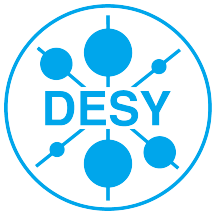
ATLAS Data Preservation

Brief summary of plans and activities of the collaboration

Roger Jones (Lancaster)
David South (DESY)

1st DPHEP Collaboration Meeting
CERN, June 9th, 2015

HELMHOLTZ ASSOCIATION



Data preservation: what does it mean?

- Since DPHEP began in 2008, the understanding of what data preservation has become clearer and the many areas of interest are better defined
- Data preservation is an active field for both funders and researchers

- ATLAS takes it very seriously; but the term can mean many different things
- For ATLAS it is important to distinguish between:
  - *Data preservation*
    - For internal use
    - For external use
  - *Data sharing*
    - For outreach
    - For research

- Learn from and collaborate with DPHEP and its members

Data preservation: some planning

- ATLAS has produced several documents in the last few years
- An ATLAS Data Preservation policy document, which outlines the **general principles of data preservation**: the data themselves, data formats and reproducibility of physics results
  https://indico.cern.ch/event/211843/contribution/12/material/0/0.pdf
- An ATLAS note outlining the requirements for preserving ATLAS data **for use by ATLAS**, ATL-SOFT-INT-2014-001
  https://cds.cern.ch/record/1697900?ln=en
- An ATLAS policy document on **data access** rules, based on the DPHEP preservation levels (next slide)
  https://indico.cern.ch/event/286440/contribution/7/material/0/0.pdf
- An ATLAS mandate for **analysis preservation,** task force now operating with conclusions expected this summer

Levels of data preservation

- ATLAS has broadly adopted the DPHEP classification of data by use case with decreasing complexity and end-user benefit

Preservation Model		Use Case	DPHEP arxiv:1205.4667
1	Provide additional documentation	Publication related info search	**Documentation**
2	Preserve the data in a simplified format	Outreach, simple training analyses	**Outreach**
3	Preserve the analysis level software and data format	Full scientific analysis, based on the existing reconstruction	**Technical Preservation Projects**
4	Preserve the reconstruction and simulation software as well as the basic level data	Retain the full potential of the experimental data	

- Preservation solutions at each level already exist, at least in part, but we are trying to make this more coherent
- The complexity comes from the supporting environment, software and tacit knowledge – preserve information, not data; data without context is meaningless

ATLAS strategy for level 4

- To keep the data live for the experiment and others, a choice
  - A final processing of the data with a fixed software/environment, maintain the latter forever
  - Periodically reprocess with new software

- The latter option is the chosen
  - Old data benefits from knew knowledge
  - Old data can be analysed with new tools
  - Avoids technology issues

The immediate challenge for ATLAS: level 4 preservation

> The data preserved has to be meaningful; from the ATLAS note earlier

1. It must be possible to reprocess the RAW data with the desired conditions and the new software version and the AOD[1] must be made available to users.
2. There must be software available to read and analyse the data AODs.
3. It must be possible to simulate newly generated Monte Carlo (MC) events with the geometry corresponding to the data.
4. It must be possible to digitize the MC events with the appropriate software to emulate the readout, pileup, beam conditions etc. corresponding to the data.
5. It must be possible to reconstruct the MC events in the same way as the data were reconstructed and write MC AODs.
6. It must be possible to determine the trigger efficiency for physics analysis.
7. it must be possible to retrieve any metadata required for physics analyses, e.g. the LHC beam conditions, ATLAS data taking and data quality conditions etc..

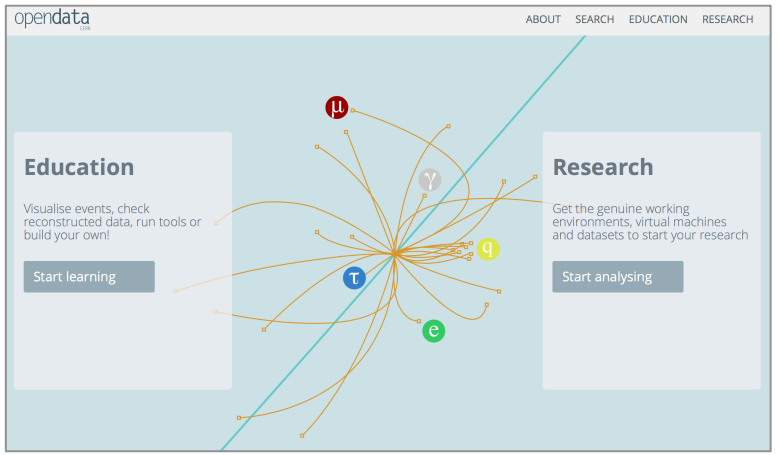
DESY

This strategy has requirements

- The RAW data must remain readable
  - You must have backward compatibility, even if you add new detectors.
  - This is difficult with some frequently changing objects, such as the trigger objects
- Reconstruction must work for old RAW data in an optimal and meaningful way
  - New software and algorithms should still work, at least nominally, with old data
  - Best-knowledge conditions need to be preserved in relevant IoVs for each year of running
- All ingredients for simulation must be available
  - New GEANT versions must be verified as describing the old detector well enough
  - Fast simulation must describe older data
  - Trigger simulation is particularly problematic, as it relies on offline software releases at the time of data taking; here old software must be used
- All of this is of course plenty of work, but the current aim is to have a coherent run-1 and run-2 dataset in 2016

Level 1, 2 data: supporting published results and outreach

- ATLAS, like all LHC experiments, has always been strong on the level 1 data
  - Subject repositories like Inspire hold the data from the paper and supplementary data supporting/augmenting the results
  - CDS holds supporting documentation
- Several level 2 outreach datasets and tools
  - 2 fb^{-1} of Higgs data (4 lepton and 2 photon modes)
- Some are now imported into the CERN open data portal http://opendata.cern.ch/



- The Kaggle Higgs challenge is an interesting case that is both outreach and also has aspects of level 3 (but is MC only) https://www.kaggle.com/c/higgs-boson
  - Recently added to the open data portal
- ATLAS is currently planning to expand our presence on the open data portal
  - Dedicated dataset(s) as well as accompanying Monte Carlo and tools to examine the data
  - It is currently not ATLAS policy to dedicate resources to release level 3 data to the public

DESY

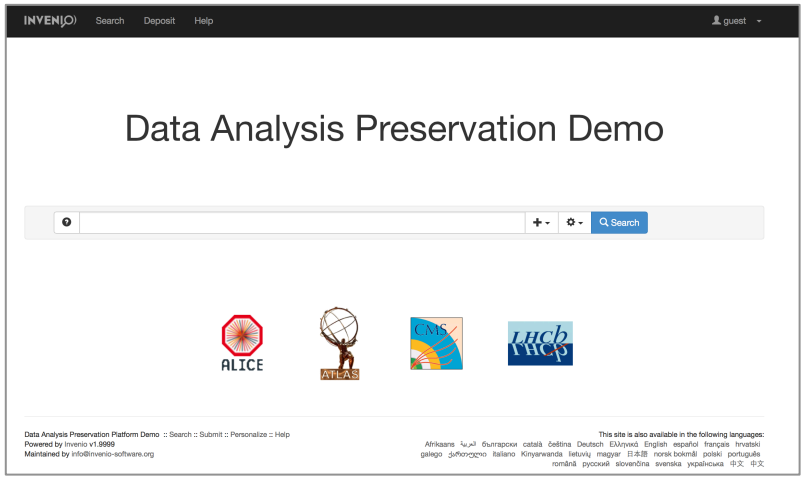
Analysis preservation: some unfortunate jargon

- An important distinction is made by the following synonyms:
- *Replicability*
  - Repeat the high level analysis procedure with new data, evolved software, calibrations etc.
  - Implies a high degree of forward-porting of tools
  - This is more towards what we called level 4 **data preservation**
- *Reproducibilty*
  - Redo an analysis with the same tools, software, data etc
  - Tools like VMs help, for a finite lifetime
  - The same results should emerge – but what required tolerance?
  - This more towards the idea of **analysis preservation**
- What is *not* primarily understood as Analysis Preservation is:
  - The full data preservation programme described in the internal note
  - The release of ATLAS data and/or software for use by non-ATLAS members for outreach purposes and involvement of ATLAS in the CERN-IT hosted open-data portal

Analysis preservation

- It is clearly desirable to be able to preserve an ATLAS analysis for the future, to fully encapsulate what was done into an easy to understand and deploy package for the collaboration
- Dedicated panel now investigating this, reporting back to ATLAS shortly
  - The case for Analysis Preservation
    - Discussion on use cases, benefits, as well as arguments presented by the funding agencies, as well as experience from other experiments
  - What is required from ATLAS to do Analysis Preservation?
    - Define the standard set of metadata and resources where it is to be harvested from,
    - User-level tools/information to be considered and when it should be done
    - The availability of data and MC files should also be addressed
  - Tools ATLAS can use for Analysis Preservation
    - The DAPF portal, benefits of ATLAS interaction with this project
    - Interaction with RECAST

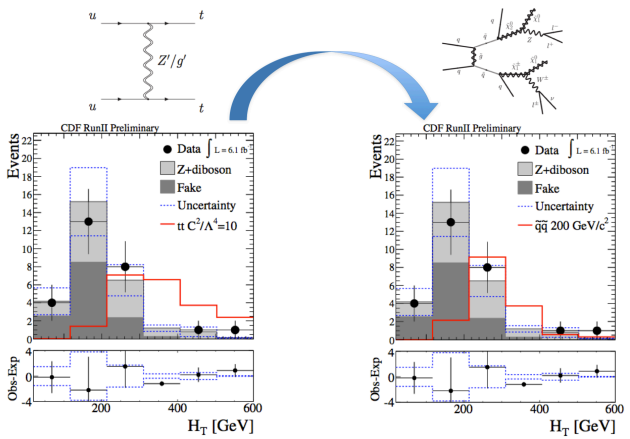
http://data-demo.cern.ch



DESY

What about RECAST?

- Recast developed by Kyle Cranmer et. al as a way of the encapsulating the full, original analysis to investigate a new model not initially considered



  - Series of analyses available on website, which collects and organises the requests and responses http://recast.perimeterinstitute.ca/?q=analyses-catalog
  - Preserves analysis information with all corrections applied
  - May be the most robust means of reuse by non-ATLAS members

- Recast is clearly applicable to Analysis Preservation
  - It provides a fully encapsulated version of the analysis as it was defined: the analysis is not re-run as such, but re-interpreted
  - Ideal for first attempt at integration into the analysis preservation framework
  - Can imagine a model where input to RECAST is harvested from the DAPF portal: use reverse to help define what must be contained within such a portal

DESY

Summary

- ATLAS must preserve its data in a meaningful way
  - This is a challenge, but we believe we are converging on a multi-faceted strategy
- Current focus for the *Data Preservation* is on forward porting run-1 to run-2 standards, following the DPHEP level 4 model
  - Is clearly difficult, for many reasons, but would create a coherent dataset and environment
- *Analysis Preservation* presents challenges - and opportunities
  - Panel set up to evaluate this concept and what it means for ATLAS
  - Close collaboration with CERN-based DAPF portal required, and we expect the recommendations of the panel to be based around this
  - RECAST also fits in nicely here
- Further outreach and open access data in preparation, proposal written
  - Dedicated data format, released alongside relevant MC, tools and exercises
  - The CERN-based open data portal is seen as the ideal host for this initiative