LOFAR Time-series data format ICD

Lars Bähren, Andreas Horneffer, Joseph Masters

March 22, 2007

Contents

1 Introduction

1.1 Purpose and scope

This document describes the internal structure of and the interface to the LOFAR time series data. Time series data – i.e. the digitized electric field strength, as received by the individual LOFAR dipoles – represent the primary input data to the UHECR (Ultra-High-Energy Cosmic Rays) analysis pipeline(s) and have to be considered as the most basic for in which the received radio signals are present within the LOFAR system.

1.2 Context and motivation

The fundamental difference between data analysis for LOFAR/CR w.r.t. the other KSPs is the fact, that processing starts from the raw digitized time-series data delivered by the individual dipoles of the LOFAR telescope. This approach is required to provide the necessary time-resolution – essentially down to the

time-interval at which the analog signal is sampled – to detect, identify and investigate the radio pulses from Extensive Air-Showers (EAS) originating form high-enery cosmic rays.

Based on a number of considerations[1] we have chosen the HDF5 data format as common wrapper for the standard LOFAR data products (or at least a considerable fraction thereof). The goal is to create along with the definitions of the standard data product also an infrastructur which will enable LOFAR users to access and manipulate such data – this document therefore also serves as reference for the implementation with the Data Access Library (DAL).

1.3 Relationship with other interfaces

The functionality to write, access and inspect time-series data will not come from a single software components (as delivered by the USG), but requires coverage by the following modules:

- Data Acces Library (DAL) – to provide read/write access to the data, as physically located on storage media (single hard-drive, RAID array, GRID, etc.)
- Data Visualization Library (DVL) – display the structure of the data set or parts of its contents; this ranges from the graphical representation of the internal hierarchical structure of the data set to the display of e.g. the stored time-series or the antenna beamshape.
- Meta-Database Interface (MDI) – to provide read/write access to data associated with the time-series data, required for inspection, processing, etc.

A detailed description of the expected interaction of above mentioned modules is given in section 3 (*Interfaces*).

2 Organization of the data

2.1 Requirements

The data format needs to be able to handle data volumes as different as a CR event with 1ms worth of data from a few antennas only to a full dump of 1 second worth of data from all LOFAR antennas in a consistent and efficient way.

UHEP:

Time series data from the formed beam	
2 pol $\times$ 2^{27} samples $\times$ 8 Bytes/sample:	2.0 GB/event
Raw time series data from individual dipoles	
77 stations $\times$ 96 antennae $\times$ 2 pol $\times$ 2^{17} samples $\times$ 2 Bytes/sample:	3.6 GB/event
Total:	5.6 GB/event

VHECR:

Raw time series data from individual dipoles	
32 core stations $\times$ 96 antennae $\times$ 2 pol $\times$ 2^{17} samples $\times$ 2 Bytes/sample:	1.6 GB/event

HECR:

Similar, but now only one station is involved	
96 antennae $\times$ 2 pol $\times$ 2^{17} samples $\times$ 2 Bytes/sample:	48 MB/event

TS-mode

Similar to VHECR but full raw data from TBB	
77 stations $\times$ 96 antennae $\times$ 2 pol $\times$ $2 \cdot 10^8$ samples $\times$ 2 Bytes/sample:	5.9 TB/event

The event rate is uncertain and estimated at one triggered event per station per 10 minutes.

[1] Wouldn't it be better to have a reference here?

2.2 Metadata

Metadata is the auxiliary data stored along with the time series data need to provide all the necessary information for automated processing of the data. This data can either be stored directly in the data set or it can be stored in an external database. In the latter case the data set must contain a pointer to the correct entry in that database (e.g. the antenna-id is needed to get the antenna position).

- DAQ mode, including Samplerate, Filters, etc.
- Timing information:
  - Trigger time relative to recorded data segement
  - Timing of the data streams relative to the trigger
  - Timing of the data streams relative to each other with sub-sample accuracy; This can be implicit, e.g. all data streams of a station start at the same time. Fields that can be in stored in an external database.
- List of RFI sources identified by the station calibration, including **direction**, **center frequency** and **peak strength**.
  → What does this actually mean: the properties of the single channel containing the highest signal level or the parameters obtained from fitting e.g. a Gaussian to a segment of the spectrum?
- dispersion measure of the ionosphere (at this point in time and space)
- health information about the antennas

System monitoring and system health. Information on the status of the various (hardware) components at a LOFAR station will be stored inside the PVSS database; a description of the datapoint-types and datapoints can be found in the `MAC/Deployment/data/PVSS` branch of the LOFAR code repository (see Tab. 1 for an excerpt).

In order to later store certain system health information along with the other data, parameters need to be subscribed to at the definition of the observation.

2.3 Structure and contents of the data

2.3.1 Station group

Given the different modes planed for cosmic ray observation, a single LOFAR station appears to be the natural choice for a first grouping of time-series data from the individual dipoles; for that matter we consider the **station group** (Tab. 2) as a basic module within the data structure.[2] Creating a snapshot of multiple stations, or even the full LOFAR array, thus will result in a set of station groups – which in turn might be collected into another superstructure.

The main purpose of this group is to serve as a common container for the separate sub-tables, which take up data from the station calibration and the trigger algorithm; the main motivation for this design is to be able to more efficiently distribute the contents of the data set. Especially the calibration information might not physically reside in the same location as the rest of data – calibration information might be interactively extracted from a calibration data-base, whether being a central one or a local snapshot.

The following entries will be found in the station group (Tab. 2, p. 6):

- One of the most obvious informations to store is the **name of the telecope** (`TELESCOPE`) with which the observation was carried out – i.e. *LOFAR*.
- `OBSERVER` holds the **name(s) of the observer(s)**.
- If the observation is carried out within the context of a specific **project**, then its name will be stored in `PROJECT`. If no specific project is defined, the variable simply should be set to *LOFAR/CR*.

[2] Though from initial perception the described structure well can be perceived as a table, the HDF5 internal data model is that of a group; in order to stick as closely as possible to the libraries naming conventions, we therefore use the name *group* instead of *table*.

Database entry	Field	Format
CalCtlr	connected	bool
	obsname	string
	antennaArray	string
	filter	string
	nyquistzone	int
	rcus	string
ObservationControl	claimPeriod	int
	preparePeriod	int
	startTime	string
	stopTime	string
	subbandList	string
	beamletList	string
	bandFilter	string
	nyquistzone	int
	antenneArray	string
	receiverList	string
	sampleClock	int
	measurementSet	string
	stationList	string
	inputNodeList	string
	BGLNodeList	string
	storageNodeList	string

Table 1: Excerpt from the list of entries into the PVSS database. The definitions of the datapointtypes and datapoints can be found in the `MAC/Development/data/PVSS` branch of the LOFAR code repository.

- `OBSERVATION_ID` is the **unique identifier** for the LOFAR observation.
- The **observation mode** (`OBSERVATION_MODE`) at which the data where recorded.

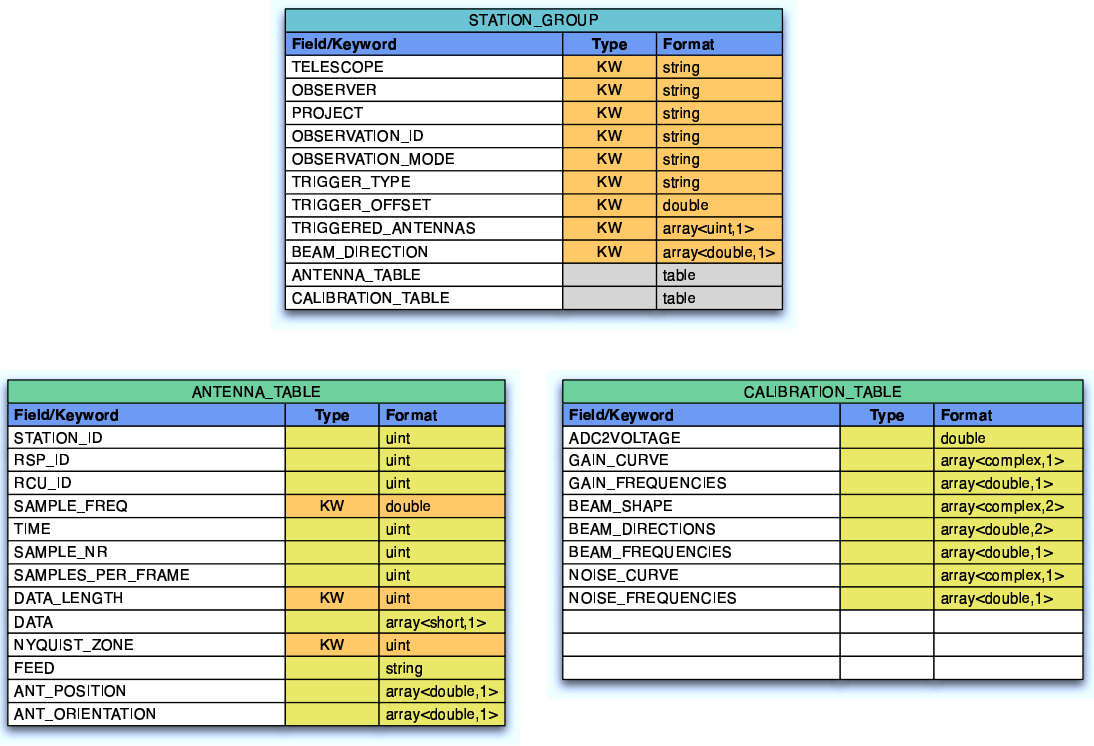
Even though there exist multiple LOFAR observation modes for cosmic rays, all have in common a (multi-level) pulse-detection and trigger-generation algorithm; the control parameters of the trigger algorithms as well as its output (in case a trigger condition was derived) needs to be stored.

- The **type of the CR trigger** (`TRIGGER_TYPE`) will depend on the observation mode, i.e. VHECR, HECR, UHEP; either the original trigger was generated at antenna, station or array level (for more details see the description of the CR processing pipelines).
- `TRIGGER_OFFSET`
- In the VHECR mode the first-level trigger is generated at dipole level; thereby we need to store the information for which **set of antennas** (`TRIGGERED_ANTENNAS`) the trigger condition was fullfilled. The shape of the vector will depend on the number of antennas for which the triggger condition was fullfilled – the maximum length is the number of antennas within a station (96).
- `BEAM_DIRECTION`

2.3.2 Antenna table

The **antenna table** (Tabb. 3) collects data on a per-dipole basis[3] – starting from the identifiers required for the unambiguous identification of an individual dipole within the full LOFAR network to the actual sampled wave-form of the EM-field at the position of each antenna feed.

[3]Please keep in mind here, that we clearly distinguish between *antenna* and *dipole/feed*: using the feed-based approach as underlying the Measurement-Equation, an antenna can consist of multiple feeds (or dipoles).



STATION_GROUP		
Field/Keyword	**Type**	**Format**
TELESCOPE	KW	string
OBSERVER	KW	string
PROJECT	KW	string
OBSERVATION_ID	KW	string
OBSERVATION_MODE	KW	string
TRIGGER_TYPE	KW	string
TRIGGER_OFFSET	KW	double
TRIGGERED_ANTENNAS	KW	array<uint,1>
BEAM_DIRECTION	KW	array<double,1>
ANTENNA_TABLE		table
CALIBRATION_TABLE		table

ANTENNA_TABLE		
Field/Keyword	**Type**	**Format**
STATION_ID		uint
RSP_ID		uint
RCU_ID		uint
SAMPLE_FREQ	KW	double
TIME		uint
SAMPLE_NR		uint
SAMPLES_PER_FRAME		uint
DATA_LENGTH	KW	uint
DATA		array<short,1>
NYQUIST_ZONE	KW	uint
FEED		string
ANT_POSITION		array<double,1>
ANT_ORIENTATION		array<double,1>

CALIBRATION_TABLE		
Field/Keyword	**Type**	**Format**
ADC2VOLTAGE		double
GAIN_CURVE		array<complex,1>
GAIN_FREQUENCIES		array<double,1>
BEAM_SHAPE		array<complex,2>
BEAM_DIRECTIONS		array<double,2>
BEAM_FREQUENCIES		array<double,1>
NOISE_CURVE		array<complex,1>
NOISE_FREQUENCIES		array<double,1>

Figure 1: Organization of the group holding the data per LOFAR station (Tab. 2); besides holding a few parameters common to all antenna elements within the station, this table serves as a container for three sub-tables, containing the raw antenna data (Tab. 3), calibration information (Tab. 4) and the output of the trigger algorithm which has initiated the data dump.

– `STATION_ID`, `RSP_ID` and `RCU_ID` are directly taken from the frame structure used in the communication between RSP and TBB [4]. The three identifiers in combination allow for an unambiguous identification of an individual dipole within the LOFAR network; depending on the range of value of the individual numbers the unique ID may be constructed via e.g.

$$N_{\mathrm{ID}} = 10^4 \cdot N_{\mathrm{Station}} + 10^2 \cdot N_{\mathrm{RSP}} + N_{\mathrm{RCU}} \tag{1}$$

– The combination of the two fields `TIME` and `SAMPLE_NR` gives an absolute time reference for the first sample in the `DATA` field. The `TIME` field gives a time offset in seconds from a certain start moment, where the LCU is completely free at chossing a time system, such as UNIX time[4]. For a constant sampling frequency (`SAMPLE_FREQ`) the timing for the remaining set of samples can derived via [4]

$$t[n] = t_{\mathrm{TIME}} + \left(t_{\mathrm{SAMPLE\_NR}} + n\right) \cdot 1/\nu_{\mathrm{SAMPLE\_FREQ}} \tag{2}$$

where n is the index for a sample in the `DATA` vector.

– Each frame of data transferred between RSP and TBB has the same fixed length (`SAMPLES_PER_FRAME`), but a frame may hold any number of samples that will fit in the payload area of the frame.

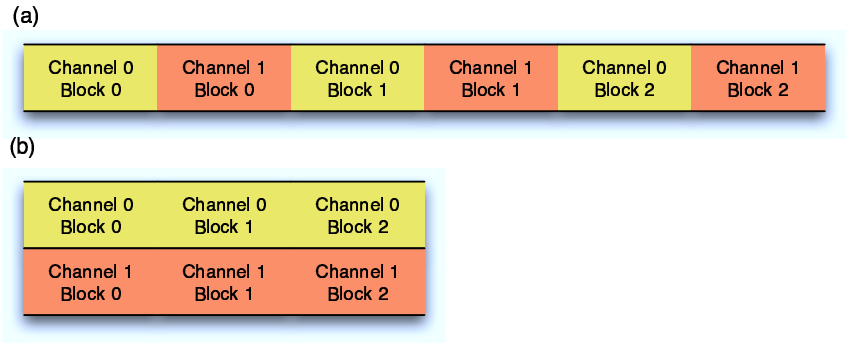
– `DATA` stores the raw ADC output for an individual signal path/dipole, consisting of `DATA_LENGTH` samples for a single dump of TBB data; the length of this data vector will vary depending on the observation mode (see list in Sec. 2.1).
It should be kept in mind how the data of a pair of dipoles is organized in the data frame structure

[4]Referring to the UINX time, this field would hold the number of seconds since 1970

Field/Keyword	Format	Unit	Description
TELESCOPE	KW/string	—	Name of the telescope
OBSERVER	KW/string	—	Name(s) of the observer(s)
PROJECT	KW/string	—	Project code/name
OBSERVATION_ID	KW/string	—	Unique identifier for the observation
OBSERVATION_MODE	KW/string	—	Observation mode (i.e. Mode 1: 30–90MHz, Mode 2: 120–190MHz etc.)
TRIGGER_TYPE	KW/string	—	The kind of trigger that triggered this observation (UHEP/VHECR/HECR?)
TRIGGER_OFFSET	KW/double	sec	Trigger time relative to the reference time
TRIGGERED_ANTENNAS	KW/array<uint,1>	—	$[N_{\mathrm{trig.Ant}}]$ List of the triggered antenna (VHECR mode); vector of variable length.
BEAM_DIRECTION	KW/array<double,1>	deg	[2] Direction of the station beam, on which the CR observation was piggy-backing.
ANTENNA_TABLE	table	—	Table: entries extracted from the frame structure used in the transfer between RSP and TBB
CALIBRATION_TABLE	table	—	Table: calibration information as delivered through the online (station) calibration

Table 2: Fields in the station data group (STATION_GROUP). The main purpose of this group of to serve as a common container for the separate sub-tables, which take up data from the TBB, the station calibration and the trigger algorithm. Shapes of vector and matrices are given in [ ]-brackets in the description. See text for detailed explanation on the individual fields in the table.

used in the transfer between TBB and RSP (see Fig. below). In order to achieve maximum data packing density, the 12-bit encoded signals from the RCU-channels are placed front-to-back within the payload segment of the data frame. This arrangement should be reorganized before storage into the time-series data set!



- the **position** (ANT_POSITION) of each receiving element, $\vec{x} = (x_1, x_2, x_3)$
- the **orientation** (ANT_ORIENTATION) of each receiving element, $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$. Of course we should try to minimize the number of parameters used for reference frame transformations; if we – and this seems reasonable – can assume the ground of a single LOFAR station as flat, we can reduce the rotation parameters to a single angle decribing a possible rotation of the antenna w.r.t. the (x, y) axes of the station coordinate reference frame.

2.3.3 Calibration table

The **calibration table** collects all the information required for the proper calibration of the recorded data. Since all cosmic ray observation modes make use of the data as they are available directly after the

Field/Keyword	Format	Unit	Description
STATION_ID	uint	—	Data source station identifier
RSP_ID	uint	—	Data source RSP board identifier
RCU_ID	uint	—	Data source RCU board identifier
SAMPLE_FREQ	double/KW	Hz	Sample frequency in MHz of the RCU boards
TIME	uint	sec	Time instance in seconds of the first sample in the payload
SAMPLE_NR	uint	—	Sample number of the first payload sample in current seconds interval in transient mode [4]
SAMPLES_PER_FRAME	uint	—	Total number of samples in the payload of the original TBB–RSP frame structure
DATA_LENGTH	uint/KW	—	The number of samples per dipole which actually stored into the data set; this might as well be different from the number of samples in a data frame
DATA	array<short,1>	counts	[DATA_LENGTH] Raw ADC output (the time-series)
NYQUIST_ZONE	uint/KW	—	Nyquist zone in which the data are sampled
FEED	string	—	Type of feed for this dipole
ANT_POSITION	array<double,1>	m	[3] Antenna position w.r.t. the station center, $\vec{x} = (x_1, x_2, x_3)$
ANT_ORIENTATION	array<double,1>	deg	[3] Orientation of the antenna w.r.t. the station reference frame, $\vec{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$

Table 3: Fields in the ANTENNA_TABLE subtable; each listed field corresponds to a column in the table, where the number of rows corresponds to the number of dipoles. The first set of values is adopted directly from the frame structure used for data transfer between TBB and RSP [4].

digitization step, the further processing incorporated into the data products delivered for other observation modes will need to be applied as part of the offline-analysis.

- The first conversion step in the processing will be to turn the recorded ADC counts into voltages; for this matter we need the scale (ADC2VOLTAGE) relating both quantities. These values are expected to be rather stable as function of time, though dependency external or system factors cannot be excluded.
- The **complex electronic gain** (GAIN_CURVE) of each receiving element as function of frequency (bandpass), $\mathbf{G}_{\text{gain}} = G(\nu)$; this array<complex,1> will be multiplied (after interpolation, if required) to the output of the Fourier transform of the dipole voltage time-series. Keep in mind, that the frequency range might be covered by non-equidistantly separated points, such that a array<double,1> per antenna will be required for the GAIN_FREQUENCIES.
- BEAM_SHAPE holds the complex **element beam pattern** as function of direction and frequency $\mathbf{G}_{\text{beam}} = G(\vec{\rho}, \nu)$. In order to correctly interpret the data – and, if necessary, interpolate – the corresponding **beam directions** (BEAM_DIRECTIONS) and **frequency values** (BEAM_FREQUENCIES) are required.
- NOISE_CURVE holds the system noise as function of frequency, $\mathbf{N}_{\text{system}} = N(\nu)$, where the frequencies are stored in NOISE_FREQUENCIES

Since CR data processing will require to perform all the calibration steps normally applied after passing the data through the transient buffer board (TBB) within the LOFAR system, availability of station calibration is vital; most aforementioned information are listed as central products/parameters to the LOFAR station calibration [5].

Field name	Format	Units	Description
ADC2VOLTAGE	double	Voltage	Factors for conversion from raw ADC counts to voltages
GAIN_CURVE	array<complex,1>	—	$[N_{\text{Channels}}]$ Complex electronic gain as function of frequency
GAIN_FREQUENCIES	array<double,1>	Hz	$[N_{\text{Channels}}]$ Frequency values for which the gain electronic gains are provided
BEAM_SHAPE	array<complex,2>		$[N_{\text{Directions}}, N_{\text{Channels}}]$
BEAM_DIRECTIONS	array<double,2>	deg	$[N_{\text{Directions}}, 2]$ Directions, for which the beam-shape is provided
BEAM_FREQUENCIES	array<double,1>	Hz	$[N_{\text{Channels}}]$ Frequencies for which the beam-shape is provided
NOISE_CURVE	array<complex,1>	—	$[N_{\text{Channels}}]$ system noise as function of frequency
NOISE_FREQUENCIES	array<double,1>	Hz	$[N_{\text{Channels}}]$ Frequencies for which the system noise is sampled

Table 4: Columns in the CALIBRATION_TABLE subtable; ; each listed field corresponds to a column in the table, where the number of rows corresponds to the number of dipoles.

2.4 Open questions

1. Are there modes forseens, in which the total LOFAR array is being split up into sub-arrays, which operate in different modes, e.g. the Compact Core using the high-band antennas, while the remote stations are observing with the LBAs? In such a case the *range of application* of some of the metadata keywords would change; in order not to shift keywords within the data structure we therefore will end up with redundant information, depending of the specific observation mode. The latter though will not pose a major problem, since this redundance will show up in non-datasize critical keywords, such e.g. OBSERVTAION_MODE or NYQUIST_ZONE.

2. How do the various components running at station level react to **holes or bit errors in the data streams** coming from the RCUs? Data frames should not be dropped in any case, but flagged if an error is detected; the peak detection running on the TBB should not come to halt but continue analysing the incoming buffer frames. A possible false alarm triggered by a single channel most likely will be cancelled out in the subsequent coincidence stage.

   → Check with station engineering group how RSP and TBB would behave in the case described above. If necessary ask for adjustment!

3. Which parameters will be there to describe the **trigger condition**?

4. What about distance dependency of the **element beam pattern**? Calibration on astronomical sources naturally will provide the far-field beam pattern, but be also might be interested in the transition from near-field to far-field – is this accessable via the EM simulations?
   Will the beam pattern be provided as mean element beam pattern or in fact as beam pattern for each individual dipole? If deviations from the mean pattern are larger than 1%, the latter is required.

5. Which data points will be generated for and distributed by the system health monitoring? While indeed there is a list of parameters available from the database description files (e.g. RSPBoard.dpdef), are more detailed explanation of the stored contents is required in order to make a proper selection of values which are about to be stored long with a CR time-series dataset.

6. In which form/format will the information on the dispersion measure be provided?

3 Interfaces

3.1 Interface requirements

t.b.d.

3.2 Relation to existing workpackages

3.2.1 Data Access Library (DAL)

The DAL must be able to handle retrieval of the data representations as used throughout the data analysis pipeline:

- ADC values as function of time
- Voltage as function of time
- Raw FFT of the voltage time-series
- Calibrated FFT
- RFI-filtered FFT
- Cross-Correlation Spectra (either from raw or calibrated FFT)
- Visibilities

From its design it becomes clear, that most likely the functionality of the DAL will be limited to the basic handling of the data I/O, whereas generation of the derived data products most likely is to become the responsibility of another processing layer.

While in most usage scenarios all the valid data from a LOFAR station will be read in to be processed together, their might be the need to select data from antenna across the borders of a LOFAR station: this will require the ability to access data based on the geographical locations, such as e.g.

- all antennas within a N Kilometer radius of the shower core
- all antennas within a sector of M degrees opening angle towards a given direction w.r.t. the shower core
- all antennas located in a ring of $N_1 < R < N_2$ around the position of the shower core

3.2.2 Data Visualization Library (DVL)

Past experience with the software for the LOPES experiment has shown, that is is crucial to provide the user with a variety of way to graphically inspect the data. This not only includes visualization of the time-series/FFT/etc. data themselves, but also displaying the various data with the wider context of the experimental setup (e.g. geographical distribution of the antennas w.r.t. to the particle detector setup)

- Display of the standard data products (also see documentation of the LOPES–`Tools DataReader`): ADC, Voltage, FFT, Calibrated FFT, RFI-filtered FFT, Cross-Corr. Spectra, Visibilities
- Display of the (intermediate) data products generated from the input data, e.g. dynamic spectra, multi-dimensional skymaps
- Flags and weights associated with the data, e.g. filter curves, antenna gain curves, etc.
- Station layout, i.e. positions of the (selected/excluded) antennas
- antenna power level distribution over the area of the station/array (this is very similar to the type of event display as known from particle physics experiments)
- mapping of the (local) RFI via an (Azimuth,Frequency) plot centered on the position of a certain station

- geographical locations of identified RFI sources; such a plot should also indicate the frequency band of the RFI (via label or bar etc.)
- geographical distribution/location of antennas which generated a trigger signal, failed, etc.

  → VR setting, combining geographical information (e.g. map of the Netherlands) with the location of localized CR-/TS-events

  ⇒ in a cave-like setup we actually can perform a fly-through, which would be ideal for outreach purposes!
- cummulative geographical distribution of detected CR events
- total power per LOFAR station (geographically distributed)
- overlay of CR data with information from other sensors (e.g. weather radar images, temperatures, etc.)

A number of the before-mentioned displayes should be interactive, in the sense that the user should be able to perform data selection from the graphical display (e.g. by drawing a circle around the core of the CR air shower, thereby selecting the antennas included in the data analysis step).

3.3 Open Questions

Glossary of terms

ADC	Analog to Digital Conversion
DAL	Data Access Library
DVL	Data Visualization Library
EAS	Extensive Air-Shower
HECR	High-Energy Cosmic Rays
LOPES	LOFAR Prototype Station
RCU	Receiver Unit
RSP	Remote Station Processing Board
TBB	Transient Buffer Board
TS	Thunderstorm
UHECR	Ultra-High-Energy Cosmic Rays
UHEP	Ultra-High-Energy Particle
USG	User Software Group

References

[1] A. J. Boonstra & M. Tanigawa (2006) LOFAR Interference Mitigation Approach and Measurement Results, LOFAR-ASTRON-RPT-104

[2] HDF5 home page, `http://hdf.ncsa.uiuc.edu/HDF5`

[3] A. Horneffer (2006) PhD Thesis, Uni. Bonn

[4] W. Poiesz (2006) TBB Design Document, LOFAR-ASTRON-SDD-047

[5] S. J. Wijnholds (2006) Station calibration, Pitfalls and possibilities 2006, LOFAR-ASTRON-MEM-217

[6] M. Wise et al (2006) LOFAR User Software Plan, ASTRON, Dwingeloo