LOFAR Data Format ICD
TBB Time-Series Data

Document ID: LOFAR-USG-ICD-001

Version 2.02.12

SVN Repository Revision: 9366

L. Bähren, K. Anderson, A. Corstanje, A. Horneffer, J. Masters

SVN Date: 2012-01-10

Contents

Change record

0.1 Version 1.x

Notes for Version 1.0 → Version 1.1

1. *Reorganization of the placement of attributes describing basic properties of the observation from which the dataset has originated.*
   Some part of the information which originally was supposed to be stored within the Station groups should be shifted upwards to the root group of the file.

   old: /
   ```
   |-- STATION_001                    ... Group
   |   |-- TELESCOPE                  ... Attribute      ... string
   |   |-- OBSERVER                   ... Attribute      ... string
   |   |-- PROJECT                    ... Attribute      ... string
   |   |-- OBSERVATION_ID             ... Attribute      ... string
   |   |-- OBSERVATION_MODE           ... Attribute      ... string
   |   `
   |-- STATION_002                    ... Group
   |   |-- TELESCOPE                  ... Attribute      ... string
   |   |-- OBSERVER                   ... Attribute      ... string
   |   |-- PROJECT                    ... Attribute      ... string
   |   |-- OBSERVATION_ID             ... Attribute      ... string
   |   |-- OBSERVATION_MODE           ... Attribute      ... string
   |   `
   ```
   new: /
   ```
   |-- TELESCOPE                      ... Attribute      ... string
   |-- OBSERVER                       ... Attribute      ... string
   |-- PROJECT                        ... Attribute      ... string
   |-- OBSERVATION_ID                 ... Attribute      ... string
   |-- OBSERVATION_MODE               ... Attribute      ... string
   |-- STATION_001                    ... Group
   |   |
   |   `
   |-- STATION_002                    ... Group
   |   |
   |   `
   ```

2. *Proper handling of coordinates*
   The motivation for this changes are in the fact, that for proper representation of a direction or position information more but just the numerical value is required – thus the additional information should be stored within the file as well, as compared to simply assuming the stored data adhere to a certain convention (which on the other hand they still should). A possible coordinate reconstruction scheme is shown in the figure below.

   old: /
   ```
   |-- STATION_001                    ... Group
   |   |-- BEAM_DIRECTION             ... Attribute      ... array<double,1>
   |   |-- 001000000                  ... Dataset
   |   |   |-- ANTENNA_POSITION       ... Attribute      ... array<double,1>
   |   |   |-- ANTENNA_ORIENTATION    ... Attribute      ... array<double,1>
   |   |   |-- SAMPLE_FREQUENCY       ... Attribute      ... double
   |
   ```
   new: /
   ```
   |-- STATION_001                    ... Group
   |   |-- STATION_POSITION_VALUE     ... Attribute      ... array<double,1>
   |   |-- STATION_POSITION_UNIT      ... Attribute      ... string
   |   |-- STATION_POSITION_FRAME     ... Attribute      ... string
   ```

```
|   |-- BEAM_DIRECTION_VALUE           ... Attribute       ... array<double,1>
|   |-- BEAM_DIRECTION_UNIT            ... Attribute       ... string
|   |-- BEAM_DIRECTION_FRAME           ... Attribute       ... string
|   |-- 001000000                      ... Dataset
|   |   |-- ANTENNA_POSITION_VALUE     ... Attribute       ... array<double,1>
|   |   |-- ANTENNA_POSITION_UNIT      ... Attribute       ... string
|   |   |-- ANTENNA_POSITION_FRAME     ... Attribute       ... string
|   |   |-- ANTENNA_ORIENTATION_VALUE  ... Attribute       ... array<double,1>
|   |   |-- ANTENNA_ORIENTATION_UNIT   ... Attribute       ... string
|   |   |-- ANTENNA_ORIENTATION_FRAME  ... Attribute       ... string
|   |   |-- SAMPLE_FREQUENCY_VALUE     ... Attribute       ... double
|   |   |-- SAMPLE_FREQUENCY_UNIT      ... Attribute       ... string
|
```

3. *Sensible (default) values*
   One of the – at least temporary – problems is, that some of the values to be stored within the HDF5 file are not available (yet) at the time of creating the file on disk. Therefore at least sensible default values/settings should be used to enable correct interpretation of the dataset; e.g. position information at a later point will be retrieved from the central parameter database, but until then the attributes should be filled already with placeholder values which agree with the conventions of the Measures framework. If a value in fact is undefined, it should be clearly marked as such by using `UNDEFINED` (in case of a string valued attribute).

```
/
|-- STATION_001
|   |-- STATION_POSITION_VALUE         ... array<double,1>   ...  {x,y,z}
|   |-- STATION_POSITION_UNIT          ... string            ...  "m"
|   |-- STATION_POSITION_FRAME         ... string            ...  "ITRF"
|   |-- BEAM_DIRECTION_VALUE           ... array<double,1>   ...  {0,90}
|   |-- BEAM_DIRECTION_UNIT            ... string            ...  "deg"
|   |-- BEAM_DIRECTION_STRING          ... string            ...  "UNDEFINED"
|   |-- 001000000
|   |   |-- ANTENNA_POSITION_VALUE     ... array<double,1>   ...  {x,y,z}
|   |   |-- ANTENNA_POSITION_UNIT      ... string            ...  "m"
|   |   |-- ANTENNA_POSITION_FRAME     ... string            ...  "ITRF"
|   |   |-- ANTENNA_ORIENTATION_VALUE  ... array<double,1>   ...  {x,y,z}
|   |   |-- ANTENNA_ORIENTATION_UNIT   ... string            ...  "m"
|   |   |-- ANTENNA_ORIENTATION_FRAME  ... string            ...  "ITRF"
|   |   |-- FEED                       ... string            ...  "UNDEFINED"
|   |   |-- NYQUIST_ZONE               ... uint              ...  1
|   |   |-- SAMPLE_FREQUENCY_UNIT      ... string            ...  "Hz"
|
```

Version 1.1 → Version 1.2

1. *Parametrization of coordinates*
   While the basic scheme for encoding and later reconstruction of the coordinate information remains unaltered, a minor detail was overlooked in the previous revision: in the general case one can not assume, that all values of a coodinate are given in the same physica unit. While e.g. for a simple direction – as described by two angles – the units are identical, this is not longer the when e.g. describing a position on the surface of the Earth (e.g. the position of the telescope). The simplest example for the latter case is the WGS84 system: in this a position is described by two angles and a height relative to the model geoid – therefore using `[deg,deg,m]` as units – hence the required modification in the data format to take this into account.

   old: `/`

```
|-- STATION_001                      ... Group
|   |-- STATION_POSITION_UNIT        ... Attribute       ... string
|   |-- BEAM_DIRECTION_UNIT          ... Attribute       ... string
```

```
|   |-- 001000000                       ... Dataset
|   |   |-- ANTENNA_POSITION_UNIT       ... Attribute       ... string
|   |   |-- ANTENNA_ORIENTATION_UNIT    ... Attribute       ... string
|
```

new:

```
/
|-- STATION_001                         ... Group
|   |-- STATION_POSITION_UNIT           ... Attribute       ... array<string,1>
|   |-- BEAM_DIRECTION_UNIT             ... Attribute       ... array<string,1>
|   |-- 001000000                       ... Dataset
|   |   |-- ANTENNA_POSITION_UNIT       ... Attribute       ... array<string,1>
|   |   |-- ANTENNA_ORIENTATION_UNIT    ... Attribute       ... array<string,1>
|
```

As a consequence the values actually stored as attributes can look something like:

```
/
|-- STATION_001
|   |-- STATION_POSITION_VALUE         ... array<double,1>    ...  {10,-6,50}
|   |-- STATION_POSITION_UNIT          ... array<string,1>    ...  {"m","deg","deg"}
|   |-- STATION_POSITION_FRAME         ... string             ...  "WGS84"
|   |
```

Version 1.2 → Version 1.x

```
BEAM_WIDTH_VALUE
BEAM_WIDTH_UNIT
```

Version	Date	Sections	Description of changes
1.00.00	yyyy-mm-dd	all	Attribute reorganization, “sensible default values”.
1.01.00	yyyy-mm-dd	all	Coordinate parametrization.
1.02.00	2008-09-08	–	Beam parameters.
1.03.00	2008-09-08	–	–/–
1.04.00	2009-07-08	all	Reorganization wrt. common ICD format.
1.05.00	2009-09-15	3, 4.5	Updated high-level structure; new section on trigger table.
1.06.00	2010-01-26	1, 4, 6	Added summary of data volumes; update to trigger and calibration information; open questions now as table.
1.07.00	2010-02-03	4.5	Cleaning up of section on trigger table; extended list of table columns and description. Removed mentioning of Data Visualization Library (DVL).
1.08.00	2010-03-16	all	Merging various comments on the previous version of the ICD. Completely reworked section on station trigger.
1.09.00	2010-04-14	4.3, 6	Added attribute to station group. Added proposal for how to store calibration information as part of the station and dipole group; suggesting to replace simple dipole dataset by dipole group to take up both time-series data and calibration data.
1.10.00	2010-04-20	4/4.1	Refactor Root Group Sec. 4.1 → sec. 4.1.1, 4.1.2.
1.11.00	2010-05-14	3, 6.2	Integration suggested changes to dipole data structure into the main document.
1.12.00	2010-06-29	**??**, **??**	Rewrite of section describing storage of calibration information. Added references.
2.00.00	2010-07-08	Cover	Changed ‘revision‘ to ‘version‘; updated this version number to 2.00.00 for LOFAR ICDs 1 through 7 to put them on the same version numbering scheme.

continued on next page

continued from previous page

Version	Date	Sections	Description of changes
2.01.00	2010-07-14	??, ??	Added *Acknowledgements* section. Adjusted version numbering scheme to include zero-padding. Added *clock offset* attribute. Adusted type of fit parameters in the *Station Trigger Group*.
2.01.01	2010-12-07	all	Using LaTeX package `hyperref` for references, enabling better navigation through the document and access to external resources.
2.01.02	2011-01-05	cover	the document Id is HARD. not svnInfo.
2.01.03	2011-02-17	all	Adding document title to page footer.
2.01.04	2011-02-28	all	Fixed data volume estimates, fixed naming conventions for Groups and Datasets to UpperCamelCase.
2.01.05	2011-03-10	all	Maintain list of references through BibLaTeX database.
2.01.06	2011-03-01	4	Added root group to hold trigger-specific information.
Version upon which the current data-reader is based* .			
2.02.01	2011-04-20	all	Strings and string arrays: useage now consistent.
2.02.02	2011-05-12	all	Shifted all 2.01 and earlier changes to the'detailed change log' in the appendix.
		3	Added a 'data flow' placeholder, and switched the order of the 'overview' and 'heirarchical structure' subsections.
		4	Adjusted the 'coordinates subgroup' section in table 1.
		4	Removed the group-type for 'DumpMetaData' and replaced with 'string'.
		4	Added 'LORA' as a possible 'DUMP_TYPE' value.
		4	Ensured all Fields/Keywords are caps only.
		4	Changed dimensionality of `BEAM_SHAPE_DATA` to 3 (was 1). Everything is now 'BEAM_SHAPE'/'BeamShape'/'beam shape'.
		3 & 4	Station Trigger Group moved to the 'root' directory.
		all	Removed all but the first 'posix-style hierachy' from the document.
		3,4	Moved VHECR-specific component of 'station trigger group' to the root-level metadata.
2.02.07	2011-06-09	0,5, A	Re-organised appendices and included full change-log.
2.02.08	2011-07-06	all	Matching up group type attributes and notation; consolidation of labels to refer to standard sections and tables.
2.02.09	2011-09-15	all	Rework after review.
2.02.10	2011-12-21	0	Added data type section.
		4	Added metadata introduction.
2.02.11	2012-01-03	4	Added 'Manual' as keyword for trigger type.
2.02.12	2012-01-03	4	Update of optional keywords which are now in italic.

*NOTE: P. Schellart used this document to design a data-reading program as of March/April 2011. Hence, I am designating everything from May onwards as 'v2.02'. Only changes from v2.01 to v2.02 are noted here — everyhting else lies in the 'detailed change log' in the appendix.

Version numbering scheme In order to track the evolution of the format specification documents the following numbering scheme has been adopted:

```
<major version>.<minor version>.<patch version>
  [0..]        .  [0..99]      .  [0..99]
```

where

- the `<patch version>` is getting incremented on changes to the document, which do not affect the actual contents of the file (such as when changing attribute names and such), e.g. correcting/augmenting descriptions, adding examples, etc.
- The `<minor version>` tracks minor changes to the actual content of the file, such as renaming, adding or removing attributes.
- The `<major version>` indicates major changes with in the file format, such as reorganization of the internal hierarchical structure or official release to the public.

Standard Data Types. The following table describes the short-form name for each type used throughout the rest of this document, it's logical meaning in the context of the astronomical data product, and the physical storage which must be allocated to it within the HDF5 data model. Future versions of this document may augment these types but will not remove support for existing types.

Name	Logical Type	Physical Storage Allocation
`short`	Integer; range -2^{15} to $2^{15}-1$	16 bit signed two's complement integer
`int`	Integer; range -2^{31} to $2^{31}-1$	32 bit signed two's complement integer
`unsigned`	Integer; range 0 to $2^{32}-1$	32 bit unsigned integer
`float`	Single precision floating-point	IEEE 754-2008 [**?**] "binary32" floating point (1 bit sign, 8 bit exponent, 23 bit mantissa)
`double`	Double precision floating-point	IEEE 754-2008 "binary64" floating point (1 bit sign, 11 bit exponent, 52 bit mantissa)
`complex<type>`	Complex form of `type`	Compound type; the part at the lower memory location is real
`bool`	Boolean true/false	32 bit signed two's complement integer; non-zero denotes "true"
`string`	Text	Null-terminated string of 8 bit bytes. The lower 128 values interpreted as ASCII encoded characters
`array<type,N>`	Array of `type` with rank `N`	

Note that data may be written with either "big-endian" or "little-endian" byte ordering; either is valid within the context of this document.

Notation.

Symbol	Description
a, A	Italic lower and upper case chracters denote scalars.
$\mathbf{a}$	Bold lower case characters denote column vectors.
$\mathbf{A}_{[L,M]}$	Bold upper case characters denote matrices; (optional) if given $[L, M]$ denotes the shape.
a_i	Element i from vector $\mathbf{a}$.
A_{ij}	Element (i, j) from matrix $\mathbf{A}$.
$[name_0] \equiv$ `['Time']`	Array of rank 1, storing a single string-type value

1 Introduction

1.1 Purpose and scope

This interface control document (ICD) describes the internal structure of and the interface to the LOFAR time series data. Time series data – i.e. the digitized voltage output, 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 form in which the received radio signals are present within the LOFAR system.

1.2 Context and motivation

The fundamental difference between analysis for LOFAR Cosmic Ray (CR) data with respect to other LOFAR Key Science Projects (KSP) 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 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 infrastructure 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 Applicable documents

Table **??** lists all the LOFAR ICDs. Most of the ICDs are for the various LOFAR data types, while ICD numbers 002 and 005 are general and applicable to all the data-format-oriented ICDs. Please note that the data and header information is written in Little-endian format within the HDF5 files.

Reference	Title	Description
ICD-001 [?]	TBB Time-Series Data	Digitized voltage output, as received by the individual LOFAR dipoles.
ICD-002 [?]	Representations of World Coordinates	Definition of how to represent and store meta-data that serve to locate a measurement in some multidimensional parameter space.
ICD-003 [?]	Beam-Formed Data	Hosting structure for LOFAR Beam-Formed data.
ICD-004 [?]	Radio Sky Image Cubes	Primary data product of the imaging pipeline.
ICD-005 [?]	File Naming Conventions	Conventions for the naming scheme applied to LOFAR standard data products.
ICD-006 [?]	Dynamic Spectrum Data	Hosting structure for dynamic spectrum data, i.e. intensity as function of time and frequency.
ICD-007 [?]	Visibility Data	Hosting structure for LOFAR UV Visibility data, primary output of interferometer operations.
ICD-008 [?]	RM Synthesis Cubes	Hosting structure for LOFAR Rotation Measure Synthesis Cubes output data.

Table 1: List of all the LOFAR Interface Control Documents. ICDs 001, 003, 004, 006, 007 and 008 describe different LOFAR data formats, while ICDs 002 and 005 are general and applicable to add the other ICDs.

2 Overview

This document is structured as follows: Section **??** will describe fundamental overall structure, including a statement of the primary data product format, HDF5. These conventions will also include names, meaning, and physical units that may be used to generate and interpret the data files. Section **??** will present a detailed specification for the data, including a description of the structure of a LOFAR TBB data naming conventions, units, physical quantities. Section 5 will provide a detailed description of the group and dataset structures contained within the LOFAR TBB Time-series file format, as well as meta-data in the form of HDF5 dataset headers.

3 Organization of the data

3.1 High level LOFAR TBB Times-series file structure

A LOFAR TBB Times-series data file will adhere to the following guidelines:

A LOFAR TBB Times-series data file will be defined within the context of the HDF5 file format. A LOFAR TBB Time-series HDF5 file structure will comprise a primary group, a "root group" in HDF5 nomenclature, which may be considered equivalent to a primary header/data unit (HDU) of a standard multi-extension FITS file. This primary group will consist only of header keywords (attributes in HDF5 nomenclature) describing general properties of an observation, along with pointers to contained subgroups. Those subgroups will comprise an arbitrary number of "StationGroups" (see sec **??**), where a Station Group will contain data and meta-data produced by an individual LOFAR station.

Figure **??** shows the basic organization of the dataset within the HDF5 format. The hierarchical structure essentially follows the hierarchical structure of LOFAR itself, i.e. in a top-down approach from array through stations down to individual dipoles. The grouping of multiple antennas/dipoles into a station is mirrored by the collection of dipole datasets into a station group.

3.2 Overview of TBB Groups

> **Comment:**
> Expand the descriptions of the groups - don't just put in references?

1. **File Root Group** (`ROOT`). The root level of the file contains the majority of associated meta-data, describing the circumstances of the observation. These data attributes include observation time (start and end), frequency window (high band vs. low band, filters) and other important characteristics of the dataset. See section **??** for further details.
2. **System Logs Group** (`SYS_LOG`). This is a catch-all envelop encapsulating information about all the system-wide steps of processing which are relevant to the entire observation, such as parameter sets and processing logs.
3. **Station Trigger Group**. (`VHECR_TRIGGER`). This group collects parameters generated by the (station-level) trigger algorithm. See section **??** for further details.
4. **Station Group** (`STATION_{NNN}`). This group serves as a common container for the separate sub-tables, which take up data from the station calibration and the trigger algorithm. See section **??** for further details.
5. **Station Calibration Group** (`STATION_CALIBRATION`). This group collects all the information required for the proper calibration of the recorded data. See section **??** for further details.
6. **Dipole Dataset Group** (`DIPOLE_DATA`). This group collects data on a per-dipole basis 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. See section **??** for further details.

3.3 Hierarchical structure

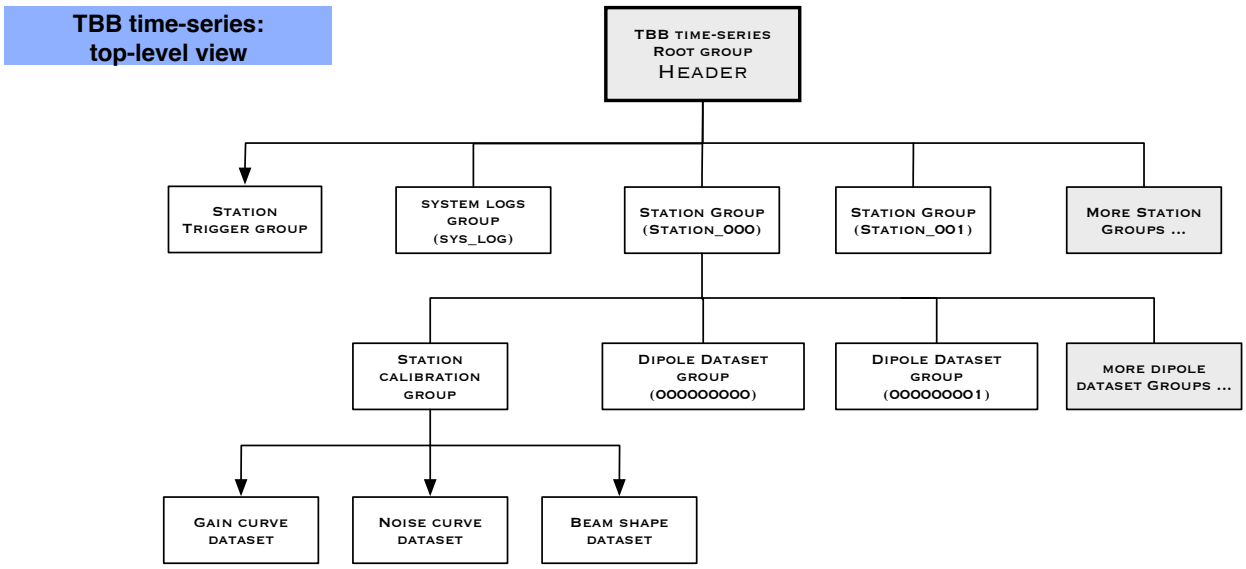

Figure 1: Hierarchical structure of a TBB times-series dataset; the internal organization of the data follows the hierarchical organization of the LOFAR system.

```
ROOT
|-- SYS_LOG                          ...  Group
|-- STATION_{NNN}                    ...  Group
|   |-- STATION_TRIGGER              ...  Group
|   |   |--TRIGGER_METADATA          ...  Group
|   |-- STATION_CALIBRATION          ...  Group
|   |   |-- GAIN_CURVE               ...  Group
|   |   |   |-- COORDINATE_0         ...  Group
|   |   |   `-- DATA                 ...  Dataset     [1D, complex<double>]
|   |   |-- NOISE_CURVE              ...  Group
|   |   |   |-- COORDINATE_0         ...  Group
|   |   |   `-- DATA                 ...  Dataset     [1D, complex<double>]
|   |   `-- BEAM_SHAPE               ...  Group
|   |   |   |-- COORDINATE_0         ...  Group
|   |   |   |-- COORDINATE_1         ...  Group
|   |   |   `-- DATA                 ...  Dataset     [3D, complex<double>]
|   |-- {NNNMMMLLL}                  ...  Dataset     [1D, short]
```

This structure can be represented through HDF5 as a POSIX-style hierarchy:

> **Comment:**
> The following table is pretty incomplete with respect to the above! Also, put this in caps as appropriate.

```
ROOT/
ROOT/SYS_LOG
ROOT/STATION_000
ROOT/STATION_000/STATION_TRIGGER
ROOT/STATION_000/STATION_TRIGGER/TRIGGER_METADATA
ROOT/STATION_000/STATION_CALIBRATION
```

```
ROOT/STATION_000/STATION_CALIBRATION/GAIN_CURVE
ROOT/STATION_000/STATION_CALIBRATION/GAIN_CURVE/COORDINATE_0
ROOT/STATION_000/STATION_CALIBRATION/GAIN_CURVE/DATA
ROOT/STATION_000/STATION_CALIBRATION/NOISE_CURVE
ROOT/STATION_000/STATION_CALIBRATION/NOISE_CURVE/COORDINATE_0
ROOT/STATION_000/STATION_CALIBRATION/NOISE_CURVE/DATA
ROOT/STATION_000/STATION_CALIBRATION/BEAM_SHAPE
ROOT/STATION_000/STATION_CALIBRATION/BEAM_SHAPE/COORDINATE_0
ROOT/STATION_000/STATION_CALIBRATION/BEAM_SHAPE/COORDINATE_1
ROOT/STATION_000/STATION_CALIBRATION/BEAM_SHAPE/DATA
ROOT/STATION_000/000000000
ROOT/STATION_000/000000001
ROOT/STATION_000/000...
ROOT/STATION_001
ROOT/STATION_001/STATION_TRIGGER
ROOT/STATION_001/STATION_CALIBRATION
ROOT/STATION_001/001000000
ROOT/STATION_001/001000001
ROOT/STATION_001/001...
...
```

4 Detailed Data Specification

LOFAR metadata are stored via Attributes in the HDF5 file, which are similar to FITS header keywords. This section details the metadata within each HDF5 Group and Sub-Group. Attribute names, data types, default values, units and descriptions are summarized for each Group, in a table. Attribute/Keyword names listed using plain text are considered as required in the HDF5 file; Attribute/Keyword names listed in *italics* are considered as optional in the HDF5 file.

4.1 The Root Group

The LOFAR file hierarchy begins with the top level `'Root Group'`. This is the file entry point for the data, and the file node by which navigatation of the data is provided. The `Root Group` will comprise a set of attributes that describe the underlying file structure, observational metadata, the LOFAR TBB data, as well as providing hooks to all groups attached to the `Root Group`.

This section will specify two sets of attributes that will appear in the `Root Group`: a set of Common LOFAR Attributes (CLA) that will be common to all LOFAR science data products, and a set of attributes that are specific to LOFAR TBB Time-series data. Though these attributes will all appear together in the `Root` attribute set, they are separated in this document in order to demarcate those general LOFAR attributes that are applicable across all data, and those attributes that are TBB-specific.

In other words,

```
Root Attributes = Common LOFAR Attributes (CLA) + Supplemental TBB Root Attributes.
```

The Common LOFAR Attributes are the first attributes of any LOFAR file root group.

4.1.1 Common LOFAR Attributes

This section will specify a set of attributes that will be common to LOFAR science data products. These "common LOFAR metadata" will appear as attributes at the root level of all LOFAR data files. *All* LOFAR data products, including TBB Time-series *inter alia,* will share a common set of metadata root-level attributes. These common LOFAR metadata are to be the first set of attributes of any LOFAR file root group.

Table **??** lists the Common LOFAR Attributes (CLA) which can be found in LOFAR observation mode data types: Beam-Formed, Transient Buffer Board (TBB) dumps, Time-series, and Sky Images within the files' root header. These Attributes are required to be in the Root Group; if a value is not available for an Attribute, a `'NULL'` maybe used in its place.

General LOFAR Group	Value	Description
Root	`'Root'`	Top-level LOFAR group type.
System Log	`'SysLog'`	System log files, parsets.
TBB	`'TBB'`	Transient Buffer Board Group.
TBB Group Subgroups	**Value**	**Description**
Station group	`'StationGroup'`	station data group.
Dipole dataset	`'DipoleDataset'`	Individual Dipole Dataset.
Trigger table	`'TriggerTable'`	Trigger algorithm table parameters.
Calibration table	`'CalibrationTable'`	Table (station) calibration data.
Source group	`'Source'`	This is a Source List group.
Processing History group	`'ProcessHist'`	This is a Processing History group.
Coordinates Group	`'Coordinates'`	This is a Coordinates group.
Coordinates Group Subgroups	**Value**	**Description**
Time coord group	`'TimeCoord'`	Describes a time axis.
Direction coord group	`'DirectionCoord'`	Describes a direction group.
Spectral coord group	`'SpectralCoord'`	Describes a frequency group.

Table 2: LOFAR TBB Time-series Group Types

- `GROUPTYPE` - The first Attribute in every group must be the attribute `GROUPTYPE`. Since the CLA are in the root header, the value in the CLA for (`GROUPTYPE`) = `'Root'`. The options for the group type are listed in Tab. **??**, grouped by category.
- `FILENAME` – Name of this file
- `FILEDATE` – File creation date, i.e. time at which the initial version of the file has been created.
- `FILETYPE` – is the **file type** for the LOFAR observation. This descriptor, which will also appear in LOFAR data filenames (see Table **??** below, or refer to [**?**]) of the LOFAR data file, indicates the kind of LOFAR data contained.
- `TELESCOPE` - **name of the telecope** 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 ID will be stored in `PROJECT_ID` and title within `PROJECT_TITLE`. Additional attributes provide further detailed information, such as the name of the project's principal investigator (`PROJECT_PI`), the name(s) of the co-investigator(s) (`PROJECT_CO_I`) as well as means to contact the project (`PROJECT_CONTACT`). If no specific project is defined, the variables simply should be set to `'LOFAR'`.
- `OBSERVATION_ID` – is the **unique identifier** for the LOFAR observation.
- The observation's start time is listed in the following formats:
  - Modified Julian Day (`OBSERVATION_START_MJD`) using NNNNNN.NNNNNNN format,
  - International Atomic Time (`OBSERVATION_START_TAI`) using yyyy-mm-ddThh:mm:ss.sssssssss format and
  - Coordinated Universal Time (`OBSERVATION_START_UTC`) using yyyy-mm-ddThh:mm:ss.sssssssssZ format.

Field/Keyword	Type	Value	Description
GROUPTYPE	string	'Root'	LOFAR Group type (this is a 'root' group)
FILENAME	string	—	File name
FILEDATE	string	—	File creation date, i.e. time at which the initial version of the file has been created. YYYY-MM-DDThh:mm:ss.s
FILETYPE	string	—	File type
TELESCOPE	string	'LOFAR'	Name of the telescope
OBSERVER	string	—	Name(s) of the observer(s)
PROJECT_ID	string	—	Unique identifier for the project
PROJECT_TITLE	string	—	Title of the project
PROJECT_PI	string	—	Name of Principal Investigator
PROJECT_CO_I	string	—	Name(s) of the Co-investigator(s)
PROJECT_CONTACT	string	—	Contact details for project
OBSERVATION_ID	string	—	Unique identifier for the observation
OBSERVATION_START_MJD	double	—	Observation start date (MJD)
OBSERVATION_START_TAI	string	—	Observation start date (TAI)
OBSERVATION_START_UTC	string	—	Observation start date (UTC)
OBSERVATION_END_MJD	double	—	Observation end date (MJD)
OBSERVATION_END_TAI	string	—	Observation end date (TAI)
OBSERVATION_END_UTC	string	—	Observation end date (UTC)
OBSERVATION_NOF_STATIONS	int	—	nof. stations used during the observation
OBSERVATION_STATIONS_LIST	array<string,1>	—	List of stations used during the observation
OBSERVATION_FREQUENCY_MAX	double	—	Observation maximum frequency
OBSERVATION_FREQUENCY_MIN	double	—	Observation minimum frequency
OBSERVATION_FREQUENCY_CENTER	double	—	Observation center frequency
OBSERVATION_FREQUENCY_UNIT	string	'MHz'	Frequency units of this observation
OBSERVATION_NOF_BITS_PER_SAMPLE	int	—	Number of bits per sample in the incoming data stream from the stations to CEP/BlueGene.
CLOCK_FREQUENCY	double	—	Clock frequency, in units of CLOCK_FREQUENCY_UNIT; valid values for LOFAR are 160.0 MHz and 200.0 MHz.
CLOCK_FREQUENCY_UNIT	string	'MHz'	Clock frequency unit
ANTENNA_SET	string	—	Antenna set specification of observation
FILTER_SELECTION	string	—	Filter selection (see description)
TARGET	string	—	Single or list of observation targets/-sources
SYSTEM_VERSION	string	—	Processing system name/version
PIPELINE_NAME	string	—	Pipeline processing name
PIPELINE_VERSION	string	—	Pipeline processing version
ICD_NUMBER	string	—	Interface Control Document number
ICD_VERSION	string	—	Interface Control Document version/issue number
NOTES	string	—	Notes or comments

Table 3: Common LOFAR Attributes (CLA)

File Type	Value	Description
UV Vis	'uv'	LOFAR visibility file w/correlation UV information.
Sky cube	'sky'	LOFAR Image cube w/RA, Dec, frequency and polarization
RM cube	'rm'	Rotation Measure Synthesis Cube w/ axes of RA, Dec, Faraday Depth, polarization.
Near-field image	'nfi'	Near Field Sky Image w/ axes of position on the sky (x, y, z), frequency time, polarization.
Dynamic Spectra	'dynspec'	Dynamic Spectra w/ axes of time, frequency, polarization.
Beamformed data	'bf'	Beam-Formed file w/ time series data with axes of frequency vs time.
TBB dump	'tbb'	TBB dump file, raw time-series: (1) intensity as a function of frequency, or (2) voltage vs time.
Instrument Model	'inst'	Parameters describing gain and other instrument characteristics for calibation.
Sky Model	'lsm'	List of sources, either point sources or shapelets.

Table 4: Overview of standard LOFAR data products and the corresponding file type attribute value.

- The observation's end time is listed in the following formats:
  - Modified Julian Day (OBSERVATION_END_MJD) using NNNNNN.NNNNNNN format,
  - International Atomic Time (OBSERVATION_END_TAI) using yyyy-mm-ddThh:mm:ss.sssssssss format and
  - Coordinated Universal Time (OBSERVATION_END_UTC) using yyyy-mm-ddThh:mm:ss.sssssssssZ format.
- OBSERVATION_NOF_STATIONS – Number of stations used for this observation
- OBSERVATION_STATIONS_LIST – A list of stations used for this observation
- OBSERVATION_FREQUENCY_MAX – Upper frequency limit of observation data
- OBSERVATION_FREQUENCY_MIN – Lower frequency limit of observation data
- OBSERVATION_FREQUENCY_CENTER – Center frequency of the covered frequency range, given as the geometric mean of maximum and minimum frequency:

$$\begin{aligned} \nu_{\text{center}} &= (\nu_{\min} + \nu_{\max})/2 \\ &= (\texttt{OBSERVATION\_FREQUENCY\_MIN} + \texttt{OBSERVATION\_FREQUENCY\_MAX})/2 \end{aligned}$$

  Given the possibilities of rather non-regular coverage in frequency space, ν_{center} is formost intended as orientation during the initial inspection of the data sets' properties; for precise information on the sampling in frequency space, one is refered to the Spectral coordinate as part of the Coordinates group.
- OBSERVATION_FREQUENCY_UNIT – When TELESCOPE is 'LOFAR', all observation frequency units will be 'MHz'.
- CLOCK_FREQUENCY – The clocking frequency used for the observation. For LOFAR, this will be one of '160' or '200'.
- CLOCK_FREQUENCY_UNIT – For LOFAR, this will be 'MHz'
- ANTENNA_SET – The **antenna set** configuration used during the observation; see Table **??** below for a list of recognized values.
- FILTER_SELECTION – The **filter selection** (frequency bandwidth) used during the observation. The metadata need to reflect the frequency band in which the data have been recorded; see Table **??** below for a list of recognized values.

Antenna set	Description
'LBA_INNER'	48 antennas of the INNER LBA configuration (see figure 2)
'LBA_OUTER'	48 antennas of the OUTER LBA configuration (see figure 2)
'LBA_SPARSE_EVEN'	Intersection of INNER-SPARSE configurations
'LBA_SPARSE_ODD'	Intersection of OUTER-SPARSE configurations
'LBA_X'	X component, ALL LBA antennas.
'LBA_Y'	Y component, ALL LBA antennas.
'HBA_ZERO'	HBA antennas 0-23 in Core stations, all HBA's in the other stations.
'HBA_ONE'	HBA antennas 24-47 in Core stations, and all HBA's in the other stations.
'HBA_DUAL'	Both HBA antenna (sub)fields in the Core stations, which set up an identical beam/pointing on each of those (sub)fields. On CEP, those (sub)fields are treated as seperate stations. On non-core stations, the whole HBA field is used and one beam is made.
'HBA_JOINED'	ALL HBA antennas in ALL stations types. For Core stations, this will result in a "weird" beamshape.

Table 5: Overview of antenna set configurations.

Filter-band, [MHz]	Attribute value
10 – 70	'LBA_10_70'
30 – 70	'LBA_30_70'
10 – 90	'LBA_10_90'
30 – 90	'LBA_30_90'
110 – 190	'HBA_110_190'
170 – 230	'HBA_170_230'
210 – 250	'HBA_210_250'

Table 6: Overview of filter-band selections and corresponding attribute values.

- TARGET - User-supplied target name holds a single source name or a list of the observed sources/targets. This field can also state that the observation was 'All-sky' or reference a grid number/identifier as part of an all-sky survey.
- SYSTEM_VERSION lists the name and (if available) version of the processing system used for carying out the observation and creating the data.
- PIPELINE_NAME and PIPELINE_VERSION list name and version of the pipeline by which the data have been processed to the recorded state.
- ICD_NUMBER and ICD_VERSION list name/number and version/issue of the Interface Control Document (ICD) to which the data abide by.
- The NOTES attributes acts as generic area for notes and comments.

4.1.2 Additional TBB Time-series Root Attributes

As explained at the beginning of Sec. **??** above, the root group of a **TBB Times-series data file** will contain a set of attributes, which can be broken down into two subsets: 1) a set of Common LOFAR Attributes (CLA) that will be common to all LOFAR science data products, and 2) a set of attributes that are specific to LOFAR Sky Image data (see Table **??**). With the Common LOFAR Attributes already listed in section **??** above, this section will focus on the second subset of root group attributes, as they are specific to a **TBB Times-series data file**.

Note that currently `TRIGGER_METADATA` is defined simply as an unformatted string. It is anticipated that in the future, this will become a group with it's own structure unique to each `TRIGGER_TYPE` (see Table **??**).

Field/Keyword	Type	Value	Description
`SYS_LOG`	Group	—	Container for system-wide log object.
`STATION_{NNN}`	Group	—	Station group collecting the data from an individual LOFAR station; the three-digit numeral is derived from the `STATION_ID` which itself is stored within the *Dipole dataset* structure.
`TRIGGER_TYPE`	`string`	—	String indicating the reason for data return.
`TRIGGER_DATA`	Group	—	Group collecting the parameters associated with the specific TRIGGER_TYPE trigger and the output parameters generated by it.

Table 7: Additional attributes and objects attached to the root group of a TBB time-series data file.

Value of `TRIGGER_TYPE`	Description
`'Unknown'`	Unknown/unrecognised reason for data return.
`'Blind'`	Data return triggered arbitrarily, i.e. off no signal.
`'VHECR'`	Single-station VHECR trigger.
`'VHECRMulti'`	Multi-station VHECR trigger.
`'LORA'`	Trigger from the LORA particle detector.
`'FRATS'`	Fast Radio Transients trigger.
`'UHEP'`	Trigger from Ultra-High Energy Particle mode (a.k.a 'Nu-Moon').
`'Lightning'`	Triggered in order to capture a lightning strike.
`'Manual'`	Manually triggered.

Table 8: Values of TRIGGER_TYPE — names for TRIGGER_DATA groups are '[TRIGGER_TYPE]_TRIGGER'.

'Unknown' Trigger Group This is a catch-all group for any undefined data dumps – if for any reason the dump type is unrecognised, it should default to this. In such a case, the group will contain only an unformatted string of arbitrary length, to allow the triggerer to insert any necessary info without having to have their own format defined in this ICD (see Table **??**).

Field/Keyword	Type	Value	Description
`GROUPTYPE`	`string`	`'UNKNOWN_TRIGGER'`	This is a group for a trigger of unknown type.
`METADATA`	`string`		Unformatted string for arbitrary trigger information.

Table 9: Contents of the UNKNOWN_TRIGGER group.

VHECR Trigger Group The `VHECR_TRIGGER` group collects parameters generated by the (station-level) trigger algorithm, which in the case of an 'internal' VHECR trigger (here just called a 'VHECR trigger') will be responsible for causing the dump of the TBB data. Table **??** shows the attributes within this group.

– Even though the most common origin for the station trigger will be the trigger algorithm running at the station LCU, other scenarios are possible which will cause a dump of a station's TBB data. `TRIGGER_SOURCE` will record the source within the system, at which the trigger was generated.

Field/Keyword	Type	Value	Description
GROUPTYPE	string	'VHECR_TRIGGER'	This is a VHECR-specific group.
TRIGGER_SOURCE	string	'LCU'	Source within the system, at which the trigger was generated. By default this will be the trigger algorithm running at the LCU.
TRIGGER_TIME	int	—	Timestamp (in seconds since 1970) for the trigger.
TRIGGER_SAMPLE_NUMBER	int	—	Sample number inside the second recorded through TIME.
PARAM_COINCIDENCE_CHANNELS	int	48	The number of channels neede to detect a coincidence.
PARAM_COINCIDENCE_TIME	double	1e-6	The time-range in seconds, during which triggers are considered part of a coincidence.
PARAM_DIRECTION_FIT	string	'simple'	Do a direction fit?
PARAM_ELEVATION_MIN	double	30	Minimum elevation (in degreees) to accept a trigger.
PARAM_FIT_VARIANCE_MAX	double	100	Maximum variance ("badness of fit") of the direction fit to still accept a trigger.
COINCIDENCE_CHANNELS	int	—	Number of channels that took part in the coincidence.
RCU_ID	array<int,1>	—	Number of the RCUs that took part in the coincidence.
TIME	array<int,1>	—	Timestamps in seconds since 1970.
SAMPLE_NUMBER	array<int,1>	—	Sample numbers inside the second marked by TIME.
PULSE_SUM	array<int,1>	—	Sum of all the samples during the pulse.
PULSE_WIDTH	array<int,1>	—	Width of the pulse in samples.
PULSE_PEAK	array<int,1>	—	The largest value (peak value) a sample had during the pulse.
PULSE_POWER_PRE	array<int,1>	—	Power before the onset of the pulse: value of the mean at the start of the trigger.
PULSE_POWER_POST	array<int,1>	—	Power before the onset of the pulse: value of the mean at the end of the trigger.
NOF_MISSED_TRIGGERS	array<int,1>	—	Number of missed triggers (+1) since the last trigger for this channel.
FIT_DIRECTION_AZIMUTH	double	—	Direction fit result for the Azimuth angle.
FIT_DIRECTION_ELEVATION	double	—	Direction fit result for the Elevation angle.
FIT_DIRECTION_DISTANCE	double	—	Direction fit result for the distance of curvature.
FIT_DIRECTION_VARIANCE	double	—	Variance ("badness of fit") of the direction fit.

Table 10: Attributes attached to a VHECR_TRIGGER group.

– The time information for the trigger is encoded the same way as the data from the individual dipoles: the combination of `TRIGGER_TIME` (full seconds since 1970) and `TRIGGER_SAMPLE_NUMBER` (sample number within the second) will provide the absolute time.

The remainder of the attributes can be divided into three groups:

1. *Trigger algorithm setup parameters.*
   – `PARAM_COINCIDENCE_CHANNELS` marks the number of channels needed to detect a coincidence. The actual number of antennas which were part in the coincidence then is recoded through `COINCIDENCE_CHANNELS`.
2. *Trigger algorithm output parameters.*
   – `COINCIDENCE_CHANNELS` is the number of channels/dipoles, that took part in the coincidence; this number will be equal or larger as `PARAM_COINCIDENCE_CHANNELS`.
   – `RCU_ID` holds a list of RCU, which have taken part in the coincidence.
   – `TIME` holds a list of the timestamps in seconds since 1970, for the RCUs which have been taken part in the coincidence.
3. *Fit results* based on the output parameters of the trigger algorithm.
   – `FIT_DIRECTION_AZIMUTH` and `FIT_DIRECTION_ELEVATION` are the fit results for the direction of arrival

4.2 The 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** (Table **??**) as a basic module within the data structure.[1] 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 (Table **??**, p. **??**):

– `GROUPTYPE` identifies the group as a `StationGroup`.

– While an internal identifier for the station is provided through the station ID, for better diagnostics the actual name of the station will be required; therefore `STATION_NAME` will store the actual name of the station, e.g. `CS001`, `RS201` or `DE602`.

– The **position of the LOFAR station** is reconstructed from the three attributes
  - `STATION_POSITION_VALUE` – numerical value of the station position coordinates
  - `STATION_POSITION_UNIT` – physical units associated with the numerical values for the station position
  - `STATION_POSITION_FRAME` – identifier for the reference frame within which the station position is provided

– The **direction of the station beam** on top of which the observation potentially has been running in piggy-back mode:

[1] 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*.

Field/Keyword	H5Type	Type	Description
GROUPTYPE	Attr	string	LOFAR group type, StationGroup.
STATION_NAME	Attr	string	The name of the station, e.g. CS001 or RS201.
STATION_POSITION_VALUE	Attr	array<double,1>	[3] Numerical value of the station position coordinates.
STATION_POSITION_UNIT	Attr	array<string,1>	Physical units associated with the numerical values for the station position.
STATION_POSITION_FRAME	Attr	string	Identifier for the reference frame within which the station position is provided.
BEAM_DIRECTION_VALUE	Attr	array<double,1>	[2] Numercial value of the station-beam direction.
BEAM_DIRECTION_UNIT	Attr	array<string,1>	Physical units associated with the numerical value of the station-beam direction. Mandatory for HBA station data.
BEAM_DIRECTION_FRAME	Attr	string	Identifier for the reference frame within which the station-beam direction is provided. Mandatory for HBA station data.
CLOCK_OFFSET_VALUE	Attr	double	(Relative) Station clock offset.
CLOCK_OFFSET_UNIT	Attr	string	Physical unit for the station clock offset.
TRIGGER_OFFSET	Attr	double	Trigger time – in seconds – relative to the reference time.
NOF_DIPOLES	Attr	int	The number of dipoles, for which data are embedded within this group.
STATION_CALIBRATION	Group	—	Calibration information as delivered through the online (station) calibration.
{NNNMMMLLL}	Dataset	array<short,1>	Dataset containing the actual raw samples read out from the transient buffer; the name of the dataset is constructed from the STATION_ID (NNN), RSP_ID (MMM) and RCU_ID (LLL).

Table 11: Fields in the station data group (StationGroup). 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.

- BEAM_DIRECTION_VALUE – numerical value of the station-beam direction
- BEAM_DIRECTION_UNIT – physical units associated with the numerical value of the station-beam direction
- BEAM_DIRECTION_FRAME – identifier for the reference frame within which the station-beam direction is provided

– TRIGGER_OFFSET

– NOF_DIPOLES is a counter for the number of dipoles, for which data are embedded within this group.

Even though there exist multiple LOFAR observation modes for TBBs, 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, need to be stored.

4.3 Station Calibration Group

The **Station Calibration Group** collects all the information required for the proper calibration of the recorded data. Since all TBB observation modes make use of the data as they are available directly after the 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. Even though the bulk of the file volume is taken up by the time-series data of the individual dipoles, most of the calibration information are larger than suitable to be stored as simple attributes; as a consequence the STATION_CALIBRATION group itself will consist of a collection of sub-groups (mainly acting as containers for attribute-type metadata) and datasets (storing the actual calibration values). The basic hierarchical structure is depicted in Figure **??**, with explanations found in Table **??**.

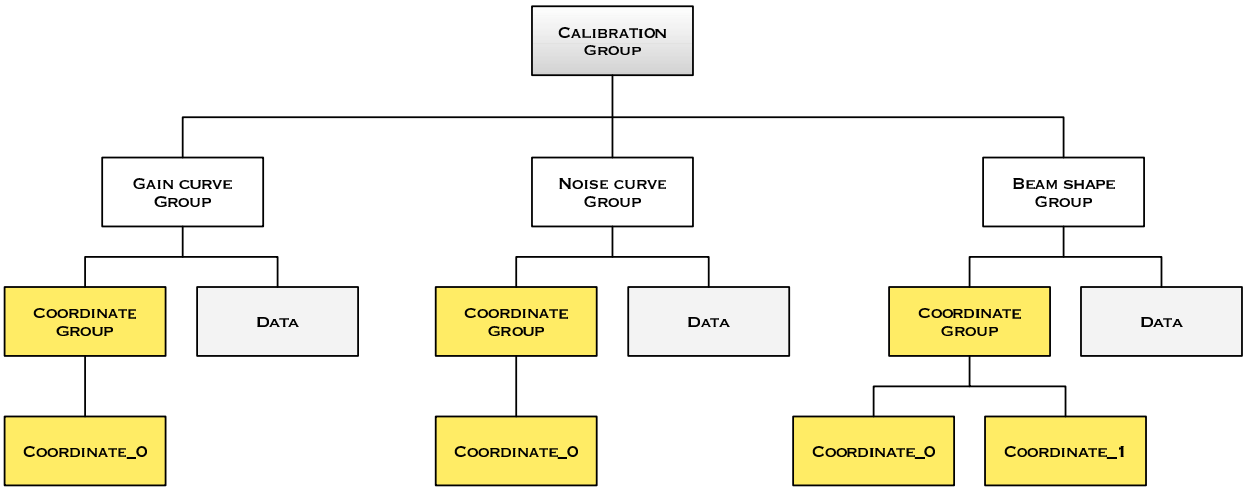

Figure 2: Hierarchical structure of the station calibration group.

Even though one could consider storing information such as the gain curve as simple attributes attached to the calibration group there are a number of arguments which speak against such course of action: a) An attribute can only be accessed via the object. b) For practical reasons, an attribute should be a small object, no more than 1000 bytes. c) The data of an attribute must be read or written in a single access, selection is not allowed.

Each of the sub-groups for an individual calibration quantity, will consist of a part describing the physical coordinates (e.g. a frequency axis) and an array storing the actual calibration values:

– GAIN_CURVE : complex electronic gain, $\mathbf{G}_{j,\mathrm{gain}} = G_j(\nu)$, for receiving element j as function of frequency (bandpass); this array will be multiplied (after interpolation, if required) to the output of the Fourier transform of the dipole voltage time-series.

Field/Keyword	H5Type	Type	Description
GROUPTYPE	Attr	string	LOFAR group type, StationCalibration.
GAIN_CURVE	Group		Sub-group storing physical coordinates and values of the complex electronic gain, $\mathbf{G}_{j,\mathrm{gain}} = G_j(\nu)$.
NOISE_CURVE	Group		Sub-group storing physical coordinates and values of the complex electronic noise, $\mathbf{G}_{j,\mathrm{noise}} = G_j(\nu)$.
BEAM_SHAPE	Group		Sub-group storing physical coordinates and values of the complex element beam pattern as function of direction and frequency $\mathbf{G}_{\mathrm{beam}} = G(\rho, \nu)$.

Table 12: Structure of the **Station calibration group**; besides a number of top-level attributes, the group acts as a container for the actual calibration data, which are stored inside individual sub-groups.

Field/Keyword	H5Type	Type	Description
GROUPTYPE	Attr.	string	LOFAR group type, GainCurve.
GAIN_CURVE_UNIT	Attr.	string	
NOF_COORDINATES	Attr.	int	Number of coordinate objects attached to this group.
NOF_AXES	Attr.	int	Number of coordinate axes represented by the attached coordinate groups.
COORDINATE_TYPES	Attr.	array<string,1>	Type of the attached coordinate objects.
COORDINATE_0	Group		Coordinate object describing the frequency axis.
GAIN_CURVE_DATA	Dataset	array<complex<double>,1>	

Table 13: Attributes, groups and dataset attached to the gain curve calibration group.

- NOISE_CURVE : complex electronic noise, $\mathbf{G}_{j,\mathrm{noise}} = G_j(\nu)$, for receiving element j as function of frequency (bandpass); this array will be multiplied (after interpolation, if required) to the output of the Fourier transform of the dipole voltage time-series.
- BEAM_SHAPE holds the complex **element beam pattern** as function of direction and frequency $\mathbf{G}_{\mathrm{beam}} = G(\rho, \nu)$. In order to correctly interpret the data – and, if necessary, interpolate – the corresponding **beam directions** (BEAM_DIRECTIONS) and **frequency values** (BEAM_FREQUENCIES) are required.

4.4 Dipole Dataset

The **Dipole Dataset** (Table **??**) collects data on a per-dipole basis[2] – 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.

Comment:
Add attribute to store values of the analoge beamformer for the HBA tiles.

- STATION_ID, RSP_ID and RCU_ID are directly taken from the frame structure used in the communication between RSP and TBB [**?**]. The three identifiers – in combination with ANTENNA_SET 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

[2]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).

Field/Keyword	H5Type	Type	Description
GROUPTYPE	Attr.	string	LOFAR group type, NoiseCurve
NOISE_CURVE_UNIT	Attr.	string	
NOF_COORDINATES	Attr.	int	Number of coordinate objects attached to this group.
NOF_AXES	Attr.	int	Number of coordinate axes represented by the attached coordinate groups.
COORDINATE_TYPES	Attr.	array<string,1>	Type of the attached coordinate objects.
COORDINATE_0	Group		Coordinate object describing the frequency axis.
NOISE_CURVE_DATA	Dataset	array<complex<double>,1>	

Table 14: Attributes, groups and dataset attached to the noise curve calibration group.

Field/Keyword	H5Type	Type	Description
GROUPTYPE	Attr.	string	LOFAR group type, BeamShape.
BEAM_SHAPE_UNIT	Attr.	string	
NOF_COORDINATES	Attr.	int	Number of coordinate objects attached to this group.
NOF_AXES	Attr.	int	Number of coordinate axes represented by the attached coordinate groups.
COORDINATE_TYPES	Attr.	array<string,1>	Type of the attached coordinate objects.
COORDINATE_0	Group		Coordinate object describing the directional axes.
COORDINATE_1	Group		Coordinate object describing the frequency axis.
BEAM_SHAPE_DATA	Dataset	array<complex<double>,3>	The beam-shape data.

Table 15: Attributes, groups and dataset attached to the beam shape calibration group.

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_NUMBER 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 choosing a time system, such as UNIX time[3]. For a constant sampling frequency (SAMPLE_FREQ) the timing for the remaining set of samples can derived via [?]

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

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. As typically the number of samples requested from the TBB will be larger but the number of samples fitting into a single frame, the resulting dataset will accumulate the contents from multiple frames.

– 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.

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

Field/Keyword	Type	Description
GROUPTYPE	string	The type of this group, *DipoleDataset*.
STATION_ID	unsigned	Data source station identifier.
RSP_ID	unsigned	Data source RSP board identifier.
RCU_ID	unsigned	Data source RCU board identifier.
SAMPLE_FREQUENCY_VALUE	double	Sample frequency in MHz of the RCU boards.
SAMPLE_FREQUENCY_UNIT	string	Physical units of the sample frequency.
TIME	unsigned	Time instance in seconds of the first sample in the payload.
SAMPLE_NUMBER	unsigned	Sample number of the first payload sample in current seconds interval in transient mode.
SAMPLES_PER_FRAME	unsigned	Total number of samples in the payload of the original TBB–RSP frame structure.
DATA_LENGTH	unsigned	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.
NYQUIST_ZONE	unsigned	Nyquist zone in which the data are sampled.
ADC2VOLTAGE	double	Conversion factor from raw ADC sample values to voltages.
CABLE_DELAY	double	Delay the length of the cable connected to the RCU adds to the signal path.
CABLE_DELAY_UNIT	string	Physical unit associated with CABLE_DELAY.
FEED	string	Type of feed for this dipole.
ANTENNA_POSITION_VALUE	array<double,1>	[3] Antenna position w.r.t. the station center, $\mathbf{x} = (x_1, x_2, x_3)$.
ANTENNA_POSITION_UNIT	string	Physical units of the antenna position.
ANTENNA_POSITION_FRAME	string	Reference frame of the antenna position.
ANTENNA_ORIENTATION_VALUE	array<double,1>	[3] Antenna orientation w.r.t. the reference frame within which the antenna positions are defined.
ANTENNA_ORIENTATION_UNIT	string	Physical units of the antenna orientation; depending on the convention used this might be DEG or M.
ANTENNA_ORIENTATION_FRAME	string	Reference frame of the antenna orientation.
TILE_BEAM_VALUE	array<double,1>	Numerical value of the tile beam direction.
TILE_BEAM_UNIT	string	Physical units of the tile beam orientation.
TILE_BEAM_FRAME	string	Reference frame of the tile beam.
TILE_BEAM_DIPOLES	array<unsigned,1>	Dipoles within the tiles, which were used for the beamforming.
TILE_COEF_UNIT	string	Physical units of the tile beam coefficients.
TILE_BEAM_COEFS	array<unsigned,1>	Coefficients used in the analogue dipole beamformer for this beam.
TILE_DIPOLE_POSITION_VALUE	array<double,3>	Position of the dipoles within the tile.
TILE_DIPOLE_POSITION_UNIT	string	Units of the tile dipole positions.
TILE_DIPOLE_POSITION_FRAME	string	Reference frame of the dipole positions.

Table 16: Fields in the dipole dataset; 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 [**?**]. Although the TILE_XXX keywords are optional for LBA station data, they are mandatory for HBA station data.

- the **position** of the receiving element within the station:
  - `ANTENNA_POSITION_VALUE` – numerical value of the antenna position coordinates
  - `ANTENNA_POSITION_UNIT` – physical units associated with the numerical values for the antenna position
  - `ANTENNA_POSITION_FRAME` – identifier for the reference frame within which the antenna position is provided
- the **orientation** of the receiving element w.r.t to the coordinate frame within which the position of the same element was provided:
  - `ANTENNA_ORIENTATION_VALUE`— – numerical value of the parameters describing the orientation of the antenna
  - `ANTENNA_ORIENTATION_UNIT` – physical units associated with the numerical values for the antenna orientation
  - `ANTENNA_ORIENTATION_FRAME` – identifier for the reference frame within which the antenna orientation is provided

  An agreement on the format, in which this information is provided, still needs to be made, but it is foreseeable to describe the orientation (tilt) of an receiving element in one of the two ways:
  1. Euler angles describing a rotation between the reference frame within which the position of the antenna is given and the coordinate frame based on the antenna's orientation.
  2. Normal (unit length) vector indicating the orientation of the antenna from the location defined by the antenna position.

5 Interfaces

5.1 Interface requirements

- low-level generic interface to individual elements within the dataset → provided by the DAL
- high-level interface tranversing hierarchy boundaries → some of this initially not provided by the DAL itself, but later on integrated from external code base (such as e.g. CR-Tools)
  - create/read/write attributes
  - (sliced) reading of channel data
  - creation of a new (empty) dataset

5.2 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.)

5.3 Relation to existing workpackages

Data Access Library (DAL) While in most usage scenarios all the valid data from a LOFAR station will be read in to be processed together, there 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

A Discussion & open questions

A.1 Open questions

The following table presents an overview of (some of the) known open questions regarding the format definition:

Item	Description	Status
01	Are there modes forseen in which the total LOFAR array is being split up into sub-arrays operating in different modes? 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. `OBSERVATION_MODE` or `NYQUIST_ZONE`.	L. Baehren
02	How to handle multiple HBA tile beams per station? Perhaps via subgroups of the dipole group?	L. Baehren
03	Is there indeed a separate value available which describes the conversion from ADC counts to voltages, or is this part of the gain calibration? If the latter is the case, then how to get a voltage time-series?	L. Baehren
04	Currently the trigger meta-data has no defined format. In the future, there will be a specific format for each trigger mode, i.e. an 'UHEP' group etc. In the meantime, the current format is to be used as a long character string that can be used flexibly.	C. James

A.2 Future enhancements

Though the file format definition is not intended to undergo considerable changes once gaining release status, there will be future enhancements to reflect new insights and address noted short-comings.

1. Some additional metadata are required when the observation is done using the HBA [input thanks to Maaijke Mevius]:
   a) At station level the 4x4 numbers which give the relative position of the dipoles in a tile, given in HBADeltas.conf. We do not really need those for analysis, but if you want to simulate the data (or do some sort of selfcal) it might be good to have them.
   Since these numbers are the same for all tiles within a station, I think they can be stored in the station metadata.
   b) The pointing of the tile beam (2 angles) should be added. I believe it is possible to have different tiles pointing in different directions, thus it should be stored per antenna.
   c) By 'tiles', do we mean literal tiles, or tiles/dipoles? [CWJ]

B Anticipated Data volumes

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-mode:

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 $\times$ 48 antennae $\times$ 2 pol $\times$ 2^{17} samples $\times$ 2 Bytes/sample:	1.8 GB/event
Total:	3.8 GB/event

VHECR-mode:

Raw time series data from individual dipoles	
48 antennae $\times$ 2 pol $\times$ 2^{17} samples $\times$ 2 Bytes/sample:	25 MB/event

HECR-mode:

Similar, but now only one station is involved	
48 antennae $\times$ 2 pol $\times$ 2^{17} samples $\times$ 2 Bytes/sample:	25 MB/event

TS-mode:

Similar to VHECR but full raw data from TBB	
77 stations $\times$ 48 antennae $\times$ 2 pol $\times$ $2 \cdot 10^8$ samples $\times$ 2 Bytes/sample:	3.0 TB/event

The event rate is uncertain, and for VHECR is estimated at one triggered event per station per 10 minutes. With 32 stations active this would amount to 110 GB per day of observing time.

C Requirements

C.1 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

C.1.1 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 Table **??** 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.

C.2 Visualization

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

Database entry	Field	Format
CalCtlr	connected	unsigned int
	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 17: 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.

- 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).