Preservation e-Infrastructure IG

What has it done for us?

Jamie.Shiers@cern.ch

RDA Plenary 5

San Diego, March 2015

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

Goals

- *The aim of the Preservation e-Infrastructure Interest Group is to reach wide agreement on the e-Infrastructure services which are needed to help repositories to preserve their data holdings, to ensure the interoperability of service implementations, and to build trust of service providers.*

- *Such distributed services supporting interoperability, including those that support continued usability, authenticity, accessibility, retrievability, visualization and replication, should allow the repositories to simplify, share the cost of, and improve, their preservation activities.*

Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics

www.dphep.org

Abstract

Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. An inter-experimental study group on HEP data preservation and long-term analysis was convened as a panel of the International Committee for Future Accelerators (ICFA).

- ~100 page document that clearly summarises the situation (2012) – but how to address it?

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**

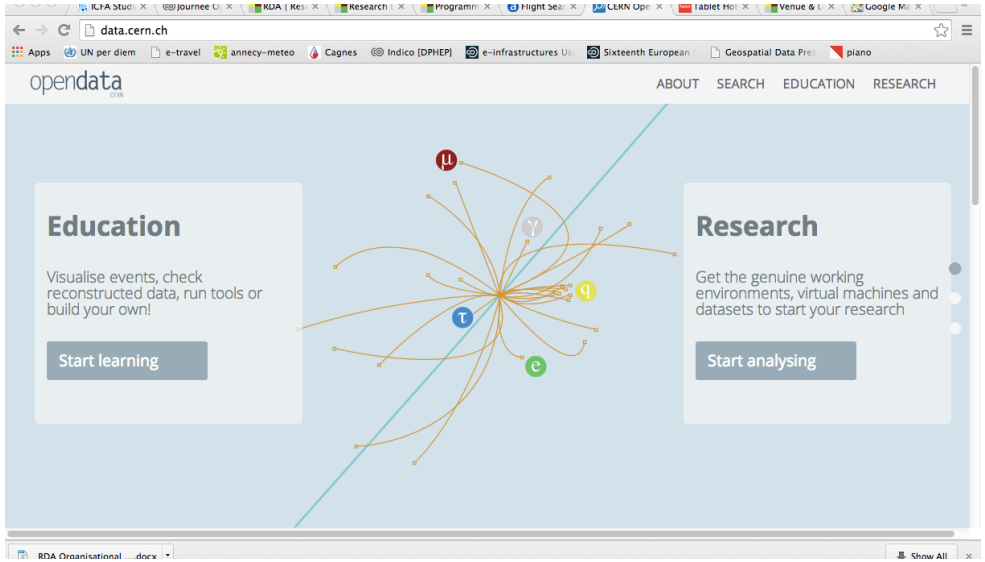
http://science.energy.gov/funding-opportunities/digital-data-management/

- ***"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).

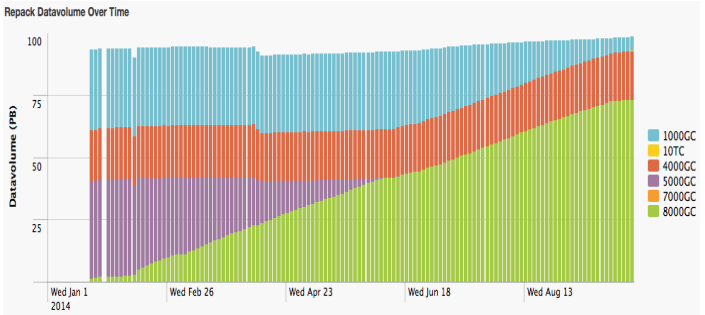
   At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.



DS8 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)

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**

Data Preservation plans

- The ALICE collaboration is committed to develop a long term program for Data Preservation to serve the triple purpose of
  i. preserving data, **software** and **know-how** inside the Collaboration,
  ii. **sharing** data and associated software and documentation with the larger scientific community, and
  iii. **give access to reduced data sets and associated software and documentation to the general public for educational and outreach activities.**

Without PeIG? RDA?

- We have clearly benefited a great deal from the knowledge and experience of individuals and projects
- Without this we would still have made progress but it would (likely) have been much more “introverted” in approach

The bottom line: we have saved – perhaps years – in achieving our goals AND defining the strategy