Some input to Open Data discussion

Reinterpretation17 workshop, Fermilab, 17.10.2017

Achim Geiser, DESY Hamburg
Jesse Thaler, MIT

subjective personal list

- issues already (partially) addressed in the presentations

- additional issues from the presenters

- additional issues from the audience

Some input to Open Data discussion

- What can be achieved with Open Data in HEP/at LHC?
  (in general, and in particular in the context of this workshop)
- What are the challenges?
- What are the specific use cases? (are there any dangers?)
- How can we (CMS) improve further? (with rather limited person power)
- How can we motivate you to try and to contribute to further improvements?
  (by feedback and/or action)
- What are the corresponding milestones?
- (How) do/can Open Data contribute to long term data preservation?
- What can be done to enhance the support for and acceptance of Open Data results in the community? (is there consensus about this goal?)
- Opinions concerning the extended vision? (~10% scientific increase with ~1% effort?)

- **Additional issues from the audience**

Backup

Feedback to community

Jet Substructure Studies with CMS Open Data

Aashish Tripathee, Wei Xue, Andrew Larkoski, Simone Marzani, Jesse Thaler

Apr 19, 2017 - 35 pages

MIT-CTP-4890
e-Print: arXiv:1704.05842 [hep-ph] | PDF

Contains section with Advice to community, Challenges, and Recommendations

Releases of 2011 CMS data+MC "exciting"
-> properly evaluate detector systematics

Conclusions: "We hope our experience motivates the LHC collaborations to further their investment in public data release and encourages the particle physics community to exploit the scientific potential of open datasets"

Some response to Challenges

General: we are doing our best with limited available person power, but it will never be perfect.

- *Scattered documentation.* Though the CMS Open Data uses an old version of CMSSW (v4.2 compared to the latest v9.0), there is still plenty of relevant documentation available online. The main challenge is that it is scattered in multiple places.

- *Lack of validation examples.* When working with public data, one would like to validate that one is doing a sensible analysis by trying to match published results. While example files were provided, none of them (to our knowledge) involved the complications present in a real analysis, such as appropriate trigger selection, jet quality criteria, and jet energy corrections. Initially, we had hoped to re-

Actions:

Improvement of Open Data web interface ongoing (with CERN scientific information service).

Prepared addendum to Open Data web instructions for research (to be merged into new web interface when ready).

Dedicated archival of relevant `old' CMS twiki pages being investigated.

Added several new public analysis and validation examples since the MIT group started their analysis, continously adding more

Some response to Challenges

- *Information overload.* The AOD files contains an incredible wealth of information, such that the majority of official CMS analyses can use the AOD format directly without requiring RAW or RECO information. While ideal for archival purposes, it is an overload of information for external users, especially because some information is effectively duplicated. The main reason we introduced the MOD file format was to restrict our access only to information that was essential for our analysis. This can be compared to the Mini-AOD format currently being developed by CMS to address a similar problem [225].

- *Presence of superfluous data.* As described on the Open Data Portal, one has to apply a cut to only select validated runs. This meant that of the initial 20 million events, only 16 million were actually usable. That said, this turns out to be a relatively small issue compared to trigger inefficiencies, which to our knowledge is not mentioned on

Actions:

- Run II data might (also) be released in Mini-AOD format in future releases.
- We encourage external users to develop strongly reduced formats specifically adapted to their needs (like many analysis groups within CMS do), maintain these themselves (person power!), and make them publicly available if relevant to others

-> e.g. MOD format by MIT group

CMS provides the original data sets. Application of quality and selection cuts for specific analyses is unavoidable. Users have to learn that ☺

But documentation how to do this efficiently is being improved. Feedback is highly welcome.

Some response to Recommendations

Action:

- *Continue to release research-grade public data.*

Yes! Release of 2012 data ongoing

- *Continue to provide a unique reference event interpretation.* A key feature of the CMS Open Data is the presence of PFCs, which provides a unique reference event interpretation with four-vector-like objects. From our experience, this seems to be the right level of information for an outside user. If

Continuing to provide Particle Flow objects ☺ (AOD and later MiniAOD). Whether they are `unique' is a matter of judgement

- *Provide validation examples.* We mentioned above the potential value of having centralized documentation about open data. Even more important than documentation, though, is having example analyses performed using open data. Explicit code helps emphasize analysis steps that might be missed by novices, including trigger selection, prescale factors, jet calibration, and luminosity extraction.

Yes! (also see challenges).

Example how to treat jet corrections has been added.
Example how to treat trigger details being added.
Detailed luminosity information has been added.

generically: make it simpler

Understandable wish but (currently) unrealistic.
Procedure is to release what exists and was used within CMS (person power!). Anybody wishing and available to make it simpler is wholeheartedly encouraged to contribute personally.