
CERN

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

HEP LTDP Use Case & EOSC Pilot

What is HEP Data?

What Services are Required?

Long-term Funding & ROI

Jamie.Shiers@cern.ch

What is a data repository?

- **For us (HEP), it is much more than just a "bit repository"**
  - And even that probably has several components
    - **Long-term archive (tape)**; **cache(s) for production & analysis (disk)**; "**Open Access**" area (not necessarily "**immediate Open Access**")
    - What data is accessed when, by whom, access patterns
- **It includes also documentation, software (+environment in which it runs), "knowledge"**
- These are probably supported by different services - some of which may already be "remote" - that evolve on different timescales
- ***Something*** is changing all the time!
  - **If you believe in transparent and seamless migrations you probably don't have a sustainable sustainability plan (or have never done a migration)**

- **Sustainable: financially** + **technically** + "**logically**" (holistically?)

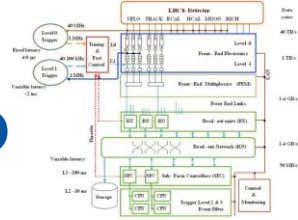
What else is a data repository?

- **Well, its not just about data**
- **Its also about compute + networking**
  - You need to ingest the data, process it, serve it, "curate it" and so forth
  - Rule of thumb for WLCG Tier2s: enough external and internal bandwidth to read through data 3 times before replacing it (not always met!)
- Experience with **LEP** (1989-2000) and **LHC** (2010 - 2040/50), plus Tier1 / Tier2 sites gives us experience in **time** and **space**
  - +JADE/DESY 1979 - 1986: data still being analysed!
- A factor of **80** in (un)reliability between best and worst sites in terms of "bit preservation" (HEPiX)
  - **But not in cost... (just a factor between largest & smallest)**

Funding, Value & ROI

- **CERN** is funded through is 22(+) member states
- **Projects** are (additionally) funded by institutes and other partners
  - Projects = experiments + detectors + accelerators
  - EU (and other) money for R&D (+EDG/EGEE/HNSciCloud/...)
- **Value is well understood in terms of scientific output, educational outreach as well as impact on society + technology in general**
- An STFC study of the Tevatron (FNAL) showed that it was roughly **cost neutral** just counting its scientific output and with a **x 10 ROI** including technology spin-offs
- The "Higgs discovery" was accompanied by a significant rise in applications to STEM in Higher Education
- For large projects (e.g. WLCG / LHC) very close "scrutiny" of requests (and eventual funding) by the Funding Agencies

What Makes HEP Different?



- We **throw away** most of our data before it is even recorded – "triggers"
- Our detectors are **relatively stable** over long periods of time (years) – not "doubling every 6 or 18 months"
- We make "**measurements**" – not "**observations**"
- Our projects typically last for **decades** – we **need** to keep data usable during at least this length of time **(but not necessarily "forever")**
- We have **shared** "data behind publications" for more than 30 years… (HEPData)

What does DPHEP do?

- DPHEP has become **a Collaboration** with signatures from the main HEP laboratories and some funding agencies **worldwide**.

- It has established a "**2020 vision**", whereby:
  - All archived data – e.g. that described in DPHEP Blueprint, including LHC data – should be easily **findable** and fully usable by the **designated communities** with clear (Open) access policies and possibilities to annotate further;
  - Best practices, tools and services should be well run-in, **fully documented** and **sustainable**; built in common with other disciplines, based on standards;
  - There should be a DPHEP **portal**, through which data / tools accessed;
  - Clear **targets & metrics** to measure the above should be agreed between Funding Agencies, Service Providers and the Experiments.

What is HEP data?

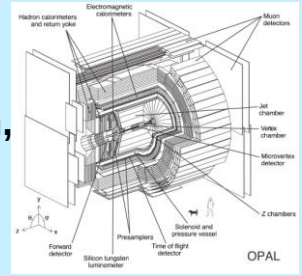
Digital information

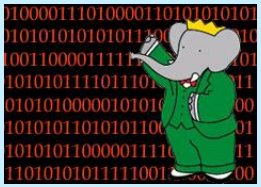
The data themselves, volume estimates for preservation data of the order of a few to 10 PB

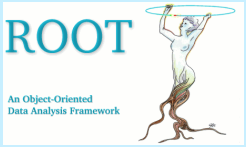
Other digital sources such as databases to also be considered

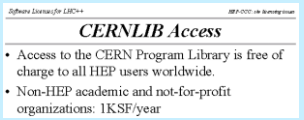
Software

Simulation, reconstruction, analysis, user, in addition to any external dependencies



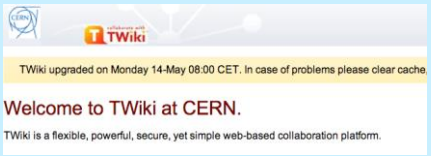






Meta information

Hyper-news, messages, wikis, user forums..



Publications

arXiv.org

INSPIRE

Durham University — Science & Technology Facilities Council

HEPDATA: REACTION DATA Database

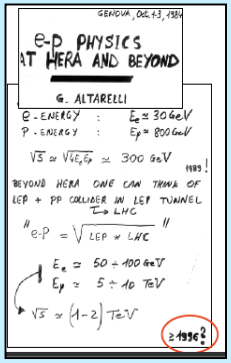
...containing numerical values of HEP scattering data such as total and differential cross sections, fragmentation functions, structure functions, and polarisation measurements, from a wide range of experiments. It is compiled by the **Durham Database Group (UK)** with help from the COMPAS group (Russia,) and is updated at regular intervals.

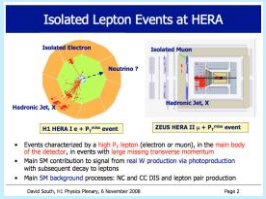
Journal of High Energy Physics

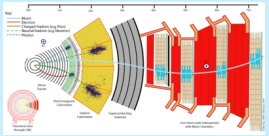
A refereed journal, written, run and distributed by electronic means

Documentation

Internal publications, notes, manuals, slides







Expertise and people

DESY

DPHEP models of HEP data preservation

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	

- These are the original definitions of DPHEP preservation levels from the 2009 publication
  - Still valid now, although interaction between the levels now better understood
- Originally idea was a progression, an inclusive level structure, but now seen as complementary initiatives
- Three levels representing three areas:
  - Documentation, Outreach and Technical Preservation Projects

DESY

HEP LTDP Use Cases

1. **Bit preservation** as a basic "service" on which higher level components can build;
   - *"Maybe CERN does bit preservation better than anyone else in the world" (David Giaretta)*
2. **Preserve data, software, and know-how** in the collaborations; Basis for reproducibility;
3. **Share data and associated software** with (wider) scientific community, such as theorists or physicists not part of the original collaboration;
4. **Open access** to reduced data sets to general public (LHC experiments)

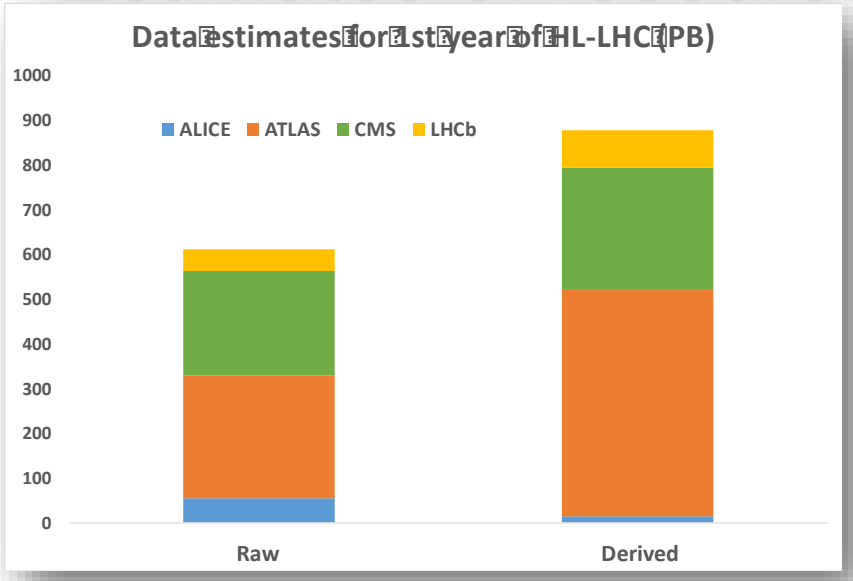
- **These match very well to the requirements for DMPs**

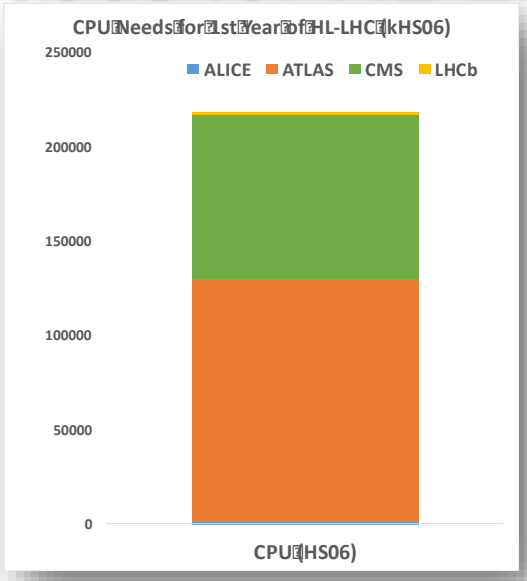
CERN Services for LTDP

1. State-of-the art "**bit preservation**", implementing practices that conform to the ISO 16363 standard
2. "**Software preservation**" - a key challenge in HEP where the software stacks are both large and complex (and dynamic)
3. Analysis **capture and preservation**, corresponding to a set of agreed Use Cases
4. Access to **data behind physics publications** - the HEPData portal
5. An **Open Data portal** for released subsets of the (currently) LHC data
6. A **DPHEP portal** that links also to data preservation efforts at other HEP institutes worldwide.

➢ **Each of these is a talk topic in its own right!**

Estimates of resource needs for HL-LHC





Data:

- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB

CPU:

- x60 from 2016

Technology at ~20%/year will bring x6-10 in 10-11 years

- Simple model based on today's computing models, but with expected HL-LHC operating parameters (pile-up, trigger rates, etc.)
- At least x10 above what is realistic to expect from technology with reasonably constant cost


CERN