Pilot project on the public CMS data

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.
- In addition to the simplified data sets (Level 2), CMS will publish part of the reconstructed data (Level 3) and the software to analyse them (first release during LS1).
- A pilot project using these data has been included in a larger project of Finnish Ministry of Education to bring research data in open use.
- The goals of the pilot project are
  - setup and test the access mechanisms to the CMS open data
  - develop the infrastructure and interfaces for access and use
  - develop high-school level open source teaching applications, which can be used as examples for further development.

Level 3 data – access and use?

- Comparison: the current Level 2 outreach samples
  - data available from CMS – for re-use, they need to be put in context
  - large amount of work done and nice tools made available to achieve this.
- The public Level 3 data and analysis software will be the **same** as used and preserved internally.
- The tools to access the data will 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 (and in consequence), **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.
  - The pilot project on open data in Finland (teaching resources on Level 3 data) will provide resources an excellent testing ground for open data release
  - Re-use of the existing outreach tools and experience.

Fulfilling the CMS commitment on open data – which way to go?

- Easy way:
  - CMS makes AODs(*) public, provides examples of use in CMSSW(**).
- Longer way:
  - Agree on a common outreach format (e.g. physics object list as 4-vectors)
  - CMS makes AODs public with a filtering program which can extract the outreach format from AODs
  - Build first applications on public outreach format
  - More complex applications can use additional information from AODs (by configuring the filter program)
    - Foresee space for such information in the common format
- Take the longer way! Benefits:
  - Applications, infrastructure, resources general – not CMS-specific.
  - External funding.

(*) Analysis Object Data: File format used in CMS analysis
(**) CMSSW: CMS Sofware for reconstruction and analysis

Why the interest from the Ministry of Education?

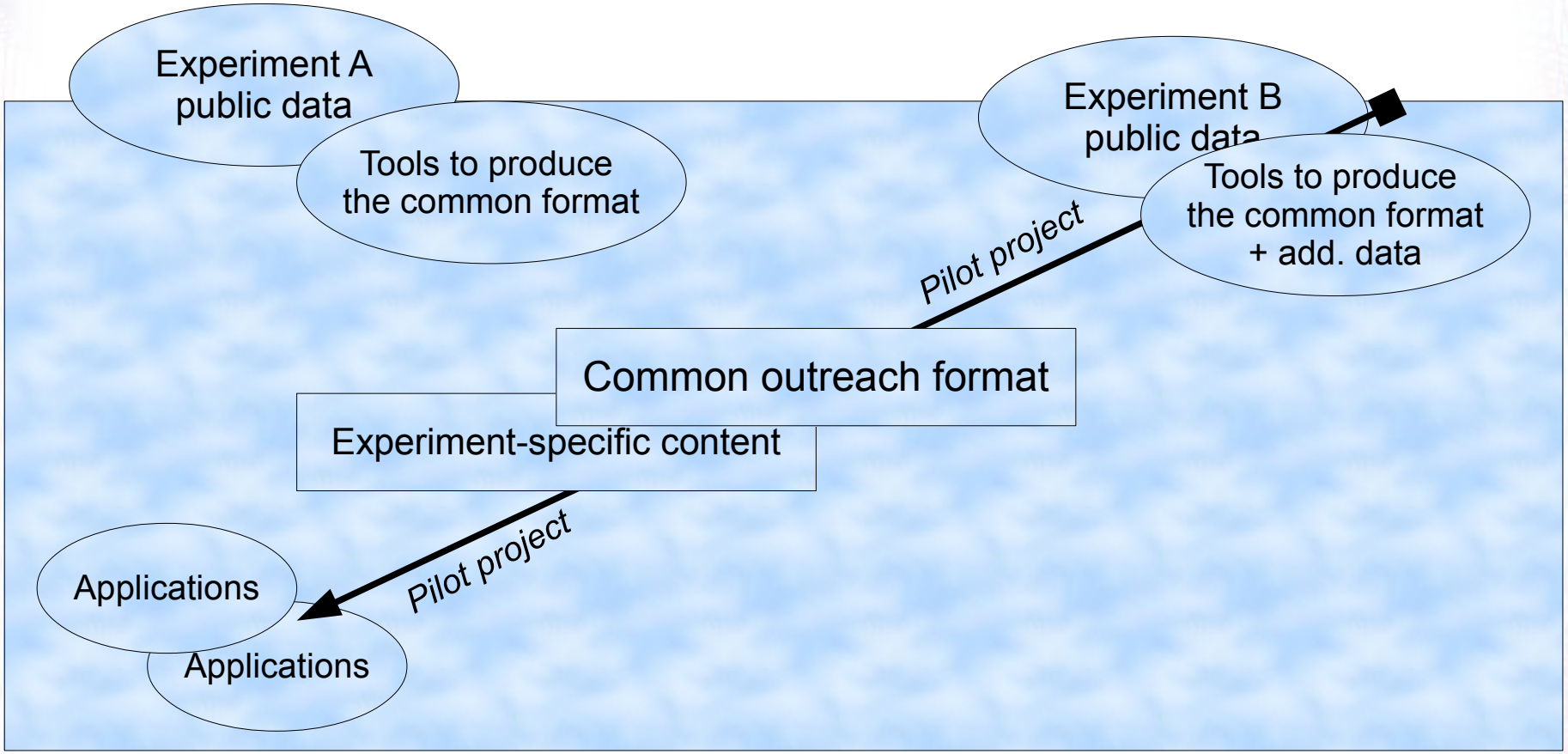
- Fits to the current trends:
  - a national plan for developing the availability and preservation of data resources to be used in research.
- Serves as a pilot:
  - gain experience on practical aspects of opening data resources
  - build general infrastructure
  - has immediate target audience in high schools
- Is attractive:
  - because of high-profile LHC data
  - can serve as an example to other science branches
  - can be used to demonstrate data analysis, statistical methods, « big science » in addition to pure particle physics
  - can attract high school students to science.

Pilot project

- **Workpackage 1: Didactics**
  - teaching applications: define contents of and pedagocical goals, survey the requirements and constraints of the target groups
    - Helsinki University – faculty of physics (physics didactics)
- **Workpackage 2: Interfaces**
  - interfaces between CMS environment (packed as a VM, provided by CMS) and the platform and storage provider (CSC(*))
  - GUI for the applications
    - CSC and Aalto University – Dept. of Media Technology
- **Workpackage 3: Data formats**
  - use and evaluate the prototype for the common HEP outreach format
- **Workpackage 4: Documentation**
  - provide adequate documentation on CMS data for external use

(*) CSC: IT Centre of Science, Finland

Simplified scheme



Considerations on the common format

- For this project, it is important to have **a** common format.
  - Applications do not exist yet, so they can adapt to any format.
- To start with, we could have
  - list of reconstructed physics objects and vertices.
- For event display purposes, the **constituents** of the physics objects are needed
  - Tracks, hits, segments, clusters...
  - Can we accomodate these experiment-specific objects in the common format (or leave a location for them)?
- For more advanced applications, we would like to retrieve additional information – e.g. **properties** of the physcis objects – from CMS AODs
  - isolation, corrections...
  - Can we foresee to include such objects (and eventually more advanced concepts like efficiencies, misidentication rates, even if we would not have them yet...) in the common format?

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).
- A pilot project on use of CMS open data in high school will test the open access chain and act as a driving force for the data preservation to happen.
- To succeed, CMS wants to act towards a common project open to all experiments
  - agree on common formats.
- The open approach combined with high-profile data attracts interest from outside CMS and enables external funding and expertise.