



High Energy Physics

Data Preservation An Open Data Portal pop up

Brief Overview

Funding agencies today require (FAIR) Data Management Plans, explaining how data acquired or produced will be preserved for re-use, sharing and verification of results.

The preservation of data from CERN's Large Hadron Collider poses significant challenges: not least in terms of scale. The purpose of this demonstrator is to show how existing, fully generic services can be combined to meet these needs in a manner that is discipline agnostic, i.e. can be used by others without modification.

Objectives

The DPHEP science demonstrator wants to deploy services that tackle the following functions:

- » Trusted and certified digital repositories where data is referenced by a Persistent Identifier (PID);
- » Scalable "digital library" services where documentation is referenced by a Digital Object Identifier (DOI);
- » A versioning file system to capture and preserve the associated software and needed environment; A virtualised environment that allows the above to run in Cloud, Grid and many other environments.

The final goal is to use non-discipline specific services combined in a simple and transparent manner (e.g. through PIDs) to build a system capable of storing and preserving Open Data (at a scale of 100TB or more).

What is the Problem?

The data from the world's particle accelerators and colliders (HEP data) is both costly and time consuming to produce - that from the LHC is a particularly striking example and ranges in volume from several hundred PB today to tens of EB by 2035 or so.

HEP data contains a wealth of scientific potential, plus high value for educational outreach. Given that much of the data is unique, it is essential to preserve not only the data but also the full capability to reproduce past analyses and perform new ones. This means preserving data, documentation, software and "knowledge".

There are numerous cases where data from a past experiment has been re-analyzed: we must retain the ability in the future.

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.

To tackle this problem, an International Study Group, DPHEP, (formally adopted by the International Committee for Future Accelerators - ICFA) was established in late 2008. This panel produced a "Blueprint Report" in 2012 that was input to the European Strategy for Particle Physics update. The Blueprint report can be found at: <http://arxiv.org/pdf/1205.4667>

DPHEP - an International Collaboration for Data Preservation and Long Term Analysis in High Energy Physics.

The latest Status Report from the DPHEP Collaboration, covering the years 2013 - 2015 inclusive, can be found at: <http://arxiv.org/abs/1512.02019>

DATA PRESERVATION HIGH ENERGY PHYSICS

DPHEP Intern Long



EOSC_{pilot}

The European Open Science
Cloud for Research Pilot Project

Main achievements

Some limited success was achieved with the individual services identified (Zenodo, CVMFS, a Trustworthy Digital Repository), but it was not possible to integrate them into an usable service.

Recommendations for the implementation

The EOSC Pilot integrates services from three well-established e-infrastructures, mentioned above. Equivalent services are used in production by the CERN Open Data Portal, which is available via anonymous access over the Internet worldwide.

While it was possible to upload a documentation file into the EUDAT B2SHARE test instance and while software from the LHC experiments is stored in the RAL CVMFS instance, there have been significant delays in finding a site that could act as a TDR for this pilot.

There were numerous misunderstandings regarding the scope, duration and scale of the demonstrator; no bulk upload of existing “Open Data” was achieved, anonymous access was not addressed, nor were the 3 services successfully integrated.

Partners of the SD

CERN



// Showing that these three building blocks could be used together in the context of the EOSC Pilot to deliver a discipline-agnostic “open data portal” would have been a powerful vindication of the Pilot. More work needs to be done to address the stumbling blocks, particularly around the area of data (“bit”) repositories. The CERN production version runs at the PB scale, so it is known that this can be achieved! //

Contacts

✉ Jamie.Shiers@cern.ch



EOSCPilot.eu has received funding from the European Commission's Horizon 2020 research and innovation programme under the Grant Agreement no 739563.