Data Catalog Update

January 15, 2015

Metadata

- Dump everything possibly useful for the 2014 run
  - https://confluence.slac.stanford.edu/display/hpsg/Proposed+Data+Catalog+Metadata
  - Full DAQ configuration
    - Thresholds, Coincidence Windows, Enabled Trigger Bits, …
  - Shift-Takers' comments
    - Target Type, Beam Current, DAQ Rate, Trigger Description, …

Extraction of Metadata for Raw Data

- Python scripts written and tested
  - For the 2014 commissioning run it is cumbersome:
    - First, try to parse it from EVIO
    - Else, from the run spreadsheet
    - Else, get it manually from logbooks (overriding mistakes in above)
  - Output:
    - A huge spreadsheet for testing
    - Command-line for registering in the catalog
  - For the next run, we will have almost all from EVIO

```
metadata={}
metadata['Run']=runno
metadata['FileNumber']=filno

# Determine metadata (order is crucial):
ERM.GetMetadataFromDAQ(runno,metadata)
ERM.GetMetadataFromRunSpreadsheet(runno,metadata)
ERM.GetMetadataFromLogbook(runno,metadata)

# add Pass-specific metadata:
ERM.GetPassMetadata(filename,metadata)
```

```
# create command-line options for metadata:
mtdopts=''
for key in metadata.keys():
  if type(metadata[key]) is str:
    mtdopts += ' --define s%s=\"%s\"'%(key,metadata[key])
  elif type(metadata[key]) is int:
    mtdopts += ' --define n%s=%d'%(key,metadata[key])
  elif type(metadata[key]) is float:
    mtdopts += ' --define n%s=%.3f'%(key,metadata[key])
  else:
    sys.exit('Wierd Metadata:  '+metadata[key])
```

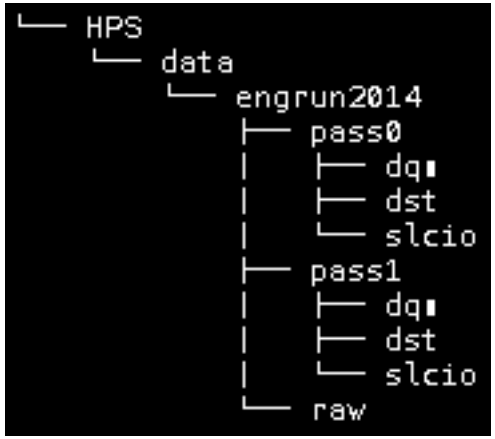
Crawling

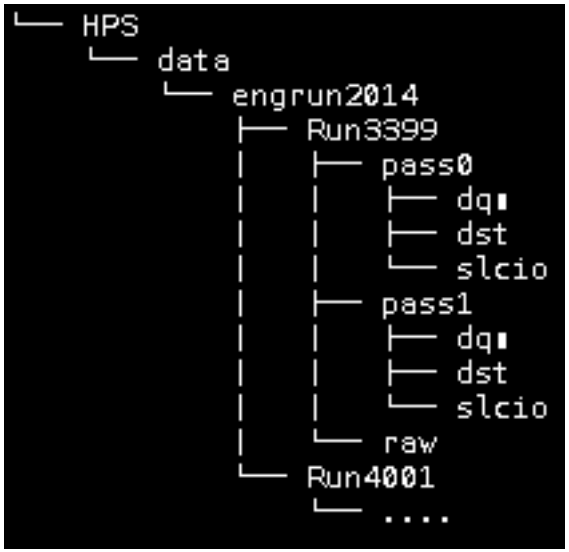
- Finds new files to add to the data catalog
- We can automatically get a list of files already in the catalog, so just compare existing list of those on disk/tape with those already in the catalog.
  - This is already part of the python scripts.
  - The previous ideas were to keep a timestamp file or update on-the-fly when they get processed.

Structure

- Ideally, for maintenance reasons, we should assign metadata to folders and have it inherited recursively
  - Still some server issues with this.

- Many possible folder layouts (can even exist simultaneously).
  - E.g. organize first by run, or pass …
  - Slight modification to scripts once decided (waiting on above issue resolution).
  - For search functionality it doesn't matter much.



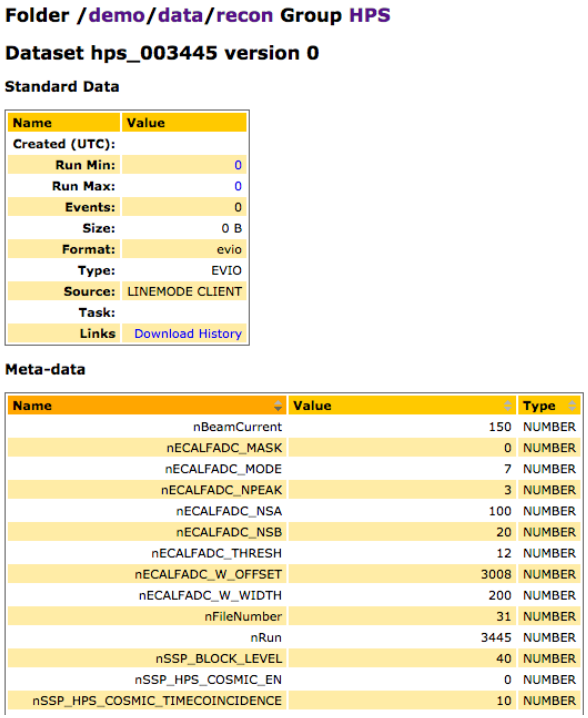


Usage

- Command Line
  - Filtering, Choose which metadata to display

```
> $dc --filter 'nECALFADC_MODE==7' --filter 'sTarget=="5 um Pt"' --display sDescription /demo/data/recon
/mss/hallb/hps/data/hps_003445.evio.31  Single + Pairs , 1 cluster 2 cluster ,
```

- Web Interface
  - Limited functionality?



Folder /demo/data/recon Group HPS

Dataset hps_003445 version 0

Standard Data

Name	Value
Created (UTC):	
Run Min:	0
Run Max:	0
Events:	0
Size:	0 B
Format:	evio
Type:	EVIO
Source:	LINEMODE CLIENT
Task:	
Links	Download History

Meta-data

Name	Value	Type
nBeamCurrent	150	NUMBER
nECALFADC_MASK	0	NUMBER
nECALFADC_MODE	7	NUMBER
nECALFADC_NPEAK	3	NUMBER
nECALFADC_NSA	100	NUMBER
nECALFADC_NSB	20	NUMBER
nECALFADC_THRESH	12	NUMBER
nECALFADC_W_OFFSET	3008	NUMBER
nECALFADC_W_WIDTH	200	NUMBER
nFileNumber	31	NUMBER
nRun	3445	NUMBER
nSSP_BLOCK_LEVEL	40	NUMBER
nSSP_HPS_COSMIC_EN	0	NUMBER
nSSP_HPS_COSMIC_TIMECOINCIDENCE	10	NUMBER

Issues

- Server
  - Couple server issues/bugs already fixed in 2015.
  - Assigning metadata to a “folder” does not seem to work in that it is not inherited recursively.
    - At present, could proceed and register all files independently, each with their own independent metadata
- Client
  - Command-Line API requires ssh’ing into a computer at SLAC
  - While web-interface does not have filtering ability?

ToDo

- Make the catalog, finally
- Still need to extract first/last event# and start/end time for each file (put this in EvioToLcio)
- Put script at SLAC to register local files (and test)
- Documentation
- Minor updates to scripts for next run