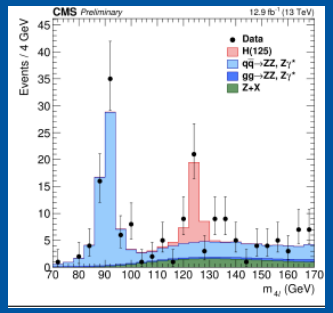
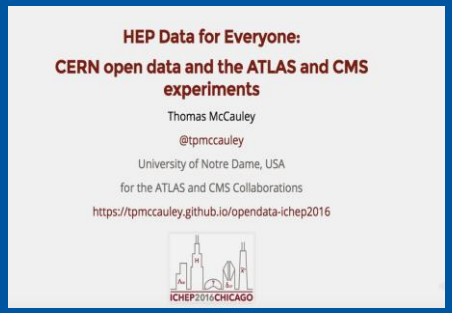
EOSCpilot WP4: Use Case 5

WLCG: large-scale, long-term data preservation and re-use of physics data





Material for

- EOSCpilot kick-off: https://indico.cern.ch/event/605222/
- OECD w/s on repositories: https://indico.cern.ch/event/577772/
- iPRES 2016: https://indico.cern.ch/event/448571/

CERN, WLCG & the EOSCpilot

- Proposed Use Case is based on **LTDP** for WLCG (and other) HEP experiments
  - 100TB / 100PB / 10EB for 3 decades +

- Services: either **fully generic** or else based on **generic** components
  - "Certified" bit preservation, INSPIRE / Zenodo, CernVM[FS]

- **Deploy (some of) these on EOSCpilot**
- **And / or offer as generic services**
- **And / or benefit from other generic services**
- **And / or link to other EOSC-related projects**

What is HEP data?

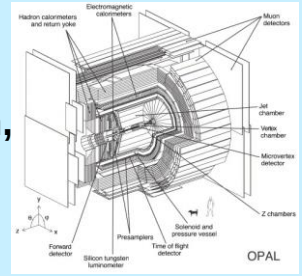
Digital information

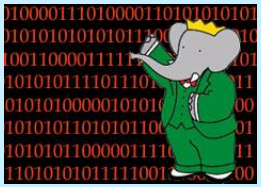
The data themselves, volume estimates for preservation data of the order of a few to 10 PB

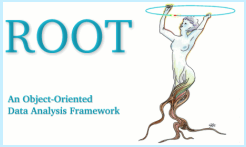
Other digital sources such as databases to also be considered

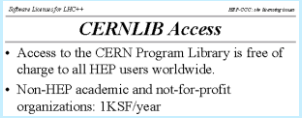
Software

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



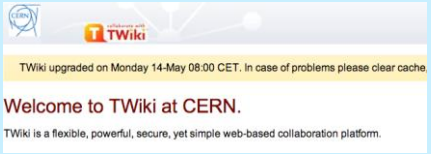






Meta information

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



Publications

arXiv.org

INSPIRE

Durham University — Science & Technology Facilities Council

HEPDATA: REACTION DATA Database

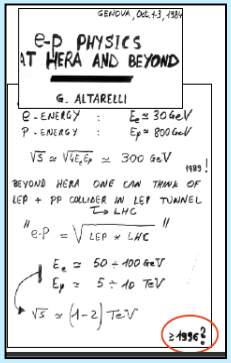
...containing numerical values of HEP scattering data such as total and differential cross sections, fragmentation functions, structure functions, and polarisation measurements, from a wide range of experiments. It is compiled by the **Durham Database Group (UK)** with help from the COMPAS group (Russia,)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

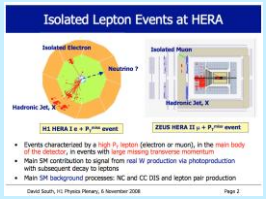
Journal of High Energy Physics

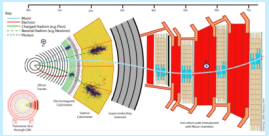
A refereed journal, written, run and distributed by electronic means

Documentation

Internal publications, notes, manuals, slides







Expertise and people

DESY

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?)

1. 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 **(later OPERA and others? …)**
6. A **DPHEP portal** that links also to data preservation efforts at other HEP institutes worldwide.

Update at DPHEP workshop in March 2017

iPRES 2016
13th International Conference on Digital Preservation //
Bern // October 3 – 6, 2016

2. 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.

First presented to ICFA in February 2013

2. How do we measure progress / success?

- **Practice:** through Open Data releases
  - Can the data really be (re-)used by the Designated Community(ies)?
  - What are the support costs?
  - Is this sustainable?
- **Theory:** by applying state of the art "preservation principles"
  - Measured through ISO 16363 (self-) certification and associated policies and strategies
  - Participation in relevant working & interest groups

One, without the other, is probably not enough. The two together should provide a pretty robust measurement...

2. ISO 16363 Certification

- Discussed already at 2015 PoW and several WLCG OB meetings
- Proposed approach:
  - An **Operational Circular** that describes the organisation's commitment to the preservation of scientific data & general principles (**draft exists**);
  - **Data Management Plans** by **project** where needed to refine embargo periods, designated communities etc.
  - A **Preservation Strategic Plan** covering a 3-5 year period
    - **DPHEP Blueprint (2012) and Status Report (2015) can be considered the first & second in such a series**
- This should cover the "**holes**" we have wrt section 3 of ISO 16363
- Needs to be done in close collaboration with experiments and other LTDP service providers: **start with a Workshop in 2017**
  - **Tentative dates: March 13 - 15 2017**

ISO 16363 metrics
Organisational Infrastructure (3)

3.1	Governance & Organisational Viability	Mission Statement, Preservation Policy, Implementation plan(s) etc.
3.2	Organisational Structure & Staffing	Duties, staffing, professional development etc. [APT etc.]
3.3	Procedural accountability & preservation policy framework	Designated communities, knowledge bases, policies & reviews, change management, transparency & accountability etc.
3.4	Financial sustainability	Business planning processes, financial practices and procedures etc
3.5	Contracts, licenses & liabilities, "rights"	For the digital materials preserved...

1. 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 **(later OPERA and others? …)**

6.A **DPHEP portal** that links also to data preservation efforts at other HEP institutes worldwide.

Update at DPHEP workshop in March 2017

iPRES 2016
13th International Conference on Digital Preservation //
Bern // October 3–6, 2016

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**

Requirements from Funding Agencies

- To integrate data management planning into the overall research plan, all proposals submitted to the **Office of Science** for research funding are required to include a **Data Management Plan** (DMP) of no more than two pages that describes how data generated through the course of the proposed research will be **shared and preserved** or explain why data sharing and/or preservation are not possible or scientifically appropriate.
- At a minimum, DMPs must describe how data sharing and preservation will enable **validation of results**, or how result could be validated if data are not shared or preserved.

H2020: Annex 1 (DMP Template)

The DMP should address the points below…

1. Data set reference and name
   - Identifier for the DS to be produced
2. Data set description
   - Description; origin; nature & scale; to whom useful; underpins publication? similar data?
3. Standards and metadata
   - Reference to standards *of the discipline*
4. **Data sharing**
   - How will it be shared? Embargo periods? Mechanisms for dissemination, s/w and other tools for re-use, access open to restricted to groups, where is repository? Type of repository?
5. **Archiving and preservation**
   - Description of procedures, how long will it be preserved? End volume? Costs? How will these be covered?

Workshop on Active Data Management Plans

- Agenda, talks, videos, conclusions
- Includes more detailed talks about HEP data preservation & Open Data releases

CMS data levels and open data

- CMS has approved a data preservation, re-use and open access policy, which defines the approach to access to them at various levels:
  - Level 1 - Open access publication and additional numerical data
  - Level 2 - Simplified data for outreach and education
  - Level 3 - Reconstructed data and the software to analyze them
  - Level 4 - Raw data, and the software to reconstruct and analyze them.

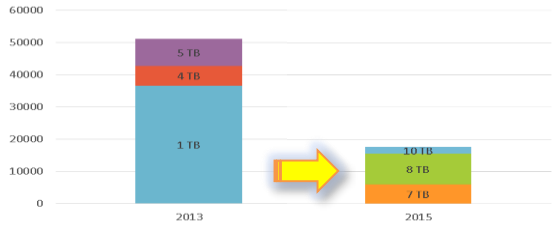
CMS Open Data

- CMS continues publishing and promoting levels 1 & 2.
- CMS data releases at level 3 - reconstructed data:
  - November 2014: 28 TB of 2010 collision data
  - April 2016: $>$ 100 TB of 2011 collision data and $>$ 200 TB of simulated data.

▸ CMS data preservation, re-use and open access policy

Bit Preservation: Steps Include

- Regular media **verification**
  - When tape written, filled, every 2 years…
- Controlled media **lifecycle**
  - Media kept for 2 max. 2 drive generations
- **Reducing** tape mounts
  - Reduces media wear-out & increases efficiency
- Data **Redundancy**
  - For "smaller" communities, a 2nd copy can be created: separate library in a different building (e.g. LEP – **3 copies at CERN**!)
- **Protecting** the physical link
  - Between disk caches and tape servers
- Protecting the **environment**
  - Dust sensors! (Don't let users touch tapes)



Constant improvement: reduction in bit-loss rate: 5 x 10^{-16}

Software Preservation

- HEP has since long shared its software across international collaborations
  - **CERNLIB – first started in 1964 and used by many communities worldwide**
- Today HEP s/w is O(10^7) lines of code, 10s to 100s of modules and many languages! (**No standard app**)

➢ **Versioning filesystems and virtualisation** look promising: have demonstrated resurrecting s/w 15 years after data taking and hope to provide stability 5-15 years into the future

➢ Believe we can analyse LEP data **~30 years** after data taking ended!

- **Does anyone have a better idea?**

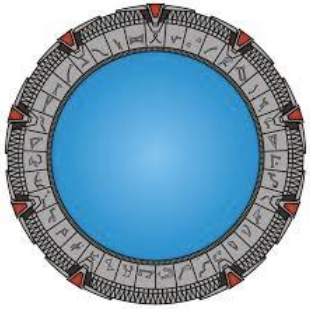
Analysis Preservation

- The ability to reproduce analyses is not only required by Funding Agencies but also essential to the work of the experiments / collaborations
- Use Cases include:
  - **An analysis that is underway has to be handed over, e.g. as someone is leaving the collaboration;**
  - **A previous analysis has to be repeated;**
  - **Data from different experiments have to be combined.**

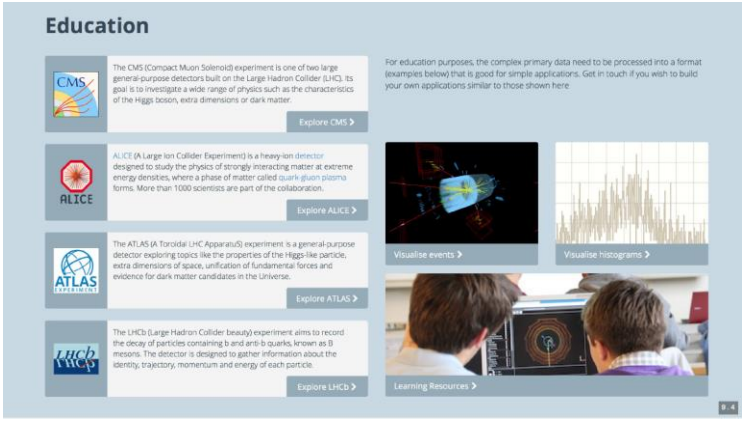
- Need to capture: metadata, software, configuration options, high-level physics information, documentation, instructions, links to presentations, quality protocols, internal notes, etc.

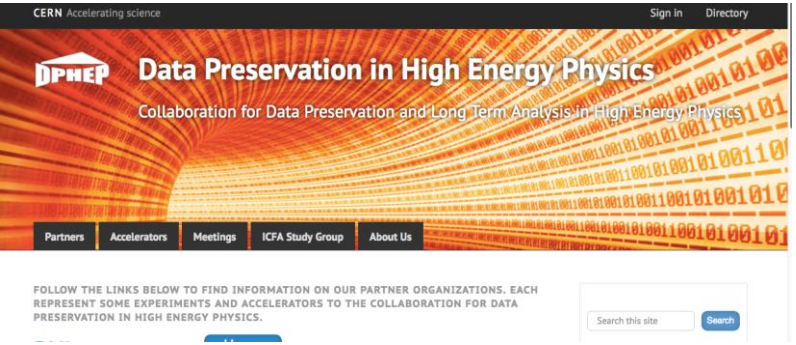
➢ At least one experiment (ALICE) would like **demonstrable reproducibility** to be part of the publication **approval process!**

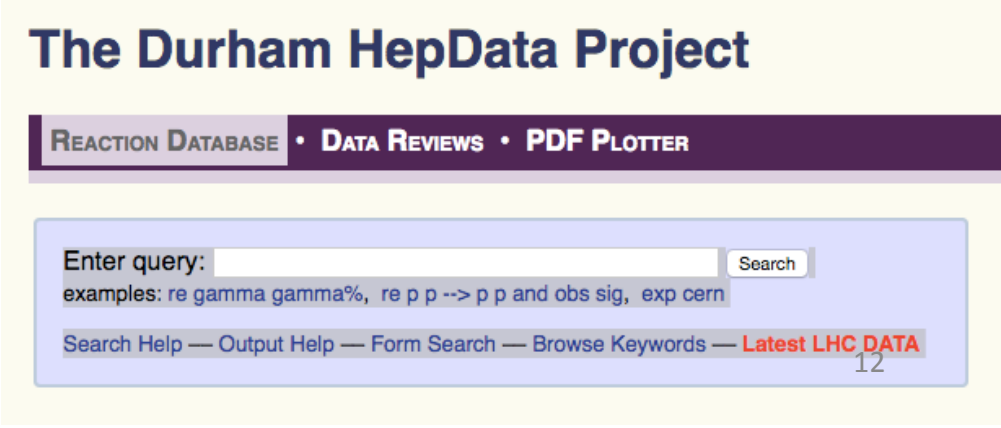
Portals



- No time to discuss in detail but clearly address the challenges of making the data "**discoverable**" and "**usable**" (if not necessarily F.A.I.R.)







"ESFRI" DMP w/s series

- Funding agencies typically require DMPs described how data will be preserved, shared and re-used

➢ **Idea: use DMPs - extended with a few key elements (e.g. data caching, movement etc.) - to look for synergies across different projects**

- Includes SKA, HL-LHC and many others: "synergies" translates to money and time saved!

- The first 1-2 workshops could be held at CERN and then hopefully rotate around large projects / institutes on a 12 - 18 month schedule

- **Is this the same as the ASTERICS-OBELICS workshops?**

  No, because it has a tighter focus but suggests that there is / would be interest in this area

CERN, WLCG & the EOSCpilot

- Proposed Use Case is based on **LTDP** for WLCG (and other) HEP experiments
  - 100TB / 100PB / 10EB for 3 decades +

- Services: either **fully generic** or else based on **generic** components
  - "Certified" bit preservation, INSPIRE / Zenodo, CernVM[FS]

- **Deploy (some of) these on EOSCpilot**
- **And / or offer as generic services**
- **And / or benefit from other generic services**
- **And / or link to other EOSC-related projects**

EOSC pilot

The European Open Science Cloud for Research Pilot Project