Lancaster University

ATLAS Data preservation

April 2015

Roger Jones for the ATLAS Collaboration

Data Preservation: What does it mean?

- *Data preservation is an active field for funders and researchers*
- ATLAS takes it very seriously; but the term can mean all things to all men
- ATLAS is clear to distinguish between:
  - Data preservation
    - For internal use
    - For external use
  - And Data sharing
    - For outreach
    - For research
- Learn from: JADE, LEP, Babar, Tevatron & the HERA
  - Build this into you model from the start

Data Preservation: Planning

Lancaster University

- *As a consequence, ATLAS has produced several documents*
- 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 policy document on **data access** rules, based on the DPHEP levels (next slide)
  https://indico.cern.ch/event/286440/contribution/7/material/0/0.pdf
- An ATLAS note outlining the requirements for preserving ATLAS data **for use by ATLAS**https://cds.cern.ch/record/1697900?ln=en, ATL-SOFT-INT-2014-001
- An ATLAS mandate for **analysis preservation**, task force currently operating

Data preservation at a high level

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

Preservation Model		Use Case	
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

Level 1 & 2 data – supporting published results and outreach

Lancaster University

- ATLAS 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
- We have many outreach datasets and tools
  - 2 fb^{-1} of Higgs data (4 lepton and 2 photon modes)
- Some are now imported into the CERN opendata portal
  - http://opendata.cern.ch/education/ATLAS?ln=en
- 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

The immediate challenge

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..

ATLAS strategy for level 4 and after

Lancaster University

- To keep the data live for the experiment and others, a choice
  1. A final processing of the data with a fixed software/environment and maintain the latter forever
  2. Periodically reprocess with new software
- The latter option is the chosen
  - Old data benefits from knew knowledge
  - Avoids technology issues
  - Old data can be analysed with new tools
- In addition, we are exploring recasting solutions, establishing where it is appropriate
  - Preserves analysis information with all corrections applied
  - May be the most robust means of reuse by non-ATLAS members

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
- The reconstruction must work for old RAW data in an optimal and meaningful way
- A best-knowledge (BK) tag of the conditions database needs to be preserved for each year of running
  - The BK tag must be migrated with technologies
  - If new software needs new conditions, it must be derivable from the older conditions or dummy
  - Downstream conditions must be derivable in an automated way

Simulation requirements

- All ingredients for simulation must be supported in the BK tag
  - New Geant versions must be verified as describing the old detector well enough
  - Fast simulation must describe older data
  - Digitization will evolve with time (e.g. effects of radiation damage) and must be appropriate to the period simulated
  - Pileup and suitable minimum bias events need to reflect the period (e.g. μ-profile)
  - Trigger simulation is particularly problematic, as it relies on offline software releases at the time of data taking; old software must be used

Analysis-level use cases: Reproducibility & Replicability

The jargon is not obvious to an English speaker, but an important distinction is captured by the following

- Reproducibility:
  - Redo an analysis with the same tools, software, data etc
  - The same results should emerge – but what required tolerance? And for what lifetime
  - The is a form of analysis preservation
  - Tools like VMs help, for a finite lifetime
- Replicability:
  - Repeat the high level analysis procedure with new data, evolved software, calibrations etc.
  - Implies a high degree of forward-porting of tools

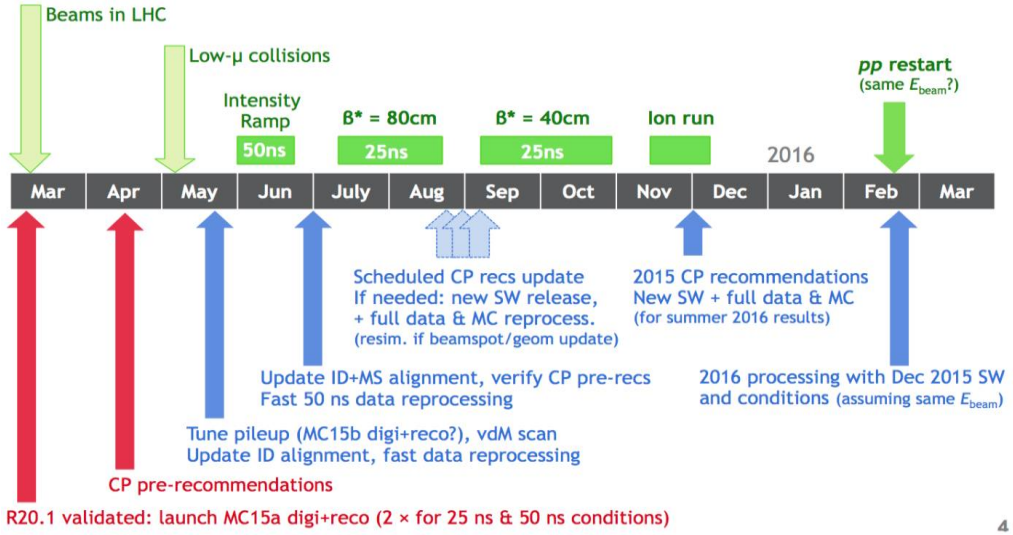
Analysis Reproducibility

Lancaster University

- Superficially simple
  - Most information is already recorded
  - Metadata in Atlas Metadata Information system, job transforms
  - Software in SVN
  - Documentation in Glance and CDS
- Practically very difficult
  - How long will a given VM system last?
  - How well can you separate from the hardware?
  - How well can every nuance be captured?
- How much is this a requirement?
  - Alluded to in funder policies, but not explicit.
  - A very useful form of documentation

Analysis Replicability

- Requires forward porting of software, tools, databases, adaptation to new data formats as discussed earlier
- For how long? Forever or until a major format change?
  - A clear division may happen, where run 1 data (e.g. AOD -vs- xAOD) and software quickly become difficult to use
  - Current schedule would reprocess the full 2015 Run 2 data in latest version at the end of the year
  - Reprocess all Run 1 in 2016
- Tools like Recast may be a better route for external reuse



Replicability – Metadata, Combined performance

- All tools reading metadata must continue to be able to read the old metadata
  - This includes in-file metadata; this is part of the RAW data readability and reconstructability requirement
- Data Quality information must be present for older data
  - Largely remains unchanged form tag to tag
  - Sometimes new software requires data features that render part of old data to change DQ status

- Combined performance groups cannot continually rerun to get recommendations for each version
  - Tools to derive them must be available, easier with new xAOD

Validation

- All levels of preservation require robust validation
- This must be made as automated and efficient as possible
- Every development of software, conditions of geometry to be validated by a central validation group

Conclusions

- ATLAS must preserve data in a meaningful way.
  - This is challenge.
- Current focus is on forward porting
- Analysis preservation presents challenges (and opportunities)
  - We are working through use cases, have trial solutions and will recommend a strategy by the summer
  - This will almost certainly involve the CERN portals under development
- This is all of potential use for Run 2