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, 8th of March 2018

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 to bring it up to date.
  - 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

Container name change

- Change of **container names** (in effect, making additional ones)**:**
  - 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/

- Successfully implemented and in production, no complaints that we are aware of.

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.

AMI tag chain length

- In the nomenclature document (and ADC tools) these are certain hard limits set:
  - the total length of a dataset is limited to 255 characters,
  - the total length of a container is 150 characters for an official dataset.
- These values are used in all the production tools:
  - e.g. Rucio has a hard limit of 255 characters for both the datasets and containers.
- **The AMI tag chain rule has been updated to 50 in the document draft:**
  - AMI tag value limit increased to 5 digits (i.e. < 100000), e.g. **r10003** to match the reality.
- In theory we violate unitarity (our own rules) with this:
  - If we go for > 10000 tag IDs, then we have 6 digits/tag, with 7 underscores and 8 steps gives: "**e25340_e45984_s33126_s34136_r59781_r96778_p35384_p35385**"
  - Or **55 chars for the AMI tag chain** - which could be the next limit, but at some point we will need to stop.
    - Rucio has a **hard limit of 50 characters for pure AMI chain (i.e. without the tid and sub part)**. Going above puts later stage processing at higher risk, which we would like to avoid as much as possible.
    - It however can be done in if needed (but carefully).
  - **A dedicated meeting held on this topic…**

Discussing the solution

- **To keep the 50 char limit we can:**
- **Rename** the datasets in addition to containers, i.e. drop the merging tags from dataset names or other modification. e.g.
  - from:
    “mc16_13TeV.
    364194.Sherpa_221_NNPDF30NNLO_Wtaunu_MAXHTPTV280_500_CFilterBVeto.merge.HITS
    .e5340_e5984_s3126_s3136”
  - to:
    “mc16_13TeV.
    364194.Sherpa_221_NNPDF30NNLO_Wtaunu_MAXHTPTV280_500_CFilterBVeto.merge.HITS
    .e5340_M_s3126_M”
    - where the “_M” is the merging label that.
    - The provenance chain and related job config is available in AMI and ProdSys, if needed.
- **Enforce** the 4 digit limit in the (some) tag values and **change letters** as needed (rXXXX -> zXXXX).
- **Something else?**

Further steps for AMI tag field length

- **Conclusions of the discussion:**
  - The currently implemented length of the 'AMI tag chain' field of 50 in our production tools (in the last version of document it is still written as 32) should suffice for the duration of the current MC16 campaign.
  - In case an urgent intervention is needed because this length needs to be extended, this can be changed in the DB for all the required components.
    - In practice, this could then violate the total dataset length but ProdSys can catch this at the time of task definitions if needed - the balancing would then probably mean a reduction of the 'physics_short' field.
- One should look for a **clever** way to optimize the dataset name length to be ready in the next years for the next major campaign (like mc18).
  - Alexei has kindly agreed to organize a proposal together with other experts on how one can do this in practice in ProdSys, Rucio, AMI etc in an optimal way.

Special PanDA datasets

- Included for documentation completeness:
  - … so nobody forgets that "_sub0473840848" is there for a reason…

- These datasets are used by PanDA for data motion and internal bookkeeping. They have a short lifetime (typically 2 weeks) and are made with the hidden metadata in order not to be exposed to normal users.
- Their naming convention is as follows:
  p**anda.<task ID : 8 bytes>.<timestamp : 5 bytes>.<data type>.<random string : 36 bytes>_dis<unique number : 10 bytes>**
  in the panda scope for dispatch datasets, and
  <tid dataset name>_sub<unique number: 8 bytes >
  in the scope of the tid dataset for sub datasets. The lengths of the fields are also indicated.
- The <tid dataset name> denotes the name of the original dataset and the <data type> follows the standard naming convention of ATLAS data type,

Optimizing/formalizing the “Physics Short”

- The general format of the MC dataset is at present:
  MC15.<dsid>.**<physics_short>**
  - with <physics_short> being defined by:
    <generators>_<tune+pdf>_<process>

    mc15_13TeV.361602.PowhegPy8EG_CT10nloME_AZNLOCTEQ6L1_WZlvvv_mll4.evgen.EVNT.e4054

  - There is a long history of this field overrunning the database constraints in production tools (**currently at 60 characters**):
    - It is a very useful field for physicists looking for basic “metadata” of a dataset.
      - It however should not describe everything about the process (there is AMI…).
      - **In the past the overruns were always accommodated by the ADC experts.**
        - With the increase of the dataset lengths, we are now hitting a limit.
- **DCC has asked the PMG conveners to prepare a short formal description for the nomenclature document update.**
  - What we currently have in terms of the physics short format requirements is defined and written up by the PMG:
  - https://twiki.cern.ch/twiki/bin/view/AtlasProtected/AtlasMcProductionMC15#Dataset_JO_input_naming_scheme
  - **The draft of the proposal has been written by PMG conveners here:**
  - https://docs.google.com/document/d/1-_F7DDM76YxIxG_sLfJ8wRAI8Nn-wLK_KQSVVOam8jw/edit

  - The changes, once approved in the required forums, will be implemented for the next MC campaign.

Comments from DataPrep

- Jonas Strandberg in Feb. 2017 on CDS, implemented in the draft by Elizabeth:
- **Conceptual question:**
  - 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”.

Next steps

- Changing these documents is not something monumental but it is still a somewhat involved procedure, getting the agreement from all involved parties and OAB.
  - Some important points are being updated and detailed in the new version now.
    - We will circulate it once we have all the components in.
- Thus, if you have further comments, suggestions etc, now is the time!
  - Please contact Elizabeth and myself …