Data Preservation: Status of and Strategy for Certification in WLCG

Jamie.Shiers@cern.ch

WLCG GDB

April 2016

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

2020 Vision for LT DP in HEP

- *Long-term – e.g. FCC-y timescales: "disrupt-ive 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, Reusable**)

Agree 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:**

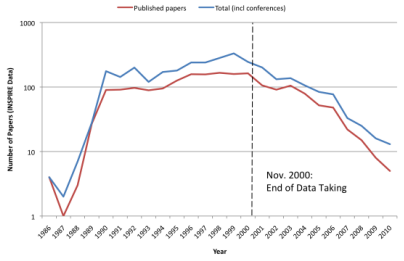
1. **DMPs should describe whether and how data generated in the course of the proposed research will be shared and preserved.**

   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).

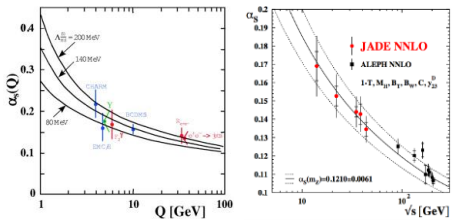
   At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

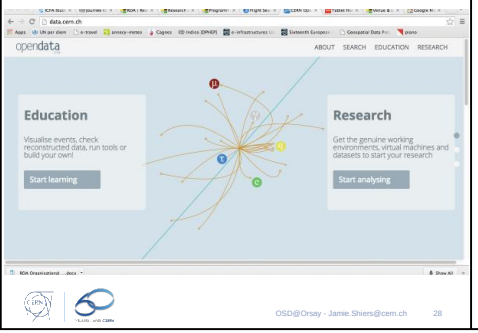
U.S. DEPARTMENT OF ENERGY | Office of Science

1 – 'Long Tail' of Papers



2 – 'New' Theore+cal 'Insights'

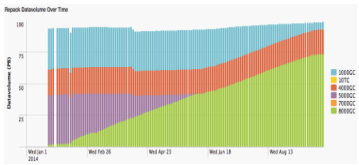




DSS Repack

CERN IT Department

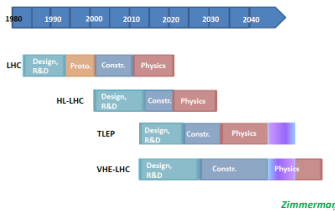
http://indico.cern.ch/event/CERN-ITTF-2014-09-26



- 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 calcula- on s but confirms our gut feeling that investment in fundamental science pays off

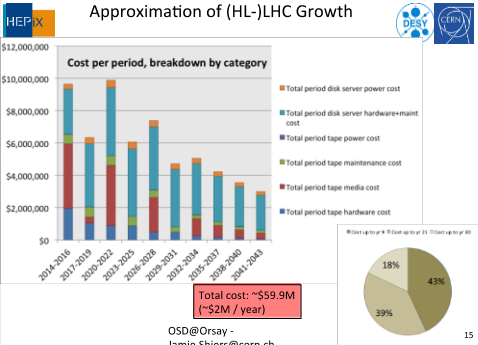
I think there is an opportunity for someone to repeat this exercise more rigorously

cf. STFC study of SRS Impact

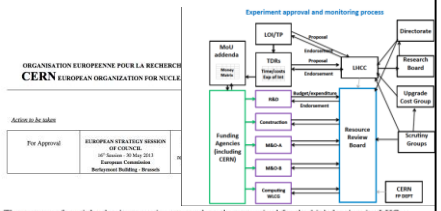
http://www.stfc.ac.uk/2428.aspx

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? End 2014

- **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**

Outline

- Role of certification: increase trust; respond to FAs; help ensure long-term sustainability
- Which model to follow?
- Where are we now?
- Plan

EU Trusted Digital Repository Framework

- A hierarchy of 3 aimed at increasing TRUST in digital repositories

1. The Data Seal of Approval (DANS – entry level);
2. Externally reviewed and publicly available self-audit based on ISO 16363 or DIN 31644;
3. Full external certification.

EU Trusted Digital Repository Framework

- A hierarchy of 3 aimed at increasing TRUST in digital repositories

1. The Data Seal of Approval (DANS – entry level);
2. **Externally reviewed and publicly available self-audit based on ISO 16363 or DIN 31644;**
3. Full external certification.

EU Trusted Digital Repository Framework

- A hierarchy of 3 aimed at increasing TRUST in digital repositories

1. The Data Seal of Approval (DANS – entry level);
2. **Externally reviewed and publicly available self-audit based on ISO 16363 or DIN 31644;**
3. **Full external certification?**

Why ISO 16363?

- **WLCG is not "entry level"**

- If we started with DSA I doubt we would ever go further

- **ISO 16363 actually matches quite well our existing practices – DSA is "too thin" for Tier0 but might be considered for Tier1s**
  - **Two processes to follow**
  - **We have already followed ISO 16363 training…**

Status

- A wiki has been created, accessible (R/W) to members of the DPHEP-IB
- So far, this concerns only the Tier0
- (Target is a draft of Tier0 self-certification prior to iPRES 2016, Bern in October)

The Metrics

- Grouped into 3 areas:

1. **Organisational infrastructure**
2. **Digital Object Management**
3. **Risk Management**

- **1** & **3** need to be addressed for all sites
- **2** can be done at the level of WLCG as a whole

Status & Plan

- I have drafted responses to many of the metrics in the areas **1** & **3** above (for CERN…)
- These need to be completed / reviewed by technical experts
  - As per procedure attached to agenda
- We then need a more formal review:
  - WLCG MB? OB? Higher?
  - **Quite some overlap with "experts" and MB…**

Issues

- **In a number of areas a formal strategy / document is expected**
  - Having such strategy documents would improve long-term sustainability
  - But will take time: some probably need to be at level of Scientific Policy Committee (or above?)

- **There are some differences in OAIS assumptions and our practices**
- These are most obvious in **Digital Object Management**
  - There are concepts in OAIS that are foreign to us…
  - … but would have value particularly in the long term

➢ **Proposal: review these once the first set of metrics has been completed, i.e. after iPRES / CHEP…**
  - In particular, in our environment, this will require close discussions with the experiments

Timeline

Year	What
2015	Training on ISO 16363 at CERN • Tier0 and some Tier1 representatives
2016	First draft of self-certification for CERN
2017	Ditto for Tier1s **Formalisation of procedures identified as missing (CERN)**
2018	**Further steps (e.g. external audit) prior to next ESPP update**
202x 203x	Repeat as required e.g. following major organisational or strategy changes

Summary

- **Even some experts consider ISO 16363 daunting**
- **But in fact, we already address many of the metrics as part of "business as usual"**
- **This exercise ties them together in terms of Long-Term Data Preservation**
- **It should help ensure that LTDP is a reality – in the long term**

➢ **I will be contacting people in the short-term to help!**