Open data and scientific reproducibility

Victoria Stodden
School of Information Sciences
University of Illinois at Urbana-Champaign

Data Science @ LHC 2015 Workshop
CERN
Nov 13, 2015

Closing Remarks: Open Data and Roundtable on Data Access

1. Goal: by 2020 all experiments who have declared they will share data, will be able to do it; and it will be discoverable (hard!)
2. why? there are several use cases
3. open data is undefined: access to closed or more open formats.
4. right now DMPs are required in the experiments

History of LEP

- when LEP was approved, card readers were still being used…
- July 14 1989 (first beams), software wasn't yet ready
- in 1992, technology was the cray, then unix machines, then the grid. cards were obsolete!
- How is important in the short term but tech changes mean we focus on the what and the why.

ALICE

- goal of data preservation is reproducibility, allow reprocessing of full chain of analysis (not public but AOD provided)
- and allow reanalysis by others
- There are 4 levels to data release: data available on 3rd party platform; 2: simplified formats made publicly available; 3: data with high levels of abstraction made available (10% after 5 years (starting now), 100% after 10 years). 4: raw data made available (only members of collaboration)

ALICE

- attribution important
- no liability
- data released with tools for analysis

ATLAS

- from management: Executive Board approved proposal to released about 1 fb-1 of the 2012 data in a limited format, along with simple tools on the Data Portal, sometime early 2016.
- goals is education and outreach, not really to carry out new science

ATLAS

- Open Data group started about a year about and used for eduction (students use data, extract a signal, etc)
- 4.10^6 events, 7GB
- most studies are monte carlo and most users dont need the full raw data.
- discussions underway for longterm preservation. also for bit-level preservation.
- documentation? data comes with tools but needs more documentation which may be an iterative process with expression of user needs. A forum can help.

CMS

- Similar 4 levels: 1: open access publications and additional numerical data; 2: simplified data for outreach and education; 3: reconstructed data and tools to analyze. 4: full raw data.
- Now: data, tools, instructions, examples.
- Challenge is knowledge preservation and meta-data, especially context at the time of data analysis.
- building open data benchmarks (high level validation code) to compare with other results later.

LHCb

- Level 1: results are public. data associated with results made available; 2: outreach/education (samples for masterclass exercises); 3: reconstructed data (50% of data 5 years after the data are collected; 100% 10 years after). 4: not permitting access to raw data because of the complexity of data processing and data size.
- Challenge data to be included in open data

BaBar

- goal: preserving raw data through computing structure for analysis.
- a wiki for real-time documentation of data usage and analysis, framework.
- data stored on tape.
- must join collaboration to access data (become a BaBarian), propose your new theory and get it approved.
- they export framework, review for simplification, and will create a data portal.

Open Data @CERN

- announcement of portal about a year ago made a big impact
- Reddit AMA
- extending code for new analysis

Recast

- Saving parameters and estimates of machine learning models. e.g. Higgs discovery.
- not enough information in the papers for reproducibility - needs to be addressed.

My Questions

- Workflows and tools to capture context during analysis.
- Links to publications
- versions (bit-level preservation?)
- ambitious plans for raw data access (except LHCb; CMS a subset of the data): driven by funding agencies and institutes, incremental approach to avoid catastrophes.. Monte Carlo produces much greater amounts of data.

My Questions

- feedback loop: how can users report bugs, or contribute anything substantial?
- presumably non-collaboration users won't have access to hardware.
- long term support? post project?
- is there any coordination across the projects? (should there be?)
- time lag between publication and 5 year embargo period - so how to link figures to raw data/software? Data DOIs in the publication? What abt snapshots of tools, with DOIs?

My Questions

- transparency in the process of creating data access? a model to other large collaborations.
- links between github and the open data portal? snapshots/versioning?
- incentives for preservation? integration of librarians into the discovery process? Could also be a model for other projects.

My Questions

- Open Data @CERN as a prototype for discovering how the public uses the data, what is most useful for longterm preservation.
- integration between Open Data @CERN tools back into the research pipeline. Connection between analysis tools and pipelines, and availability in Open Data @CERN. Comparisons of results by independent efforts (even within CERN).
- Can these pipelines be shared? for example including parameters and model fitting information. Zenodo/Open Data @CERN?