Deliverable 2.2:

Data selection and Flagging modules

Authors Samuele Galeotta
Daniele Tavagnacco
Andrea Zacchei

Date Sep 14st, 2018

Work Package WP2 - Data Selection and Flagging

DocId [INAF_BP02-02-2.0]

Revision History

Version	Authors	Date	Changes
1.0	Samuele Galeotta Daniele Tavagnacco Andrea Zacchei	May 31st, 2018	Initial Version
2.0	Samuele Galeotta Daniele Tavagnacco Andrea Zacchei	Sep 14th, 2018	Final Release

Contents

1 Introduction

The main goal of the Beyond Planck project is to build an end-to-end Gibbs sampler for the Planck LFI data, and use it to improve the overall calibration and fidelity of the final LFI sky maps.

Working Package 2 (WP2), *Data Selection and Flagging*, serves as starting point to select at the *timelines* level, the data to be used in the Gibbs sampler. This working package is responsible to flag the data that should be excluded from analysis, according to a predefined criterion such as: maneuver periods, gain changes in the data acquisition electronics that caused saturation, abrupt changes in voltage outputs caused by gain fluctuations, etc.

The final goal of this WP is to provide the subsequent WPs input data streams clean from instrumental effects [1] like ADC correction and 1Hz spikes.

Working Package 2, *Data Selection and Flagging*, is organized into two independent software modules:

- *Data extraction and anomalies flagging,* addressing the WP task of extracting the data from the Planck-LFI database and organizing them into portable HDF5 files. To the data extracted from the database flags for anomalies and planets transits in the observations are are applied.
- *Data selection,* in charge of preparing data as input to the Gibbs sampler (know systematics cleaned), transforming raw data in engineering value and computing pointing information.

2 Data extraction and anomalies

In the Planck project data was preserved at LFI Data Processing Centre (DPC) with a predefined structure inside three Oracle databases organized as follows:

- *Level 1 database* containing the raw data recovered from telemetry packets into time ordered streams.
- *Level 2 database* containing time ordered data streams in engineering value before and after calibration process: this will be the input to the Gibbs sampler.
- *Test database* containing all the output of the pipeline tests.

This structure is not portable because is highly dependent by the database engine (Oracle). For this reason, a different approach is needed for the scope of Beyond Planck.
After some analysis, a file based approach resulted to be more portable. In this framework, the first module of *Data Selection and Flagging* will extract all the *Level 1* raw data from the LFI database into HDF5 files. This file format is well known in the Astronomical community, is free and has I/O function and graphical interfaces already available.
The description of the extracted data HDF5 file structure is depicted in section 2.2.
During the data extraction procedure, all the anomalies in the data and Solar system planets observations are identified and flagged according to the predefined bit mask structure inherited from Planck definition. The data flag is stored as a 32 bit integer number, the meaning of each flag is described in section 2.1
The data extraction is also checking the data for possible gaps in the streams. This is done for each frequency by merging the On Board Time (OBT) of all the datastreams of the frequency and search for missing samples in the single diode stream of data. Missing data are filled with zeros and the missing data flag is associated to the sample according to the mask.
This activity was performed later in the Planck pipeline [2], but, taking into account that the gap filling will never change, moving this computation in the extraction module that will run just once for the scope of Beyond Planck, will speed up the data selection module.

2.1 Data flag structure

The data flag for each sample is stored as a 32 bit integer in the FLAG dataset of the specific diode in the HDF5 file (Section 2.2) according to the predefined bit mask structure inherited from Planck definition as shown in Figure 2.1.1.
For the Beyond Planck scope, just part of the Planck bit mask is relevant. In particular for the data selection and differentiation the used bits of the flag are:

- Bit **4**: “maneuvers”. This identify the data acquired during the satellite maneuvers, at the moment discarded in the analysis.
- Bit **11**: “start/stop gap”. This identify the samples at the begin at the end of a missing block of data.
- Bit **14**: “Bad data”. This identify the samples containing wrong data that should be excluded.
- Bit **16**: “gap”. This identify the missing data filled with zeroes. This data should be discarded because the samples are only used as placeholders to have all the datastreams with the same length and avoid computing the sample time correlation between the diodes.

The other flags used that are relevant for the other WP such ad calibration ad map-making are:

- Bit **18**: “planets”. This identify the samples acquired that contain a planet transit.
- Bit **19**,**20**: “moving objects”. Those identify the samples acquired during the transit of any object.
- Bit **22**: “special observation”, This identify sets of non-standard operations performed during the data acquisition. In particular this flag identify the “spin up” data acquisition performed by Planck-LFI.

...	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
...	**1**	_	**1**	**1**	**1**	_	**1**	_	**1**	_	_	1	_	_	_	_	1	1	**1**	_	_	_	_

4 maneuver
11 start/stop gap
14 bad data
16 gap
→ ***differentiation***

5/6 time quality
18 planet
19/20 moving objects
22 special observation
→ ***calibration***

Figure 2.1.1 Planck 32bit mask with relevant flags used for Beyond Planck

2.2 Data extraction HDF5 file structure (Level1)

The raw datasets extracted from Planck-LFI database are organized into HDF5 files. Each HDF5 file contains all the data for one Planck-LFI Operational Day (OD) of each LFI horn antenna. The OD is defined as the time between two consecutive connection to the satellite, roughly 24 hours. The total number of HDF5 is 1604 OD times 11 LFI horn antennas. The file name convention used is:

LFI_FFF_HH_**L1**_OOOO.**h5**

where:

- **FFF** is the horn frequency. For LFI is 030, 044 or 070
- **HH** is the horn number going from 18 to 28
- **OOOO** is the Planck Operational Day going from 91 to 1604

Each HDF5 file is organized into three type of groups:

- **Time**: containing all the time information organized into OBT, SCET and MJD datasets.
    - OBT: On Board Time, the clock of the satellite. The time is measured in 65536 clock values per second.
    - SCET: a TAI time computed from OBT and expressed in microseconds.

    - MJD: the Mean Julian Date associated to the SCET.
- **HHDD**: HH horn per diode (DD) acquired data. The diode notation is 00 and 01 for the Main radiometric chain, 10 and 11 for the Side radiometric chain. There are 4 of these groups in each HDF5 file, one for each diode.
- **AHF_info**: containing all the information extracted from the Attitude History File (AHF): PID, PID_PSO, PID_start, PID_stop, PID_changes, science_start.
    - PID: the unique Pointing ID, a progressive number identifying the Planck pointing.
    - PID_PSO: the PID equivalent identifier in the Attitude History File used to get the position of the satellite.
    - PID_start: the OBT when the PID begins in the OD datastream.
    - PID_stop: the OBT when the PID ends in the OD datastream.
    - science_start: the OBT when the stable pointing (no maneuver) starts after the PID_start.
    - PID_changes: the sample number where the PID changes in the whole OD datastream.

Figure 2.2.1 shows the detailed structure of a Level 1 extracted HDF5 file including the attribute keywords stored in each HDF5 used to identify the data contained. Figure 2.2.2 shows an example of a HDF5 extracted file structure opened with HDFView software [3].

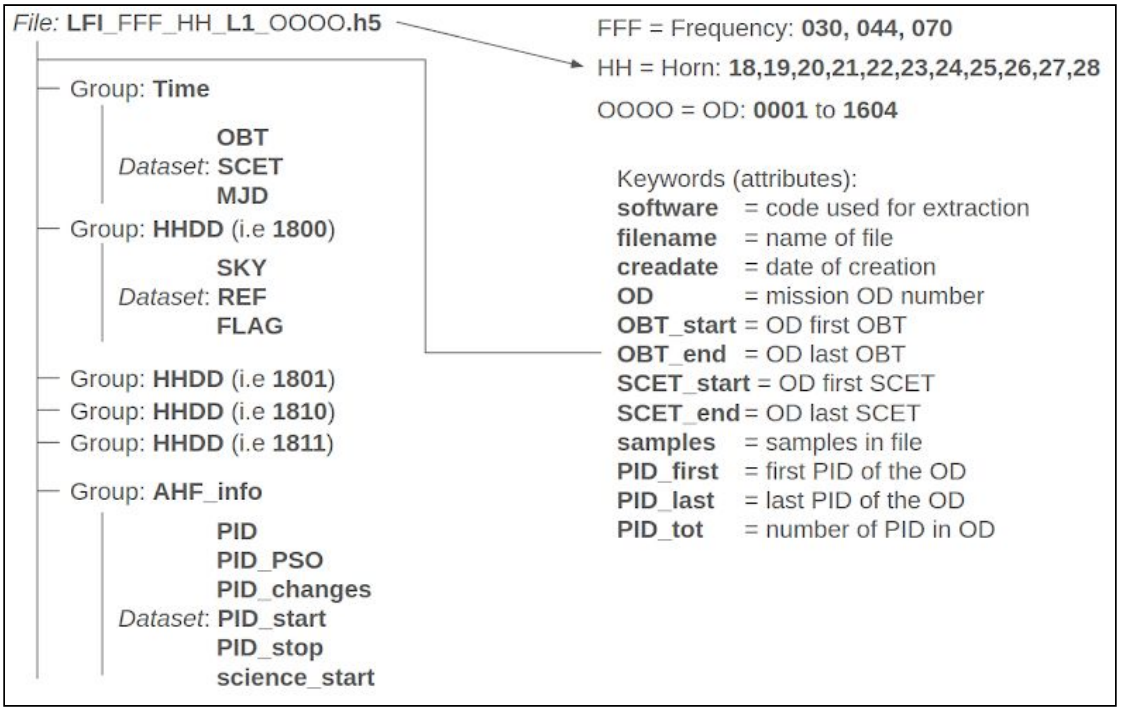

Figure 2.2.1. Raw Level 1 data HDF5 file structure

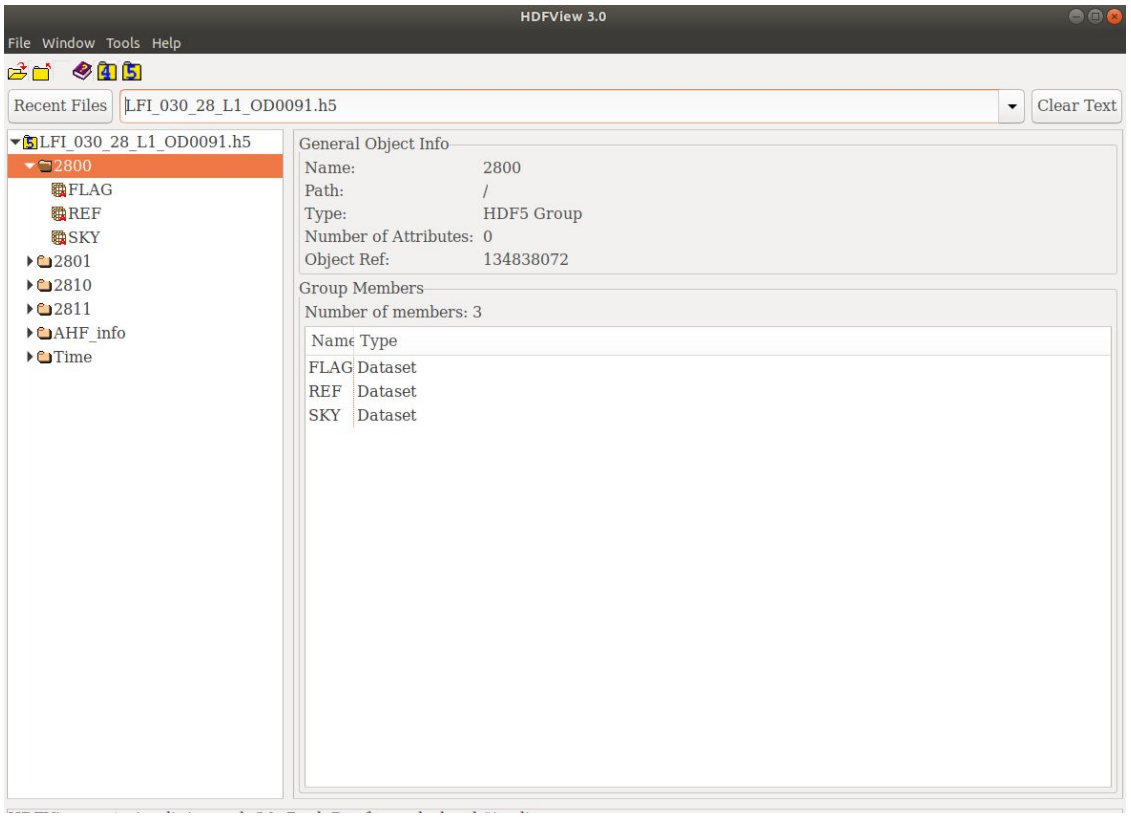

Figure 2.2.2. Level 1 data file structure with HDFview software

The extraction software should be executed in the Planck LFI DPC because it needs the Planck LFI database to extract the data. The software is available in the Beyond Planck GitLab repository, even if the module is strongly dependent from the Trieste Planck DPC software environment and its interface to the Oracle databases.
The *Data Selection* module is executed with the following command:

```
python DataExtraction.py --startOD=N --stopOD=M
```

where *N* is the starting OD to process, that should a number greater or equal 91 and *M* is the past-the-last OD to be processed.

This code is aimed to be runned once in order to produce the Level 1 raw data for the Beyond Planck project to be used as base input for the data selection part of the WP2.

3 Data selection

The *Data Selection* module uses the Level 1 raw data produced in the *Data Extraction* module as input and produce the differentiated data sets and the pointing informations for each data sample. An overview of the *Data Selection* module is shown in figure 3.1.

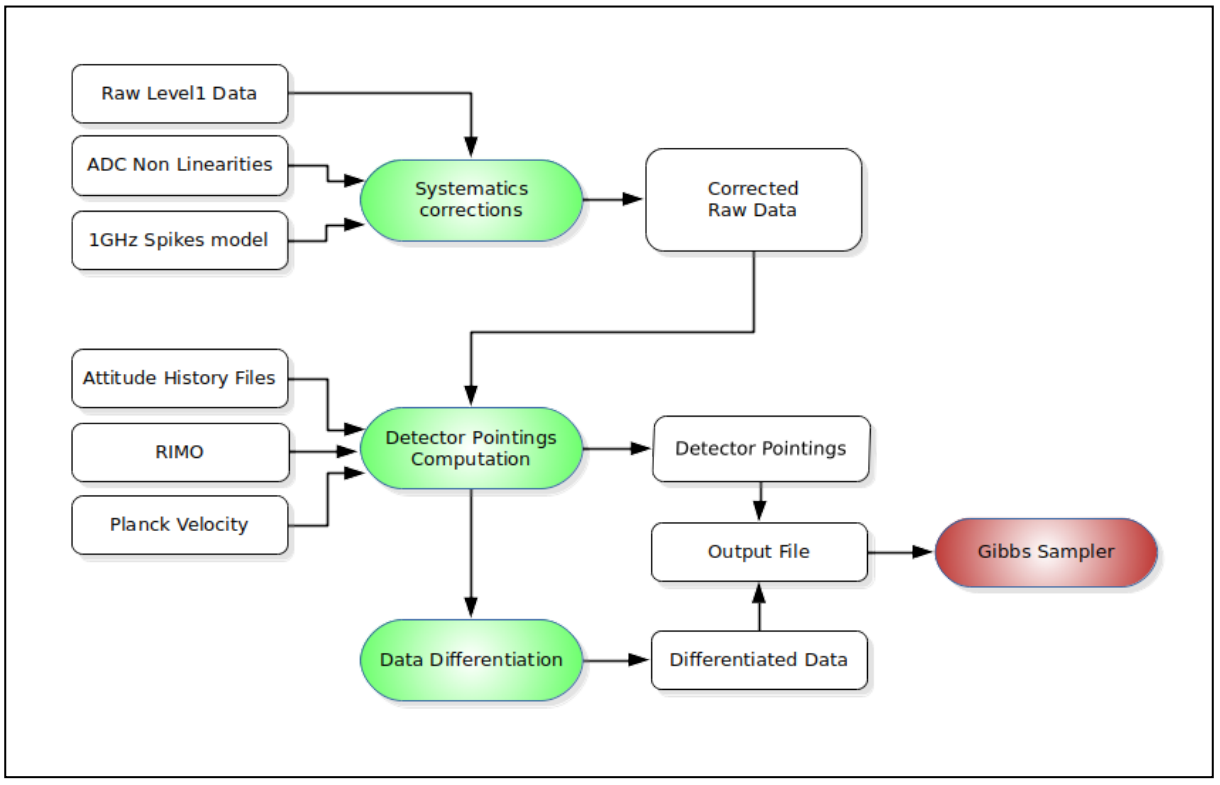

Figure 3.1. Data selection module overview

The *Data Selection* module is divided into three steps:

1. *Systematics correction*: this step transforms raw *Level 1* input data to engineering value data using the instrument models built during the Planck project and correct the data for the known instrument systematic effects [1]. In order, the two corrected systematic effects are:
    - ADC non linear response correction

- 1Hz spikes removal

2. *Detector pointing computation*: using the satellite attitude information stored in auxiliary AHF files, the instrument model and the Planck velocity file, this step computes the position in the sky (detector pointing) associated to each acquired sample.

3. *Data Differentiation*: this step is necessary to reduce the 1/f instrument noise by combining the data signal coming from the sky and the corresponding data signal coming from a reference source in the Planck satellite, as described in [2].
   To differentiate the data the Gain Modulation Factor (GMF) is computed for each PID of the 4 diode input raw data stream of a LFI radiometer. The GMF is obtained by averaging all the SKY and REF samples as:

$$GMF = \langle SKY \rangle / \langle REF \rangle$$

   The GMF factors computed are used in the process of data differentiation as:

$$DIFF_{sample} = SKY_{sample} - GMF * LOAD_{sample}$$

   The four differentiated -per-diode- datastreams are finally combined into -per-radiometer- datastreams by using a fixed weight table associated to each diode as:

$$SIGNAL_{sample} = W_0 * DIFF\_0_{sample} + (1-W_0) * DIFF\_1_{sample}$$

   The diodes are combined as:

Diode_00 + Diode_01 to obtain M datasream
Diode_10 + Diode_11 to obtain S datasream

   This is the same algorithm used for Planck-LFI data analysis.

The final output of the *Data Selection* module is a set of HDF5 files containing both the M and S differentiated data streams and the detector pointing data.
The description of the differentiated data HDF5 file typology is reported in section 3.1.

The *Data Selection* software is available in the Beyond Planck GitLab repository and it can be executed in a Beyond planck compatible computing infrastructure with the following command:

mpirun -n N DataSelection parameter.txt

where *N* is the number of parallel processes used to split the computation and *parameter.txt* is the parameter file that specify the configuration.
A README file is available in the repository with the description of the keywords used in the parameter file and the detailed description of the parameter file keywords is reported in section 3.2.
The software is completely written in C++ language and uses the same algorithms developed during the Planck project. The base algorithms have been extended by adding specific classes to handle FITS and HDF5 files to provide a portable and LFI database computing environment independent code.

The resulting differentiated and clean datasets are the main delivery of this WP and represent the base input for the calibration work package (WP 3). In Figure 2 an overview of the *Data Selection* module is reported.

3.1 Data Selection HDF5 file structure

The raw datasets extracted from Planck-LFI database are corrected for systematic effects, differentiated, detector pointing information is computed and the results are organized into HDF5 files. Each HDF5 file contains all the data for one Planck-LFI Operational Day (OD) of each LFI horn antenna.
The total number of HDF5 is 1604 OD times 11 LFI horn antennas. The file name convention used is:

LFI_FFF_HH_**L2**_OOOO.**h5**

where:

- **FFF** is the horn frequency. For LFI is 030, 044 or 070
- **HH** is the horn number going from 18 to 28
- **OOOO** is the Planck Operational Day going from 91 to 1604

Each HDF5 file is organized into three type of groups:

- **Time**: contain all the time information organized into OBT, SCET and MJD datasets.
    - OBT: On Board Time, the clock of the satellite. The time is measured in 65536 clock values per second
    - SCET: a TAI time computed from OBT and expressed in microseconds
    - MJD: the Mean Julian Date associated to the SCET
- **HHR**: HH horn per radiometer (R) acquired data. The radiometer notation is **M** for the diode 00 and 01 datastream combination and **S** for diode 10 and 11 datastream combination. There are 2 of these groups in each HDF5 file, one for each horn radiometer.
- **AHF_info**: containing all the information extracted from the Attitude History File: PID, PID_start, PID_stop:
    - PID: the unique Pointing ID, a progressive number identifying the Planck pointing
    - PID_start: the OBT when the PID begins in the OD datastream
    - PID_stop: the OBT when the PID ends in the OD datastream

Figure 3.1.1 shows the detailed structure of a differentiated data HDF5 file including the attribute keywords stored in each HDF5 used to identify the data contained. Figure 3.1.2 shows an example of a HDF5 extracted file structure opened with HDFView software [3].

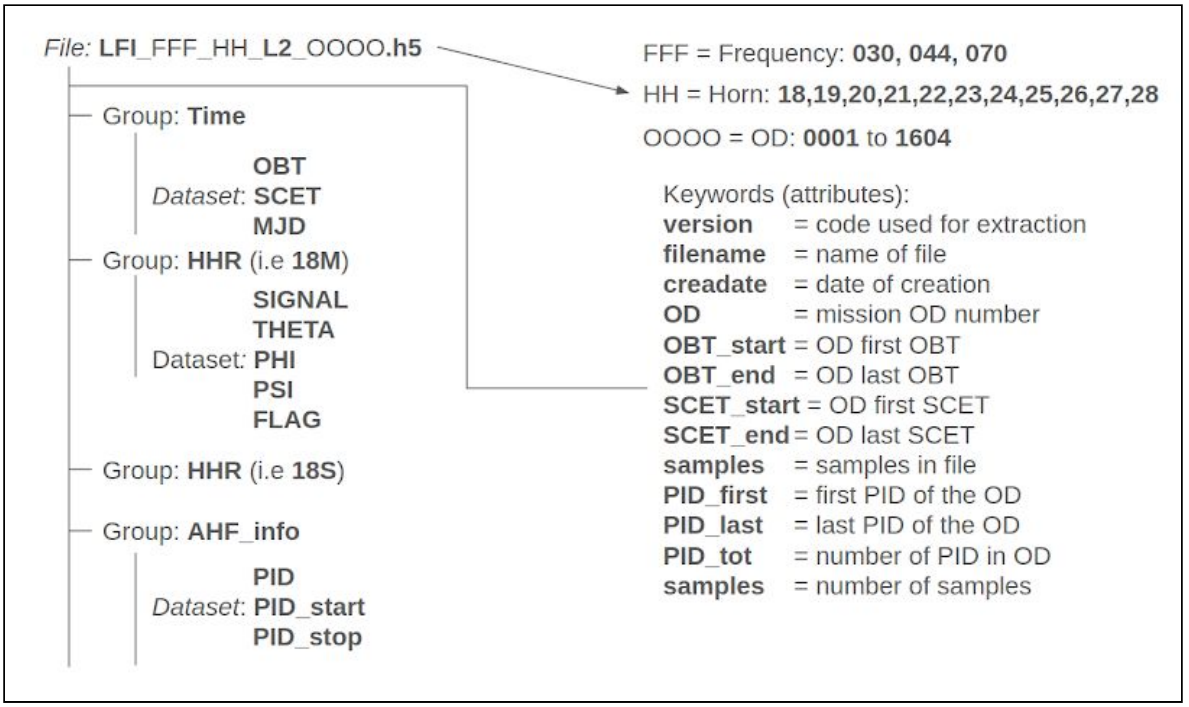

FIGURE 3.1.1. Differentiated data HDF5 file structure

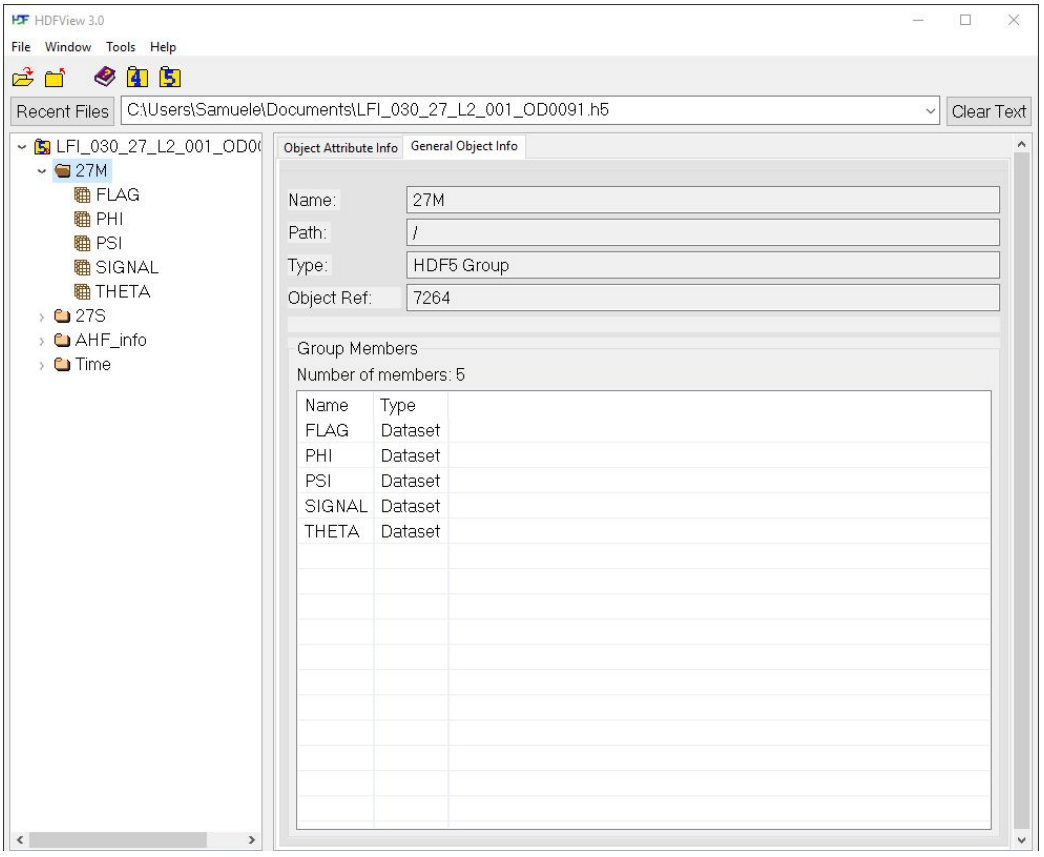

FIGURE 3.1.2. Differentiated data file structure with HDFview software

3.2 Data Selection parameter file keywords

The parameter file to run *Data Selection* contains the following parameters. For each one we provide a short description, the default value and if it is mandatory.

3.2.1 General Input

PARAMETER NAME	DESCRIPTION	DEFAULT VALUE
log_level	Logging level, in increasing order of verbosity: ERROR WARNING INFO DEBUG	INFO

star_od	Starting Operational Day (91 is the first science OD for Planck mission)	*Mandatory*
end_od	End Operational Day (1604 is the last science OD for LFI)	*Mandatory*
horn	LFI Horn to be analyzed (18 to 28)	*Mandatory*
quality_flag	Flagged data to be removed from analysis	83984

3.2.2 Auxiliary Data

PARAMETER NAME	DESCRIPTION	DEFAULT VALUE
aux_path	Complete path to the auxiliary data	*Mandatory*
velocity_file	Planck Satellite velocity taken from Horizons	satellite_velocity.fits
mask	Mask file, must be in Ecliptic coordinate and should be the same used by DataCalibration	mask_calib_30GHz.fits
weights_table	Weights for LFI diode combination	diode_weights.fits
remove_spikes	Whether or not remove spikes	*Mandatory*
spikes_basename	Basename for spikes tables following the naming convention: spikes_LFIHHR-DD.fits (HH=horn, R=M/S,	*Mandatory*

	DD=00,01,10,11). The basename ends before LFI	
adc_basename	Basename for ADC correction tables following the naming convention: adc_response_OOO_LFIHHR-DD.fits (same as spikes_basename, OOO=OD). The basename ends before OOO.	*Mandatory*
r_checkpoints_basename	Basename for table of peculiar pointing periods where to compute different Gain Modulation Factors following the naming convention: r_checkpoints_HHR.fits (same as spikes_basename). Basename ends before HH	*Mandatory*
rimo	Reduced Instrument MOdel. It contains summary of LFI characteristics (noise, beam, etc …)	LFI_RIMO_R3.31.fits

3.2.3 Attitude Files

PARAMETER NAME	DESCRIPTION	DEFAULT VALUE
ahf_path	Complete path to the Attitude History files	*Mandatory*
ahf_basename	Basename for Attitude History files following the naming convention: att_hist_high_OOOO.fits (OOOO=OD). Basename ends before OOOO	*Mandator*

3.2.4 Input Data

PARAMETER NAME	DESCRIPTION	DEFAULT VALUE
input_data_path	Complete path to Raw Level1 data	*Mandatory*
level1_data_basename	Basename for Level1 data following the naming convention: LFI_FFF_HH_L1_ODOOOO .h5 (FFF=frequency, HH=horn, OOOO=OD). Basename ends before HH	*Mandatory*

3.2.5 Output Data

PARAMETER NAME	DESCRIPTION	DEFAULT VALUE
output_data_path	Complete path to the output	*Mandatory*
level2_data_basename	Basename for L2 data following the naming convention: LFI_FFF_HH_L2_VVV_OD OOOO.h5 (same as level1_data_basename with the addition of VVV=version). Basename ends before HH	*Mandatory*
version	Version number of the output (integer)	0
subsample_basename	Basename for Level1 subsampled data (1 sample per pointing period) following the naming	*Mandatory*

	convention: LFI_subsampled_FFF_LFIHHR-DD.fits (FFF=frequency, HH=horn, R=M/S, DD=00,01,10,11). Basename ends before LFI	

4 References

[1] Planck 2013 results. III. LFI systematic uncertainties, A&A 571, A3 (2014).
[2] Planck 2013 results. II. Low Frequency Instrument data processing,A&A 571, A2 (2014)
[3] https://www.hdfgroup.org/downloads/hdfview/