Data preservation
CMS perspective

Kati Lassila-Perini, Achim Geiser

Helsinki Institute of Physics, DESY

WLCG Grid Deployment Board
CERN
November 8, 2017

Outline

Motivation

- Without active data preservation, the LHC data will soon become (or has already become) unusable
  - Do we still have the knowledge and tools to access and intepret the data collected in 2009?
- The data preservation activities "compete" with normal operation
  - and come second in this competition by default, but it is a very honorable position and does not mean that we have to give up.
- With a relative small investment, we can achieve a lot in terms of data (re)usability
  - there is nothing in the computing models and practices which makes data preservation impossible, it is all about caring of small details.
- Using common, not experiment specific resources offers an excellent test bed for data preservation.

Data management in DP perspective

Very different requirements for data management plans and practices for:

Operations and active analysis

- High performance
- High demand
- Agreed and dedicated resources
- User authentication
- Developing SW, latest OS
- Access to all data
- Experiment specific services
- Highly specialized expertise
- Active human knowledge base
- Short-medium term availability

Data preservation

- Acceptable performance
- Low demand
- Unknown resources, now
- Open access (or authentication?)
- Fixed SW versions, compatible OS
- Legacy data
- Generic, common services
- Low expertise on technical details
- Usability relies on documentation
- Long term availability

Successful data preservation = long term usability $\rightarrow$ Simplify, simplify...

Data management in DP perspective

- CMS adopted a Data preservation, re-use and open access policy in 2012
  - DOI:10.7483/OPENDATA.CMS.UDBF.JKR9
- Rather than a plan, it is a statement of intent, which only later realized in concrete measures, but it has still been very useful:
  - as a final, approved outcome of the discussions within the collaboration
  - as a document for funding agencies.
- We are now drafting a real plan with the experience we have had from DPOA activities since 2012.
  - CMS has concentrated efforts in open data through CERN Open Data portal
  - We have demonstrated availability and usability of 2010 data for which slc5 environment is no more available through normal CMS resources.
  - Not all data is open data, now starting to address all legacy data.

Data management plan - Run1 wish-list

- Cataloguing all different Run1 legacy datasets, with **the core pp**
  - 2010 ($\approx$ 36 pb^{-1} out of 40 pb^{-1} available publicly)
  - 2011 ($\approx$ 2.5 fb^{-1} out of 5 fb^{-1} available publicly)
  - 2012 ($\approx$ 13 fb^{-1} out of 22 fb^{-1}to be released this year)

  and **the corresponding MC**
  - none with the legacy SW release for 2010
  - partial set with the legacy SW for 2011 (200 TB, available publicly)
  - 2012 (1.1 PB to be released this year)

  and **the special datasets** with
  - Heavy ions: PbPb and pPb
  - pp with low beta, CASTOR, TOTEM
  - pp at different collision energies (0.9, 2.36, 2.76, 5 TeV)
- Open data tools useful and usable also for collaborators, but - ideally - we would like to make possible for CMS members to
  - generate, simulate and reconstruct new MC for legacy data
  - re-reconstruct from RAW
  - access CMS resources needed for analysis of legacy data
  - use special tools and software needed for special datasets.

CMS Open Data

- First release in CERN Open Data Portal in November 2014, second in April 2016, third in preparation
  - first research publication out: Journal, Arxiv, OA
  - and we have not had any trouble, yet.
- Full provenance information of collision and MC datasets
  - not in a user-friendly but physicist-readable format.

  Collision data records (primary datasets) with detailed data selection information

  Simulated data records with detailed production information
- Workload i.e. questions from external users
  - have reported some temporary downtimes
  - have requested additional information necessary for analysis (which has triggered action in CMS)
  - have provided solutions to some technical issues (file sharing etc)
- Data used
  - In physics research, but it takes at least as long as for CMS members
  - In other research, e.g machine learning (data format not optimal...)
  - Widely in education.

CMS Open Data - usage pattern

- Data accessible from disk (eospublic)
  - through xrootd
  - or direct download from the portal.
- Software and condition data served to a CMS OpenData VM (CernVM with CMS contextualization) from cvmfs.
  - environment similar to lxplus environment for CMS members.
- Example code (to be compiled in CMSSW environment on VM) downloaded from the portal and accessibe on github.
  - Supplementary information needed for analysis, which CMS users get from CMS specific services and db's, is provided through the portal.
- For many example analyses and most our validation examples, single desktop and xrootd access is enough
  - however, research data users probably download the datasets locally
  - and reprocess the CMS specific AOD data format to a format convenient for their use and independent of the CMS software and environment.

Assessing FAIRness of Datasets

- Findable (defined by metadata (PID included) and documentation)
  1. No PID nor metadata/documentation
  2. PID without or with insufficient metadata
  3. Sufficient/limited metadata without PID
  4. PID with sufficient metadata
  5. Extensive metadata and rich additional documentation available
-  + !!
- Thanks to CERN Open Data portal.
- You need know very well what you looking for (some improvements for search foreseen in the portal update soon).
- Metadata is very complete but not easy to read.

Assessing FAIRness of Datasets

- Accessible (defined by presence of user license)
  1. Metadata nor data are accessible
  2. Metadata are accessible but data is not accessible (no clear terms of reuse in license)
  3. User restrictions apply (i.e. privacy, commercial interest, embargo period)
  4. Public access (after registration)
  5. Open access unrestricted
-  !!!
- Thanks to CERN Open Data Portal.
- Usability of the data relies on SW and environment (provided).

Assessing FAIRness of Datasets

- Interoperable (defined by data format)
  1. Proprietary (privately owned), non-open format data
  2. Proprietary format, accepted by Certified Trustworthy Data Repository
  3. Non-proprietary, open format = "preferred format"
  4. As well as in the preferred format, data is standardised using a standard vocabulary format (for the research field to which the data pertain)
  5. Data additionally linked to other data to provide context
- 
- Consider root file format as preferred format for our "designated community", no standard vocabulary formats in HEP yet.
- Usability of the data relies on SW and environment (provided)
- NB: Definition of "interoperable" in FAIR:
  - Does it mean, in research use in HEP: Data can be analysed by people outside of the collaboration and compared to MC predictions?
  - Or going further, combined with the data from other experiments?
  - Or more generic, used in other context, such as maching learning?

Assessing FAIRness of Datasets

- Assessing the reusability of the datasets by reusing the datasets
  - in documented analysis examples
  - in benchmarks analyses reproducing some published results.

  and providing the software and instructions on the portal.
  - Available: see dimuon peak, track p_T, η spectra Validation collection
  - More coming soon: J/Psi, Upsilon, inclusive jet xsec, ridge effect
- All results obtained using windows or linux office desktop computers.
- No grid jobs, no batch jobs on farm, no CMS account needed.
- See also for further insight CMS Open Data in Research by Achim Geiser.

CERN Analysis Preservation (CAP) framework

- Data preservation has several facets:
  - the preservation and usability of the datasets (see open access)
  - the preservation of the knowledge that leads to the final outcome i.e. publication (analysis preservation).
- See an introduction in the DPHEP workshop: ▸ CERN Analysis Preservation
  - preserving what is between the primary data and the final publication.
- The CAP use-cases are well acknowledged by CMS,
  - "Yes, I have it somewhere, but I have to find it"...

  and it is addressing an area not covered by our own tools and practices.
  - And we do expect skepticism and resistance as for any other tool or practice requiring additional work.
- We are now entering to a very important interface testing phase with real data deposit.
  - We are still in an exploratory phase.
- We want to demonstrate that we can gain in efficiency with CAP.

Software sustainability

- We fully rely on CernVM and cvmfs services for availability of
  - CMS and other HEP specific software (normal operations and DPOA)
  - working environment (DPOA)
  - condition data (DPOA).
- CernVM service already provides us the possibility to access old data, and offers sustainable solutions after the end of slc5 support.
  - Needed sooner or later for slc6, slc7... i.e. a standard step in DP chain.
  - Our use cases hopefully give an early opportunity to explore and test long term solutions.
- It is mandatory that services on which data preservation relies are aware that they are providing DP services (in addition to services for normal operations).
- From the DPOA point of view, cvmfs and CernVM services are the cornerstone for the reusability of our data - and they work so smoothly that the service providers hardly get any credit - thank you!

Outlook

- CMS is a heavy user of (and a contributor to) CERN DP services
  - CERN Open data portal
  - CERN Analysis preservation
  - CernVM and cvmfs
  - HEPData (IPPP Durham, CERN).

  benefiting greatly from expertise in IT, SIS and EP-SFT services.
- CMS is reassured that the bits are being preserved.
- CMS is measuring the success of DP actions in practice through open data releases
  - but we do acknowledge the value of formal assessments through (self-) certification and associated policies and strategies.
- DP activities remain background activities for an active experiment:
  - we rely on expertise of heavily overloaded people
  - we may rely on already heavily overloaded resources
  - but we observe a positive change in attitude within the collaboration (although it does not always result in faster reaction time).