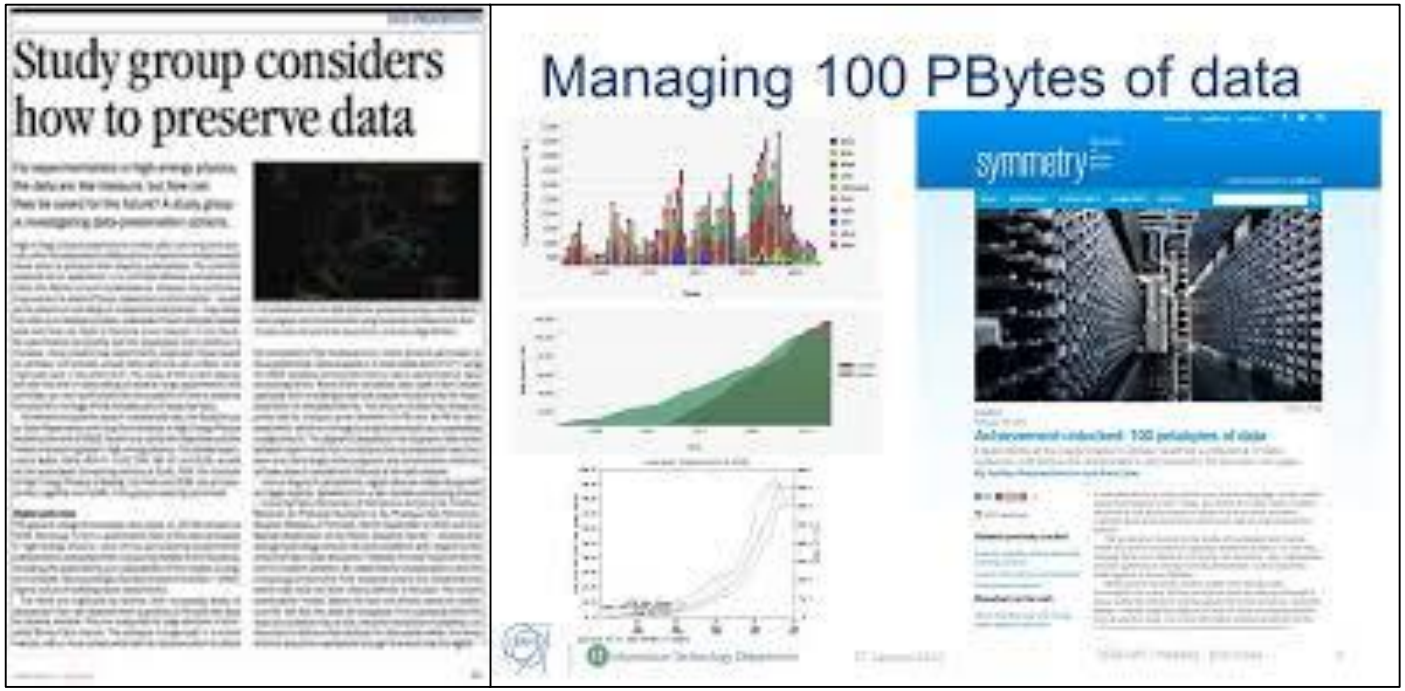
Date	Collider (e^+e^-)	Computing
1981	Approved by Council	Card readers still exist! 
1983	Civil Engineering starts	Computing at CERN in the LEP era published
1988	LEP Tunnel completed 	Data Management project requested by experiments
1989	1st beams, collisions, and results 	Was the s/w really ready? 
1992	LHC Computing starts 	Mainframes replaced Unix, later PCs
1996	LEP 2 (W pairs) starts	
2000	Final run of LEP	HEP gets bitten by Grid 

Date	Collider	
1981	Approve	s still exist!
1983	Civil Eng	t CERN in the ished
1988	LEP Tun complet	ement project y experiments
1989	First bea collision	
1992	LHC Con	computing
1996	LEP 2 (W	Unix, later PCs
2000	Final run	ten by Grid

CERN Circular Colliders + FCC



HEP has a long history of planning, financing and executing multi-decade projects



DATA PRESERVATION

2020 Vision for LT DP in HEP

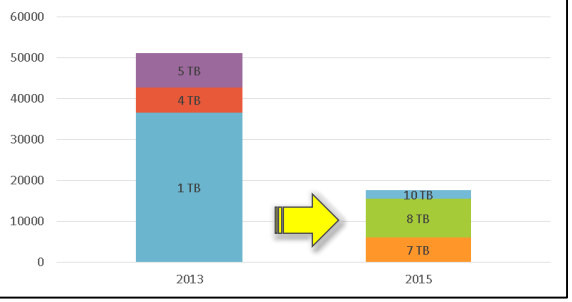
- ***Long-term – e.g. FCC timescales**: disruptive change*

  - By 2020, all **archived data** – e.g. that described in DPHEP Blueprint, including LHC data – easily **findable**, fully **usable** by **designated communities** with clear (Open) access policies and possibilities to annotate further

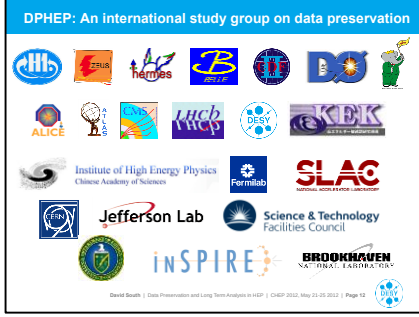
  - Best practices, tools and services well run-in, fully documented and sustainable; built in common with **other disciplines**, based on standards

  - **DPHEP portal**, through which data / tools accessed
    - **"HEP FAIRport": Findable, Accessible, Interoperable, Re-usable**

➢ **Agree with Funding Agencies clear targets & metrics**



Aspects of LT DP



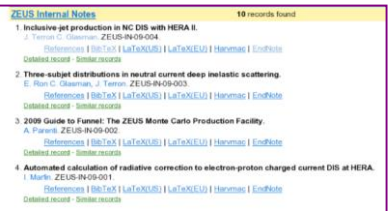
- **<u>A common approach across the main HEP labs worldwide, including:</u>**

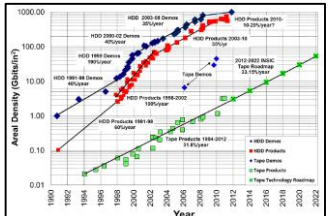
1. **Data (bit preservation) – state of the art at exascale (1PB-10PB-100PB-1EB etc);**
2. **Software (and environment) – combination of validation + virtualisation;**
3. **Documentation (I would say "knowledge") – digital library technologies + regular testing as part of training and data re-use**

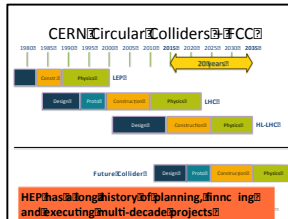
- **LEP – and other Colliders worldwide – allow us to "see into the future" and compare different options for LTDP**

➢ **Expectation for LEP is that data will be usable (and used) until ~2030 – 3 decades after end of data taking! (<u>Copy on disk + 2 on tape @ CERN!)</u>**

- Data will (should) be available much longer; "resurrection" of HEP data + software has been demonstrated but requires significant motivation + effort







http://science.energy.gov/funding-opportunities/digital-data-management/

- ***"The focus of this statement is sharing and preservation of digital research data"***

- **All proposals submitted to the Office of Science (after 1 October 2014) for research funding must include a Data Management Plan (DMP) that addresses the following requirements:**

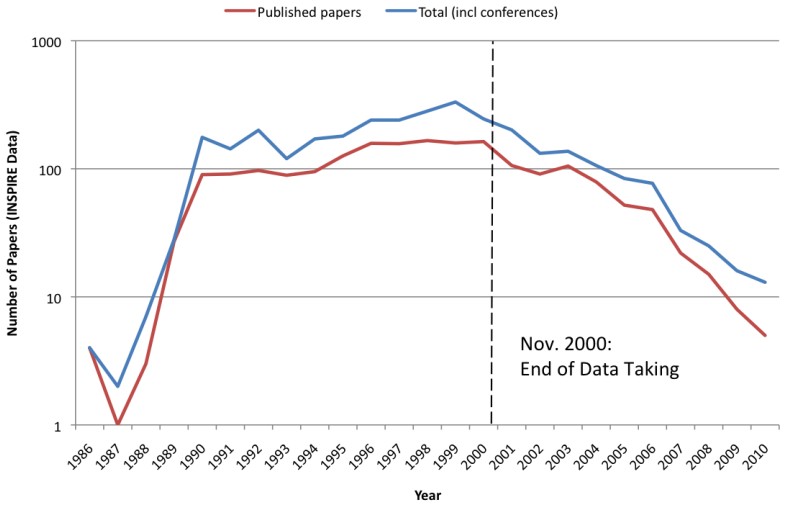
1. **DMPs should describe whether and how data generated in the course of the proposed research will be shared and preserved.**

   If the plan is not to share and/or preserve certain data, then the plan must explain the basis of the decision (for example, cost/benefit considerations, other parameters of feasibility, scientific appropriateness, or limitations discussed in #4).

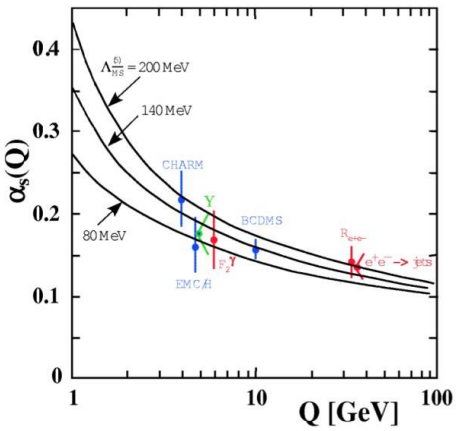
   At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

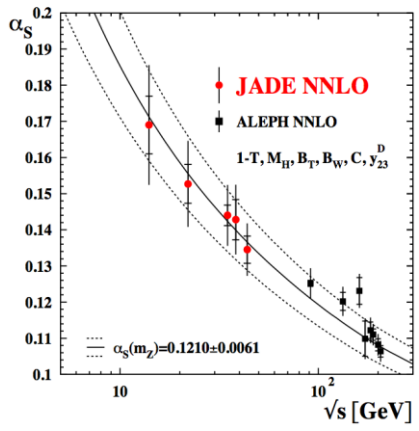
RE-USE (= FUNDING)

1 – Long Tail of Papers



2 – New Theoretical Insights

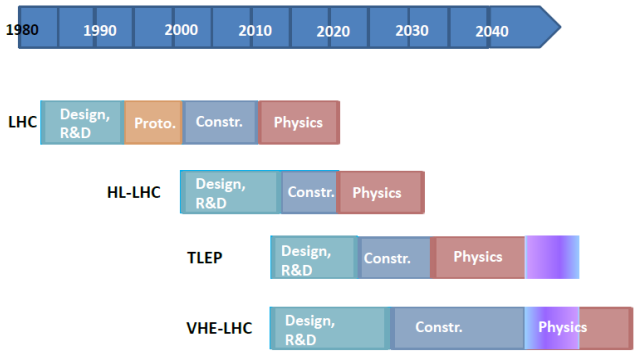




3 – "Discovery" to "Precision"

TLEP

possible long-term time line



Zimmermann

Use Case Summary

1. Keep data usable for ~1 decade
2. Keep data usable for ~2 decades
3. Keep data usable for ~3 decades

Volume: 100PB + ~50PB/year (+500PB/year from 2025)

Use Cases – “all HEP”

1. Bit preservation – basically OK (at CERN) but not a formal policy
   - Data taken by the experiments should be preserved
2. Preserve data, software, and know-how in the collaborations
   - Foundation for long-term DP strategy
   - Analysis reproducibility: Data preservation alongside software evolution
3. Share data and associated software with (larger) scientific community
   - Additional requirements:
   - Storage, distributed computing
   - Accessibility issues, intellectual property
   - Formalising and simplifying data format and analysis procedure
   - Documentation

- Open access to reduced data set to general public
  - Education and outreach
  - Continuous effort to provide meaningful examples and demonstrations
- Strategy and scope in approved policy documents for all (LHC+LEP) collaborations
  - http://opendata.cern.ch/collection/data-policies

LEP (and other?) access policies exist (L3?) – need to be uploaded & given DOI

CAP Use Cases (I) (=know-how?)

1. The person having done (part of) an analysis is leaving the collaboration and has to hand over the know-how to other collaboration members.
2. A newcomer would like join a group working on some physics subject
3. In a large collaboration, it may occur that two (groups of) people work independently on the same subject
4. There is a conflict between results of two collaborations on the same subject

CAP Use Cases (II)

5. A previous analysis has to be repeated
6. Data from several experiments, on the same physics subject, have to be statistically combined
7. A working group or management member within a collaboration wishes to know who else has worked on a particular dataset, software piece or MC
8. Presentation or publication is submitted for internal/collaboration review and approval: lack of comprehensive metadata
9. Preparing for Open Data Sharing


Lessons Learned
recognize mistakes
observe what works
document them
share them

LESSONS

1. **There are enormous benefits in working with other projects and disciplines: IMHO we have saved years (=money) AND we can also help others (if they want)**

2. **Having a Business Case and Cost Model is essential;**

3. **It is never too early to consider data preservation: early planning is likely to result in cost savings that may be significant. Furthermore, resources (and budget) beyond the data-taking lifetime of the projects should be foreseen from the beginning;**

4. ***Caveat emptor:* there are disruptive changes ahead. How does one prepare for these, particularly when a project is no longer in the active phase? (Don't get hooked on any particular technical solution – it will change!)**

1. **There are enormous benefits in working with other projects and disciplines: IMHO we have saved years (=money) AND we can also help others (if they want)**

0. You can justify it; afford it = do it!

significant. Furthermore, resources (and budget) beyond the data-taking lifetime of the projects should be foreseen from the beginning;

4. ***Caveat emptor:*** **there are disruptive changes ahead. How does one prepare for these, particularly when a project is no longer in the active phase? (Don't get hooked on any particular technical solution – it will change!)**

OUTLOOK

2020 Vision for LT DP in HEP

- Long-term (e.g. FCC-y timescales): disruptive change
  - By 2020, all archived data – e.g. that described in DPHEP Blueprint, including LHC data – easily findable, fully usable by designated communities with clear (Open) access policies 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 via "HEP FAIRport": Findable, Accessible, Interoperable, Reusable

Agreement with Funding Agencies (clear targets & metrics)

http://science.energy.gov/funding-opportunities/digital-data-management/

- *"The focus of this statement is sharing and preservation of digital research data"*
- **All proposals submitted to the Office of Science (after 1 October 2014) for research funding must include a Data Management Plan (DMP) that addresses the following requirements:**

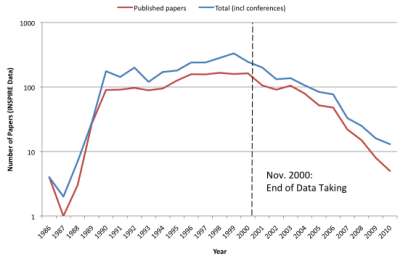
1. **DMPs should describe whether and how data generated in the course of the proposed research will be shared and preserved.**

   If the plan is not to share and/or preserve certain data, then the plan must explain the basis of the decision (for example, cost/benefit considerations, other parameters of feasibility, scientific appropriateness, or limitations discussed in #4).

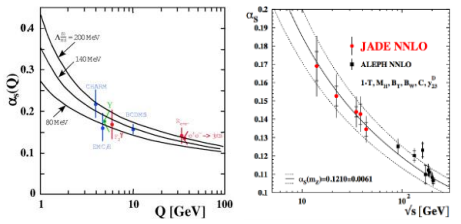
   At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

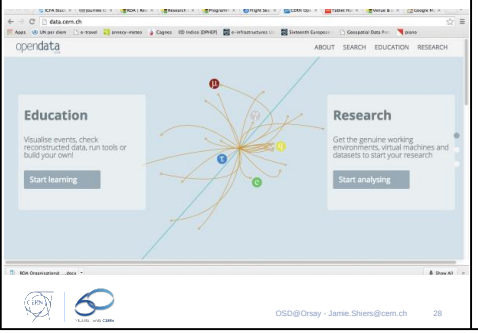
U.S. DEPARTMENT OF ENERGY | Office of Science

1 – 'Long Tail' of Papers



2 – 'New' Theoretical Insights



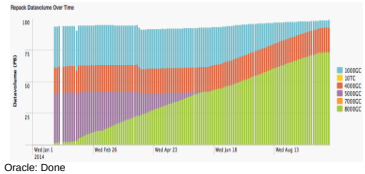


DSS

Repack

CERN IT Department

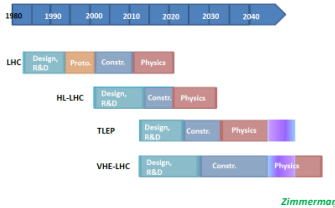
http://indico.cern.ch/event/CERN-ITTF-2014-09-26



- Oracle: Done
  - 39PB self-repacked (5->8TB), 27PB 1TB emptied
- IBM: Dec'14-Mar'15
  - 20PB of IBM 4TB to self-repack and 5.6PB 1TB tapes to empty
- All repacked media has been verified
- All problem source tapes identified and being handled (cf next slides)
- Cleanup of tape pools and (properly) establishing double copies
  - across buildings
  - complete second copies where missing (ie OPAL)

3 – "Discovery" to "Precision"

possible long-term time line



Zimmermann

Use Case Summary

1. Keep data usable for ~1 decade
2. Keep data usable for ~2 decades
3. Keep data usable for ~3 decades

Volume: ~100PB + ~50PB/year (+500PB/year from 2025)

4C Roadmap Messages

A Collaboration to Clarify the Costs of Curation

1. Identify the **value** of digital assets and make **choices**
2. Demand and choose more **efficient** systems
3. Develop **scalable** services and infrastructure
4. Design digital curation as a **sustainable** service
5. Make funding **dependent** on costing digital assets across the whole lifecycle
6. Be **collaborative** and **transparent** to drive down costs

Balance sheet – Tevatron@FNAL

- 20 year investment in Tevatron ~$4B
- Students $4B
- Magnets and MRI $5-10B } ~$50B total
- Compu>ng ' $40B''''

Very rough calculation confirms our gut feeling that investment in fundamental science pays off

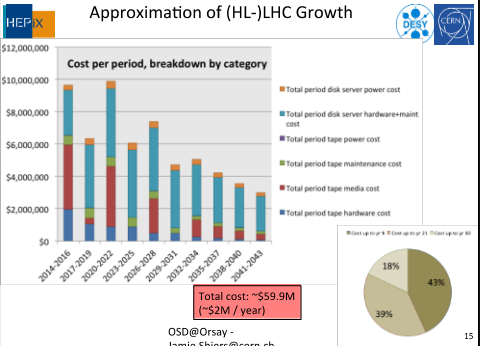
I think there is an opportunity for someone to repeat this exercise more rigorously

cf. STFC study of SRS Impact

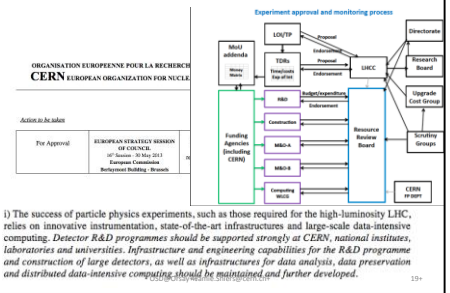
http://www.stfc.ac.uk/2428.aspx

Science & Technology Facilities Council

Approximation of (HL-)LHC Growth



Sustainability + Funding

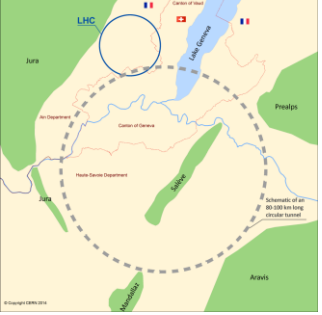


i) The success of particle physics experiments, such as those required for the high-luminosity LHC, relies on innovative instrumentation, state-of-the-art infrastructures and large-scale data-intensive computing. *Detector R&D programmes should be supported strongly at CERN, national institutes, laboratories and universities. Infrastructure and engineering capabilities for the R&D programme and construction of large detectors, as well as infrastructures for data analysis, data preservation and distributed data-intensive computing should be maintained and further developed.*

What Next?

- **Training on, and certification of, sites as "Trusted Digital Repositories"**
- **Expanding "DPHEP Portal" to other (non-LHC) experiments and external sites**
- **Supporting key experiment Use Cases / Funding Agency Requirements**
  - **Reproducibility, Open Access for Outreach, DMPs**

➢ **Ensuring everything is sustainable, documented, "standards-based" and complete**



- See DPHEP Workshop in Lisbon for more details, including:
  - Original DPHEP Blueprint (2012)
  - New status report (2015)
  - And key work items for 2016 and beyond
- https://indico.cern.ch/event/444264/

Data Preservation in High Energy Physics
The road to DPHEP

http://dphep.org