GRAND Data Management meeting

François Legrand

- Values in the data (which fields are not properly filled, how can we correct that)
- How do we manage the need for new "custom" fields/trees ?
- Issue with directory format
- Plans to reprocess and correct data (filling the du_id, feb_id etc... in root files even if wrong in the raw data)

- **Values in the data (which fields are not properly filled, how can we correct that)**

  See tables

- **How do we manage the need for new "custom" fields/trees ?**
  - During analysis some new « fields » need to be created. We need to set up a common mechanism which should be simple, flexible and versatile !

- **Issues with directory format**

Remind that :

Dataset directories name should match the following structure :

[sim|exp|mod]_[site]_[date]_[time]_RUN[run_number]_[mod]_[extra]_[serial]

And root files inside the directory should match the following structure:

[grouptreename]_[run_number]_L[analysis level]_[serial].root for run trees

[grouptreename]_[date]_[time]_[first_event_number]-[last_event_number]_L[analysis level]_[serial].root for event trees

- **Issues with directory format**

There are 2 main problems :

1) In dataset name, date_time should be the date and time of the first event (or begining of the run). But binary files comes in many files for the same run… thus gtot create the « run » directory with the date time of the first processed file and add data in that directory for the next processed files !

→ directory name depends on which file of the run is processed first which is not necessarily the first one in the run because :

- Some file may not be transfered at the same time (the transfer of the first file can fail and be resumed the next day)

- When processing bin files on the computing farm we don't know which job will start first (unless we add dependencies which can be complex to manage)

→ date_time information in the dataset name is not « predictible » nor reliable !

⇒ **So is it really usefull ?**

- **Issues with directory format**

2) When adding files into a directory, gtot will add some new event files (adc, rawvoltage etc. files) but will update (actually replace) the run files (run and runrawvoltage files). This prevent to process files from a run in parallel (because several process will replace the same run file and it will fail) !

→ We NEED to process files in parallel
(at least if we want to reprocess the 116 682 files we already have) !

For the record we have now ~10 000 runs

The biggest have 6611 files

~100 runs with more than 100 files

25 runs with more than 1000 files

- Plans to reprocess and correct data
  - Do we fill the du_id, feb_id etc... in root files even if wrong in the raw data ? If yes, then need a gtot « correcting » version !