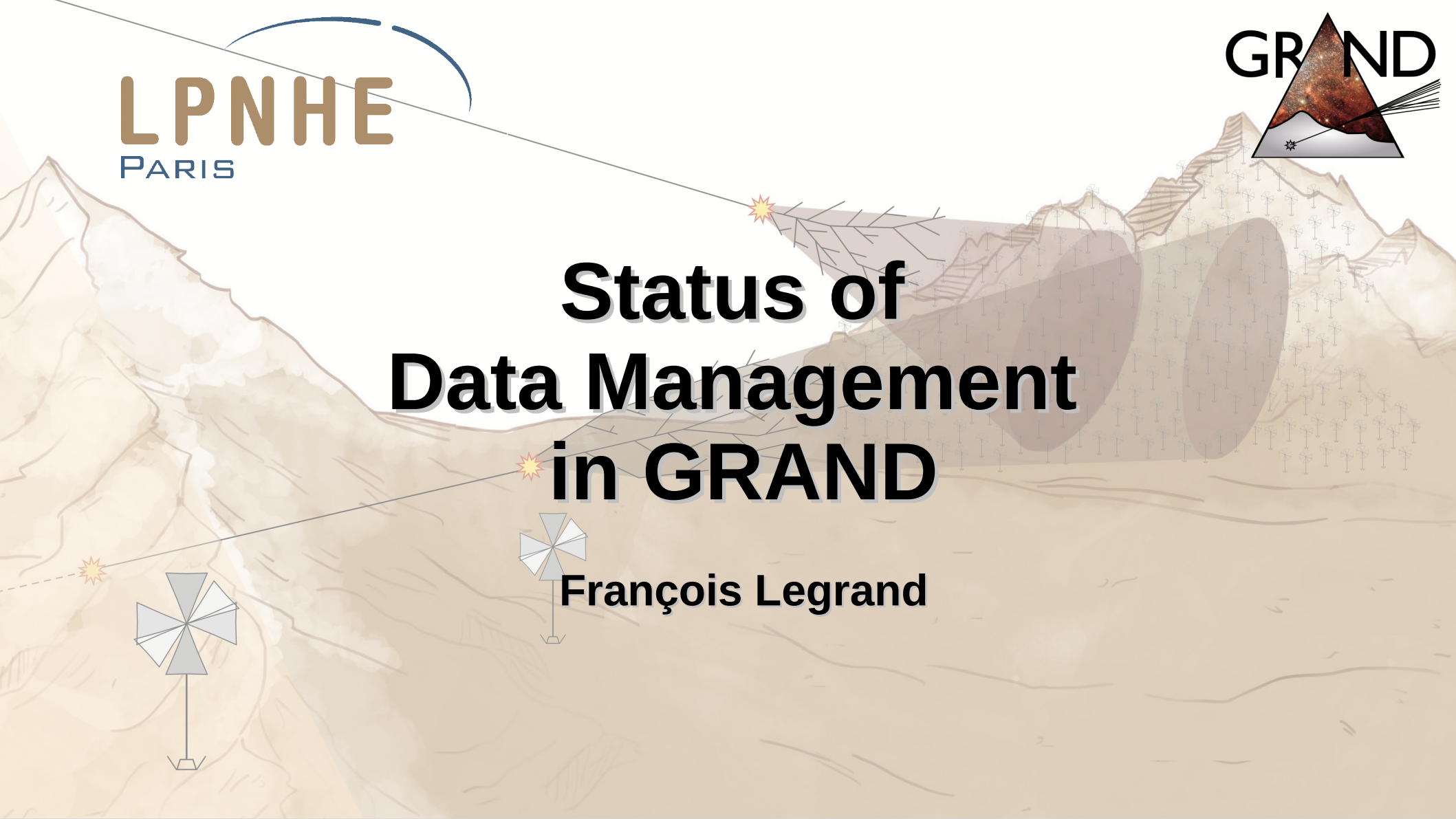


Status of Data Management in GRAND

François Legrand

Data management group

Olivier Martineau

François Legrand

Pengxiong Ma

Ramesh Koirla

Matias Tueros

Lech Piotrowski

Objectives

The objective of the group is to set up the rules and procedure concerning data in Grand and how to manage them

- Identify products (type of data,…) and flows
- Determine what should be kept and referenced
- Coordinate the reflexion about data organization (formats, naming,…) and structure (root format)
- Define storage strategy (versioning, replication/backup, security, hosting, ...)
- Define data access rules (protocols, confidentiality, referencement and access,..)
- Etc...

Present structure of data

- Data are stored @ccin2p3 in :
  - /sps/grand/data/<site>/<raw|GrandRoot>/<year>/<month>/

- Root files contains all trees and have the same name as raw files (except the extension)

  This will change (we hope before the end of this year)
  → We plan to reprocess the data to match the new rules/format

Naming rules : Raw files (will not change)

- Raw filenames follow the pattern :
  [site]_[date]_[time]_RUN[run_number]_[mod]_[extra].bin
  - site is gp80, gaa or nancay
  - date and time are YYYYMMDD HHmmss (UTC)
  - mod is CD (**C**oincidence **D**ata), MD (**M**inimal bias **D**ata or **M**onitoring **D**ata), UD (**U**nit **D**ata), TR (TriggeR data)
    - CD: data corresponding to central DAQ trigger (so called Second Level Trigger or T3), ie several DU triggers (so called First Level Trigger or T2) in coincidence
    - MD: data recorded with automatic, forced trigger (eg 20Hz or 10s)
    - UD: data corresponding to DU triggers (First Level Triggers or T2) not passing central DAQ trigger (Second Level trigger or T3).
    - TR: Trigger data
  - extra can be whatever generator think can be usefull (20db-du85, etc…)
  - It's really important to respect that format (underscores to separate fields and no underscores in the fields)

General rules for new format

- Data will not be overwritten → analysis or treatment will produce **new files**
- Datas will be stored into **directories**.
  - One directory (dataset) will correspond to an observation run or to a simulation.
  - Each directory will contains one Trun file describing the run parameters and some additional root files (at least Trunefieldsim, Tshower and Tshowersim for simulations and Tadc, Trawvoltage for observations).
  - Files containing trees with traces may be be splitted on a event number base (to limit size of files)

Naming rules : GrandRoot files

- Dataset directories name will match the following structure :
  [sim|exp|mod]_[site]_[date]_[time]_RUN[run_number]_[mod]_[extra]_[serial]
  - Serial is an extra number to distinguish between different version of a run (e.g. different processing…)
  - Different mods → different dirs, different extra → different dirs

- Root files inside the directory will match the following structure:
  [grouptreename]_[date]_[time]_[events|run]_L[analysis level]_[serial].root
  - events will be the range of events in the "event file" and run the run_number for run trees
  - Grouptreename will be the Ttree name or a group tree (in case of several Ttrees in the same file)
  - Serial will identify different versions of analysis.

Naming Analysis Level

- L0 → No Noise data
  - Sim data (efield, shower) would start at L0, and voltage and adc without noise generated from the efield would also be L0
- L1 → Data with Noise
  - Hardware data is with noise, so ADC, RawVoltage, Voltage and reconstructed Efield coming from hardware would be L1
  - ADC generated from Sim + added noise would be L1, and would correspond to the ADC from hardware. So would resulting L1 rawvoltage, voltage, reconstructed efield
- L2 → Analysis

sim : efield/shower_L0 → voltage_L0 → adc_L0 → adc_L1 (added noise) → voltage_L1 → efield_L1 (reconstructed)
obs : adc_L1 → rawvoltage_L1 → voltage_L1 → efield_L1 (reconstructed)

Present storage @ccin2p3

```
/sps/grand/data/
├─ gaa
│   ├── GrandRoot
│   ├── raw
├── gp13
│   ├── GrandRoot
│   ├── raw
├── gp80
│   ├── GrandRoot
│   │   ├── 2024
│   │   │   ├── 11
│   │   │   │   ├── GP80_20241104_112732_RUN125_MD_RAW_ChanXYZ_20dB_GP80_35DUs_10s_001.root
│   │   │   │   ├── GP80_20241104_114228_RUN127_UD_RAW_ChanXYZ_20dB_GP13_XorY_Y2FLOAT_X2Z_001.root
│   │   │   │   ├── …
│   │   │   ├── …
│   ├── raw
│   │   ├── 2025
│   │   │   ├── 02
│   │   │   │   ├── GP80_20250225_170434_RUN14_CD_20dB_9DUs_GP80_t3Test-CD-10000-1.bin
│   │   │   │   ├── GP80_20250225_182804_RUN14_CD_20dB_7DUs_GP80_t3Test-CD-10000-1.bin
│   │   │   │   ├── …
│   │   │   ├── …
```

Future storage @ccin2p3

```
/sps/grand/data/
├─ gaa
│   ├── GrandRoot
│   ├── raw
├── gp13
│   ├── GrandRoot
│   ├── raw
├── gp80
│   ├── GrandRoot
│   │   ├── 2024
│   │   │   ├── 11
│   │   │   │   ├── exp_GP80_20241104_112732_RUN125_MD_RAW_ChanXYZ_20dB_GP80_35DUs_10s_001
│   │   │   │   │   ├── run_125_L0_0000.root
│   │   │   │   │   ├── adc_20241104_112732_0-1000_L1_0000.root
│   │   │   │   │   ├── runrawvoltage_125_L0_0000.root
│   │   │   │   │   ├── …
│   │   │   ├── …
│   ├── raw
│   │   ├── 2025
│   │   │   ├── 02
│   │   │   │   ├── GP80_20250225_170434_RUN14_CD_20dB_9DUs_GP80_t3Test-CD-10000-1.bin
│   │   │   │   ├── …
│   │   │   ├── …
```

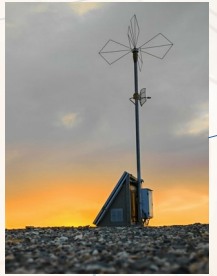
Automation

- Automatic conversion into GrandRoot format
- Results of transfer and conversion registered into DB
- Root files referenced into the DB
- DB website updated (https://database.grand-observatory.org)
- Root files copied on disk into Irods (long term storage)
- All is done in slurm using prod_grand account (so files are read only for the collaboration)
- Each month (on 15th) all raw files from the previous month are tared (in a structured archive) and pushed on tapes (with 2 different copies) into Irods (long term archiving)

G@A

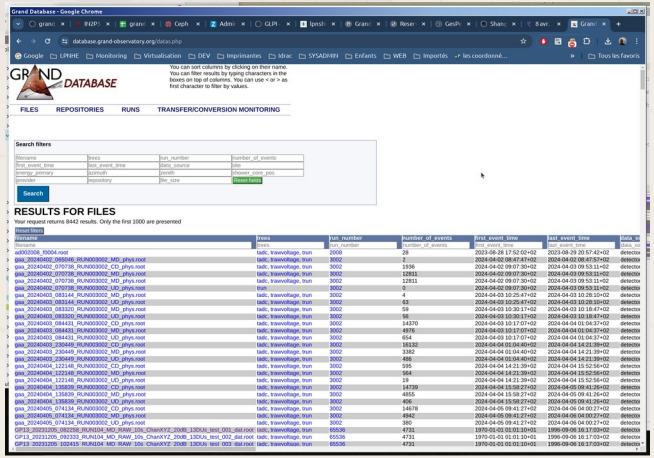
Automatic transfer

CCIN2P3



GP80

Automatic transfer



Next steps

- Reprocess data to match the new directory structure (hopefully before the end of the year)
- Register simulations into the database
- Versioning of files and code
- Setup a second storage site (China ?)
- Define rules and protocols to add processed files into the “official” data repository (L2 data)
- Define some validation process (check files integrity,…)

Points to discuss

- MD data : MD files can be Monitoring Data (10s trigger) or Minimal bias Data (10-1kHz trigger)… to implement a proper monitoring we may need to identify the monitoring data !

  → **Should we separate these files with a different mod name ?**

- New format implies to have all the files from a single run in the same directory.
  - If runs last for a long time (and produce) a lot of data we may end with thousands of files in the same directory → should be hard to manage (at least to be human readable) !
  - Data are supposed to be stored in a structure by month (and date is part of the directory name). Also the archiving is done on a month base. This should be problematic if runs extends over different months

  → **Should we limit the duration of a run and systematically change run number when changing of month ?**

Points to discuss

- **Where to store single events extracted from different events/runs ?** (i.e real UHECR) extracted from the original directories. But for this things, run numbers, dates, event ranges have no sense, as events can be form completely different periods of time

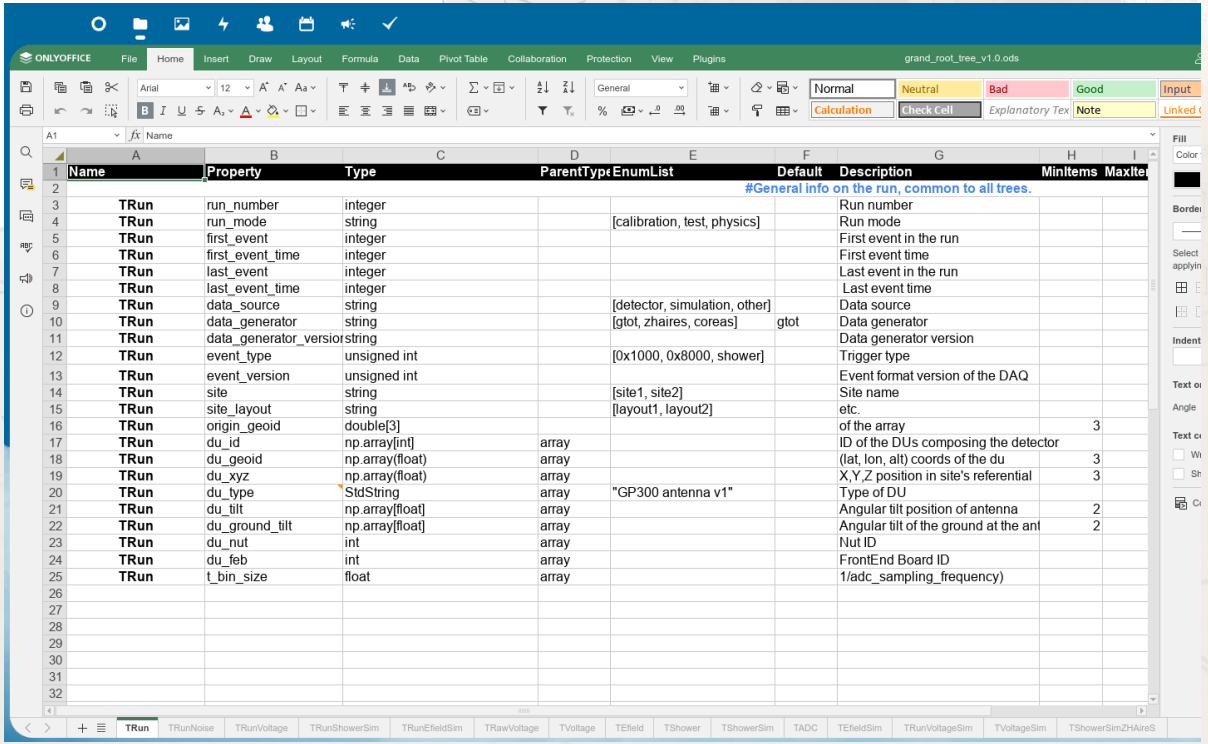
  - Create one dedicated directory and add files on the fly (may grow too much in the future) ?
  - Create several dedicated directories and group by period of time (month?) ?
  - Leave each event it its original directory and create simlinks in one/several directory ?
  - Simply “flag” these events in the database ?
  - ...

Thanks

Questions ?

Root files structure

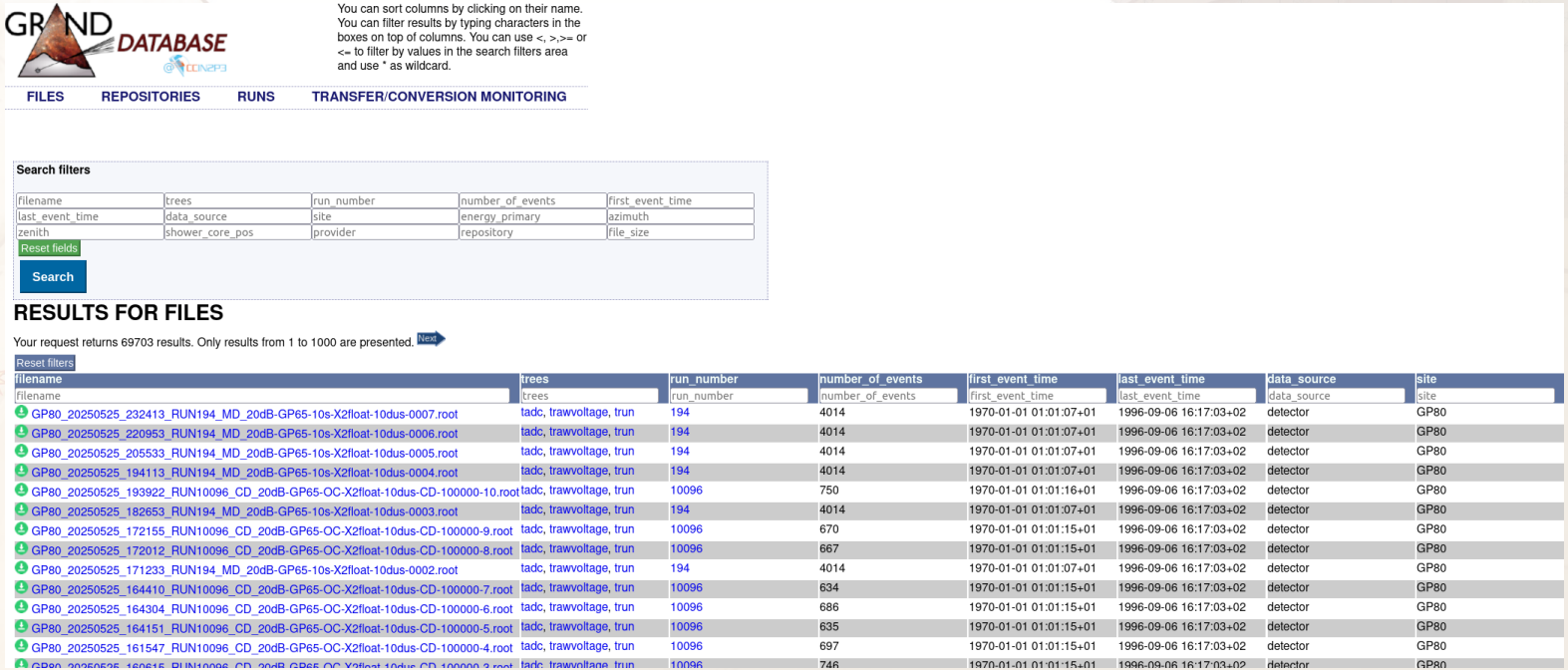
- Root files structure is available at : https://box.in2p3.fr/index.php/s/ipmk4XZP87pjRnr
- We try to keep it up to date



Name	Property	Type	ParentType	EnumList	Default	Description	MinItems	MaxIte
						#General info on the run, common to all trees.		
TRun	run_number	integer				Run number		
TRun	run_mode	string		[calibration, test, physics]		Run mode		
TRun	first_event	integer				First event in the run		
TRun	first_event_time	integer				First event time		
TRun	last_event	integer				Last event in the run		
TRun	last_event_time	integer				Last event time		
TRun	data_source	string		[detector, simulation, other]		Data source		
TRun	data_generator	string		[gtot, zhaires, coreas]	gtot	Data generator		
TRun	data_generator_versior	string				Data generator version		
TRun	event_type	unsigned int		[0x1000, 0x8000, shower]		Trigger type		
TRun	event_version	unsigned int				Event format version of the DAQ		
TRun	site	string		[site1, site2]		Site name		
TRun	site_layout	string		[layout1, layout2]		etc.		
TRun	origin_geoid	double[3]				of the array	3	
TRun	du_id	np.array[int]	array			ID of the DUs composing the detector		
TRun	du_geoid	np.array(float)	array			(lat, lon, alt) coords of the du	3	
TRun	du_xyz	np.array(float)	array			X,Y,Z position in site's referential	3	
TRun	du_type	StdString	array	"GP300 antenna v1"		Type of DU		
TRun	du_tilt	np.array[float]	array			Angular tilt position of antenna	2	
TRun	du_ground_tilt	np.array[float]	array			Angular tilt of the ground at the ant	2	
TRun	du_nut	int	array			Nut ID		
TRun	du_feb	int	array			FrontEnd Board ID		
TRun	t_bin_size	float	array			1/adc_sampling_frequency)		

Database referencement

- https://database.grand-observatory.org
  - Accessible using mattermost credential
  - Data can be searched and downloaded



GRAND DATABASE

You can sort columns by clicking on their name. You can filter results by typing characters in the boxes on top of columns. You can use <, >,>= or <= to filter by values in the search filters area and use * as wildcard.

FILES REPOSITORIES RUNS TRANSFER/CONVERSION MONITORING

Search filters

filename	trees	run_number	number_of_events	first_event_time
last_event_time	data_source	site	energy_primary	azimuth
zenith	shower_core_pos	provider	repository	file_size

Reset fields

Search

RESULTS FOR FILES

Your request returns 69703 results. Only results from 1 to 1000 are presented. Next

Reset filters

filename	trees	run_number	number_of_events	first_event_time	last_event_time	data_source	site
GP80_20250525_232413_RUN194_MD_20dB-GP65-10s-X2float-10dus-0007.root	tadc, trawvoltage, trun	194	4014	1970-01-01 01:01:07+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_220953_RUN194_MD_20dB-GP65-10s-X2float-10dus-0006.root	tadc, trawvoltage, trun	194	4014	1970-01-01 01:01:07+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_205533_RUN194_MD_20dB-GP65-10s-X2float-10dus-0005.root	tadc, trawvoltage, trun	194	4014	1970-01-01 01:01:07+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_194113_RUN194_MD_20dB-GP65-10s-X2float-10dus-0004.root	tadc, trawvoltage, trun	194	4014	1970-01-01 01:01:07+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_193922_RUN10096_CD_20dB-GP65-OC-X2float-10dus-CD-100000-10.root	tadc, trawvoltage, trun	10096	750	1970-01-01 01:01:16+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_182653_RUN194_MD_20dB-GP65-10s-X2float-10dus-0003.root	tadc, trawvoltage, trun	194	4014	1970-01-01 01:01:07+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_172155_RUN10096_CD_20dB-GP65-OC-X2float-10dus-CD-100000-9.root	tadc, trawvoltage, trun	10096	670	1970-01-01 01:01:15+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_172012_RUN10096_CD_20dB-GP65-OC-X2float-10dus-CD-100000-8.root	tadc, trawvoltage, trun	10096	667	1970-01-01 01:01:15+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_171233_RUN194_MD_20dB-GP65-10s-X2float-10dus-0002.root	tadc, trawvoltage, trun	194	4014	1970-01-01 01:01:07+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_164410_RUN10096_CD_20dB-GP65-OC-X2float-10dus-CD-100000-7.root	tadc, trawvoltage, trun	10096	634	1970-01-01 01:01:15+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_164304_RUN10096_CD_20dB-GP65-OC-X2float-10dus-CD-100000-6.root	tadc, trawvoltage, trun	10096	686	1970-01-01 01:01:15+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_164151_RUN10096_CD_20dB-GP65-OC-X2float-10dus-CD-100000-5.root	tadc, trawvoltage, trun	10096	635	1970-01-01 01:01:15+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_161547_RUN10096_CD_20dB-GP65-OC-X2float-10dus-CD-100000-4.root	tadc, trawvoltage, trun	10096	697	1970-01-01 01:01:15+01	1996-09-06 16:17:03+02	detector	GP80
GP80_20250525_160615_RUN10096_CD_20dB-GP65-OC-X2float-10dus-CD-100000-3.root	tadc, trawvoltage, trun	10096	746	1970-01-01 01:01:15+01	1996-09-06 16:17:03+02	detector	GP80