Preparations for the public release of high-level CMS data

Tom McCauley (University of Notre Dame)
Kati Lassila-Perini (Helsinki Institute of Physics)
for the CMS collaboration

thomas.mcccauley@cern.ch
@tpmcccauley
CMS (Compact Muon Solenoid) is one of the two general-purpose experiments at the LHC.

Over 300 papers published describing searches for SUSY and exotica, measurements of QCD, electroweak, top, b, forward, and heavy-ion physics, as well as the discovery of the Higgs boson.

Collected ~ 28 l/fb of proton-proton collision data at COM energies up to 8 TeV.

Nearly 3000 physicists and ~800 engineers from over 40 countries.

http://cern.ch/cms
CMS data levels in preservation and open access

- **Level 1**: Open access journals used; additional information provided (e.g. data behind figures)

- **Level 2**: Simplified, selected data for education and outreach

- **Level 3**: Reconstructed data; software and documentation needed for analysis

- **Level 4**: Raw data; software and documentation needed for reconstruction and analysis

CMS data preservation, re-use, and open access policy:
https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=6032
Open data and its usage

• For use in masterclasses* (aimed at high-school level) CMS has released small, selected data samples

• CMS has approved the release of a large set of reconstructed data for public use

• Challenge (one of them) faced by public: need physics and specialized computing knowledge in order to extract meaningful results

• How to get value for the public from open data? Initially focus on a concrete use-case with already-proven success: education

* See E&O session tomorrow
How does one get from...

AOD file in ROOT format, needing CMS software to properly read it
an event in a browser-based online event display (written in JavaScript)
...or to an online histogram application written in JavaScript with simple JSON input
Public data release

• To be released: ~ 30 TB of 2010 p-p collision data at 7 TeV (tens of 1/pb, a tiny fraction of total collected) in CMS AOD format (suitable for analysis, as the “A” implies)

• Accompanied by: i) simple, ready-to-use online applications like histogram tools and event display, ii) a virtual machine with CMS software environment with example analysis code

• Possibility of development of outreach/education applications by the public
Data re-use

- Data will be released under the Creative Commons CC0 waiver: http://creativecommons.org/publicdomain/zero/1.0

- Data will be identified with persistent data identifiers (DOIs)

- It is expected that third parties will cite the CMS public data through these identifiers

- Publications based on public data by members of the experiment regulated
Data preservation and open access services at CERN

- DP&OA services in prototyping phase at CERN, with involvement from IT, GS, and PH
- Services to be developed with input from CMS and other LHC experiments
- CMS public release will rely on these services for:
  1. legacy data storage and metadata archives,  
  2. access and distribution of data,  
  3. an entry portal for tools and instructions,  
  4. security and bandwidth restrictions for public access
Common open data portal for outreach and education

• LHC experiments releasing public data have strong incentives to use common services at CERN: i) can benefit from wide expertise at CERN in data storage, archiving, and transfer, ii) can avoid overlapping, experiment-specific solutions, iii) can have a platform with common tools, and give external people the possibility to access them and build applications.

• Possible functional areas of a common portal hosting applications:

[Diagram showing various functional areas such as applications, data catalogue, discussion forums, and common tools.]
Outlook

• We plan to make the CMS public data release coincide with the CERN 60th anniversary in September

• Pilot implementation of basic open data services at CERN will be in place

• Excellent opportunity to understand what public data implies in the terms of expectations of external users, the impact on the experiments, and the resources needed

• Excellent opportunity to demonstrate to the public and the funding agencies that suggestions (and increasingly requirements) concerning DP&OA are being taken seriously

• We are now preparing the examples to accompany the data release: basic online tools, in-principle usable for other experiments as well as simplified analysis examples
Acknowledgments

- Thank you: collaborators on CMS and at CERN
- Thank you: the organizers