Data Preservation: The LHC Experiments

Roger Jones, David South, Mihaela Gheata, Predrag Bucic,
Kati Lassila-Perini, Silvia Amerio, Frank Berghaus, Jamie Shiers

Objectives

- Preserve data, software, and know-how in the collaborations
  - Foundation for long-term DP strategy
  - Analysis reproducibility: Data preservation alongside software evolution
- Share data and associated software with larger scientific community
  - Additional requirements:
    - Storage, distributed computing
    - Accessibility issues, intellectual property
  - Formalising and simplifying data format and analysis procedure
  - Documentation
- Open access to reduced data set to general public
  - Education and outreach
  - Continuous effort to provide meaningful examples
- Bit preservation
- Strategy and scope in approved policy documents for all collaborations
  - http://opendata.cern.ch/collection/data-policies

Analysis Reproducibility

- Target: Collaboration
- Published analysis metadata coming with required provenance
    - Long term preservation of analysis ingredients for re-use and reproducibility
    - Analysis and production software stored as tags in version control systems (git or subversion)
        - Binary builds of tag made available via cvmfs
        - Rebuild software in the future or store binaries with environment in a virtual machine
    - Exercise first within collaboration then gradually expose to sharing platforms: https://data-demo.cern.ch/
- Reproducibility further requires:
    - Operating system and software framework, conditions databases, analysis macros, and documentation
    - Need to separate analysis from production DB to allow packaging into a VM environment
- Projected storage requirements: O(10TB) per analysis
    - Could be virtual

Scientific Community

- Fraction of analysis level data released
  - For some experiments so far
  - Provides Virtual Machine with required software environment
    - Connects to cvmfs and database services
    - Task: Separation of DB needs for analysis and production
  - Available via open data portal: http://opendata.cern.ch
  - Need independent access and storage
  - Want simple, well documented data access methods (HSF)
- Should only release single version of data
  - May change with reprocessing etc.
- Envisioned to share O(1PB) of data per experiment (2010-2012)
  - CMS gives open access to AODs via the open data portal
  - ATLAS has plans to allow open access to data via a Kaggle challenge
  - ALICE planning to release 10TB of 2010 data
  - LHCb plan to release their data in 2018

Education & Outreach

- First effort: CERN Master Class program
  - Access to limited data set with for high-school students and teachers
    - Simple data format
    - Could use full AOD set
  - Available via open data portal:
    - http://opendata.cern.ch
  - Demonstrator program with interactive event display
- Provides access to data, software tools, and documentation
  - Out of the box procedure: download and run graphical user interface without complications and environment settings
- Portal access allows users to write independent demos
  - Based on released data and existing examples
- Small hardware requirements: O(1TB) storage

Data/Bit Preservation

- RAW data (bits) should be preserved
- Site perspective:
  - Memorandum of Understanding for the tier0 and tier1’s
  - CERN’s currently preserves all bits in the data store
  - Tier1’s migrate to new tapes
    - Each defines their own intervals
  - Validation by regular reading of data tapes
    - Possibly include running physics code on data
- Experiments responsible for distribution across WLCG
- Want to schedule a training seminar:
  - http://www.iso16363.org/

Summary

- Large overlap between needed tools, services, and support
    - Many pieces already exist
    - Need some flexibility to accommodate all experiments
- LHC experiments are already collaborating on these use cases
- Report status and progress every ~6 months to GDB
    - Schedule dedicated topical meetings for in-depth discussion as needed