CMS : Data preservation and access policy plans

Kati Lassila-Perini / Helsinki Institute of Physics
on behalf of the CMS collaboration
and the CMS Data Preservation and Access Task force

- CMS collaboration
  - Organization of the experiment
  - Analysis workflow
  - Physics motivation for data preservation
- CMS Data preservation and access policy task force
  - Mandate, schedule and goals
  - CMS approach : how to proceed.

CMS Collaboration

People

3000 scientists
In 159 institutes
In 36 countries

Data

Data recorded at CERN T0
Distributed to T1's and
accessed from T2's worldwide

Analysis

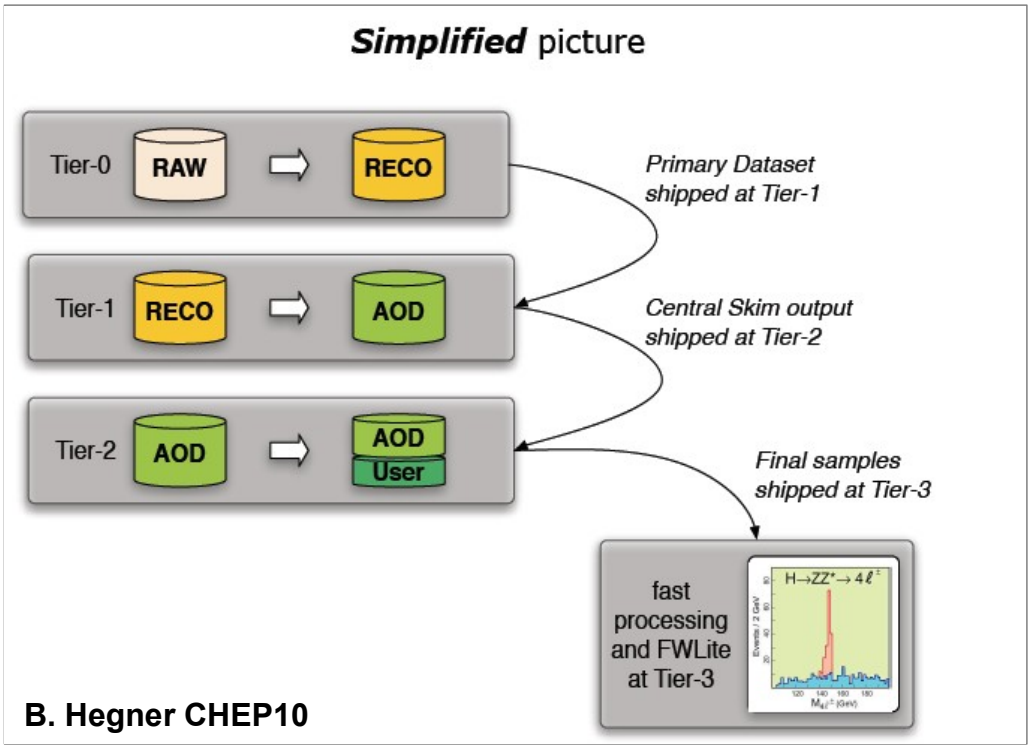
Over the grid worldwide
Using common software
and tools

CMS governance – extracts from the CMS Constitution

The CMS Collaboration Board is the governing body of the experiment and makes all major decisions within the Collaboration. CMS Institutions are represented on the CMS Collaboration Board.

- The Collaboration Board has created **a Committee to oversee the publication of CMS papers**, including Notes, and to ensure their high quality. It proposes rules and guidelines, or amendments to these, for CMS publications, which must be approved by the Collaboration Board.
- **Responsibility for producing and publishing physics results** is delegated to the Physics Coordinator and the CMS Publications Committee respectively by the Spokesperson.
- Opportunity to participate in all stages of the production and publication of CMS data results must be open to all members of the collaboration. It is the responsibility of the Spokesperson, Physics Coordinator and the Chairperson of the Publications Committee to ensure that this is so.
- For physics results there are two complementary aspects in the approval procedure: **the "physics approval"** process, steered by the Physics Coordinator and Physics Coordination, and **the "publication approval"** process, steered by the Publications Committee. The two processes are linked via the Analysis Review Committee (ARC), which is appointed jointly by the Physics Coordinator and the Chairperson of the Publications Committee.
- **All material that involves the use of CMS raw data or CMS software which is to be made available outside CMS must be approved.**
- **Publications using CMS data, by a limited list of authors from within CMS or elsewhere, may only be based on information already published by CMS.**

CMS analysis workflow



Guidelines for analysis software to facilitate the review and approval of physics results

- Use the standard data samples or skims.
- Use the standard physics object reconstruction, identification and cleaning.
- Use official tools.
- Have your code reviewable.

- See also the summary talk by Liz Sexton-Kennedy
  - http://indico.cern.ch/getFile.py/access?contribId=6&sessionId=1&resId=0&materialId=slides&confId=116485

Physics motivation for data preservation in CMS

- The CMS Collaboration is foreseen to continue for at least 20 years, given the present schedule of the LHC project.
- There is however a strong physics case to discuss data preservation now, in order to allow easy access to data collected in previous years,
  - at different centre-of-mass energy,
  - or with lower trigger thresholds.
- Examples of use of these data are
  - precision measurements with new or improved theoretical calculations,
  - cross checks for discoveries made at higher energy/higher luminosity,
  - studies related to new models of physics beyond the standard model.

CMS Data Preservation and Access Task Force

- Set by the CMS Collaboration Board:
- Motivation:
  - Need for a CMS policy on Data Preservation and Access in order to be prepared for future requests from funding agencies and general public.
- Mandate:
  - Produce **a policy and a plan**, **in coordination with CERN and other LHC experiments**, to be approved by the Collaboration.
- Timetable:
  - Intermediate status report in September (CMS week)
  - Approval in November (CB meeting).

Goals of the task force

- The goal of the task force is to review how the CMS practices confirm to the eventual requirements of other stake-holders and to prepare a policy (« what ») and a plan (« how ») in terms of the public release of the data
  - « what » : a political statement
  - « how » : a practical plan using CMS terms.
- Data preservation and access beyond anything we do now has a cost – but the consequences will be useful to the experiment
- At this point of the experiment lifetime, we are in a position to define a viable and successful policy.

CMS policy : guiding principles

- Any policy and practices defined for the data preservation and open access should be useful to CMS collaborators internally.
- Do not reinvent the wheel :
  - Valuable work done in the framework of DPHEP.
  - CERN expertise in Open Access issues.
- Use common (DPHEP) language with the other HEP experiments :
  - Important for funding agencies.
- Our preferred solution :
  - Part 1 : A general policy common to the LHC experiments
  - Part 2 : A CMS-specific resource-loaded implementation plan.

CMS Policy : How to proceed

- Elaborate the details of each data preservation model level with **current CMS practices**

  Level 1. Provide additional documentation

  Level 2. Preserve data in a simplified format

  Level 3. Preserve the analysis level software and data format

  Level 4. Preserve the reconstruction and simulation software and basic level of data

- Identify what is needed **to extend** what we are already doing to a public data access model and to assure the long-term functionality.
- Estimate the required resources.

CMS policy in concrete terms

- CMS already complies with the levels 1 & 2
  - Level 1 : Additional data attached in hepdata / inspire
  - Level 2 : Simplified data sets provided (0.5 FTE)
    - Quarknet and I2U2 : data in ig format (common project with ATLAS, ALICE)
- Levels 3 & 4 : Concrete questions for a concrete approach :
  - The LHC runs 2011 and 2012 followed by a long shutdown. Identify the requirements for the two **internal** use-cases :
    - Level 3 : You want to rerun your analysis on the 2011 reconstructed data after the long shutdown.
      - NB : all run 2011 data re-reco'ed and stored with one CMSSW version.
    - Level 4 : You want to re-reconstruct the data from the 2011-2012 run and compare to an improved MC.
      - NB : RAW data in CMS guaranteed to be stored and readable.

CMS Policy : very preliminary considerations for Level>2

- Our current practices are not incompatible with a viable data preservation model :
  - no need to change our computing paradigm but good amount of work needed.
- For each analysis, details are recorded an internal Analysis Note :
  - exact data sets and MC samples, certified runs, triggers, analysis SW version.
- The analysis software is required to be « reviewable ».
- With additional 1-2 FTE resources we could preserve all the necessary information
  - the AOD data, the connected analysis software, configurations, corrections, conditions, how-to, forward porting to the SW version of re-reco'ed data

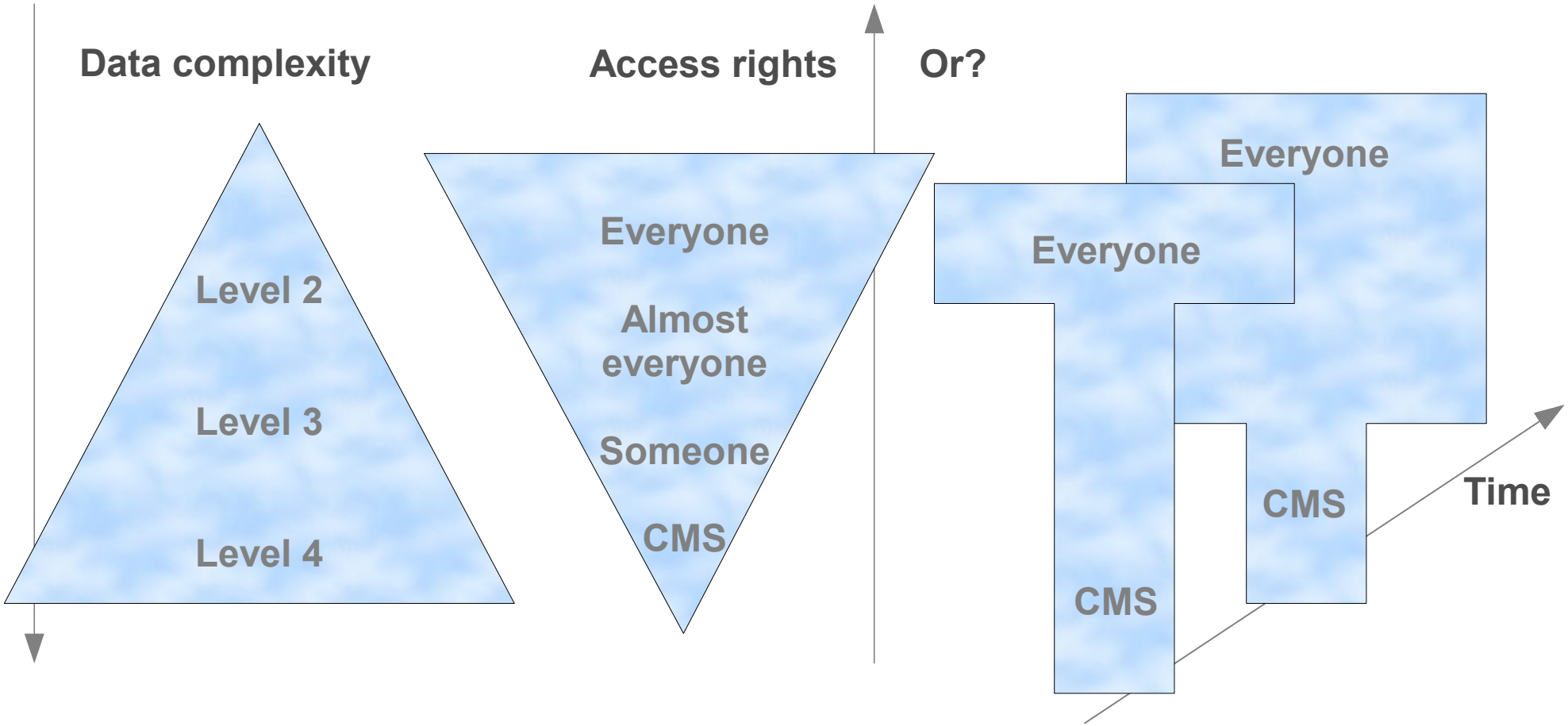
  to reproduce published physics results and to redo the analysis.
  - High time to start – CMS has 74 physics papers !
- For long-term preservation, a data archivist position is mandatory.

CMS Policy : miscellaneous considerations

- Archival support
  - CMS specific storage/archival for data, software and documentation may pose problems for the long-term preservation.
- Documentation
  - Data and software are useless if you do not know how to use them.
  - Currently, the limited CMS resources used in a way to maximize the outcome – training new users, documenting common tools.
    - We have been quite successful.
  - However, the collaborative tools (twiki) in use not optimal for version dependent documentation.
    - For data preservation purposes, we need to identify the **minimal** set of instructions and store it close to the source.

CMS Policy : Open access

We need to address the questions : **what**, **who** and **when** ?



Conclusions

- CMS will draft a Data Preservation and Open Access policy and plan by September 2011.
- The CMS and CERN managements very supportive.
- At the moment we are analysing the CMS practices and discussing how do they comply with
  1. long-term data preservation
  2. reuse
  3. open access.
- Our mandate is to prepare the policy in coordination with CERN and other LHC experiments.
  - We are open to discuss and share ideas !
- It is a great opportunity for us to join DPHEP now !