

Open Data HSF Data Analysis Meeting

Monday 3rd July

Attending: Alex Held, Nick Smith, Nicole Skidmore, Giovanni Guerrieri, Achim Geiser, Allie Hall, Carlos Usme, cburgard, Daniel Galway, Dillon Fitzgerald, Eduardo Rodrigues, Grodon Watts, James Frost, Kati Lassila-Perini, Leonardo Toffolin, Liz Sexton-Kennedy, Mason Profitt, Oksana Shadura, Pablo Saiz, Tibor Simko, Vangelis Kourlitis, Zacharias Zacharodimos, Daan Rosendal, Elise Le Boulicaut

22 attendees + 5 attendees in IT 28/R-015

Open data at ATLAS

Speaker: Giovanni Guerrieri

- Transferable expertise a goal - ML
- Set of Jupyter notebooks (eg. SWAN) to perform analyses in different frameworks (incl. RDataFrame)
- Completely offline options

Kati: Plan to release research level data?

- Yes, will have this as well as outreach (education) data
- Timescale - none so far

IT 28/R-015: How different is research and education data?

- Size has to change as physics scope is different for research data
- More data and complexity - may need a rethink

Nicole: Do SWAN notebooks offload to Lxbatch?

- On K8s cluster but yes

Dillon: Does ATLAS have its own Online portal ?

- ATLAS has own portal which links to CERN one

Open data at CMS

Speaker: Kati Lassila-Perini

- CMS Open Data is actively used for external studies - never had any problems
- Yearly Open Data workshops
- Analysis preservation by analysts is key

Nicole: Can data be held back if CMS analysis is ongoing?

- Yes done on agreement basis

Vangelis: What are the example workflows?

- Can run with
 - Docker containers with specific CMS software (for conditions etc.)
 - VM
- Examples with RDataFrame available

Dillon: Is there an Inspire label for CMS Open Data?

- Request people who use CMS Open Data cite it

Dillon: Appreciate feedback system for metadata etc.

Liz: Kati has done heroic job

- Default for collaboration board is “follow our policies”
- Holding back data is exception
- Senior members of collaboration have to write OD policies in grants etc. so must act on them
- This has allowed Analysis Grand Challenge (AGC)

Alex: Concur this is crucial for AGC

Open data at LHCb

Speaker: Dillon Fitzgerald

- MC samples on demand
- LHCb data formats not scalable for Open Data going forward
 - NTuple Wizard chosen method to give external users access to data in form of custom tuples
 - Offshoot of Analysis Productions - how LHCb members get their tuples
- No experience yet for how LHCb Open Data will be used - excited about opportunities!

Achim: Documentation and examples for ntuples from Wizard

- Wizard still in development
- For Run 1 data on Open Data portal in (m)DST format so need DaVinci - still need to include example workflows
- Plan for full analysis workflow examples

Alex: How do we control (potentially high) computing resources for Wizard

- To be discussed
- All requests ultimately submitted by an LHCb person - have final control

Achim: Will other people's tuples be available to all? Will there be standard tuples on OD Portal?

- Yes to both in principle
- Plan is for tuples to be preserved with DOI infinitum, but still need to make a formal decision

Nick: Did CERN discuss storage vs computing resources with you? As you are favouring using computing resources over storage resources with the NTuple Wizard. Trade off between the two?

- Wizard uses grid computing resources
 - CERN Open Data provides storage not computing resources
 - Might have to change
- Don't know how popular tool will be - might change considerations

Discussion

Alex:

- Common thread is person power
- Are there joint effort areas?

Dillon:

- How did you find people to work on OD?

Kati:

- Service work credits help
- Can contribute most of my time - others fractional
- OD can control their own deadlines
- It is interesting work - future ideas
- Long term commitment is an issue as the OD process is long

Alex:

- How do the credits work?

Kati:

- All in CMS must do service work and this is included along with shifts etc.
- LHCb has no credit system - purely voluntary

Dillon:

- The fact that NTuple Wizard can be used by LHCb members might help gain person power

Giovanni:

- OD work does count to credit system
- Try to make OD work useful to ATLAS work eg. ML and Git skills

- This can help gain people

Dillon:

- OD is in a WG with analysis preservation
- The two are very connected
- How are other collaborations handling the organisation?

Kati:

- Data preservation and open access group exists and reports to collaboration board and in constitution
- Common tools is a better home for analysis preservation as it has to happen "live"
- CMS requires code is reviewable for analysis preservation

Alex:

- No analysis preservation group
- Exists within physics WGs
- Analysis required to be on ReCast
- No link to OD

Liz:

- Funding calls for generic data preservation like the OD portal at CERN
- Needs to be a willingness of funding agencies to fund experimental specific initiatives

Kati:

- Funding always comes via a person's institution
- Universities should see this as a valuable contribution

Alex:

- Experiments are accountable via statements made in funding applications
- Does this have real power?

Liz:

- Depends on the country
- NSF was one of the first to support here, also Helsinki Institute of Physics
- Have not seen ripple effects of OD policy
- Hope EU funding agencies will require funding proposals to discuss data preservation policies
- ECFA panel - heading in this direction?

Additional comments:

Dillon:

- If we organize another meeting next year, can you get in touch before hand? I think it would be useful to have a meeting like this some months after the Ntuple Wizard is released, but the release date is not yet set in stone. Thank you!

DPHEP report referenced by Liz: <https://arxiv.org/abs/2302.03583>, section 3.4 discusses cost/benefit and the “data preservation miracle”