CMS Data Policy

Kati Lassila-Perini

Coordinator for data preservation and open access in CMS

- The CMS data preservation, re-use and open access policy, was approved by the CMS Collaboration Board in March 2012.
- Defines the CMS approach for data preservation at different levels of complexity
  - Level 1 – publications and additional data
  - Level 2 – simplified data for outreach and education
  - Level 3 – reconstructed data and sw, doc for analysis
  - Level 4 – raw data and sw, doc for reconstruction and analysis

Goals of the policy

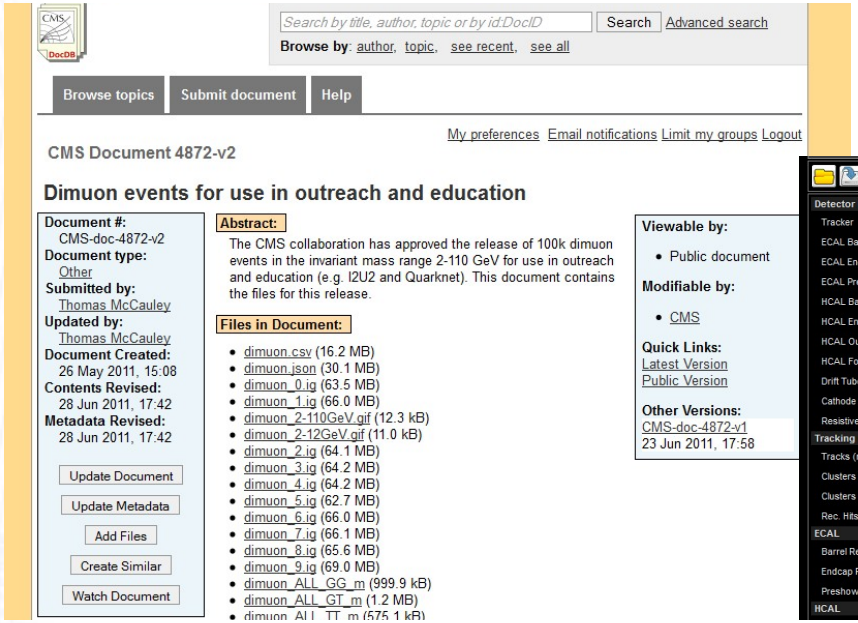
- Ensure that the data are preserved and stay usable in long-term.
- Promote the **current practices:**
  - **Level 1**
    - provide additional numerical data together with the publications
      - e.g. numerical values in tables and figures uploaded in HEPDATA and linked through INSPIRE
    - provide results in form in which they can be easily exploited by external public
      - e.g. experimental measurements in Rivet format for comparison with different event generator packages.
  - **Level 2**
    - provide selected simplified data sets for outreach and education.
- **New**: **Level 3** – Open a part of the collision data for public use
  - after a suitable delay allowing CMS to fully exploit their scientific potential.

Level 1 – Additional data

- Our concerns:
  - follow the best practices, but how to define them?
  - how to measure the added value of providing these data
    - monitor usage, count citations.
- For the moment, the additional data are
  - small in size
  - straight-forward to intepret
  - easy to produce
  - connected to a specific publication.
- See very nice examples in the next talk on HEPData.

Level 2 – Outreach data

- The current Level 2 outreach samples
  - data available from CMS-specific data base
  - usage instructions provided by outreach initiatives within and external to the CMS collaboration.



- Can we store these data outside CMS?

Level 3 data – What is different?

- The first public release of reconstructed CMS data* (part of 2010 samples and the corresponding analysis software) is foreseen during the long shutdown 2013-2014.
- Level 3 reconstructed data may not fit to the « additional data » category
  - large in size
  - requires analysis software and documentation to intepret them
  - not connected to a specific publication.
- The use of the data:
  - Data will be released under Creative Commons CC0 waiver
    - allows re-use by anyone, under the responsibility of final users.
  - Data will be identified with persistent data identifiers, and it is expected that the third parties cite the public CMS data through these identifiers.
  - It is required that the CMS collaborators will not publish independent studies based on the public data.

* In the same format which is used for final analysis in CMS

Level 3 data – access and use?

- Comparison: the current Level 2 outreach samples
  - data available from CMS, but for re-use, they need to be put in context.
- The public Level 3 data and analysis software will be the **same** as used and preserved internally.
- The tools to access the data will – at least, initially – be the **same** as already in use internally.
- The (virtualized) computing and analysis environment will be the **same** that is already in use for CMS computing internally.
- However, **to make the data usable**, we must provide an interface between the CMS world and the rest of the world and it will require a considerable effort.
  - Discussions ongoing for an externally funded pilot project on open data in Finland (teaching resources on Level 3 data) which would provide an excellent testing ground for open data release.

Outlook

- With the data policy, CMS has taken an important step towards open science
  - commitment to preserve the data at an early stage of data-taking
  - commitment to make results re-usable by a wide community
  - commitment to give open access to a part of the data.
- This brings along challenges
  - limited resources for anything on top of the physics program
  - unknown use-cases (quality and quantity).
- The implementation of the policy is now starting, and the CMS collaboration is looking forward to see the interest and benefit raised by this initiative.
- To succeed, CMS relies on external information platforms for access to its data
  - close collaboration is needed and it has already started
  - new approaches/opportunities for public reconstructed data?