Data Preservation, Curation and Stewardship
Review of use cases from CMS

Kati Lassila-Perini

Helsinki Institute of Physics

on behalf of the DPOA team
DPHEP discussion
January 8, 2015

Use cases

- As defined in the ALICE document
  - preserving data, software and know-how inside the Collaboration
  - sharing data and associated software and documentation with the larger scientific community
  - give access to reduced data sets and associated software and documentation to the general public for educational and outreach activities.
    - NB: through the open data release, CMS gives access to the data in the original format also to the general public, with example sets of reduced data and instructions.

Service needs: open data

- Scope: long-term preservation of data and software with a compatible working environment allowing for diverse uses cases
  - either scientific, or educational and outreach usage.
- Services through the CERN Open Data portal: opendata.cern.ch
  - all records (data, software, instructions) preserved in the portal independenlty from the experiment-specific tools
  - common service for all experiments
- "Open data" =
  - storage of the open legacy data (currently at CERN eospublic) and access to them (through xrootd), including the conditions database
  - compatible CMS software accessible from github
  - compatible CMS working environment as CernVM image
- Open CPU resources are not offered for external use: can some limited resources be envisaged?
  - Example: https://cmsopendata.ifca.es/

Preserve the usability

- Long-term preservation means migrations
  - storage media
    - well-known, familiar procedures?
  - data format
    - from complex AOD format to somewhat more "ready-to-use" MiniAOD (in development for Run2)?
    - in a (common) four-vector type format?
  - working environment
    - how long will the current CMS Opendata CernVM (slc5) survive?

  and they will require validation.
- Need to extract a minimal validation set for any migration needs in the future.
  - Select and store reference figures and plots, and the corresponding workflow
  - Foresee a validation testbed.

Service needs: analysis preservation

- Scope: long-term preservation of "ingredients" of an analysis for re-use and reproducibility
  - the level of reproducibility to be defined and can vary: from raw data to results - from final ntuples to final plots
- Services through CERN analysis preservation portal (in development)
  - all records (data, software, instructions) preserved in the portal independenlty from the experiment-specific tools
  - common service for all experiments
- "Ingredients of an analysis" =
  - record all processing steps and corresponding selections
  - store the sw and instructions, ideally also the datasets for each step
    - recording of the datasets will be limited by the availability (some superseded by later reprocessings) and storage space
  - we will start investigating and exercising with concrete use-cases and prototype services.