Nomenclature document update

Elizabeth Gallas, Borut Kersevan
U. Oxford, Jozef Stefan Institute

egroup: atlas-data-curation@cern.ch
Twiki: https://twiki.cern.ch/twiki/bin/view/AtlasComputing/DataCuration

CERN, 18th of September 2017

Status and plans

- Elizabeth is the custodian of the ATLAS Dataset Nomenclature document (thanks!!) :
  - https://cds.cern.ch/record/1070318?ln=en (ATL-COM-GEN-2007-003, last revision January 2016).
- The motivation for bringing this to your attention:
  - An update is needed for revising the ATLAS Dataset Nomenclature document for "naming convention to place DAOD merge datasets".
  - There are also some further items to be incorporated.
- Incremental updates only - need to keep the doc **consistent and precise!**

ATLAS Note

Report number	ATL-GEN-INT-2007-001 ; ATL-COM-GEN-2007-003
Title	**ATLAS Dataset Nomenclature**
Author(s)	Albrand, S ; Chapman, J ; Cote, D ; Fiorini, L ; Gallas, EJ ; Garonne, V ; Gwenlan, C ; Laycock, P ; Klimentov, A ; Malon, D *Show all 27 authors*
Affiliation	(CERN)
Imprint	21 Nov 2007. - 16 p.
Note	The scope of the document covers:- (1) Monte-Carlo datasets (2) Real Data Datasets. a. Primary b. Super datasets (including relational event collections) (3) User datasets (4) Group datasets (5) Conditions datasets (6) Application Internal datasets
Subject category	Detectors and Experimental Techniques
Accelerator/Facility, Experiment	CERN LHC ; ATLAS
Free keywords	Dataset ; Nomenclature ; Project ; Version ; Name ; Character Set ; Data Type
Abstract	This document describes the dataset nomenclature for ATLAS datasets. This Version 5 is the update after the Metadata Task Force Report of 2014 and subsequent clarifications with experts through 2015.

Email contact: solveig.albrand@lpsc.in2p3.fr ; elizabeth.gallas@physics.ox.ac.uk

Changes in detail

- Main points:
  - Derived ntuples/DxAOD **containers** consistently get a new label '**deriv**' (already in the document!).
  - The **production/AMI tags corresponding to merging** are dropped from the container name in all production stages.
- Merge Step. Derived datasets - Follow up from Oct 5th meeting:
  - NTUP dataset example:
    mc16_13TeV.
    301057.PowhegPythia8EvtGen_AZNLOCTEQ6L1_DYtautau_4500M5000.**merge**.NTUP_PILEUP.e3649_e5984_a875_r9364_**r9315**_p3288_**p3126**_tid12232943_00
    - production container:
      mc16_13TeV.
      301057.PowhegPythia8EvtGen_AZNLOCTEQ6L1_DYtautau_4500M5000.**merge**.NTUP_PILEUP.e3649_e5984_a875_r9364_**r9315**_p3288_**p3126**/
    - container according to new convention
      mc16_13TeV.
      301057.PowhegPythia8EvtGen_AZNLOCTEQ6L1_DYtautau_4500M5000.**deriv**.NTUP_PILEUP.e3649_a875_r9364_p3288/

- In production, no complaints that we are aware of.
- The question to be discussed could be, if we rename all the datasets as well?

6.1.3 prodStep

This is a string which gives the last production step which was used to create the data. Although this field is not needed to define the dataset, since all the production steps are encoded in the AMITag field, it renders the name more immediately understood.

Table 6-1: The currently approved list of production steps.

prodStep	Description	Input Format	Output Format	AMITag character (rule (4))
evgen	MC Event generation	None	EVNT	e
simul	MC Event simulation	EVNT	HITS	s
digit	MC Digitization	HITS	RDO, RAW	d
recon	Reconstruction	HITS, RDO, RAW	ESD, (x)AOD, TAG	r (ProdSys grid reco)
				f (Tier0 first pass reco)
				c (calibration processing)
deriv	Group Production	EVNT, ESD, (x)AOD	NTUP, (x)AOD	
merge	Merge after processing	Same as Output	Same as Input	f, r, c (noted above)
	merging from ProdSys			t
	merging at Tier 0			m
skim	RAW data skimming	RAW	RAW	
daq	RAW data acquisition	SFO	RAW	No tag (was previously o)

(1) Pre-Run 2 prodStep values are described in older versions of this document. This document reflects the proposed values for future datasets.

- The prodStep = 'merge' is being generally integrated into steps with more meaningful names but it is still found for some final data formats.

Comments from DataPrep

- Jonas Strandberg in Feb. 2017 on CDS:
- **Conceptual:**
  - Page 7, final bullet point: Do we have any capability in principle to call datasets e.g. "mc15a_13TeV" if we run several campaigns with the same hits?
  - Or is it impossible to change the project tag (for good reasons) for downstream tasks?
- **Technical/formal:**
  - The document states that "The last two characters (numeric) denote ...", is it a "must" or a "should" that the last two digits are numeric?
  - Page 8, first bullet point: When we have had non-integer collision energy, we have used e.g. “2p76TeV” since the "." is not allowed. This could be stated explicitly here.
  - Page 11, item #2: Should we change "physics coordinators" to "data preparation coordinators”?
  - Page 13, first example: It would be nicer to have an example of a run-2 dataset name rather than a run-1 example (NTUP_COMMON => DAOD_SUSY3 or sth).
  - Table 8-2, physicsShort name: It would be good to mention PMG, as they are responsible for this.
  - Table 8-2, dataType name: "group production" -> "derivation production coordinator”.

Anything else?

- Changing these documents is not something monumental but it is still a somewhat involved procedure, getting the agreement from all involved parties and OAB.
- So preferably we should collect changes periodically and incorporate them.
- Thus, if you have comments, suggestions etc, now is the time!
  - Please contact Elizabeth and myself …