EOSC Pilot – DPHEP
Long-Term Sustainability demonstrator

Giuseppe Fiameni
Giovanni Morelli
Overview

- Introducing DPHEP
- The Data Preservation use-case
- Mapping to EOSC services
- Status of deployment
  - What was easy
  - What was challenging
- Outlook
Partners: BNL, CERN, CSC, DESY, Fermilab, IHEP, IN2P3, INFN, IPP, KEK, SLAC, STFC...

The collaboration aims to create a natural forum for the high energy physics community to foster discussion, archive consensus, and transfer knowledge on technological solutions and the diverse governance applying to the preservation of data, software, and know-how in the high energy physics community.

“Active” since 2009 - workshops, study groups etc..

- High level but also many pragmatic and practical people ensuring that things get done

What is (HEP) data? ...it's not just «the bits»

Digital Information
The data themselves, volume estimates for preservation data of the order of a few to 10 EB

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

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

Publications

Documentation
Internal publications, notes, manual, slides

Expertise and people

www.eoscpiilot.eu

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563
The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563.

- Robust, stable services over several decades
- Data preservation and re-use over similar timescale
- Need to support transparent data migrations
- Data growing, 100TB, 100PB… Exabytes…
  - One well defined DMP could be the same (now and tomorrow)
  - And today’s data volumes may be trivial for tomorrow’s storage
Goal: Demonstrate “best practices” regarding data management in the arena of LTDP, “open” data (sharing and re-use) - how we can realize this on the EOSC.

- PID for data and metadata stored in TDRs
- DOIs for documentation
- Expose and Archive the SW + environment

SD Goal => Equivalent to CERN Open Data Portal but using EOSC resources, thus allowing this solution to be opened to other communities.
Mapping the use-case to the service

<table>
<thead>
<tr>
<th>Service</th>
<th>HEP</th>
<th>EOSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trustworthy Digital Repository (TDR)</td>
<td>CERN Castor+EOS</td>
<td>EUDAT TDR (part of CDI)</td>
</tr>
<tr>
<td>PID/DOI systems</td>
<td></td>
<td>EUDAT B2Handle</td>
</tr>
<tr>
<td>Digital Library</td>
<td>CERN Document Server</td>
<td>EUDAT B2Share (Zenodo)</td>
</tr>
<tr>
<td>Software + Environment</td>
<td>CVMFS + CernVM</td>
<td>CVMFS + CernVM Tested on EGI FedCloud</td>
</tr>
</tbody>
</table>

Mix of EGI and EUDAT services/resources required - good to show interoperation between e-infrastructures.
The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563
Data Archive - Solution

WebDav
Initial option

B2STAGE
HTTP API
Final solution

HTTP REST API

Web Interface

iRODS

10 TB

GSI

PID

B2SAFE
Replica rule

Long Term
Preservation
Software and Environment:
- CVMFS instance working
- CernVMs tested on FedCloud and OpenStack/Vagrant

Document Server:
- B2SHARE - Documents uploaded to test instance

Trustworthy Digital Repository:
- CERN/CINES ingestion OK; CINES/CINECA replica OK, WebDav access OK
- Discussions regarding roles/requirements of communities and providers mainly done
- Service to open data still to be deployed

Conclusion: most of the boxes ticked, BUT some difficult aspect is still being tackled!
Deployment – levels of challenges

Relatively Easy: FedCloud (on demand service)

Relatively Easy: B2SHARE Document store (on demand service)

Medium Challenge: CVMFS (human interactions)

Significant Challenge: Archive solution (intense human interactions)

Specification of requirements is based on bi-lateral human interaction (on both sides). The requirement of Open Accessible archived data implies additional knowledgeable human resources.
Conclusion – so far...

General:

- Most of the pieces put in place
- Good example of a use-case for EOSC (not just technically)
- Community have learned quite a bit too, especially w.r.t the archiving of data

Next steps

- Consolidate the Data Preservation workflow (CERN/CINES interaction)
- Allow access to the replica by HTTP REST API @CINECA
- Harmonize the entire workflow take into account metadata also