Update from CMS
(or notes from yesterday's presentations and discussion)

Kati Lassila-Perini, Achim Geiser

Helsinki Institute of Physics, DESY

DPHEP Collaboration meeting
CERN
March 14, 2017

Outline

1. Data management plan
2. Open data
3. FAIR
   - Find
   - Access
   - Interoperate
   - Reuse
4. Analysis preservation
5. Software sustainability
6. Outlook

Data management plan

- Necessary to distinguish: data management plans and practices
  - for operations and active analyses
  - for data preservation

  have very different requirements.
- 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.

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 we want to ensure that CMS continues to be able
  - generate, simulate and reconstruct new MC for legacy data
  - re-reconstruct from RAW
  - access CMS resources with old VMs
  - 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
  - and we have not had any trouble, yet.
- Extensive upgrade for 2016 release with 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
  - In education, for High-school teachers in Florida (Project CODER), at CERN (HST), Finland (course material in jupyter notebooks)

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....
- 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.
- To make use of the the data you access you rely on software and environment (which are provided but not part of this assessment)....

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
- 
- We can consider root file format as preferred format for our "designated community".
- We do not have further standard vocabulary formats in HEP yet.
- To make use of the the data you access you rely on software and environment (which are provided but not part of this assessment)....

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:
    - ★ dimuon peak, track p_T, η spectra Validation collection
  - Soon available
    - ★ J/Psi, Upsilon, inclusive jet cross section, ridge effect
  - Work ongoing
    - ★ W/Z cross section, b tagging
- All results obtained using windows or linux office desktop computers at DESY by Achim and collaborators.
- No grid jobs, no batch jobs on farm, no CMS account needed.

CERN Analysis Preservation framework

- See the introduction by Suenje and Tibor on Monday
  - CERN Analysis Preservation
- The CAP use-cases are well acknowledged by CMS, and it is addressing an area not covered by our own tools and practices.
  - 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 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).
- See the introduction by Jakob on Monday

  ▸ Software preservation
- We are very interested to study the dependencies of our workflows with tools like umbrella and parrot

  ▸ Haiyan's presentation
- From the DPOA point of view, cvmfs and CernVM services are a cornerstone for the reusability of our data - and works so smoothly that it hardly gets any credit - thank you!

Outlook

- Presentations in this workshops are very relevant to CMS DP.
- 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.
- DPHEP is an excellent forum to
  - discuss common projects which are essential for long-term preservation
  - share the expertise with others and learn from long-term experience
  - raise the awareness of effort and funding needed to DP.