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

http://dphep.org

refs:

arXiv:1205.4667 (the blueprint)

arXiv:0912.0255

Preliminary Views of HSF

Marcello Maggi (INFN)

On behalf of DPHEP

presented by Andrea Valassi (CERN)

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

- ICFA panel since 2009
  (led by Cristinel Diaconu, IN2P3)
- DPHEP Project Manager appointed in October 2012
  Jamie Shiers, CERN
- Collaboration agreement July 2014
  Signed by CERN, DESY, HIP, IHEP, IN2P3, IPNS, MPP
- First DPHEP Collaboration meeting for spring 2015
  - 2 (~10) DPHEP project (study-group) workshops
  - Several Implementation Board meetings
  - Participation to many conferences and workshops: visibility and new opportunities

Definition

Digital Data are affected by digital obsolescence

Medium
Access
Semantic

Preservation means

bit preservation
access preservation
knowledge preservation

Data Preservation is

Long Term Data Sharing facing disruptive changes

Why DP in HEP

Experiments are (in practice) not reproducible:
LEP, Tevatron, HERA, b-factories, LHC,...
are huge investments and
Larger communities can get involved (open access) during data taking and in longer term

Observations are unique:
Astrophysics, Earth Science, in fact pioneered DP:
FITS, Open Access Policies, OAIS

The "enemy"

Each single experiments design "proprietary" SW and data format for resource optimization...
Not needed in Long Term

LEP decade: Factor 300 increase of CPU power
SHIFT50 DEC Alpha had 320 Cern Units = 2.5 SpecHEP
Today 1 machine is 64 core & 560 SpecHEP

A fraction of a smartphone does the job...

DPHEP Objectives -1

Preserve data, SW, and know-how in the collaborations

- **Foundation** for long-term **DP strategy**
- **Analysis reproducibility**: Data preservation alongside software evolution

Share data and SW with larger scientific community

- **Additional** requirements:
  - Storage and distributed computing
  - Accessibility issues, intellectual property
- **Formalising** and simplifying data format and analysis procedure
- **Documentation**

DPHEP Objectives -2

Open access to reduced data set to general public

- **Education** and **outreach**
- **Continuous effort** to provide meaningful examples and demonstrations

Bit preservation

- **Data** taken by the experiments should be preserved

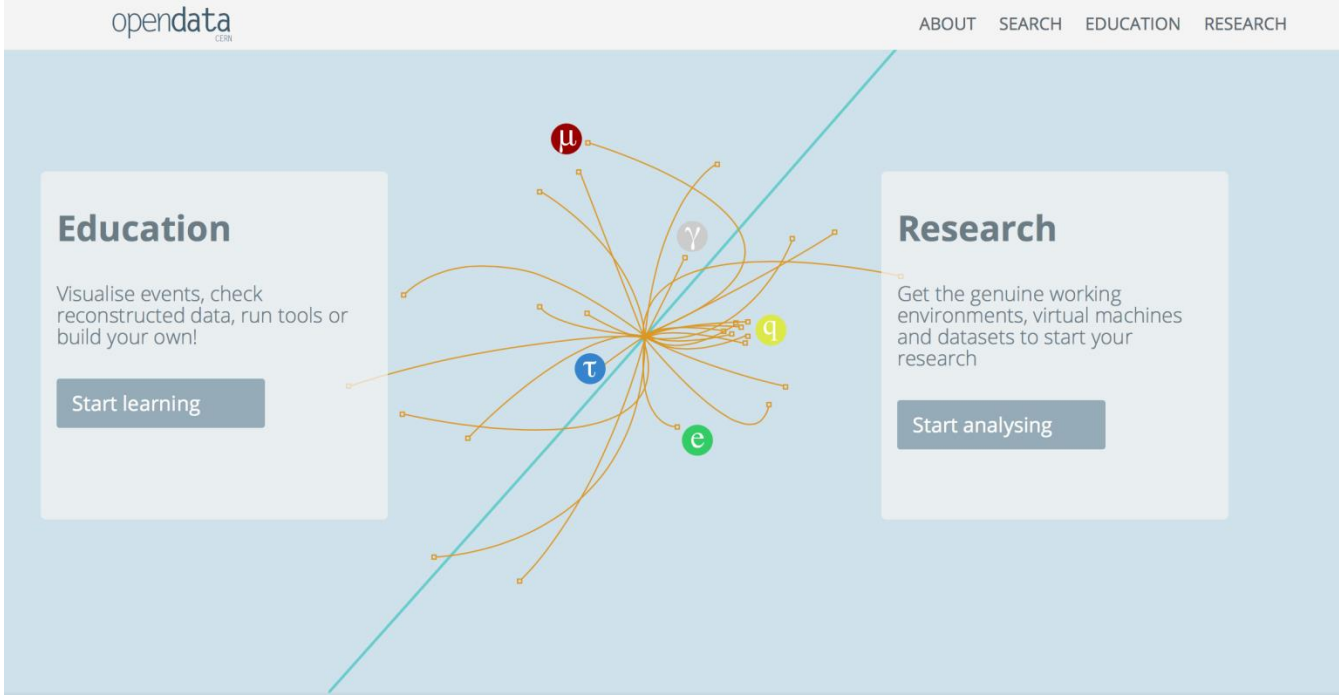
Strategy and scope in approved policy documents for all LHC collaborations

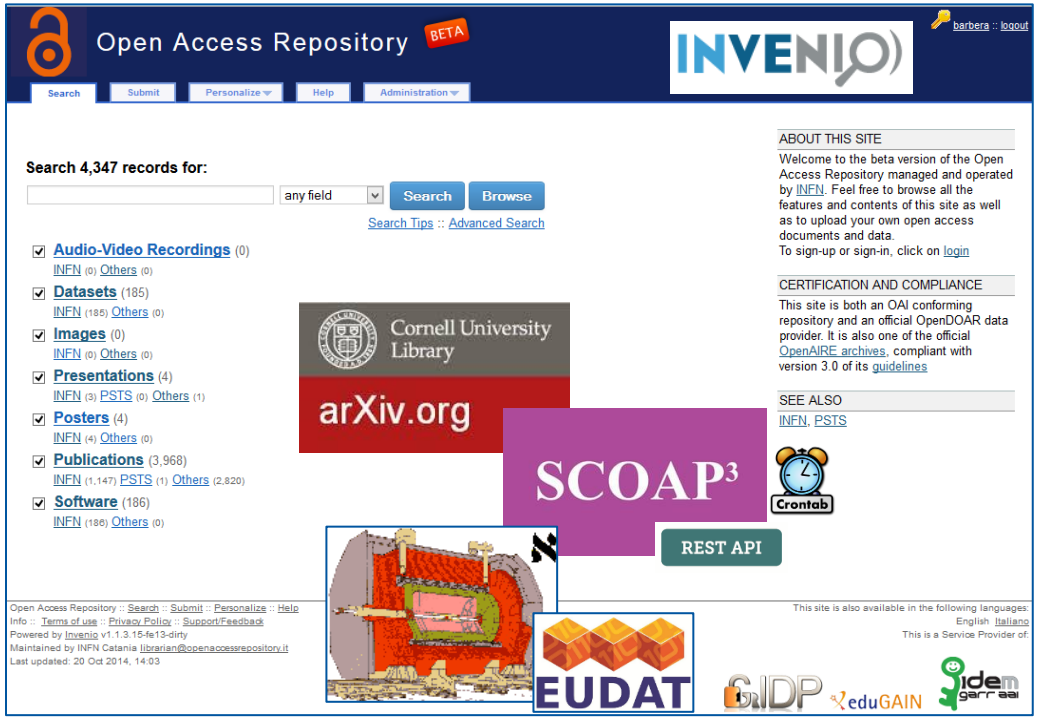
http://opendata.cern.ch/collection/data-policies

many other initiatives (DASPOS, PREDON, etc.)

Technologies:

INVENIO
CERNVM
CERNVM-fs
+…





Services:

OpenData CERN
OpenAccessRepository INFN

The views

- **DPHEP Collaboration** provides the framework for inter-experiment and inter-laboratory cooperation on data preservation

- **HSF** is the opportunity to extend
  - the collaboration around the main technological pillars of the DP framework: INVENIO CERNVM CERNVM-fs, …
  - And the SW collaboration for Common **Data Analysis** framework on preserved data & **Validation**

Relevance of DP to HSF and vice versa

(some ideas from myself and other HSF startup team members)

- HSF gives visibility to common software projects by many experiments
  - DPHEP hosts common software developments (data analysis/validation framework)

- HSF and DP are both concerned with the long-term evolution of software
  - They both recommend using data format standards and following best practices
  - Software sustainability, maintenance and documentation for future users are vital
  - Continuous porting to new platforms/compilers/externals and disruptive changes are an issue for both (DP may also use frozen configurations and virtualised environments)
  - ("Re[peatable|producible|computable] research" in Neil's presentation yesterday)

- Both HSF and DPHEP recommend Open Access policies
  - For all of data, software source code and publications
  - And both require the implementation of open access repositories for all such categories (the implementation itself being a candidate for sharing knowledge and developments)

- HSF and DP are both committed to engage with non-HEP communities
  - Other scientific communities and possibly beyond
  - DPHEP, DASPOS, PREDON, RDA are projects with several non-HEP links for DP