Data Preparation Related Tools

B. Heinemann

UC Berkeley and
Lawrence Berkeley National Laboratory

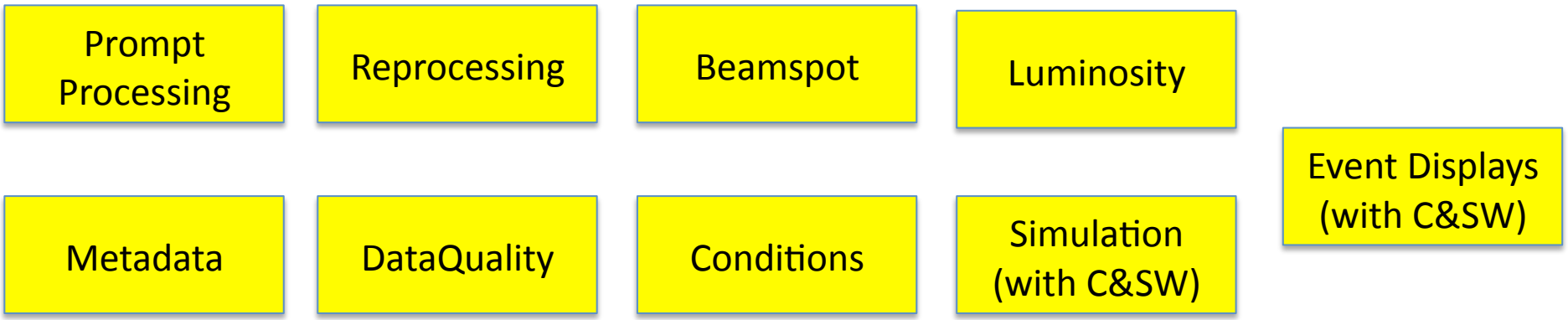
SCSWT Workshop, Tbilisi/Georgia, Oct. 2010

Outline

- Introduction
- Data Periods and Physics Containers
- AtlRunQuery and DataSummary
- GoodRunList, LumiCalc and AMI
  - See also talks by E. Gallas and S. Albrang
- Summary

Data Preparation

- https://twiki.cern.ch/twiki/bin/viewauth/Atlas/DataPreparation
- Coordinators: Jamie Boyd and BH



- General email list for information: atlas-data-preparation@cern.ch
  - Open subscriptio (~500 people are subscribed)
- Help mailing list: atlas-data-preparation-help@cern.ch
  - Contains all the responsibles (~20 people) and should always give an answer

Data Periods and Physics Containers

- DataPeriods: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/DataPeriods
  - Major period indicated by letter (e.g. period "data10_7TeV.periodF")
    - Minor sub-periods have number: F1, F2,…
    - Constraint: F=F1+F2
    - Otherwise periods do not overlap in content, i.e. the same run is *never* included in F and G (or in F1 and F2) but always in F and F1
  - Creation of periods takes into account varying detector conditions, conference schedules etc.
    - Decided by hand (DP and DQ conveners with input from Run Coordination and Systems)
- Physics Containers: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/PhysicsContainers
  - One physics container per dataset and per period, e.g. `data10_7TeV.periodF.physics_Egamma.ESD.PhysCont.t0pro04_v01`
  - Purely created for convenience of users
  - The content of a PhysicsContainer corresponds to a given version of the reconstruction software
  - Only contain runs that are potentially useful for physics analyses
  - Also displayed in AMI
  - Usually the starting point of analysis nowadays

DataPeriods

Period	Subperiod	Run Range	Comment	Luminosity (nb^{-1})	GRL
A	n.a.	152166-153200	• unsqueezed stable beam data (beta*=10m): typical beam spot width in x and y is 50-60 microns.	0.4	DetStatus-v03-repro04-01
B	B1-B2	153565-155160	• first squeezed stable beams (beta*=2m): typical beam spot width in x and y is 30-40 microns. • this period is further subdivided into 2 sub-periods: B1-B2	9	DetStatus-v03-repro04-01
C	C1	155228-155697	• very similar to period B but higher luminosity due to more bunches in machine	9.5	DetStatus-v03-pass1-analysis-2010C
	C2	156682	• different configuration for L1Calo in eta region near crack at 1.45: previously L1Calo was sampling only from the Endcap while now it is sampling from the barrel in this difficult region. See explanation by Trigger Coordination for details.		
			• • •		
H	H1	166466-166850	• 233 colliding bunches in ATLAS • the CSC problem that affected the last two runs in periodG6 got fixed • run 166850 was the run where a Van-der-Meer scan was done at the CMS, LHCb and ALICE experiments and ATLAS ran with a rather special trigger menu. The luminosity of this run is only 7.5 nb^{-1} and a large bandwidth was given to a special high-multiplicity trigger.	6500	DetStatus-v03-pass1-analysis-2010H
	H2	166856-166964	• new RPC timing calibration, see this elog entry	2800	not yet
I	I1	167575-?	• 295 colliding bunches in ATLAS • the RPC window was unstretched in the L1 trigger, i.e. now only one BC contributes to L1 decisions	>1000	not yet

Which Data to use with which MC

- https://twiki.cern.ch/twiki/bin/view/AtlasProtected/DataMCForAnalysis

Name	Datasets	Release	AMI tag for data	Physics Container	AMI tag for MC
May2010 reprocessing ("repro04")	As April2010, and adding 2010 7 TeV data up to run 155160	15.6.9.8	r1297 (runs 141748-154817) and r1299 (runs 154822-155160), change of DB release only between the two 'r' tags	repro04_v01	**ESD** (if requested): r1302 (default setup) r1303 (particle ID and cal. hits in ESD for Jet/Etmiss samples) r1304 (GEO-10-01-00 - extra material - sample reconstruction) r1305 (samples without the EM constant term for e/gamma group) r1307 (GEO-10-00-00, no track slimming, for MB) r1336 (GEO-10-00-00, MC_lumi1E31_simpleL1Calib_physics_prescale for some Exotics samples) r1312 (GEO-10-02-00 in digi and GEO-10-00-00 in reco)

RunQuery and DataSummary

- AtlRunQuery (http://atlas-runquery.cern.ch/)
  - A lot of different kinds of information about a small subset of runs
    - Start/end, bunch scheme, trigger info, luminosity, stream rates, dataset sizes, DataQuality information, pileup,…
    - Takes a very long time when e.g. using all 2010 collision runs (or times out)
- Atlas-DataSummary (http://atlas-datasummary.cern.ch)
  - Static information plotting many variables useful for understanding conditions in single run
    - Luminosity, backgrounds, beam currents, bunch scheme, data-taking efficiency…
- Neither is (or should be) used directly for physics analysis but both are useful to see what went on in particular run
  - Mostly used by experts to understand features of given run

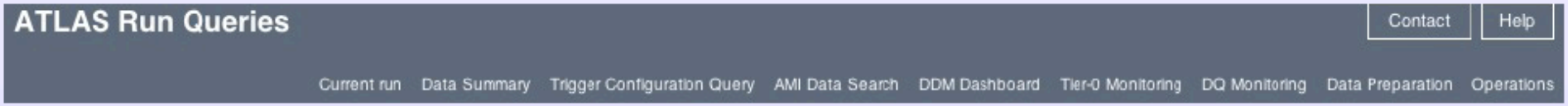
AtlRunQuery

by A. Hoecker and B. Stelzer

(slides taken from J. Walder)

Accessing the Run Query Tool

- The Run Query Tool is written in python,
  - Can be access in two main ways:
  - Simple (yet powerful web interface)
    - http://atlas-runquery.cern.ch/
  - Allows interactive search queries



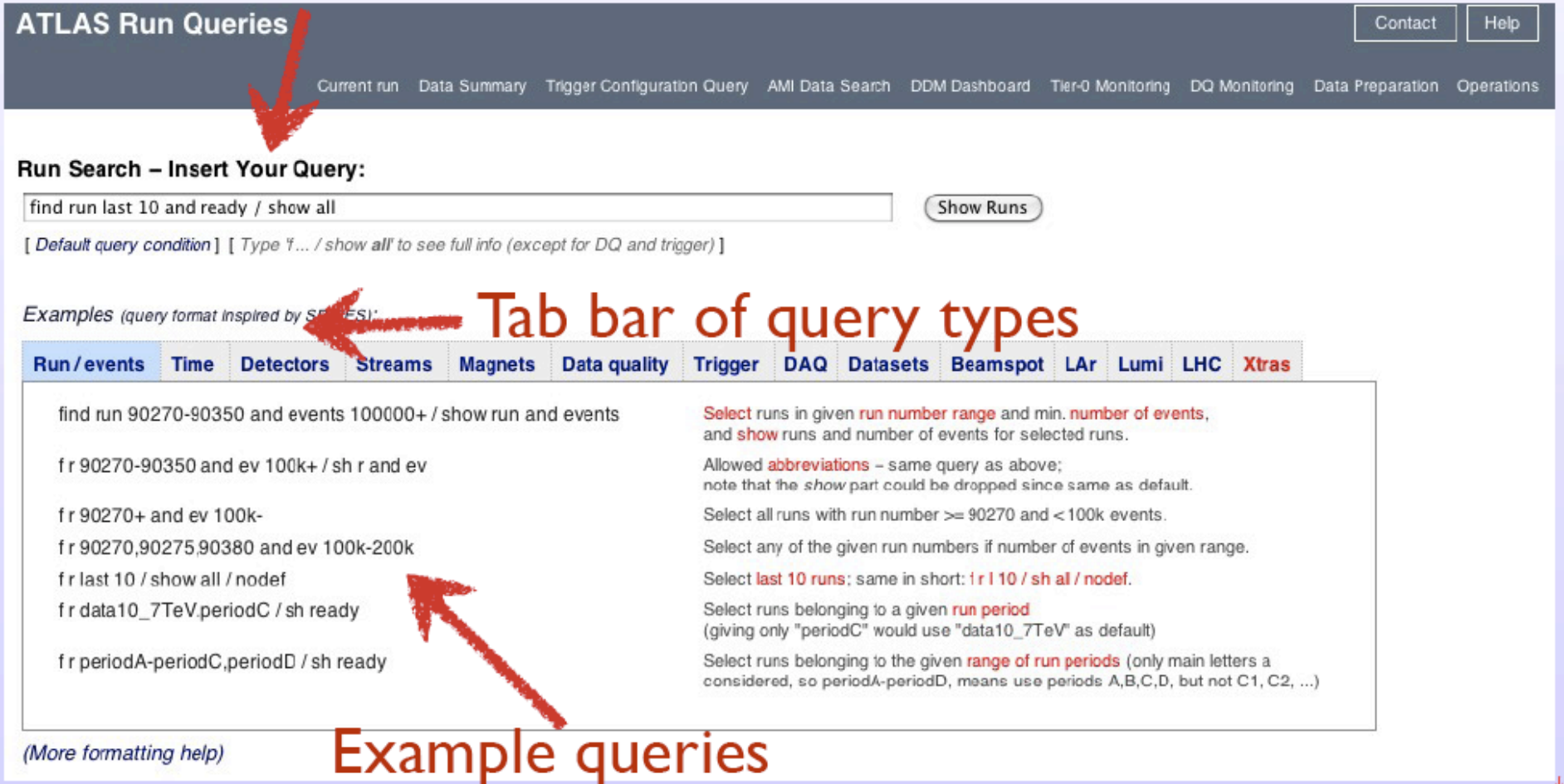
  - Command line python (need to setup recent Athena release).
    - AtlRunQuery.py '<arguments>'
    - Enter the same search criteria as in the web interface.

Web based queries

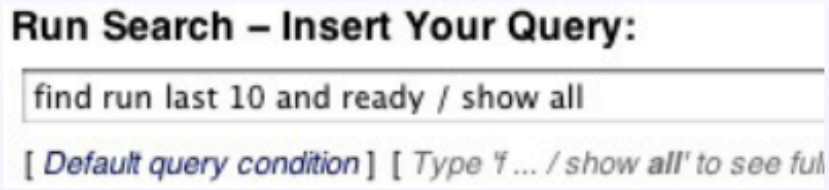
http://atlas-runquery.cern.ch/

Help

Query Box



Query Structure



- Query box designed to be familiar to Spires method of text input
- Query is separated into a matching part and display part
  - 'find' / 'display'

- Query words can be shorted when unambiguous.
- Eg. find run and events -> f r and ev

Example Queries – Physics data

- Find runs where stablebeams have been set
  - find run last 10 and lhc stablebeams t / show all

Detailed links

						LHC and online luminosity information			
Run	Links	#LB	Start and endtime (CEST)	#Events	Ready for physics	LHC Fill	Stable beams	Beam energy and intensities	Online del. Luminosity [ATLAS_PREFERRED]
166964 (in calib loop)	DS, RS, BS, AMI, DQ, ELOG, DCS:SoR/ EoR	487 (117 s)	Sun Oct 17 2010 18:15:17 – Mon Oct 18, 10:06:32	1,220,426 (21.4 Hz)	LB 1-318: 0 LB 319-333: 1 LB 334-487: 0	LB 1-156: 1429 LB 157-370: 1430 LB 371-487: 1431	LB 1-310: FALSE LB 311-336: TRUE LB 337-487: FALSE	Maximum intensities: Beam 1: 342 e11 protons Beam 2: 337 e11 protons Maximum beam energy: 3500 GeV	Entire run: 365.7 nb^{-1} Stable beams: 320.2 nb^{-1} Peak lumi: 1.5e+02 e30 $cm^{-2}s^{-1}$ Peak <μ>: 3.21 Approx. lifetime: 14 h
166927 Period: H, H2 (in calib loop)	DS, RS, BS, AMI, DQ, ELOG, DCS:SoR/ EoR	131 (112 s)	Sun Oct 17 2010 06:11:56 – 10:17:36	2,850,495 (193.4 Hz)	LB 1-6: 0 LB 7-104: 1 LB 105-131: 0	1427	LB 1-109: TRUE LB 110-131: FALSE	Maximum intensities: Beam 1: 237 e11 protons Beam 2: 235 e11 protons Maximum beam energy: 3500 GeV	Entire run: 625.5 nb^{-1} Stable beams: 625.5 nb^{-1} Peak lumi: 57 e30 $cm^{-2}s^{-1}$ Peak <μ>: 1.23 Approx. lifetime: 23 h
166925 Period: H, H2	DS, RS, BS, AMI, DQ, NEMO, ELOG, DCS:SoR/ EoR	35 (114 s)	Sun Oct 17 2010 04:15:31 – 05:22:15	765,755 (191.3 Hz)	LB 1-4: 0 LB 5-35: 1	1427	TRUE	Maximum intensities: Beam 1: 241 e11 protons Beam 2: 241 e11 protons Maximum beam energy: 3500 GeV	Entire run: 201.2 nb^{-1} Stable beams: 201.2 nb^{-1} Peak lumi: 62 e30 $cm^{-2}s^{-1}$ Peak <μ>: 1.33 Approx. lifetime: 22 h

- For physics, require "Ready for Physics" flag
  - find run last 10 and ready / show all

Displaying Data Quality

- find run last 5 and ready / **show dq**

Good Lumi Blocks

Run	Links	#LB	Start and endtime (CEST)	#Events	PIXB (SHIFTOFL)	PIX0 (SHIFTOFL)	PIXEA (SHIFTOFL)	PIXE (SHIFTOFL)	SCTB (SHIFTOFL)	SCTEA (SHIFTOFL)	SCTEC (SHIFTOFL)
166964 (in calib loop)	DS, RS, BS, AMI, DQ, ELOG, DCS:SoR/ EoR	487 (117 s)	Sun Oct 17 2010 18:15:17 – Mon Oct 18, 10:06:32	1,220,426 (21.4 Hz)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
166927 Period: H, H2 (in calib loop)	DS, RS, BS, AMI, DQ, ELOG, DCS:SoR/ EoR	131 (112 s)	Sun Oct 17 2010 06:11:56 – 10:17:36	2,850,495 (193.4 Hz)	G LB 1–6: R LB 105–131: R	G LB 1–6: R LB 105–131: R	G LB 1–6: R LB 105–131: R	G LB 1–6: R LB 105–131: R	G LB 106–131: R	G LB 106–131: R	G LB 106–131: R

Bad Lumi Blocks

- If you require the pixel detector to be in the run
  - find run 166927 and det pix / show dq
- And if you want periods with 'good' pixel data quality (say yellow or green).
- find run 166927 and dq pix y+ / show dq
  - Other systems pix, sct, csc, lar

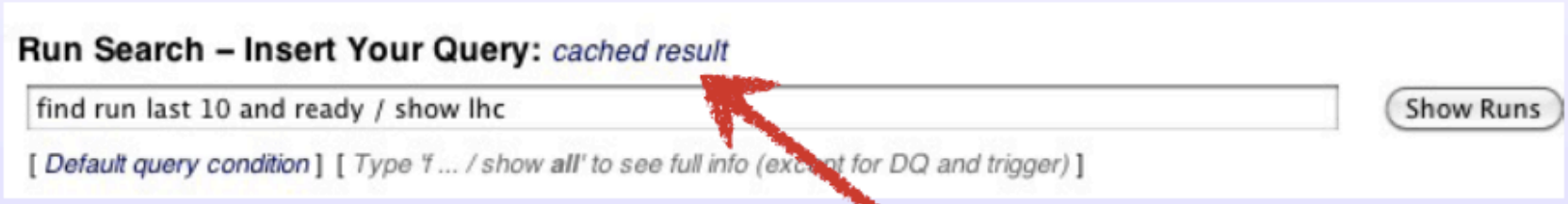
Other Queries

Query: *find run 90270-90350 and events 100000+ / show …*

f r 89465-93000 and duration 2m+	runs that had a duration of at least two minutes
f r 90270+ and ev 100k-	runs with run number >= 90270 and < 100k events
f r 90100-90150 and det any sct	runs with at least one SCT in and show all participating detectors
f r 90270-90350 and st *IDCos* 10k+ and st *RPC*	runs where sum of *IDCos* streams has more than 10000 events and one *RPC* stream is enabled
f r 90270-90350 and mag s and not mag t	runs with solenoid on and toroid off
f r 90270-91350 and dq lar g	runs where all LAr-EM are "green"
f r 90270-90350 and dq em y+ and dq pixb y+	runs where LAr-EM and Tile DQ flags are at least "Y" and Pixel-B is "G"
f r 90270-90350 and dq any pix n.a.	runs where at least one Pix qualitity flag is unset
f t 10.9.2008-13.9.2008 and ptag data08_1beam*	runs with single-beam project tag, between 10th and 13th Sep 08
f r 91890-92070 and tr EF_e5*	runs with triggers matching pattern EF_e5* [not case sensitive]

Result Caching

- Some searches require queries across many systems and runs, first query may take some time and place demands on the server.
  - Each query can now be cached.
    - Send the cached query in email links etc..



Cached link

http://atlas-runquery.cern.ch/query.py?q=arq_101018192855pznq

- Cache will remain in database for as long as it is 'popular'

Some less well-known favorites of mine

- `find run XXX / show lhc all`
  - Detailed bunch structure
- `find run XXX / show da`
  - Details of all datasets
- `pileup 1.5E32 71.5 295`
  - Pileup for L=1.5E32 $cm^{-2}s^{-1}$, xsec=71.5 mb, 295 coliding bunches
  - Result is 3.2 for this particular query
- `find run XXX / show bs`
  - Shows beamspot positions and width as stored in COOL versus the luminosity block number

GoodRunList Maker, LumiCalc, AMI

- GoodRunList: https://atlasdqm.cern.ch/grl/
  - Critical for all physics and performance analyses
  - Should only be made with above webpage
    - AtlRunQuery can also make them but this is not supported by the DQ group (integrity cannot be guaranteed)
- Luminosity Calculation: https://atlas-datasummary.cern.ch/lumicalc/
  - Inputs are GRL and trigger
  - Should be used for any luminosity calculation
- AMI
  - Primary source of datasets, e.g. only reliable source of the number of events in AOD for given dataset
  - Should be used by physics analyses to cross-check events in ntuple
  - Provides a lot more of course (see talk by S. Albrand)

GoodRunList

By M. Baak, P. Onyisi

(slides from J. Walder)

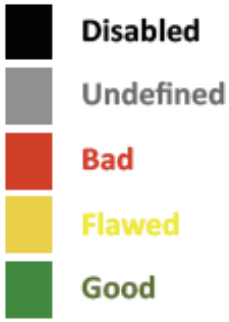
What are Good Runs Lists?

- For physics measurements, must be confident that the data we use satisfies our quality constraints.
- Number of ways data may be classed as 'bad':
  - LHC not in stable-beam (or ready for physics) mode.
  - Magnets Off or ramping
  - Sub-detectors switched off
  - Too many noisy cells.
- Provides granularity down to the luminosity block level (~2min).
- There needs to be a way to telling your analysis which sets of data to exclude:
  - The Good Runs List (GRL)
- In it's simplest form, the GRL is an xml file which specifies sets of 'good' luminosity blocks, over which your data should run.
- Can also contain blacklists, to explicitly exclude some sections (may be useful in using exclusive sets of data).
- By creating the xml file and adding it to Athena's job options, Luminosity information can be kept intact and correctly calculated.

Data Quality Flags

- Traffic light system used as indicator of quality of flag:
  - DQ traffic light system explained:

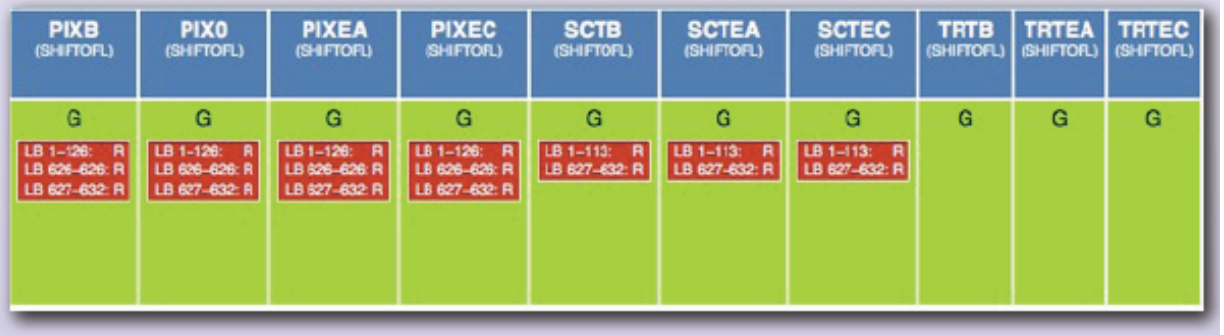
• Black (-1):	disabled	Subdetector is disabled
• Grey (0):	undefined	Very short runs, or problems DQ monitoring. Decided upon later.
• Red: (1)	bad	Not good for physics
• Yellow (2):	flawed	Use with caution. Decided upon later.
• Green (3):	good	Good for physics use.



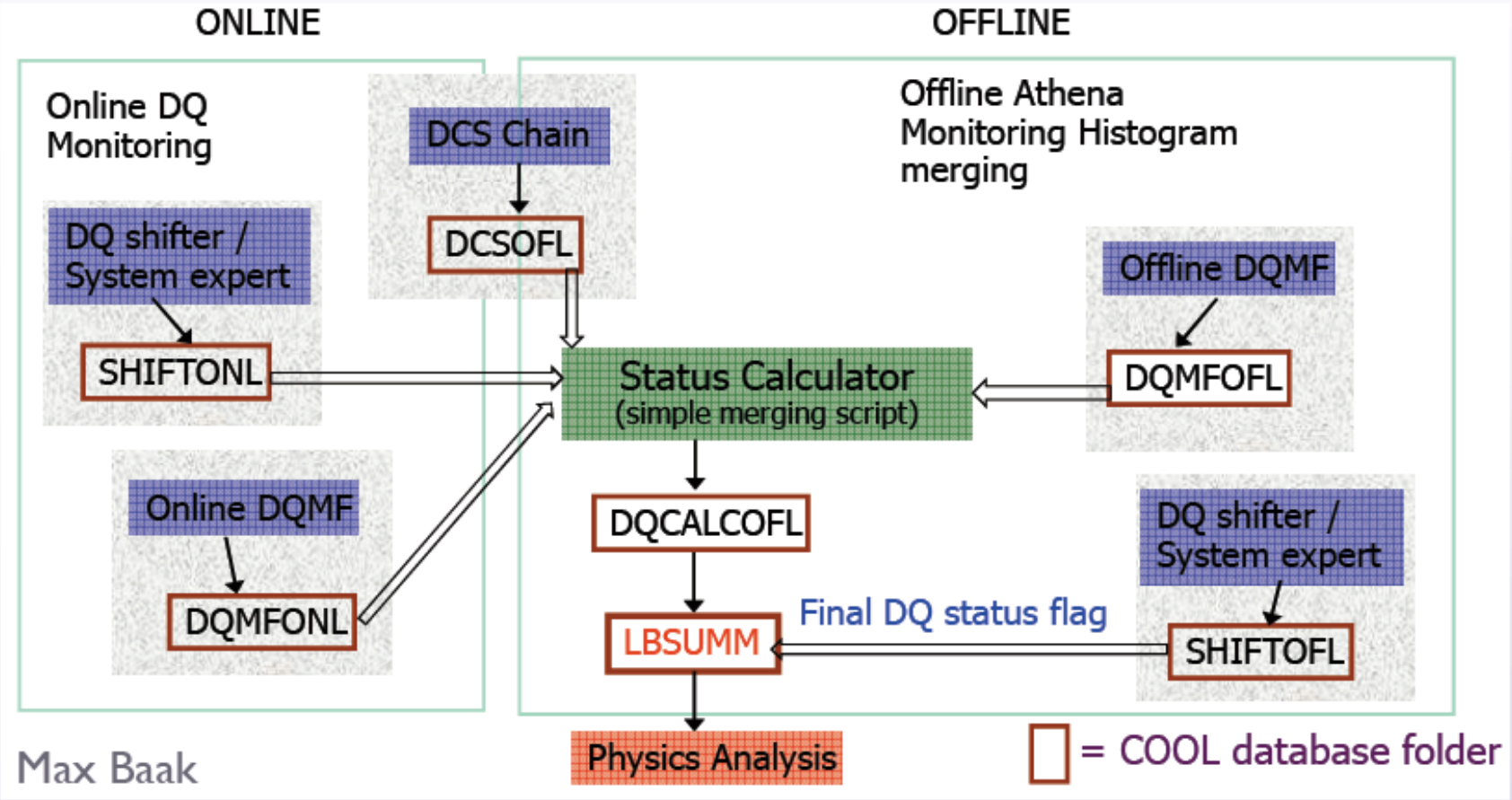
  - Order: black < grey < red < yellow < green

Max Baak, CERN

- Over 100+ flags exist:
  - Sub-detectors (pixel, SCT)
  - Trigger (L1 dead time)
  - Combined performance (eg. electrons, jets, ...)
  - Global flags.



Setting of DQ Flags

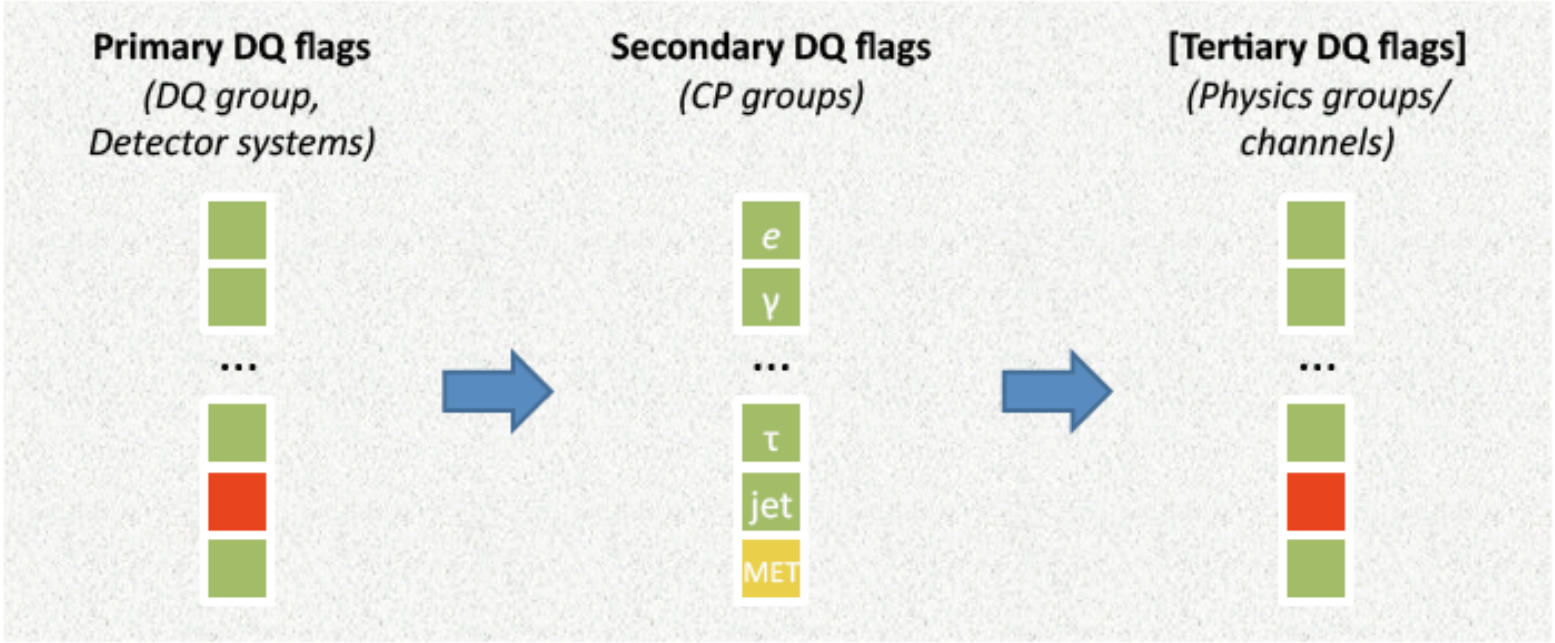


- Data Quality (DQ) flags are set mainly by a team of shifters.
  - Ultimately aim for automated flagging of data quality, where possible.

Data Quality Flags (cont.)

- A DQ hierarchy is in the process of being constructed,

  - DQ flags hierarchy consists of three levels:



  - Yellow and grey flags indicate that work (eg. calibration) is still ongoing. Eventually no yellow or grey flags.
  - Yellow flags allowed. Usage depends on physics analysis.
  - In process of being finalized.
  - No yellow flags for physics analysis!
  - (Tertiary flags not yet defined.)

Example GRL

- Common run/LB xml forma agreed-upon for good run list file.
  - Design by TAG grp
- Simple and transparent structure
  - Ranges of LBs
- Also contains versioning and metadata info
  - Name, version nr.
  - Search query used to create xml file.
  - Streams, trigger names of interest (work in progress)

```
<?xml version="1.0" ?>
<!DOCTYPE LumiRangeCollection
  SYSTEM 'http://atlas-runquery.cern.ch/LumiRangeCollection.dtd'>
<!-- Good-runs-list created by AtlRunQuery.py on 2010-07-07 14:34:52.653539 -->
<LumiRangeCollection>
   <NamedLumiRange>
      <Name>muon_7TeV</Name>
      <Version>2.1</Version>
      <Metadata Name="Query">find run 158045-158392 and partition ATLAS and db DATA and
      lhc stablebeams T and st physics_MinBias and ptag data10_7TeV and
      dq ATLGL LBSUMM#DetStatus-v03-pass1-analysis-2010D g
      and dq atltor LBSUMM#DetStatus-v03-pass1-analysis-2010D g
      and dq atlsol LBSUMM#DetStatus-v03-pass1-analysis-2010D g
      and dq pix LBSUMM#DetStatus-v03-pass1-analysis-2010D g
      and dq sct LBSUMM#DetStatus-v03-pass1-analysis-2010D g
      and dq trtb,trte LBSUMM#DetStatus-v03-pass1-analysis-2010D y+
      and dq mdt LBSUMM#DetStatus-v03-pass1-analysis-2010D g
      and dq rpc LBSUMM#DetStatus-v03-pass1-analysis-2010D g
      and dq tgc LBSUMM#DetStatus-v03-pass1-analysis-2010D g
      and dq csc LBSUMM#DetStatus-v03-pass1-analysis-2010D g  </Metadata>
      <Metadata Name="RunList">158392,158299,158269</Metadata>
      <Metadata Name="RQTSVNVersion">CoolRunQuery-unknown</Metadata>
      <Metadata Name="StreamListInfo">
         <Stream Name="physics_CosmicCalo" TotalNumOfEvents="1906748" NumOfSelectedEvents="745591"/>
         <Stream Name="physics_L1Calo" TotalNumOfEvents="11572490" NumOfSelectedEvents="7871572"/>
         <Stream Name="physics_MinBias" TotalNumOfEvents="2788076" NumOfSelectedEvents="1201539"/>
         <Stream Name="physics_MuonswBeam" TotalNumOfEvents="3629031" NumOfSelectedEvents="2817397"/>
         <Stream Name="physics_RNDM" TotalNumOfEvents="729444" NumOfSelectedEvents="272074"/>
         <Stream Name="physics_ZeroBias" TotalNumOfEvents="102922" NumOfSelectedEvents="65628"/>
      </Metadata>
      <LumiBlockCollection>
         <Run>158392</Run>
         <LBRange Start="218" End="340"/>
         <LBRange Start="358" End="358"/>
         <LBRange Start="367" End="447"/>
      </LumiBlockCollection>
      <LumiBlockCollection>
         <Run>158299</Run>
         <LBRange Start="378" End="434"/>
      </LumiBlockCollection>
      <LumiBlockCollection>
         <Run>158269</Run>
         <LBRange Start="3" End="81"/>
      </LumiBlockCollection>
   </NamedLumiRange>
</LumiRangeCollection>
```

GRL name

DQ criteria

Event count

Passed LumiBlocks in each run

Generating a Good Run List

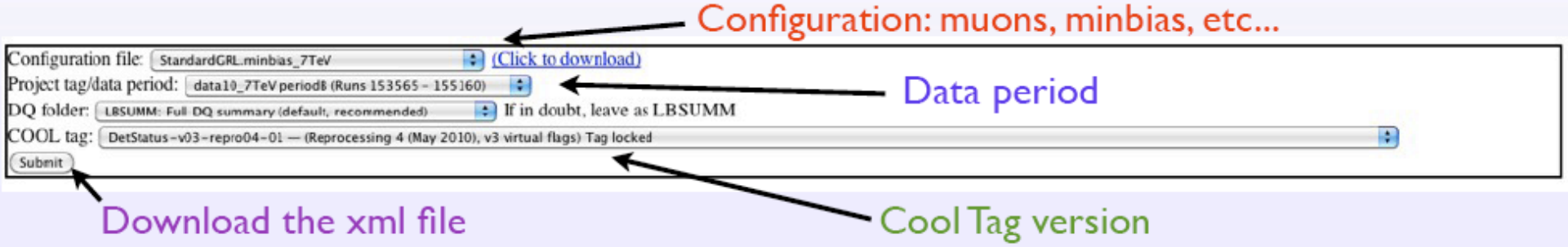
- The simplest way is to use the GRL generator page: https://atlasdqm.cern.ch/grl/
  - Web-based interface, can download the xml file.
- It is also possible to use the Run Query Tool (web page, or command line) to create an xml file,
  - However for physics analysis you should use 'official' lists.
- Ultimately it is the physics/performance groups that will generate the definition of the correct GRL for their analyses.
- For the moment, and with first data it is a good time to learn about data quality.

- Good runs lists are very closely linked to the luminosity calculation tools.

- More advanced techniques exist, but we won't exercise these at this tutorial.
  - A very good, full tutorial is available from:
    - https://twiki.cern.ch/twiki/bin/viewauth/Atlas/GoodRunsListsTutorial
    - Includes how to use GRLs within ROOT (i.e. outside of Athena).

GRL generator web page

- https://atlasdqm.cern.ch/grl/ : Simplest way to access 'official' GRLs.

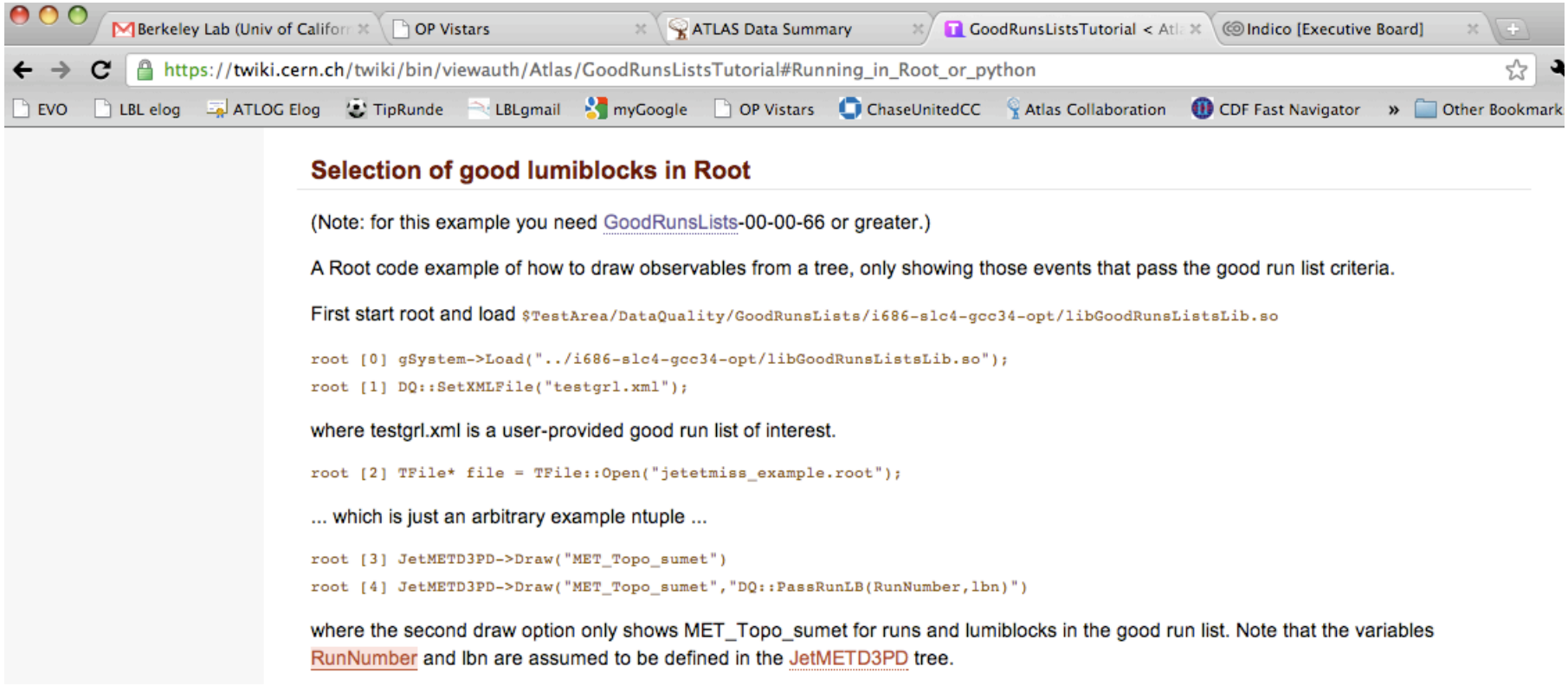


- Helpful table to correctly match the various options is also on the page:

Data	Project tag / data period	COOL tag	Input datasets to use
7 TeV, runs 152166-155160	data10_7TeV periods A,B	DetStatus-v03-repro04-01	May reprocessing (repro04_v01)
7 TeV, runs 155228-156682	data10_7TeV period C	DetStatus-v03-pass1-analysis-2010C	Tier-0 processing (t0pro04_v01)
7 TeV, runs 158045-159113	data10_7TeV periods D1-5	DetStatus-v03-pass1-analysis-2010D -or- DetStatus-v03-pass1-analysis-2010D-RPCloose	Tier-0 processing (t0pro04_v01)
7 TeV, runs 159179+	data10_7TeV period D6	DetStatus-v03-pass1 (not for final results!)	Tier-0 processing (t0pro04_v01)
900 GeV, 2009 data	data09_900GeV AllYear	DetStatus-v03-repro04-01	May reprocessing (repro04_v01)

- Once downloaded, you would copy the xml to the Athena run area.

Using a GRL in root



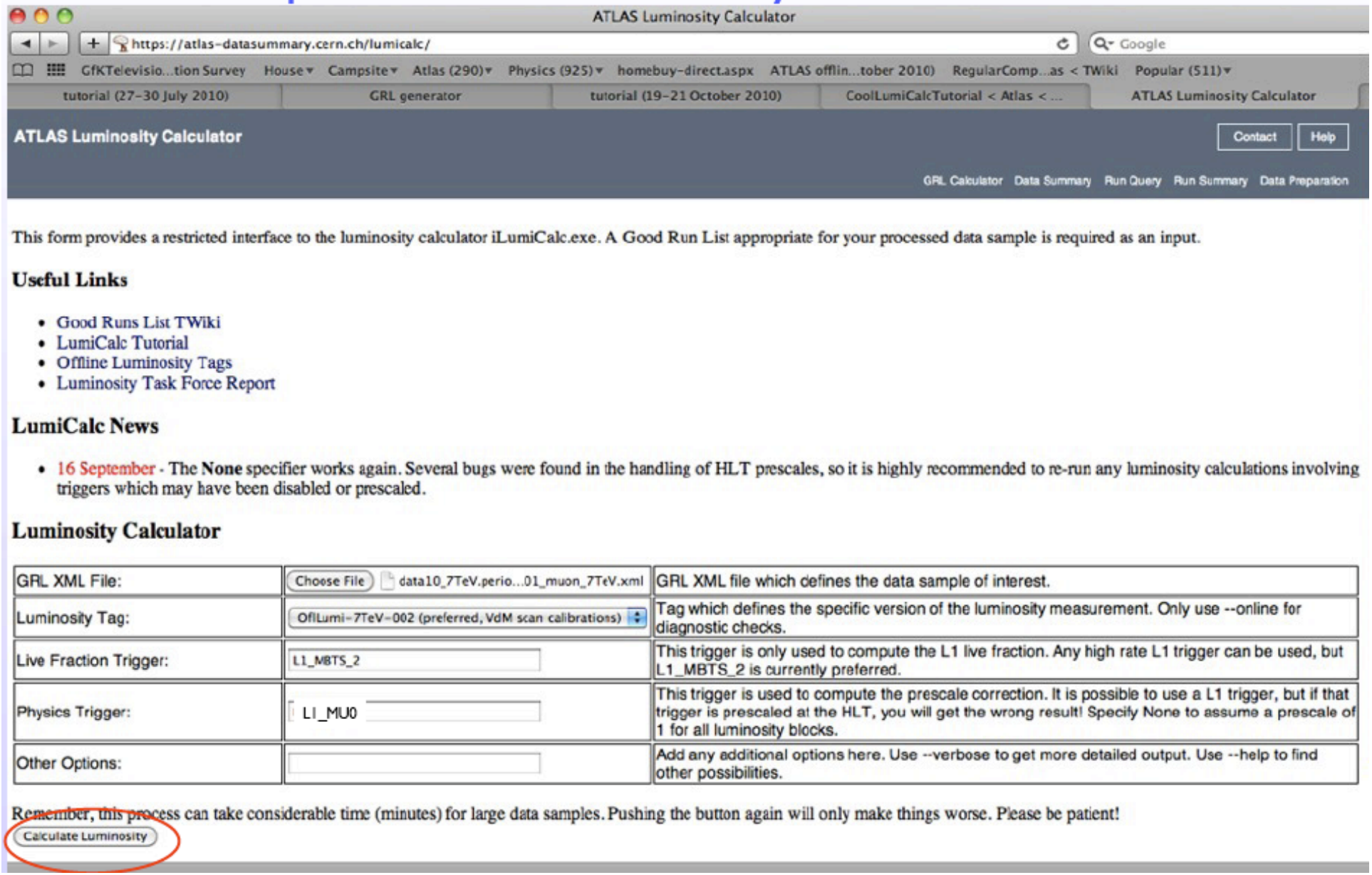
- We recommend applying the GRL at final analysis stage (i.e. not during making of ntuple)
  - GRL can change (due to mistakes or improved DQ understanding by systems/CP groups)
  - The fraction of Not-good data is rather small (no real saving in CPU)

Luminosity determination

- https://atlas-datasummary.cern.ch/lumicalc/
- The luminosity of the data you analysis depends on both the Trigger that you are using,
- and the 'amount' of data processed.
  - This 'amount' is affected by the GoodRunsLists, and needs to be used as part of the calculation.

- Fortunately, a simple web-page exists to perform the calculation for you.
  - You need to have:
    - A GoodRunsList xml file (we'll use the one from the previous example).
    - The trigger used for your analysis. Pick your favourite trigger,
      - Or, for example select EF_MU0

https://atlas-datasummary.cern.ch/lumicalc/



Output

- Output page provides equivalent command-line statement (requires athena)
- Also the total (and per-run) luminosity for the giver GRL period.

- Luminosity DB tag: OflLumi-7TeV-002
- Live fraction trigger: L1_MBTS_2
- Physics trigger: L1_MU0
- GRL file: data10_7TeV.periodB.153565-155160_LBSUMM_DetStatus-v03-repro04-01_muon_7TeV.xml was uploaded successfully to /tmp/lumifiles/81268c

Specification complete!

Environment

LumiBlockComps-00-02-05

Command Details

iLumiCalc.exe --lumitag=OflLumi-7TeV-002 --livetrigger=L1_MBTS_2 --trigger=L1_MU0 --xml=/tmp/lumifiles/81268c/data10_7TeV.periodB.153565-155160_LBSUMM_DetSta
01_muon_7TeV.xml

Command complete - CPU time: 0.0 s, Clock time: 19.4 s

Total Luminosity: 8.79698 nb^{-1}

Run Number	Number of Good LBs	Number of Bad LBs	Luminosity Delivered (nb^{-1})	Luminosity Livefraction Corrected (nb^{-1})	Luminosity Prescale Corrected (nb^{-1})	Lumi-Weighted Live Fraction (percent)	Lumi-Weighted Prescale
153565 (ARQ, Prescales)	696	0	0.665156	0.664095	0.664095	99.84	1.0
153599 (ARQ, Prescales)	107	0	0.0593819	0.0593123	0.0593123	99.88	1.0
154810 (ARQ, Prescales)	63	0	0.155331	0.152803	0.152803	98.37	1.0
154813 (ARQ, Prescales)	180	0	0.302937	0.299541	0.299541	98.88	1.0
154815 (ARQ, Prescales)	32	0	0.0678402	0.0624522	0.0624522	92.06	1.0
154817 (ARQ, Prescales)	279	0	0.53776	0.533888	0.533888	99.28	1.0
154822 (ARQ, Prescales)	189	0	0.428662	0.427019	0.427019	99.62	1.0
155073 (ARQ, Prescales)	326	0	1.19135	1.17645	1.17645	98.75	1.0
155112 (ARQ, Prescales)	499	0	3.6389	3.57969	3.57969	98.37	1.0
155116 (ARQ, Prescales)	86	0	0.522896	0.51253	0.51253	98.02	1.0
155160 (ARQ, Prescales)	264	0	1.33498	1.32921	1.32921	99.57	1.0
Total	2721	0	8.9052	8.79698	8.79698	98.78	1.0

Summary

- Find out which data exist and their approximate luminosity: DataPeriods twiki
- Find the corresponding physics containers: AMI
  - Submit job on physics containers
- Find out corresponding MC: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/DataMCForAnalysis
  - Again find corresponding dataset in AMI and submit job
- Create a GoodRunList depending on your analysis
- Calculate luminosity depending on GRL and your preferred trigger
- Analyze data and corresponding MC