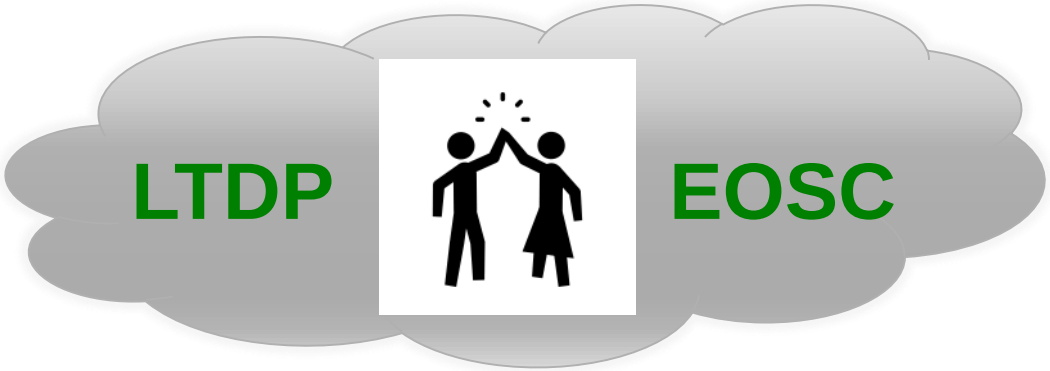

CERN

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

Long Term Data Preservation meets the European Open Science Cloud



PASIG, Oxford, 2017

Jamie.Shiers@cern.ch

https://indico.cern.ch/event/664663/

Outline

- LTDP: What is it?
- EOSC: What is it?
- How can one help the other?
- Current state of play

Long Term Data Preservation

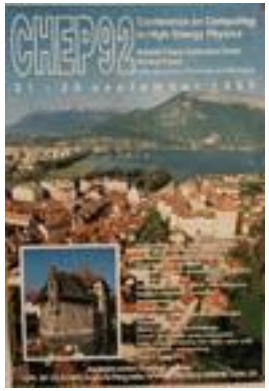
- e-IRG definition:

Long-term *is defined as a period of time long enough for there to be concern about the loss of integrity of digital information held in repositories, including deterioration of storage media, changing technologies, support for old and new media and data formats (including standards), and a changing user community.*

- DPHEP definition:
  - **"Disruptive change";**
  - **Target periods: 25 – 30 – 50 years.**

➢ **Let's look back: 25 / 30 / 50(!) years…**

In the year…



- ***T – 25:***
  - www was ***just*** emerging with the first X-based browsers…
  - CERN (HEP) had just begun migration to "distributed computing" and Unix…
  - Fortran, VMS, VM/CMS still dominant…

- ***T – 30:*** mainframe era; open reel 6250 bpi tapes (no-one in HEP even dreamed of LTDP)
- ***T – 50:***

(moon landing: T – 48)

What is the EOSC?

- EC communication from 2016:
  - *"It aims to develop a trusted, open environment for the scientific community for storing, sharing and reusing scientific data and results, the European Open Science Cloud"*
- Projects (EOSC Pilot, EOSC Hub, future calls in 2018 and beyond)
- Builds on existing services / previous projects, including EGI, EUDAT, OpenAIRE
- HLEG on EOSC: now in round 2

HLEG = High Level Expert Group

Is EOSC relevant for LTDP?

- *Research Infrastructures such as the ones on the ESFRI roadmap and others, are characterised by the very significant data volumes they generate and handle.*
- *These data are of interest to thousands of researchers across scientific disciplines and to other potential users via Open Access policies.*
- ➢ ***Effective data preservation and open access for immediate and future sharing and re-use is a fundamental component of today's research infrastructures and Horizon 2020 actions.***
- *In this context, European research stakeholders make increasing use of cloud services to effectively handle such data.*

EOSC Pilot Science Demonstrators

- SD on LTDP based on “HEP Use Case”
  - Attempt to implement something like “CERN Open Data Portal” using EOSC Pilot services
  - Follows closely recommendations from RDA Experts (“conventional wisdom”), e.g.
    1. *Digital objects in trustworthy (aka certified) repositories;*
    2. *PIDs / DOIs;*
    3. *Not just “data” but also schemas, queries, concepts etc.*
- ➢ Explicitly:
    1. *TDR for “physics data” (PIDs)*
    2. *Digital Library for documentation (DOIs)*
    3. *Software + Environment capture (CVMFS / CernVM)*

Targets & Stretch Targets

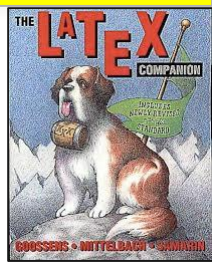
- **Given that:**
  - TDRs (believed to) exist (EUDAT?); **[100 TB scale – not 200 PB!]**
  - CernVM / CVMFS offered in production;
    - EGI InSPIRE, EGI Engage, WLCG, other HEP labs;
  - Invenio-based services: CDS, INSPIRE-HEP, Zenodo, B2SHARE, …

➢ ***What were we going to do after coffee?***

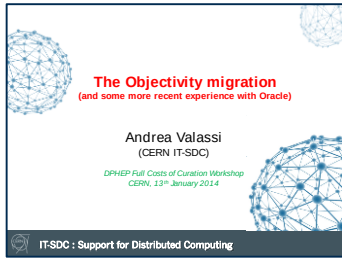
- **Stretch targets:**
  - Understand / implement F.A.I.R. in multi-disciplinary environment
    - Benefit from FAIR expertise in project
  - Understand (and potentially prototype) use of generic services by other disciplines

Migrations

- **Migrations are simply inevitable in LTDP!**



- Concrete examples in HEP:
  1. "CERNLIB" documentation – from 1995 LaTeX + PostScript / HTML3 to PDF(/A) and Digital Library
  2. Pre-LHC experiments: 200TB from ODBMS to HEP format: "triple migration" (Data, Media, S/W)
  - **Both cases took O(1 year) to perform**

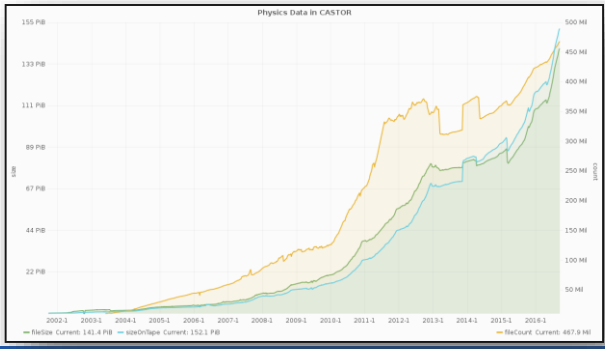


- Except in "trivial" cases, I don't see how migrations can be handled by TDR alone

Scalability Issues

- How many PIDs / DOIs does your system support?
- How many lookups per unit of time?
- Can this scale to e.g. 100PB (EB) of data, 1000s of users worldwide?
- For how long are your PIDs / DOIs guaranteed?

➢ **Are we (really) there yet?**



The Way Ahead…

- There are lots of promising ideas and even technical "solutions"
- But there are still many unanswered (and in some cases un-asked) questions

➢ **To solve these issues, realism, openness and honesty are needed**

- **And not e.g.**

  "We've been implementing FAIR for 100 years"…

- Measure success through combination of **"theory + practice"**

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...

HEP Data for Everyone:

CERN open data and the ATLAS and CMS experiments

Thomas McCauley

@tpmccauley

University of Notre Dame, USA

for the ATLAS and CMS Collaborations

https://tpmccauley.github.io/opendata-ichep2016

2012: "never", 2015: 1/2 day workshop at CERN, 2016: ICHEP presentation!

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

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.

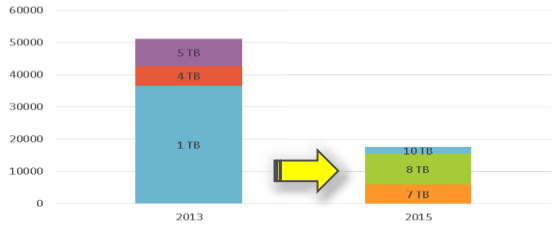
Update at DPHEP workshop in March 2017

its own right!

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

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!**

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…

Collaboration has helped us in...

1. The elaboration of a clear "**business case**" for long-term data preservation
2. The development of an associated "**cost model**"
3. A common view of the **Use Cases** driving the need for data preservation
4. Understanding how to address Funding Agencies requirements for **Data Management Plans**
5. Preparing for **Certification** of HEP digital repositories and their long-term future.

EOSC: What are the right metrics? (From e-IRG PPT)

- **As easy to use as Amazon?**
- **Cheaper (and better) than doing it in-house?**
- **A majority of ESFRIs use it as their baseline?**

➢ ***"To find dark matter, you need the EOSC"***?


CERN