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 and demonstrations
- Bit preservation
  - Data taken by the experiments should be preserved
- Strategy and scope in approved policy documents for all collaborations
  - http://opendata.cern.ch/collection/data-policies

Analysis Reproducibility

- Target: Collaboration
- Analysis and production software stored as tags in version control systems (git or subversion)
  - Binary builds of tag made available via cvmfs
  - Production: Builds are released regularly
- Reproducibility further requires:
  - Operating system and software framework, conditions databases, analysis macros, and documentation
- Published analysis metadata coming with required provenance
  - Long term preservation of analysis ingredients for re-use and reproducibility
    - Level of reproducibility? From RAW or derived data?
  - Should be possible to change analysis parameters or input data
- Exercising limited access portal for analysis reproducibility

Analysis Reproducibility

- Exercise first within collaboration then gradually expose to sharing platforms:
  - https://data-demo.cern.ch/
  - Implementations are in discussion/active development
- Projected Requirements
  - Data storage: O(10TB) per analysis
    - Could be virtual
  - Software repositories and snapshots
  - Database framework
  - Virtual machine (CernVM) and cvmfs infrastructure
  - …

Scientific Community

- Fraction of AOD-level data released
  - For some experiments so far
  - Provides Virtual Machine with required software environment
    - Connects to cvmfs and database services
    - User needs to provide compute resources, data access via xrootd
    - Large scale scientific use will require users to provide storage and compute resources
  - Available via open data portal:
    - http://opendata.cern.ch
- Needs independent access and storage
- Should only release single version of AODs
  - Released version 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

Education & Outreach

- All LHC experiments are already collaborating
- Projected Requirements:
  - Storage: O(1TB) data
  - Software
  - Distribution via web portal
  - …

Data/Bit Preservation

- RAW data (bits) should be preserved
  - Memorandum of Understanding for the tier0 and tier1's
  - CERN's currently preserves all bits in the data store
    - Example: three migrations of full (including LEP) data since LHC start
    - What is the current practice at the tier1's?
  - Should we schedule a training seminar?
    - http://www.iso16363.org/