LOFAR Data Format ICD
Beam-Formed Data

Document ID: LOFAR-USG-ICD-003.tex

ISSUE 2.04.18

SVN Repository Revision: 9374

A. Alexov, K. Anderson, L. Bähren, J.-M. Grießmeier, J.W.T. Hessels, J.S. Masters, B.W. Stappers

SVN Date: January 10, 2012

Contents

Change record

Version 1.x

Version 1.0

Document made available through USG repository on 01 Sep 2008.

Version 1.00 → Version 1.01

Bug-fixes:

- NOF_SAMPLES : attribute needs to accurately reflect the length of the vector.
- CENTER_FREQUENCY : attribute may need to be changed to the *start* frequency for each subband.
- Rename EMBAND to OBSERVATION_MODE and make this attribute part of the list of common LOFAR metadata found in every data product.

The following new or modified root-level header attributes will appear in the next version:

- CHANNELS_PER_SUBBAND (new) : i.e. 16
- CLOCK_RATE (new) : Either 160,000,000 or 200,000,000
- CLOCK_RATE_UNIT (new) : Hz
- SAMPLING_RATE : 156250 or 195312.5

  Each station divides the spectrum into 512 subbands (2 polarizations), thus 160Mhz / 512 / 2 = 156250 and 200Mhz / 512 / 2 = 195312.5
- SAMPLING_RATE_UNIT (new) : Hz
- SAMPLING_TIME (new) : 1/SAMPLING_RATE
- SAMPLING_TIME_UNIT (new) : seconds
- SUBBAND_WIDTH (new) : 156250 or 195312.5
- SUBBAND_WIDTH_UNIT (new) : Hz
- CHANNEL_WIDTH (new) :

  SUBBAND_WIDTH / CHANNELS_PER_SUBBAND
- CHANNEL_WIDTH_UNIT (new) : Hz
- TOTAL_INTEGRATION_TIME :

  SAMPLING_TIME * NOF_SAMPLES
- TOTAL_INTEGRATION_TIME_UNIT (new) : seconds

Change Record style changed starting after V1.01

Version	Date	Sections	Description of changes
1.02	—	—	Major changes in the file format from BF RAW to BF TAB file structure. Changes to the document include: · Changes in H5 structure - division by Station Beam, Pencil Beam, Stokes, Subband, Channels. · Added diagrams of BF data flow to the Blue Gene, then offline cluster, then into H5 files. · Added diagrams with BF H5 structure defined. · Added Attributes to the various Groups within H5 B5 file. · Added Coordinate Group(s). · Added text to describe new file features. · Changed the sections to match the rest of the ICDs.
1.03	—	—	Iterated over file structure. Updated ICD with comments from Pulsar Group.
1.04	—	—	Added common glossary, common file name structure include files. Fixed issues/comments from Jason (July 28, 2009) Activated SVN revision and dates to get proper information filled in at SVN checkout. Updated Observation DATE & TIME related fields in the Root Header. Added Truth Table of various sub-observation modes and how they relate to Stokes header parameter settings.
1.05	—	—	Added Example Coordinate section. Changed Linear Coordinates to combination of Linear & Tabular Coordinates per Beam. Changed all the figures to match the changes in the Coordinates scheme.
1.06	—	—	Cleaned up ICD so that names of groups are consistent across the various sections and diagrams. Updated names of groups to be more in line with the Sky Image ICD. Minor small changes for clean up (Attribute tables should have "value" instead of "units"). Removed `NYQUIST_ZONE` header keyword, as it is now replaced by `FILTER_SELECTION` and `ANTENNA_SET`.
1.07	—	—	Removed Exposure time in LST coords. Added additional comments regarding weather/temperature/humidity header Attributes. Added `BF_RAW`, `BF_OUT`, `BF_VERSION` Attributes. Updated BF Data Flow diagram.
1.08	—	—	Cleaned up list of attributes considered to go into the coordinate groups; attributes not part of the WCS parameters have been moved to the SubArrayPointing group and the Beam group. Considerable cleaning up has been done on the attributes tracking modes and processing
1.09	—	—	Taken in suggestion to change the possible values for `BF_FORMAT`.
1.10	—	—	Moved coordinates group table to external file.
1.11	—	—	Renamed groups for station and beam data, to make them better readable: `StatBeam` → `SubArrayPointing`, `PencBeam` → `Beam`
1.12	2009-12-10	—	Updates based on feedback from John Romain and James Anderson. Minor text fixes; clarifications; addition of 2nd transpose to Figure 4; precision updates to header Attributes (float → double, etc); changed BFRAW → BF RAW and BFOUT → BF TAB (Tied-Array Beam); changed STOKES I, Q, U, V references to 0, 1, 2, 3 and added option for XX, XY, YX, XX.

continued on next page

continued from previous page

Version	Date	Sections	Description of changes
1.13	—	—	Updates based on ICD meeting following up on James Anderson's comments. Moved the Coordinate Group into the Beam group; changed text and all figures to reflect this. Changed all references to time, as the start time of observation/exposure, etc. Added STATIONS_LIST and number of stations to the Beam group (also exist in Station Beam Group).
1.14	—	—	Removed the Group Type Table from the Common Header and incerted Group Type Table per ICD.
1.15	—	—	Corrected Group Type table - expanded spelling of the group names; added Beam series which contains the Stokes bfData Group Types, which were mislabeled in the previous version.
1.16	—	—	Changed the naming scheme of the BFH5 structure: 'Station Beam' changed to 'Primary Pointing Direction' and 'Pencil Beam' to 'Beam'. All text and figures have been updated.
1.17	—	—	Changed attributes with names with 'PB' to 'Beam'. Added section on expected BF file sizes to the Discussion Section (Section 6).
1.18	—	—	Updated the Discussion section with future enhancement requests. Decision to use the value of zero for all data padding/-gaps.
1.19	—	—	Moved Tracking attribute to the Beam Group. Fixed some table widths to make the large tables fit better on a single page.
1.20	2010-06-16	—	Changed ALT/AZ Attributes to be of type string. Changed `CENTER_FREQUENCY` to `FREQUENCY_CENTER`.
1.21	2010-06-21	—	Changed the author list to primary author + all others in alphabetical order.
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.
2.00.01	2010-07-13	all	Updated to include Jochen Eisloeffel's comments; updated the Change Record section to match the other ICD's by using a table instead of item list per change.
2.00.02	2010-08-06	all	Making ICD's more consistent (JMG)
2.00.03	2010-10-27	Sec 4	Removed `TOTAL_INTEGRATION_TIME` and _UNIT because all beams will have the same observation time
2.00.04	2010-12-03	Change record	Converted earlier change records to table format, to be more consistent with the other ICDs.
2.00.05	2010-12-06	4.5, 4.7	A few minor adjustments following closer inspection of format during implementation as part of the DAL: `NOF_STOKES` moved up one level to the Beam group. Renamed *Stokes groups* → *Stokes datasets*. New attribute `STOKES_COMPONENT` attachd to Stokes Dataset.
2.00.06	2010-12-06	all	Using LaTeX package `hyperref` for references, enabling better navigation through the document and access to external resources. Added note on version numbering scheme.

continued on next page

continued from previous page

Version	Date	Sections	Description of changes
2.00.07	2011-01-17	§4.9.2	Fixed Full Stokes with CoherentSum, `SIGNAL SUM = COHERENT` not `''INCOHERENT''` as had been indicated.
2.01.00	2011-01-18	4.7; ??	Enable recording number of channels to vary between frequency sub-bands; `NOF_CHANNELS` now array with `NOF_SUBBANDS` entries. Minor cleaning up to avoid overfull hboxes.
2.02.00	2011-02-10	all	Get naming scheme for beam groups and Stokes datasets consistent.
2.03.00	2011-02-22	all	Adjust naming scheme for groups a datasets according to decision made at meeting by Data Formats Group.
2.03.01	2011-03-10	all	Maintain list of references through BibLaTeX database.
2.04.00	2011-03-15	all	Updated Attribute and Group names to all capitals and to use _ for multi-word cases; updated Coordinates groups to have TIME and FREQUENCY instead of Linear and Tabular top level logic. All figures were also updated to match text. All references to 248 subbands were changed to 244 subbands, now the maximum.
2.04.01	2011-04-26	all	Changed FREQUENCY coordinate to SPECTRAL coordinate. Changed the Data Size table to match the one in the LOFAR Pulsar Paper which is more extensive.
2.04.02	2011-05-11	4.6	Added paragraph with notation conventions.
2.04.03	2011-05-25	all	Updated many sections using comments from the Data Formats Group in order to make all ICDs as similar as possible; Small typos and corrections were also made. Quotes were changed to be backquote for the latex forward quote syntax. Moved Section 6 into Appendix; moved Data Flow section into Appedix.
2.04.04	2011-07-05	all	Changed PROCESS_HIST to PROCESS_HISTORY; updated the group type names - took out underscores; added horizontal lines in tables to separate the attributes from the groups.
2.04.05	2011-07-05	4.5	Changed units from microseconds to seconds, in order to be consistent with Dynamic Spectrum ICD.
2.04.05	2011-07-05	3	Moved the HDFView figure to the appendix.
2.04.06	2011-07-06	all	Matching up group type attributes and notation; consolidation of labels to refer to standard sections and tables.
2.04.07	2011-09-02	all	Removed instances of "currently" with a date from text; added list of all ICDs as applicable documents; removed NOF_STATIONS and STATIONS_LIST from SUBARRAY_POINTING Groups - these are valid in the top-level for all SUBARRAY_POINTING Groups; the BEAM Groups can have different NOF_STATIONS in Fly's Eye Mode.
2.04.08	2011-09-02	all	Changed all 'float' types to 'double'; changed all 'bool' types to 'unsigned int'; changed all 'integer' to 'int'.
2.04.09	2011-09-25	4.3	ALT and AZ keyword datatypes from strings to double; removed CLOCK_RATE and CLOCK_RATE_UNIT from the SAP Group since the Root Group applies to the rest of the groups in the file; changed EXPTIME_START/END UTC to type double; changed OBS_DATATYPE to OBSERVATION_DATATYPE; added an S to NOF_SUB_ARRAY_POINTING.

continued on next page

continued from previous page

Version	Date	Sections	Description of changes
2.04.10	2011-09-26	all	Added PROCESS_HISTORY into SAP table; changed ProcessHist group type to ProcessHistory; changed the PROCESS_HISTORY table to better mimic ICD004 with the _PARSET and _LOG following the item they describe.
2.04.11	2011-12-05	4.4; 4.5	Changed 'unsigned int' back to 'bool' for those keys which were originally boolean types.
2.04.12	2011-12-18	4	Added input file metadata_intro.tex which introduces the optional vs required header keys.
2.04.13	2011-12-18	all	Changed a few parameters to optional in the Attributes tables: weather-related parameters along with alt/az and parsets.
2.04.14	2011-12-18	A.5	Added a new section is appendix summarizing emails regarding the HBA_JOINED discussion.
2.04.15	2011-12-19	A.5	Added additional text to HBA JOINED discussion section; moved the text into a separate latex file as import, so that it can also be used by ICD006.
2.04.16	2012-01-09	All	Added new import tex file containing the data types information, just before the Acknowledgements section.
2.04.17	2012-01-09	title page	Changed the svnInfoRevision to svnInfoMaxRevision, in order to take the sub-tex file changes into account for the latex compile.
2.04.18	2012-01-09	title page	Changed svnInfoDate to svnInfoMaxToday, similar to above change of Rev number, this is the date of last revision.

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

Acknowledgements

1. Introduction

1.1. Purpose and Scope

This Interface Control Document (ICD) sets forth a formal data interface specification for LOFAR beam-formed data products. The specification applies to data structures produced by various LOFAR processing pipelines that will be called LOFAR Beam-Formed (BF) data.

LOFAR beams can be formed where signals from antennae and/or stations are coherently added to form sets of smaller beams, each generating a time-series voltage sum [?]. Alternatively, beams can also be incoherently added to create total intensities.

This document is intended to be the formal interface control agreement between the LOFAR project, observers/users of LOFAR data products, and the eventual LOFAR science archive facility.

1.2. Context and Motivation

A LOFAR Beam-Formed data file will be the data hosting structure for LOFAR Beam-Formed data. It is therefore incumbent on the LOFAR project to define and describe the structure of the LOFAR BF file format, and how the various data types are defined and described within the context of that format.

While tables/arrays are the primary data product of the BF pipeline, they are accompanied by a number of by-products, such as processing history parameters and logs, multiple coordinate sets, etc. In the more traditional approach, where all such products are stored and managed separately, a large amount of book-keeping is required to maintain consistency. For the LOFAR project, a BF file product will be defined within the context of the Hierarchical Data Format 5, or HDF5 [1]. HDF5 allows for storage, not only of the data, but for the associated and related meta-data describing the data's contents, conditions of observations, etc.. As an "all-in-one" wrapper, the HDF5 format simplifies the management of what are expected to be very large datasets that formats such as FITS cannot pragmatically accommodate.

1.3. Applicable documents

Table 1 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.

[1]For further discussion about HDF5 as the appropriate choice for the LOFAR data products, see the publicly readable LOFAR document repository. Download a full set of LOFAR documentation from the svn repository, `http://usg.lofar.org/svn/documents/trunk usg-docs`. A "data format meeting" regarding the issue was held on Oct 8, 2008 at ASTRON.

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

LOFAR data will be presented in a number of LOFAR data formats, all of which will provide data arrays of differing dimensions, depending upon the respective observation. Dynamic Spectra, Sky Image Cubes, Rotation Measure Cubes, Near-field cosmic ray images ("CR image" in Table 2), etc., all have different dimensions and coordinate types. Table 2 illustrates the various data array dimensions that LOFAR may produce.

Image	ICD	Quantity	Axes	Units
TBB time-series	001 / [?]	$I(t)$	Time	s
BF data	003 / [?]	$I(p, \nu, \mathrm{Dec}, \mathrm{RA})$	Pol/Freq/Dir/Dir	.. /Hz/deg/deg
Sky image	004 / [?]	$I(p, \nu, \mathrm{Dec}, \mathrm{RA})$	Pol/Freq/Dir/Dir	.. /Hz/deg/deg
Dyn. Spectrum	006 / [?]	$I(p, \nu, t)$	Pol/Freq/Time	.. /Hz/s
RMSC	008 / [?]	$DF(p, \mathrm{Dec}, \mathrm{RA}, \phi)$	Pol/Dir./Dir./Faraday Depth	.. /deg/deg/rad m^{-2}
RM map	—	$RM(\mathrm{Dec}, \mathrm{RA})$	Dir./Dir.	/deg/deg
CR image	—	$I(p, \nu, r, \mathrm{El}, \mathrm{Az})$	Pol/Freq/Dist/Dir./Dir./	.. /p/Hz/m/deg/deg
CR image	—	$I(p, t, \nu, \xi_3, \xi_2, \xi_1)$	Pol/Time/Freq/Pos/Pos/Pos	.. /s/Hz/m/m/m

Table 2: Overview of the various data arrays types, associated coordinates and dimensions. Where possible a reference for the data format specification is provided.

Each data type is described in detail by an appropriate interface control document. This document pertains to, and describes only those data conforming to the LOFAR datatype "Beam-Formed Data". The dataset array in a Beam-Formed dataset will be a `ndarray` data structure, as can be created by the C-based *numarray/numpy* Python packages. The Beam-Formed data format is designed to store the time-series data produced from the Pulsar pipeline. The nominal dimensionality of a Beam-Formed `Data` group's dataset will be 2 (`N_AXIS=2`), wherein the data arrays will be defined in (C-type order) *time*, *spectral*, as shown in Table 2.

This document is structured as follows: Section 3 will describe Beam-Formed data flow from the LOFAR stations to the first BF dataset product, the fundamental overall file structure, syntax rules, i.e. file format, and semantic conventions the interface will adhere to, 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 4 will describe the detailed specification for the data, including a description of the overall structure of a Beam-Formed file, file naming conventions, units, physical quantities. Section 4 will also provide a detailed description of the group and dataset structures contained within the BF file format, along with meta-data in the form of HDF5 dataset headers.

3. Organization of the data

3.1. High level LOFAR BF file structure

A LOFAR Beam-Formed file will adhere to the following guidelines:

A LOFAR BF file will be defined within the context of the HDF5 file format. An effort was made to minimize the hierarchical depth of the file structure, in order to make the BF file as "flat" as possible. However, a BF file cannot be completely "flat" due to the inherent, scientific, logical subdivisions within the data. Therefore access to the necessary data table must be done via hierarchical tree crawling.

The BF 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.

3.2. Overview of BF Groups

Figure 1 shows the high level group structure of a BF H5 file (Version 1.0). The example shown here is for 3 Sub-Array Pointings (0, 1, NNN) each with 3 Beams (0, 1, NNN). The Beams are color coded, so that the series of sub-groups attached to each Beam Group is the same color as its' parent group (shown in green, pink and gray). The data are at the Stokes Dataset level (hierarchical level 4); the Stokes data storage containers themselves are not shown here.

Narrowing in on the Beam-Formed high level structure diagram (Figure 1), Figure 2 shows the same structure but only for one pathway (within one Sub-Array Pointing, one Beams' data storage tables/arrays) (Version 1.0).

A LOFAR BF file will comprise a `System Log Group` just below the root level which contains logs and parameter files which are relevant to the entire BF file. Additionally just below the root level, the BF file will contain an arbitrary, observation-dependent number of `Sub-Array Pointing Groups` containing N `Beam Groups` each with it's own pointing information in the header; a `Processing History Group` at this level keeps track of logs and parameter sets relevant to the Sub-Array Pointings. Each `Beam Group` will contain a `Processing History Group` (relevant to the Beams), a `Coordinates Group` as well as one or four `Stokes Datasets`. The `Stokes Datasets` contain the tabular/array data.

These main building blocks of the BF HDF5 file are:

1. **File Root-level** (`ROOT`). The root level of the file contains the majority of associated meta-data, describing the circumstances of the observation. These data attributes include time (start and end), frequency window (high band vs. low band, lters) and other important characteristics of the dataset. See Sections 4.1 and 4.1.1 for 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. See Sec. 4.2 for details.
3. **Sub-Array Pointing Groups** (`SUB_ARRAY_POINTING_{NNN}`). Each observation sub-array pointing is stored as a seperate group within the file, each containing its own beam groups. Characteristics about each sub-array pointing such as direction are stored as Attributes in group headers. LOFAR can observe with up to several hundred sub-array pointings, since you can "trade bandwidth for beams" (less data per beam in order to gain additional beams). Therefore it is unknown as to the maximum number of `Sub-Array Pointing Groups` in a BF H5 file. See Section 4.3 for details.

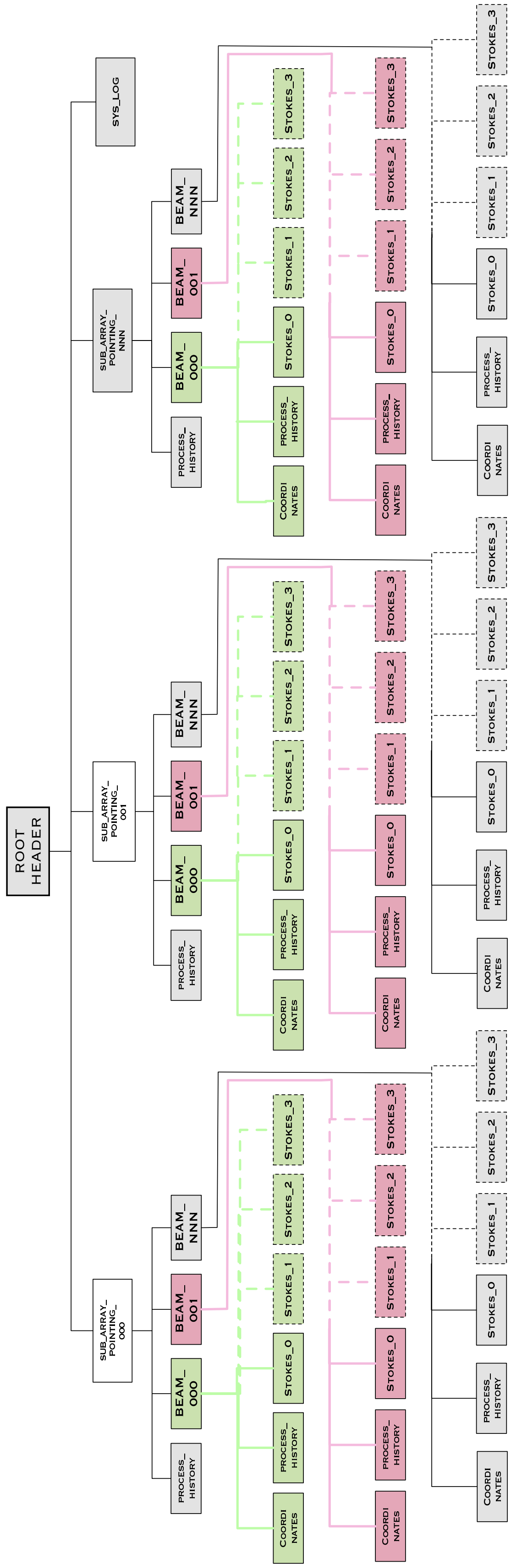


Figure 1: Beam Formed HDF5 High Level Data Structure; note, no tables/arrays are shown, but are implied at the Stokes level. (BF version 1.0)

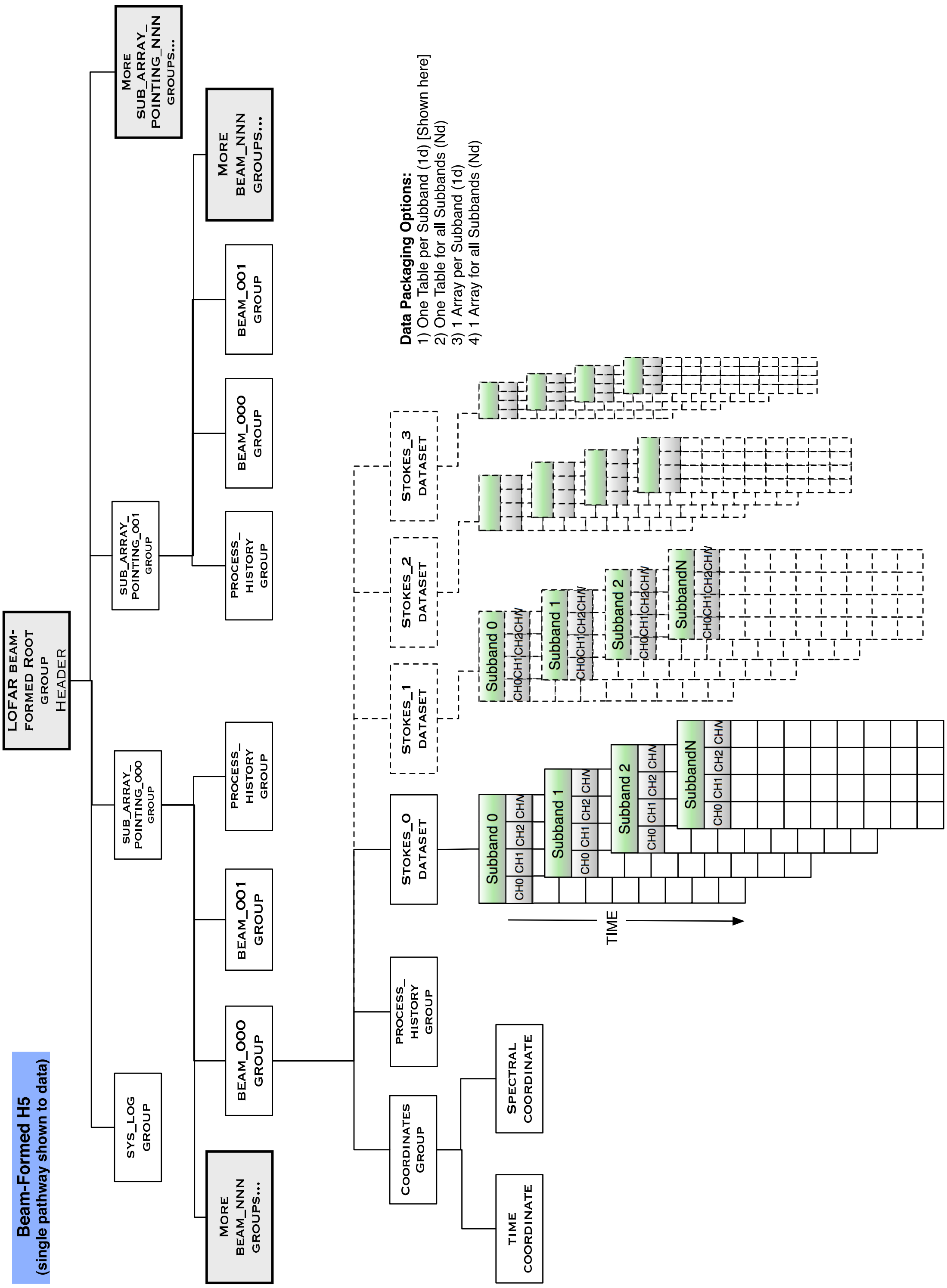


Figure 2: Segment of Beam Formed HDF5 Data Structure for 3 Sub-Array Pointings with 3 Beams each. (BF version 1.0)

4. **Beam Groups** (`BEAM_{NNN}`). Each observation beam is stored as a seperate group within the `Sub-Array Pointing Groups`, each containing its own set of `Coordinate` and `Stokes Datasets` with the data. The pointing information is given for each Beam as Attributes. The starting frequency of the subband (start frequency of all the channels) information is stored as World Coordinate System (WCS) information in the Beam Coordinates Group. Beams can be incoherently added, coherently added or not added at all. This group encompasses a variety of possible 'beams', including the most known which will be used frequently, called the Tied-Array Beam. See Section 4.5 for details.
5. **Coordinates Groups** (`COORDINATES`). Each `Beam Group` contains one `Coordinates Group`, which stores the time series relation for the data stored in the lowest (4th) hierarchical depth of the file in the `Stokes Datasets`. The `Coordinates Group` will contain 2 Coordinate types (1 Time and 1 Spectral (Frequency) Coordinate) as an associated pair. See Section 4.6 for details.
6. **Processing History Groups** (`PROCESS_HISTORY`). `Processing History Groups` can be found on the `Sub-Array Pointing Group` and `Beam Group` levels. These are catch-all envelops encapsulating information about all the steps of processing, such as parameter sets and processing logs. See Section 4.4 for details.
7. **Stokes Datasets** (`STOKES_{N}`). Each Beam has a fixed number `Stokes Datasets`, one per Stokes type (or polarizations). The Stokes type depends on the type of observation and summing of the data. Without summing there are four Stokes types in the file: I, Q, U, V within the `STOKES_0`, `STOKES_1`, `STOKES_2` and `STOKES_3` datasets, respectively. With summing, there is only one Stokes type which is the Intensity (I) stored within the `STOKES_0` dataset. The channel frequency information is stored as an attribute of the table. Sometimes, the Stokes data groups will be used to store the "raw" data, which are two 32-bit complex numbers (Xreal, Ximaginary, Yreal and Yimaginary), also stored within the `STOKES_0`, `STOKES_1`, `STOKES_2` and `STOKES_3` datasets, respectively. See Section 4.7 for details.
   For each Stokes Dataset, the subband and channel data are stored as HDF5 "Datasets" – it is at this 4th hierarchical depth that the bulk of the data reside. The data storage options are still being investigated, in order to determine the maximum efficiency of data seeks and file I/O. Of the four data storage options under consideration, we are likely to use option number 2 below:
   a) One-dimensional table per subband, where each column represents a channel observed. (Note, this option is used for most diagrams in this ICD.)
   b) One-dimensional array per subband, where each column represents a channel observed.
   c) N-dimensional single table, where dimensions N and M are subband and channel.
   d) N-dimensional single array, where dimensions N and M are subband and channel.

3.3. Hierarchical Structure of the BF H5 File

The beamformed data are organized within a hierarchical structure, which reflects upon the structure in which data are grouped during processing. This structure can be represented as a HDF5 hierarchy with all four Stokes Datasets present:

```
OBSERVATION/
OBSERVATION/SYS_LOG
OBSERVATION/SUB_ARRAY_POINTING_000/PROCESS_HISTORY
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_000
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_000/PROCESS_HISTORY
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_000/COORDINATES
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_000/COORDINATES/TIME/LINEAR
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_000/COORDINATES/SPECTRAL/TABULAR
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_000/STOKES_0
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_000/STOKES_1
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_000/STOKES_2
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_000/STOKES_3
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_001
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_001/PROCESS_HISTORY
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_001/COORDINATES
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_001/COORDINATES/TIME/LINEAR
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_001/COORDINATES/SPECTRAL/TABULAR
```

```
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_001/STOKES_0
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_001/STOKES_1
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_001/STOKES_2
OBSERVATION/SUB_ARRAY_POINTING_000/BEAM_001/STOKES_3
...
OBSERVATION/SUB_ARRAY_POINTING_NNN/BEAM_NNN/PROCESS_HISTORY
OBSERVATION/SUB_ARRAY_POINTING_NNN/BEAM_NNN/COORDINATES
OBSERVATION/SUB_ARRAY_POINTING_NNN/BEAM_NNN/COORDINATES/TIME/LINEAR
OBSERVATION/SUB_ARRAY_POINTING_NNN/BEAM_NNN/COORDINATES/SPECTRAL/TABULAR
OBSERVATION/SUB_ARRAY_POINTING_NNN/BEAM_NNN/STOKES_0
OBSERVATION/SUB_ARRAY_POINTING_NNN/BEAM_NNN/STOKES_1
OBSERVATION/SUB_ARRAY_POINTING_NNN/BEAM_NNN/STOKES_2
OBSERVATION/SUB_ARRAY_POINTING_NNN/BEAM_NNN/STOKES_3
...
```

Similarly, one can represent the same structure as such:

```
L1002977_B000_S0_P000_bf.h5      .. Group      .. Root level of the HDF5 file
|-- SYS_LOG                      .. Group      .. System-Wide Logs group
|-- SUB_ARRAY_PONTING_000        .. Group      .. First Sub-Array Pointing group
|   |-- PROCESS_HISTORY          .. Group      .. Process History Group for Sub-Array Pointing Group
|   |-- BEAM_000                 .. Group      .. First Beam group
|   |   |-- PROCESS_HISTORY      .. Group      .. Processing History Group for Beam Group
|   |   |-- COORDINATES          .. Group      .. First Coordinates group
|   |   |   |-- TIME             .. Group      .. Time Coordinates group (LinearCoord)
|   |   |   `-- FREQUENCY        .. Group      .. Frequency Coordinates group (TabularCoord)
|   |   |-- STOKES_0             .. Dataset    .. First Stokes (I) dataset
|   |   |-- STOKES_1             .. Dataset    .. First Stokes (Q) dataset
|   |   |-- STOKES_2             .. Dataset    .. First Stokes (U) dataset
|   |   `-- STOKES_3             .. Dataset    .. First Stokes (V) dataset
|   |-- BEAM_001                 .. Group      .. Second Beam beam group
|   |   |-- PROCESS_HISTORY
|   |   |-- COORDINATES
|   |   |    |-- TIME
|   |   |    `-- FREQUENCY
|   |   |-- STOKES_0
|   |   |-- STOKES_1
|   |   |-- STOKES_2
|   |   `-- STOKES_3
|   `-- BEAM_{NNN}
|-- SUB_ARRAY_PONTING_001        .. Group      .. Second Sub-Array Pointing group
|   |-- PROCESS_HISTORY
|   |-- BEAM_000
,
```

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 (`ROOT`)

The LOFAR file hierarchy begins with the top level **File Root Group** (`ROOT`). This is the file entry point for the data, and the file node by which navigation of the data is provided. The **File Root Group** will comprise a set of attributes that describe the underlying file structure, observational metadata, the LOFAR Beam-Formed Data, as well as providing hooks to all groups attached to the **File Root Group**.

This section will specify two attributes sets that will appear in the `ROOT Group`: a set of Common LOFAR Attributes (CLA) shared by all LOFAR science data products, and a set of attributes that are specific to LOFAR Beam-formed 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 Beam-Formed-specific. In other words,

```
ROOT Attributes = Common LOFAR Attributes (CLA) + Additional Beam-Formed 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 BF files, as well as other LOFAR timing and image analysis data. *All* LOFAR data products, including Beam-Formed *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 3 lists the Common LOFAR Attributes (CLA) which can be found in LOFAR Beam-Formed data products within the files' root header. They will only appear in LOFAR data products comprising TBB and Beam-formed data. 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.

For LOFAR Beam-Formed Data, FILETYPE='bf'.

- 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. 7, 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 4 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.
- 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

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 calibration.
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.

- 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:

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

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 5 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 6 below for a list of recognized 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.

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.

– 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 Beam-Formed Root Attributes

Table 8 contains the Beam-Formed Data root header Attributes. The CLA Attributes have already been listed in Sec. 4.1.1 above, therefore these are additional attributes in the root header.

4.2. The System Logs Group (SYS_LOG)

Comment:
This is where the SYS-LOG Group should be described.

General LOFAR Group	**Group Type Value**	**Description**
Root	`'Root'`	Top-level LOFAR group type
System Log	`'SysLog'`	System log files, parsets
Sub-Array Pointing	`'SubArrayPointing'`	Sub-Array Pointing group
Sub-Array Pointing Group Subgroups	**Value**	**Description**
Beam Group	`'Beam'`	This is a Beam group
Processing History Group	`'ProcessHistory'`	This is a Processing History group
Coordinates Group	`'Coordinates'`	This is a Coordinates group
Beam Group Subgroups	**Value**	**Description**
Processing History Group	`'ProcessHistory'`	This is a Processing History group
Stokes Datasets	`'bfData'`	This is a series of Datasets
Coordinates Group	`'Coordinates'`	This is a Coordinates group
Coordinates Group Subgroups	**Value**	**Description**
Time Coordinate group	`'TimeCoord'`	Time axis is described in this coord group
Spectral (Frequency) Coordinate group	`'SpectralCoord'`	Frequency axis is described in this coordinate group

Table 7: Beam-Formed Group Types.

4.3. The Sub-Array Pointing Group (SUB_ARRAY_POINTING_{NNN})

There are N Sub-Array Pointings in a BF file, where the maximum N is currently unknown and can be greater than 8. The Attributes listed in Table 9 are found in the header of each of the Sub-Array Pointing (SUB_ARRAY_POINTING_{NNN}) Groups.

4.4. The Processing History Group (PROCESS_HISTORY)

The data definition for the `Processing History (PROCESS_HISTORY)` Group is necessarily loose, and will accommodate a variety of ancillary meta-data related to or produced by the various LOFAR processing pipelines. In BF data `Processing History` Groups can be found at the Sub-Array Pointing level, which describe information for all Sub-Array Pointings, and at the Beam level, which describe information for all the Beams.

Products such as Presto log files, processing parameters sets, ephemeris used for folding the data, RFI mitigation tables, etc. In fact, and due the wide-ranging data types and free-form ASCII format of many log files that the `Processing History` Group may encompass, this group will be a catch-all envelop encapsulating information about all steps of processing should the user need such information. And it is because of this free-form nature of the meta-data that it is very difficult to define a header describing attached data when it is not yet know just what those data may include. An attempt has been made to provide by example how this will or should appear in the `Processing History` Group header (Table 10 and Figure 4.4).

There is one `Processing History` Group per Beam. The attributes will contain a brief summary of the appended processing files contained therein, with pointers to tables containing the logging data, parameter sets, etc..

[**Comment:**
a) Need to update the information of what the BF log and parset files are and how many there will be. b) Update figure to match the BF structure for the Processing History Group.]

4.5. The Beam Group (BEAM_{NNN})

There are N Beam Groups per Sub-Array Pointing Group, where the maximum N maybe around 200. The Attributes listed in Table 11 are found in the header of each of the BEAM_NNN Groups.

Field/Keyword	H5Type	Type	Unit	Description
CREATE_OFFLINE_ONLINE	Attr.	string	—	Whether the file was created 'Online' (stream to file) or 'Offline' (file to file).
BF_FORMAT	Attr.	string	—	Set to "RAW" if the file is BeamFormed RAW data; set to "TAB" if the file is BeamFormed Processed data.
BF_VERSION	Attr.	string	—	BeamFormed data format version number.
EXPTIME_START_UTC	Attr.	string	hh:mm:ss.sssssssss	Start time of obs (UTC time, 24-h clock).
EXPTIME_STOP_UTC	Attr.	string	hh:mm:ss.sssssssss	End time of obs (UTC time, 24-h clock).
EXPTIME_START_MJD	Attr.	double	NNNNNN.NNNNNNN	Start time of obs (MJD).
EXPTIME_STOP_MJD	Attr.	double	NNNNNN.NNNNNNN	End time of obs (MJD).
EXPTIME_START_TAI	Attr.	string	hh:mm:ss.sssssssss	Start time of obs (International Atomic Time).
EXPTIME_STOP_TAI	Attr.	string	hh:mm:ss.sssssssss	End time of obs (International Atomic Time).
TOTAL_INTEGRATION_TIME	Attr.	double	s	Total integration time of the observation.
OBSERVATION_DATATYPE	Attr.	string	—	Type of observation: Searching, timing, etc. —
SUB_ARRAY_POINTING_ DIAMETER	Attr.	double	arcmin	FWHM of the sub-array pointing at zenith at center frequency.
BANDWIDTH	Attr.	double	MHz	Total bandwidth (excluding gaps).
BEAM_DIAMETER	Attr.	double	arcmin	FWHM of the beams at zenith at center frequency.
WEATHER_TEMPERATURE	Attr.	array<double,1>	$^{\circ}C$	Approximate outside temperature; order must match the listing in STATIONS_LIST attribute. (* See note below)
WEATHER_HUMIDITY	Attr.	array<double,1>	—	Approximate humidity (%); order must match the listing in STATIONS_LIST attribute (* See note below)
SYSTEM_TEMPERATURE	Attr.	array<double,1>	K	System temperature for the various stations ; order must match the listing in STATIONS_LIST attribute. (* See note below)
NOF_SUB_ARRAY_POINTINGS	Attr.	int	—	Number of sub-array pointing synthesized beams.
SUB_ARRAY_POINTING_{NNN}	Group	---	—	Container for individual Sub-Array Pointing objects
SYS_LOG	Group	---	—	Container for system-wide log object

**WEATHER_TEMPERATURE, WEATHER_HUMIDITY and SYSTEM_TEMP - this info does not exist for each station; just some of the stations are hooked up with weather equipment. Additionally there is temperature information within the electronics cabinet, but once again, this info is NOT per station, and therefore cannot match the STATION_LIST one-to-one. Therefore, we may need a STATIONS_LIST_TEMPERATURE, STATIONS_LIST_HUMIDITY and STATIONS_LIST_SYS_TEMPERATURE Attributes for list matching the arrays for the weather numbers. Lars: If indeed we also want to be able to store meteorological data within the files, I would rather argue for this to be encapsulated into a separate group and not have to have a mixed set of data added to the root level of the file. Depending on for which system components which specific readings are available, this might require a bit more flexibility as is expected from the root level attributes*

Table 8: The Additional Root Header Group Attributes.

FIELD/KEYWORD	H5TYPE	TYPE	UNIT	Description
GROUPTYPE	Attr.	string	—	Group type = 'SubArrayPointing'
POINT_RA	Attr.	double	'deg'	J2000 RA of SAP at start.
POINT_DEC	Attr.	double	'deg'	J2000 DEC of SAP at obs start
POINT_ALTITUDE	Attr.	array<double,1>	'deg'	Alt. of SAP, obs start.
POINT_AZIMUTH	Attr.	array<double,1>	'deg'	Az. of the SAP, obs start.
NOF_SAMPLES	Attr.	int	—	number of time samples.
SAMPLING_RATE	Attr.	double	'MHz'	Sampling rate, per sub-array pointing, in units of SAMPLING_RATE_UNIT
SAMPLING_RATE_UNIT	Attr.	string	—	Sampling rate units
SAMPLING_TIME	Attr.	double	's'	Sampling time is 1/SAMPLING_RATE, per sub-array pointing, in units of SAMPLING_TIME_UNIT
SAMPLING_TIME_UNIT	Attr.	string	—	Sampling time units are microseconds
CHANNELS_PER_SUBBAND	Attr.	int	—	Number of channels for each subband
SUBBAND_WIDTH	Attr.	double	'MHz'	Subband width 0.156250 or 0.1953125
SUBBAND_WIDTH_UNIT	Attr.	string	—	Subband width units
CHANNEL_WIDTH	Attr.	double	'MHz'	Channel width is equal to the SUBBAND_WIDTH / CHANNELS_PER_SUBBAND
CHANNEL_WIDTH_UNIT	Attr.	string	—	Channel width units
NOF_BEAMS	Attr.	int	–	Number of beams.
PROCESS_HISTORY	Group	---	—	container for processing history for any of the SAPs
BEAM_{NNN}	Group	---	—	container for individual Beam objects

Table 9: The Sub-Array Pointing Group Header Attributes

FIELD/KEYWORD	H5TYPE	TYPE	VALUE	DESCRIPTION
GROUPTYPE	Attr.	string	'ProcessHistory'	BF group type.
OBSERVATION_PARSET	Attr.	bool		Observing parset present? (True=1=Yes/False=0=No)
OBSERVATION_LOG	Attr.	bool		Observing log present? (True=1=Yes/False=0=No)
PRESTO_PARSET	Attr.	bool		Presto parset present? (True=1=Yes/False=0=No)
PRESTO_LOG	Attr.	bool		Presto log present? (True=1=Yes/False=0=No)

Table 10: Processing History Group Attributes

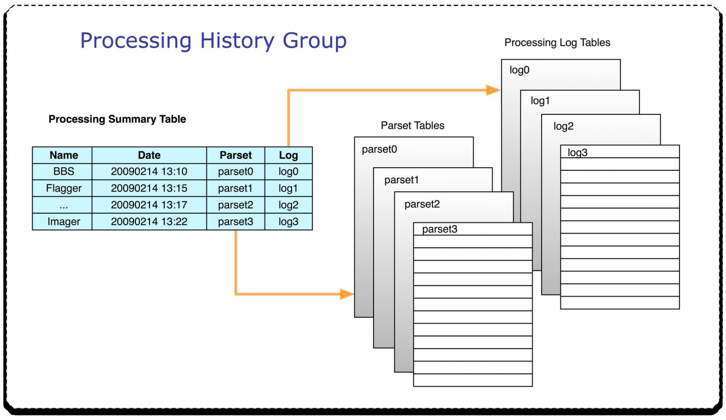


Figure 3: The processing history group (example for Sky Image ICD), nested tabulation

Field/Keyword	H5Type	Type	Unit	Description
GROUPTYPE	Attr.	string	'Beam'	Group type
TARGET	Attr.	array<string,1>	—	Single or list of observation targets/-sources observed within this Beam
NOF_STATIONS	Attr.	int	—	Number of stations used within this Beam
STATIONS_LIST	Attr.	array<string,1>	—	List of stations used for this Beam; list must match the NOF_STATIONS numerically
TRACKING	Attr.	string	–	Indication whether tracking was turned on. ('J2000' = tracking ON; 'LMN' = tracking OFF; TBD = tracking of solar system object)
POINT_RA	Attr.	double	—	J2000 right ascension of the center of the beam at start of observation, in degrees (at LOFAR core).
POINT_DEC	Attr.	double	—	J2000 declination of the center of the beam at start of observation, in degrees (at LOFAR core).
POSITION_OFFSET_RA	Attr.	double	—	RA offset of beam from sub-array pointing at start of observation (degrees).
POSITION_OFFSET_DEC	Attr.	double	—	Dec offset of beam from sub-array pointing at start of observation (degrees).
BEAM_DIAMETER_RA	Attr.	double	—	The diameter of the beam ellipse in the major axis a (RA), units of degrees, at zenith at center frequency.
BEAM_DIAMETER_DEC	Attr.	double	—	The diameter of the beam ellipse in the minor axis b (Dec), units of degrees, at zenith at center frequency.

continued on next page

FIELD/KEYWORD	H5TYPE	TYPE	UNIT	DESCRIPTION
BEAM_FREQUENCY_CENTER	Attr.	double	—	Beam center frequency - the middle frequency of all the channel frequencies.
BEAM_FREQUENCY_CENTER_UNIT	Attr.	string	'MHz'	Beam center frequency units.
FOLDED_DATA	Attr.	bool	—	Are the data folded within this Beam? (True=1=Yes/False=0=No)
FOLD_PERIOD	Attr.	double	—	The fold period if the data were folded.
FOLD_PERIOD_UNIT	Attr.	string	's'	The fold period units.
DEDISPERSION	Attr.	string	—	Where the data dedispersed incoherently (INCOHERENT), dedispersed coherently (COHERENT), or not dedispersed at all (NONE)?
DISPERSION_MEASURE	Attr.	double	—	The dispersion measure applied to the data, if data were de-dispersed
DISPERSION_MEASURE_UNIT	Attr.	string	'pc/cm^3'	The dispersion measure units.
BARYCENTER	Attr.	bool	—	Are the data barycentered? (True=1=Yes/False=0=No)
NOF_STOKES	Attr.	int	—	Number of Stokes tables/arrays (1 or 4).
STOKES_COMPONENTS	Attr.	array<string,1>	—	Stokes components for which data are attached to this group.
COMPLEX_VOLTAGE	Attr.	bool	—	Is the data in Complex Voltages (Xreal Ximg, Yreal, Yimg)? (True=1=Yes/False=0=No)
SIGNAL_SUM	Attr.	string	—	For any **Stokes** observation modes or Complex Voltages, was the signal summed coherently (COHERENT) or incoherently (INCOHERENT)?
PROCESS_HISTORY	Group	---	—	Container for Processing History for the Beam.
COORDINATES	Group	---	—	container for Coordinates Group for this entire Sub-Array Pointings' set of data.
STOKES_0	Dataset	string	—	Container for Stokes_I data or XX.
STOKES_1	Dataset	string	—	Container for Stokes_Q data or XY.
STOKES_2	Dataset	string	—	Container for Stokes_U data or YX.
STOKES_3	Dataset	string	—	Container for Stokes_V data or YY.

Table 11: Beam Group Attributes.

- TARGET – Single or list of observation targets/sources observed within this Beam
- POINT_RA & POINT_DEC – J2000 right ascension and declination of the center of the beam at start of observation, in degrees (at the LOFAR core, as the positions will be slightly different for solar system objects).
- STOKES_COMPONENTS – Providing both flexibility and readability, the Stokes components for which data are attached to the Beam Group, are stored as array<string,1>. Besides the currently envisioned choices

  STOKES_COMPONENTS = ["I"]
  STOKES_COMPONENTS = ["I","Q","U","V"]

  the attribute also has the flexibility to host other combinations and types of Stokes data, such as e.g.

STOKES_COMPONENTS = ["I","Q"]
STOKES_COMPONENTS = ["L","R"]
STOKES_COMPONENTS = ["XX","XY","YX","YY"]

4.6. The Coordinates Group (COORDINATES)

Coordinate information within a LOFAR BF file will exist in what is called a Coordinates (COORDINATES) Group, which will act as a container for at least one Coordinates Group object. There is one Coordinates Group per Beam Group in the BF data. There are two Coordinates Group objects, a TIME and SPECTRAL (frequency) coordinate set which apply to that Beam. As show in Table 12 below, this group only contains a small number of attributes by itself and is a container for the coordinate sub-groups which contain the World Coordinate Information (WCS) of the data within that Beam Group.

FIELD/KEYWORD	TYPE	DESCRIPTION
GROUPTYPE	string	Group type descriptor, Coordinates
REF_LOCATION_VALUE	array<double,1>	Numerical value(s) of the reference location
REF_LOCATION_UNIT	array<string,1>	Physical unit(s) for the reference location
REF_LOCATION_FRAME	string	Identifier for the reference system of the location; see Tab. 13 for a list of recognized values.
REF_TIME_VALUE	double	Numerical value of the reference time
REF_TIME_UNIT	string	Physical unit of the reference time
REF_TIME_FRAME	string	Identifier for the reference time system used
NOF_COORDINATES	int	N of coordinate objects
NOF_AXES	int	N of coordinate axes
COORDINATE_TYPES	array<string,1>	embedded coordinate object types
COORDINATE_{N}	Group	coordinate object container

Table 12: Components of a Coordinates group.

In this:

- NOF_COORDINATES — The number of coordinate objects/groups contained within the coordinates group (BF data, this is 2: TimeCoord and SpectralCoord).
- NOF_AXES — The number of coordinate axes associated with the coordinate objects. Keep in mind, that a coordinate can have multiple (coupled) axes: e.g. a direction coordinate is composed of two axes. In the case of BF data, NOF_AXES = 2, they are 'Time' and 'Spectral' (representing frequency).

The different axes stored within this container are described in the remainder of this section. More information on coordinates in LOFAR data files in general can be found in ICD-002 [?].

4.6.1. The Time coordinate

This group describes the time coordinate of the associated BEAM Group.
Specifially, for each LOFAR BEAM Group, the following values are used:

- COORDINATE_TYPE — The coordinate type = 'Time'.
- STORAGE_TYPE — The storage type can be either 'Linear' or 'Tabular'. For BeamFormed data, it will be 'Linear' in most cases.
- NOF_AXES — The number of coordinate axes = 1.
- AXIS_NAMES — The names of the axis = 'Time'.
- AXIS_UNITS — The units of the coordinate axis, e.g. = 'microseconds'.

Reference Position	Description	Comments
GEOCENTER	Center of the Earth.	
BARYCENTER	Center of the solar system barycenter.	
HELIOCENTER	Center of the Sun.	
TOPOCENTER	"Local"; in most cases this will mean: the location of the telescope.	
LSRK	Kinematic Local Standard of Rest: 20 km s^{-1} in the direction of GALACTIC_II $(56, +23)$.	Only to be used for redshifts and Doppler velocities, and spectral coordinate.
LSRD	Dynamic Local Standard of Rest: 16.6 km s^{-1} in the direction of GALACTIC_II $(53, +25)$.	
GALACTIC	Center of the Galaxy: 220 km s^{-1} in the direction of GALACTIC_II $(90, 0)$ w.r.t. LSRD.	
LOCAL_GROUP	Center of the Local Group: 300 km s^{-1} in the direction of GALACTIC_II $(90, 0)$ w.r.t. BARYCENTER.	
RELOCATABLE	Relocatable center; for simulations.	Only to be used for spatial coordinates.

Table 13: Recognized values for the reference frame to specify a location; values and descriptions have been adopted from the "Space-Time Coordinate Metadata for the Virtual Observatory" [?], as produced by the IVOA Data Model Working Group.

- REFERENCE_VALUE — The World Coordinate Reference value (CRVAL) will be the 0,0 location of the dataset (start time of each sample).
- REFERENCE_PIXEL — The World Coordinate Reference pixel (CRPIX) will be the 0,0 location of the dataset.
- INCREMENT — The World Coordinate increment (CDELT) is the time bin (SAMPLING_TIME).
- PC — The World Coordinate Reference scaling delta matrix is flat (1, 0) for BeamFormed data.

More detail on time coordinates for LOFAR data files can be found in Sec. 4.3.1 of ICD-002 [?].

4.6.2. The Spectral coordinate

This group describes the spectral coordinate of the associated BEAM Group.

Specifially, for each LOFAR BEAM Group, the following values are used:

- COORDINATE_TYPE — The coordinate type = 'Spectral'.
- STORAGE_TYPE — The storage type can be either 'Linear' or 'Tabular'. For BeamFormed data, it will be 'Tabular' in most cases[2].
- NOF_AXES — The number of coordinate axes = 1.
- AXIS_NAMES — The names of the axis = 'Frequency'.
- AXIS_UNITS — The units of the coordinate axis is = 'MHz'.

[2] Given the flexibility concerning the arrangement of frequency channels or subbands, the values along this coordinate axis might be linear, but do not necessarily have to be.

Field/Keyword	H5Type	Type	Value	Description
GROUPTYPE	Attr.	string	'TimeCoord'	Group Type descriptor
COORDINATE_TYPE	Attr.	string	'Time'	Coordinate Type descriptor
STORAGE_TYPE	Attr.	array<string,1>	'Linear' | 'Tabular'	coordinate storage type
NOF_AXES	Attr.	int	1	N of coordinate axes
AXIS_NAMES	Attr.	array<string,1>	['Time']	World axis names
AXIS_UNITS	Attr.	array<string,1>	['μs']	World axis units
REFERENCE_VALUE	Attr.	array<double,1>	—	Reference value
REFERENCE_PIXEL	Attr.	array<double,1>	—	Reference pixel
INCREMENT	Attr.	array<double,1>	—	Coordinate increment
PC	Attr.	array<double,1>	—	
AXIS_VALUES_PIXEL	Attr.	array<double,1>	—	Reference pixels
AXIS_VALUES_WORLD	Attr.	array<double,1>	—	Reference values

Table 14: Time Coordinate Attributes

Field/Keyword	H5Type	Type	Value	Description
GROUPTYPE	Attr.	string	'SpectralCoord'	Group type descriptor
COORDINATE_TYPE	Attr.	string	'Spectral'	Coordinate Type descriptor
STORAGE_TYPE	Attr.	array<string,1>	'Linear' | 'Tabular'	coordinate storage type
NOF_AXES	Attr.	int	1	nof. coordinate axes
AXIS_NAMES	Attr.	array<string,1>	['Frequency']	World axis names
AXIS_UNITS	Attr.	array<string,1>	['MHz']	World axis units
REFERENCE_VALUE	Attr.	array<double,1>	—	Reference value (CRVAL)
REFERENCE_PIXEL	Attr.	array<double,1>	—	Reference pixel (CRPIX)
INCREMENT	Attr.	array<double,1>	—	Coordinate increment (CDELT)
PC	Attr.	array<double,1>	—	
AXIS_VALUES_PIXEL	Attr.	array<double,1>	—	Reference pixels
AXIS_VALUES_WORLD	Attr.	array<double,1>	—	Reference values

Table 15: Spectral Coordinate Attributes

- REFERENCE_VALUE — The World Coordinate Reference value (CRVAL) will be the 0,0 location of the dataset.
- REFERENCE_PIXEL — The World Coordinate Reference pixel (CRPIX) will be the 0,0 location of the dataset.
- INCREMENT — The World Coordinate increment (CDELT) is the frequency bin (CHANNEL_WIDTH).
- PC — The World Coordinate Reference scaling delta matrix is flat (1, 0) for Dynamic Spectrum data.
- AXIS_VALUES_PIXEL — Reference pixels - List of the subbands. (See Sec. B).
- AXIS_VALUES_WORLD — Reference values - List of the equivalent frequencies of the list of subbands. (See Sec. B)

More detail on spectral coordinates for LOFAR data files can be found in Section 4.3.2 of ICD-002 [?].

4.7. The Stokes Datasets (STOKES_{N})

Within each Beam group, there are either one or four Stokes datasets; or the Stokes dataset contains BF RAW data with two 32-bit complex numbers (containing Xreal, Ximg, Yreal and Yimg). If the data are summed, then there is only one Stokes I dataset containing all the channel intensities per subbeam. If the data are not summed, then there are four Stokes tables (I, Q, U, V or XX, XY, YX, YY), one per polarization, containing all the channel intensities per subbeam. Table 16 lists the Attributes in the Stokes datasets. The general structure of the datasets is such that channels are columns and time bins are rows.

Field/Keyword	H5Type	Type	Value	Description
GROUPTYPE	Attr.	string	'bfData'	Group type.
DATATYPE	Attr.	string		Data type used (char, int, float, double).
STOKES_COMPONENT	Attr.	string	—	Stokes component stored within the dataset.
NOF_SAMPLES	Attr.	int	—	Number of bins along the time axis of the dataset.
NOF_SUBBANDS	Attr.	int	—	Number of frequency sub-bands, into which the frequency domain is being divided
NOF_CHANNELS	Attr.	array<int,1>	—	Number of channels within the subband.
bfData	Dataset	array<double,N>		

Table 16: Attributes attached to a Stokes dataset.

- GROUPTYPE is the group type of this structure, i.e. 'Data'.
- DATATYPE describes the type of the elements within the dataset.
- STOKES_COMPONENT lists the Stokes component for which data are being recorded.
- NOF_SAMPLES is the number of bins/samples along the time axis of the dataset.
- NOF_SUBBANDS is the number of frequency sub-bands, into which the frequency domain is being divided.
- NOF_CHANNELS is the number of frequency channels per sub-band; while in the most simple case all sub-bands are sub-divided into the same number of channels, in general this sub-division can vary between individual sub-bands.

4.8. The Subband/Channel tables/arrays

The Subband/Channel information will be stored as either 1-D or N-D tables or arrays. This is where the bulk of the data reside, within each Stokes group. The general structure is such that channels are columns and time bins are rows. If using 1-D tables/arrays, then each subband will be in its own table/array. If using N-D tables/arrays, then each subband will be a plane in a data cube. The maximum number of subbands is 244. As of (January 9, 2012), we have decided to try 1-dimensional Datasets (arrays) per subband, as the storage containers. Time increments are quantized and are filled for gaps; the time axis is linear. Padding of time samples will be done using zero's. For each subband, the channels are quantized and filled for gaps, so within each subband table/array, the frequency axis is linear. However, since the subbands can have large gaps (are not quantized), the overall frequency axis is not linear. The frequency and time coordinate axis information is stored in the Coordinates Group in the BF file.

For example, Subband 0 can start at 140 MHz, Subband 1 at 150 MHz and Subband 2 at 165 MHz; each Subband has the same number of channels (subdivisions), say 512, all with the same channel widths, and same number of time increments (rows), say 2000000. Therefore, there is a gap between Subbands 1 and 2 in this example. One cannot assume that the last channel in Subband (N-1) is followed directly in frequency by the first channel in the next Subband N. Gaps are accounted for in the Coordinates Group.

4.9. TiedArray Observing Sub-Modes & Data Format

4.9.1. Standard Observing Mode for TiedArray

The default data stored in each subband table is full resolution, polarized voltages from the dipoles. These voltages are stored as 16-bit complex pairs for both X and Y. In other words, each sample is stored as (X-real, X-imaginary)(Y-real, Y-imaginary) for a total of 64-bits (2 32-bit complex numbers). This is the format of the BF RAW data (see Figure 5 for data flow, and Figure 4 for HDFView data storage format.)

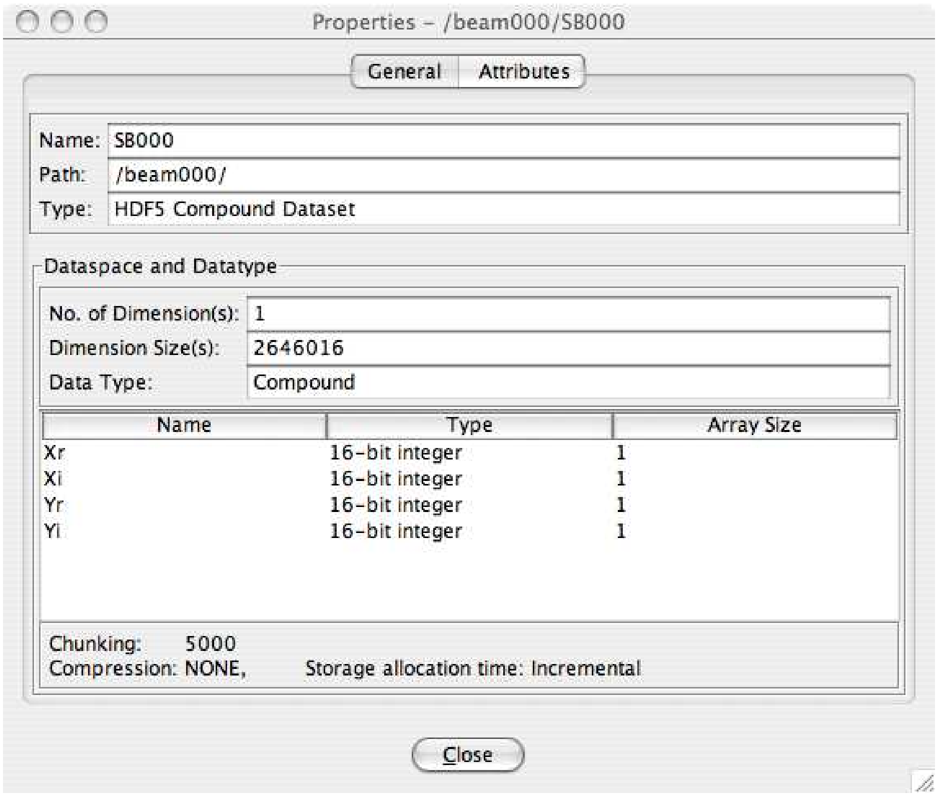


Figure 4: Subband properties as viewed through HDFView.

4.9.2. Sub-Observing Modes for TiedArray

There are several possibilities of sub-observing modes for TiedArray. Four sub-modes are listed below, but please note that this is not yet a full set of sub-modes – the list below explains the main sub-modes in words.

1. *Stokes I with IncoherentSum*

   ```
   STOKES_COMPONENTS = ["I"]
   SIGNAL_SUM = INCOHERENT
   ```

   Data taken for **Stokes I** TiedArray observing mode with Incoherent Summing. Data are being stored as one number, the total intensity per unit time. Currently this number is being stored as a single float and will eventually be stored as a 16-bit integer. Example HDFview (to be added).

2. Stokes I with CoherentSum

   ```
   STOKES_COMPONENTS = ["I"]
   SIGNAL_SUM = COHERENT
   ```

   Data taken for **Stokes I** TiedArray observing mode with Coherent Summing. Data are being stored as one number, the total intensity per unit time. Currently this number is being stored as a single float and will eventually be stored as a 16-bit integer.

3. Full Stokes with IncoherentSum

   ```
   STOKES_COMPONENTS = ["I","Q","U","V"]
   SIGNAL_SUM = INCOHERENT
   ```

   Data taken for **Full Stokes** TiedArray observing mode with Incoherent Summing. Data are being stored as 4 numbers, per unit time. Currently these 4 numbers are being stored as 4 floats and will eventually be stored as 4 16-bit integers. Example HDFview (to be added).

4. Full Stokes with CoherentSum

   ```
   STOKES_COMPONENTS = ["I","Q","U","V"]
   SIGNAL_SUM = COHERENT
   ```

   Data taken for **Full Stokes** TiedArray observing mode with Coherent Summing. It is stored as 4 numbers, per unit time. Currently these 4 numbers are being stored as 4 floats and will eventually be stored as 4 16-bit integers.

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, 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

A. Discussion & open questions

This section contains open questions, issues and examples regarding the size of the BF data as well as information on previous data structure considered for the BF H5 data sets.

A.1. BeamFormed Data Flow Overview

Figure 5 shows the overall Beam-Formed data flow, starting from the LOFAR stations, then moving to the Blue Gene (BG/P) I/O nodes, followed by data processing on the BG/P compute nodes, then streaming to the off-line cluster for packaging into a HDF5 BF file. Figure 5 can be broken up into three segments (described in more detail below), moving from left to right:

1. Data flows from the LOFAR stations to the BG/P I/O nodes and BG/P compute nodes for processing,
2. BF RAW and BF TAB (Tied-Array Beam) processed data leaves the BG/P via a socket stream,
3. BF HDF5 data file is created.

A.1.1. Segment 1 - Station bits to BG/P Processing

In segment 1 (left-hand of Figure 5) data arrive from multiple LOFAR stations onto BG/P I/O nodes, one node per station. This stage of the data is called Beam-Formed RAW (BF RAW). These data are then distributed onto BG/P compute nodes, where they are transposed from a station-based structure into a subband-based structure. Since the station-based data are transposed into subband-based divisions, each SB chunk of data on a compute node contains data from all the stations, just for a single subbands.

For each set of subbands on an I/O node, data arrive in one second chunks. As part of the BG/P BF data pipeline processing the data are further subdivided into the following divisions for each second: Channels 0 to N, Beams 0 to N, small time samples within that second, and by four Stokes parameters (I, Q, U, V). Therefore the order of the bits are: to-be-determined for each time sample, for each channel, for each subband within a one second chunk for each Beam. This stage of the data is called Beam-Formed Tied-Array Beam (BF TAB).

A.1.2. Segment 2 - BF RAW and BF TAB Data Streams

In segment 2 (middle, center of Figure 5) BF data are moved via a socket from the BG/P to the off-line cluster. Currently only the BF RAW data are being streamed to a socket off the BG/P to the off-line cluster in order to be packaged into the HDF5 format. Since these data mimic the layout on the BG/P at the early stage of processing, the BF RAW format is divided by subband, in one second chunks. Within each subband chunk are time samples of data. For example, using a 200 MHz clock, the time samples are 5.2 μs calculated as: 100MHz/512SB = 195.3125KHz $\Rightarrow$ 1/195.3125KHz = 5.2 μs. Each sample is comprised of two 32-bit complex numbers (*e.g.* Xreal, Ximg, Yreal, Yimg).

Ultimately we need the BF TAB data stream to be written to the socket in order to create the BF TAB HDF5 data file. The difference between BF RAW and BF TAB is that beams may be formed and stations may be combined; there are N time fewer time-samples in BF TAB as BF RAW, where N is equal to the number of channels. Each sample can be one 16-bit integer (INT) when there is just intensity (Stokes I), or four 16-bit INTs when all four polarizations are present (I, Q, U, V, or XX, XY, YX, YY).

A.1.3. Segment 3 - BF to HDF5 File Creation

A description of the software interface to the beam-formed data is given in the *LOFAR Data Access Library (DAL) User Manual* [?].

In segment 3 (right-hand of Figure 5) BF data are fed to the DAL which produces the BF HDF5 dataset file. The program used for this task is called “bf2h5” (Beam-Formed to HDF5). Currently only the BF RAW is written out to a HDF5 file (top right-hand of Figure 5). The red “TBD” (to-be-determined) ovule in the figure indicates that this aspect is still being actively worked on.

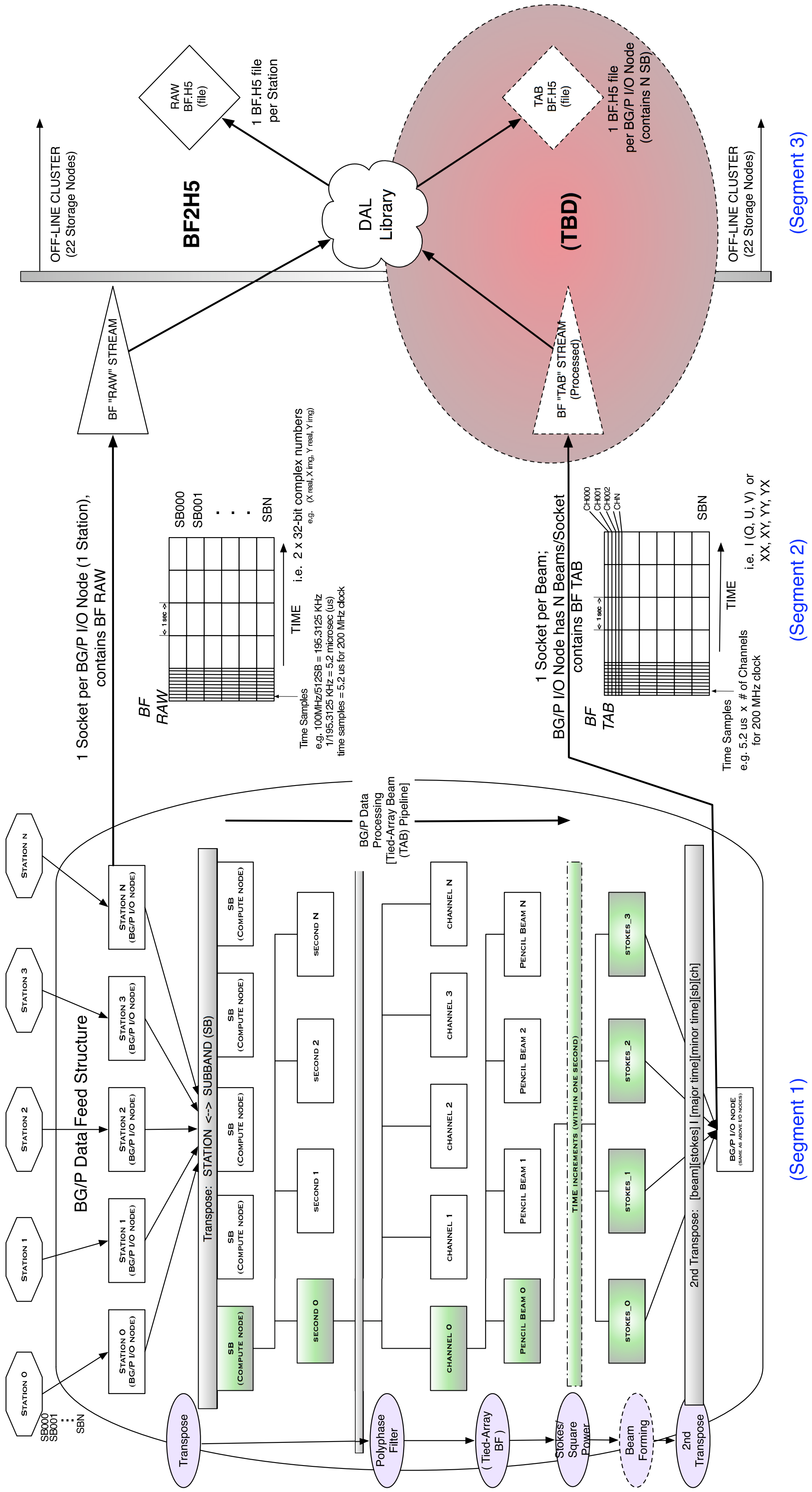


Figure 5: Beam Formed Data Flow

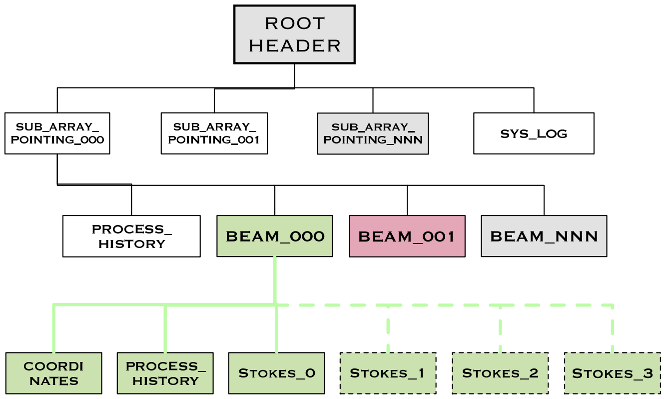


Figure 6: Single, compact, simplified view of a Beam-Formed dataset

Ultimately the BF TAB data need to be written to a HDF5 data file (bottom right-hand of Figure 5). The process entails getting sockets of the BF TAB data streams from the BG/P I/O Nodes which arrive at the Off-Line Cluster Storage I/O nodes. Each of these sockets contains one subband for multiple sub-array pointings and multiple beams. Each stream will use the DAL to convert the BF TAB to the BF HDF5 format.

Currently the DAL is neither distributed nor parallelized. Because of the large file sizes of the BF format, utilizing multiple storage and I/O nodes on the off-line cluster is a priority. In order to do so without key features in the DAL, the BF HDF5 files will be split up per observation on various nodes on the off-line cluster. There are two main BF observing scenarios: (1) Known Pulsar Observation and (2) Pulsar Survey mode. For scenario 1, each BF HDF5 file will contain a set number of subbands for an observation, a segment of the total set of subbands. For scenario 2, each BF HDF5 file will contain all the subbands for one beam of an observation, where there are N beams in an observation. Each BF HDF5 file will contain the same file structure, and therefore repeat the Root Header Attributes and hierarchy within each file for an observation.

In order to keep the BF structure figure more compact within a larger scheme Figure 6 will now represent a single BF dataset (containing N subbands, therefore, part of an observation). Figure 7 shows the BF data flow from the BG/P I/O nodes to the Off-Line Cluster Storage Nodes, in a multiple-file arrangement, for scenario 1. This makes use of multiple I/O nodes to distribute the contents of a single observation. A "master" bookkeeping file will keep track of all the files associated with one observation.

Figure 8 displays an example of a Beam-Formed file as viewed using the tool HDFView. The example has two Sub-Array Pointing Groups, which contain three Beam sub-groups (of which only the first Beam is expanded to show detail). The first Beam (`BEAM_000`) has the sub-group `Stokes_0` (Stokes_I) which contains three array datasets (Subbands 0, 1, and 2). Subband 0 (`SB_000`) is expanded on the right hand side of the figure, but only a portion of the full data array is visible; columns are channels and rows are time bins.

The "bf2h5" program will need to be updated to receive the new BF TAB data stream as well as be able to write out the new BF output structure as outlined in this ICD.

A.2. BF Data Size

The objective is to observe with LOFAR at full time resolution (5.12μs using the 200 MHz clock) with all stations, using 244 subbands, which equates to approximately 31TB/hour (8.8GB/s) of RAW data. This will be the maximum RAW data rate using LOFAR for BF observations; all other BF observations which do not require full resolution will have a data rate of less than 31TB/hr. Listed below are the current statistics on BF RAW data (non offline pipeline processed) files sizes and observing times, for the two main modes of observing: (1) known pulsar observations and (2) pulsar search mode observations (also known as "fly's eye" mode). Note that all calculations below assume 5.12μs time resolution using the 200 MHz clock.

1. To date the typical LOFAR BF data rate for observing known pulsars is 672GB/hr (186.7MB/s); when

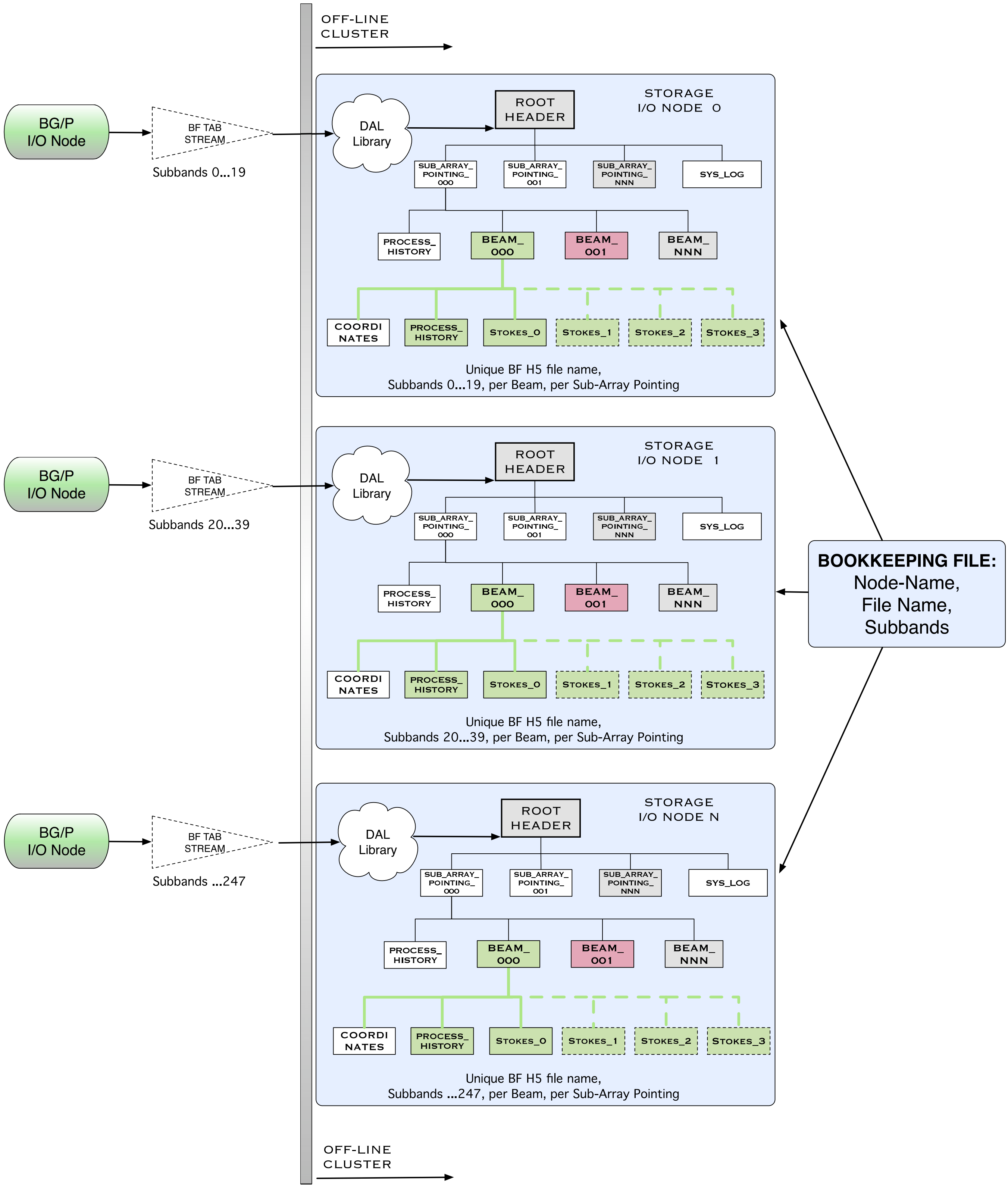


Figure 7: Beam-Formed to HDF5 Creation - multiple HDF5 files, each with some fraction of the subbands for an observation. This reflects scenario 1, where a single known pulsar was observed.

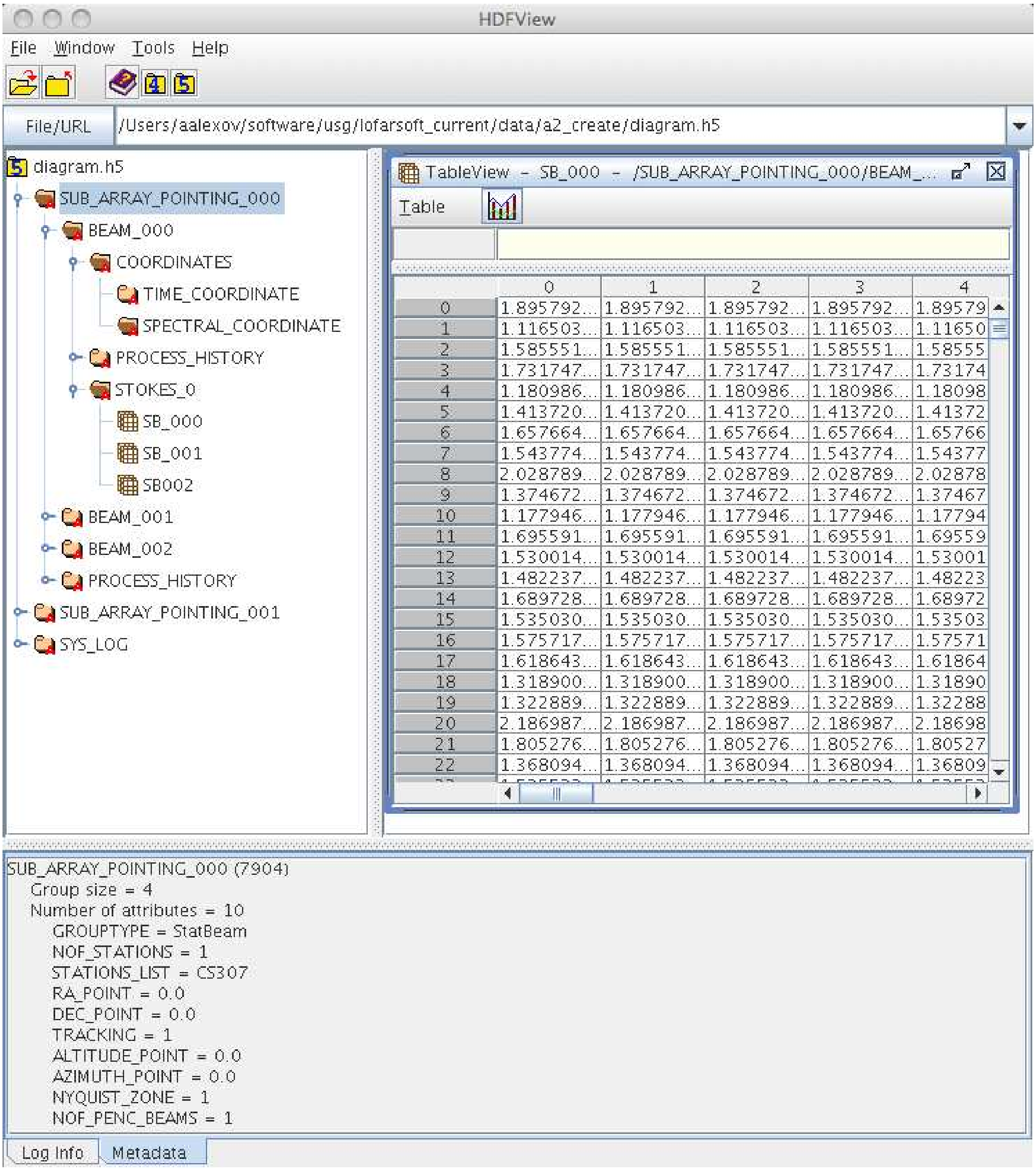


Figure 8: Beam Formed data as viewed in HDFView

using a downsampling factor of 16, this rate decreases to 42GB/hr. Most observations are using 244 subbands, with or without channelization (16-128 channels are typical), and adding the station signals. If channelization is used, time samples are added together for lower time resolution; if channelization is not used, then there are no channels and higher time resolution. Currently, the most common cases use 16 channels for channelized observations. When observing with N stations, the signals are combined to make one Beam with full bandwidth. Similarly, N stations can be set up to create M Beams (the subbands are then split into M ranges from each station) and so each Beam uses up 1/M of the bandwidth. Therefore, the total data rate remains constant for M Beams; all the data for one observation are placed in one file, no matter the number of Beams made.

2. In the pulsar search mode ("fly's eye" mode) the signals from the stations are not added, therefore the data rate scales linearly with the number of stations or the number of tied-array beams. An observation using 5 stations at full resolution in fly's eye mode is: 672GB/hr $\times$ 5 stations = 3360GB/hr (933MB/s).

For observing with the Beam Former, we suggest using 10's of minutes exposure times for HBA observations, and one hour observation times when using the LBA antennas. These observation times can be used for both known pulsar and search mode observations. It is conceivable to need exposure times as great as 12 hours for some applications.

Table 17 presents many, but still not all the possible/envisioned LOFAR beam-formed modes and their (approximate) associated sensitivity, FoV, resolution (i.e. $\delta\Omega$), data-rate, and survey FoM [?]. High-band (HBA) and low-band (LBA) sensitivities and FoMs have been normalized to that of a single 24-tile HBA sub-station or a 48-dipole Dutch LBA field respectively[3]. Quantities are quoted assuming one beam per station (48 MHz bandwidth) and 8 beams per station (6 MHz bandwidth per beam) respectively. FoV ($\propto \lambda_{\rm obs}^2$) and resolution (i.e. FWHM of the beam, $\propto \lambda_{\rm obs}$) are quoted for a central observing frequency of 150 MHz (HBA, $\lambda_{\rm obs} = 2$ m) and 60 MHz (LBA, $\lambda_{\rm obs} = 5$ m). Note that FWHM is taken to be $\alpha \times \lambda_{\rm obs}/D$, where $\alpha = 1.3$ and D is the size of a station or the maximum baseline between combined stations where applicable. As LOFAR stations consist of several square tiles, they are not perfectly circular; thus, the product of FoV and sensitivity is not constant when station size increases. We have used LBA (Inner) / LBA (Outer) / HBA station sizes of 32.3 m / 81.3 m / 30.8 m (core), 32.3 m / 81.3 m / 41.1 m (remote), and 65 m / 56 m (international, Inner/Outer mode does not apply here). Further empirical beam modeling will likely refine the value of α, and will somewhat effect the rough values quoted here. Where applicable, we assume that 24 core stations of 2×24 HBA tiles / 48 LBA dipoles, 16 Dutch remote stations of 48 HBA tiles / 48 active LBA dipoles, and 8 international stations of 96 HBA tiles / 96 LBA dipoles are available and can be recorded separately if desired. Fly's Eye mode assumes all Dutch stations - i.e. 48 HBA core sub-stations plus 16 remote HBA stations or 40 LBA fields of 48-dipoles each are used. For the "Coherent" modes, we assume the maximum number of tied-array beams required to cover the station beam, up to a maximum of 100 (per station beam), can be synthesized, and that the maximum baseline between stations is 300 m for the Superterp and 2000 m for the entire Core. The "Coherent Sum Core" mode assumes that all 48 Core sub-stations are combined coherently. The "Dutch Incoherent Sum" mode assumes that all 40 Dutch stations (24 core / 16 remote) are combined incoherently. The "Intl. Incoherent Sum" mode assumes that all 8 international stations are combined incoherently. The "Constrained Coherent Core" mode is a hybrid coherent/incoherent summation in which the two HBA sub-stations of each core station are first summed coherently at station level before these stations are in turn summed incoherently. The integration time used in each mode is assumed to be the same, though this would likely differ in practice, especially in the case of wide-field surveys. The data rates assume 16-bit samples (this could be reduced if desired), summed to form Stokes I, at the maximum possible spectral/time resolution, which for certain applications can be downgraded by a factor of a few in order to save on disk space and processing load.

Any number of downsampling factors can be used on the fly; the data size is then equal to the full resolution size divided by the downsampling factor. Note that currently the BF RAW and BF OUT data are stored as floats. The BF pipeline processed data is about half the size of the BF RAW (or BF OUT) data because they are stored as 16-bit integers.

[3] Recall that each Dutch LBA field contains 96 dipoles, only 48 of which are used in any particular observation. Unless otherwise stated, we assume the LBA Outer mode is being used. This mode gives somewhat higher gain, but reduced FoV compared with the LBA Inner mode.

Mode	Sensitivity (Norm.)	FoV (sq. deg.)	Resolution (deg)	Data Rate (TB/hr)	FoM (Norm.)
	1 beam of 48MHz / 8 beams of 6MHz				
High-Band Antennas (HBAs)					
Single HBA sub-station	1 / 0.35	18 / 147	4.8	0.3	1
Single Rem. Station	2 / 0.7	10 / 82	3.6	0.3	3
Single Intl. Station	4 / 1.4	6 / 45	2.7	0.3	9
Fly's Eye	1 / 0.35	1050 / 8400	4.8	20	56
Dutch Inc. Sum	11 / 4	10 / 82	3.6	0.3	77
Intl. Inc. Sum	11 / 4	6 / 45	2.7	0.3	73
Coherent Superterp (94 beams)	12 / 4	18 / 147	0.5	29	1382
Coherent Sum Core (100 beams)	48 / 17	0.4 / 3	0.075	31	3206
Constrained Coherent Core (29 beams)	10 / 3.5	18 / 147	0.9	9	512
Low-Band Antennas (LBAs)					
Single Core Station Outer	1 / 0.35	17 / 132	4.6	0.3	1
Single Core Station Inner	<1 / <0.35	105 / 840	11.6	0.3	<1
Single Rem. Station	1 / 0.35	17 / 132	4.6	0.3	1
Single Intl. Station	2 / 0.7	26 / 211	5.8	0.3	5
Fly's Eye	1 / 0.35	660 / 5300	4.6	12	40
Dutch Inc. Sum	6 / 2	17 / 132	4.6	0.3	40
Intl. Inc. Sum	6 / 2	26 / 211	5.8	0.3	44
Coherent Superterp (15 beams)	6 / 2	17 / 132	1.2	4.5	138
Coherent Sum Core (100 beams)	24 / 8.5	3 /	0.19	30	2460

Table 17: This table presents many, but still not all the possible/envisioned LOFAR beam-formed modes and their (approximate) associated sensitivity, FoV, resolution (i.e. $\delta\Omega$), data-rate, and survey FoM (see text).

A.3. Full Resolution and Downsampled Data

- It is very possible that future version of LOFAR beam-formed data will contain both the full resolution polarized samples, as well as a heavily down-sampled total intensity table. The total intensity table will serve as a snapshot for the subband and add very little to the total size of the file.

  Current thinking is to have two separate files, one with raw data and the other down-sampled.

  > **Comment:**
  > It is possible to expand the file structure to include both raw and down-sampled data within one HDF5 TiedArray data file. We should give this some thought and see whether two files are easy enough to keep track of instead of merging the information into one file.

A.4. Future enhancements

- The DAL will need to be expanded to include distributed/parallel HDF5 infrastructure. In the current specification in order to distribute the BF HDF5 file, N Beam-Formed HDF5 files are created on various nodes, all kept track of by a bookkeeping ASCII file. Each of these segments contains a subset of the subbands of a complete observation. A tool will be needed in order to gather the N segments and write a single, complete HDF5 file.

- James Anderson writes:

  "We need to know which antennas have been used to form a station beam in order to be able to calculate the beam shape of the station, in order to calibrate the gain of an observation target in a specific direction. So, yes, I would like a list of all RCUs, and the gain settings for each one of them for the specific beam used for the observation. Antennas, RCUs, and so on fail all of the time (the last time I checked there were 8 broken RCUs in the Effelsberg station, which certainly affects the overall beam shape of the station). And there needs to be a list of the coordinates of each antenna/tile being used as well, so that future processing software doesn't have to go looking for this in some unknown location. This is a topic also being discussed for the visibility data format too."

The RCU information/encapsulation cannot be implemented in the first version of the data format. We will add this to the "Discussion" section of the ICD as a future enhancement of the data structure. It looks like it is possible the get the RCU information from the LOFAR Health and Monitoring system within Calibration. What is unclear is whether the RCU info you need is time dependent and at what time resolution would you need it. Could you please come up with a solution to this issue with respect to what sort of information you need (example: table of on/off bits for each RCU with some minimum time resolution, etc). We will add your suggestion/solution to the ICD discussion section for future implementation.

- The time axis is linear; therefore, if there are gaps/dropouts in the data, then ZERO's will be filled in for all the time bins lost. We will add the possibility of using `NaN`s instead of zero's in the future. `NaN`s would be more accurate, but we do not currently know the implications on the HDF5 library and subsequent pulsar tools as far as using NaN values. All this needs to be investigated. If HDF5 can handle writing/reading NaN's, in a future version of the BF data we can have a switch which allows the user to specify converting NaNs on the fly to zero's if need be by other programs which cannot deal with NaN values. In summary, all gaps will be filled in with zero's; we will investigate switching from using zero's for gaps to using NaN's in the future.

A.5. Summary of HBA JOINED Discussion

The following text and email exchange summarizes topic of the HBA_JOINED antenna field discussion. This issue will be picked up in 2012 for ICD003 and ICD006 for the possibility of adding additional information to the ICDs in order to calculate the beam model per observation.

A2/JMG/AR discussed HBA_JOINED as new Antenna field; the Pulsar Working Group would like this info to create beam model in ICD003. What is needed is:

- Health status of antennas during the observation, calibration, ...
- HBA_JOINED: antenna subfields have different orientation with respect to north! This breaks a number of assumptions on antennafield homogeneity

A2 emailed Ger asking for details on where this information is stored. (see thread below).

The meeting notes are as follows:

```
---------------------------------------------------------------------------
Data Formats Splinter Meeting (HBA joined), 2011-09-02
---------------------------------------------------------------------------

1 Setting of the meeting
==========================

Present via Skype: Anastasia Alexov (A2), Jean-Mathias Griessmeier (JMG),
Adriaan Renting (AR)

Minutes: JMG

2 Notes from the meeting
==========================

ICD-003 and ICD-006 have no detailed description of antenna field,
whereas the MS description does.
```

```
But calibration (esp. polarization calibration) requires the beamshape!
However, the calibration may be done before the data are written to disk.

For comparison, MS description has a 3 table:
- stations  CS001
- antennas  CS001_HBA0, CS001_HBA1, CS001_HBA
- antenna field CS001_HBA0, CS001_HBA1 (only of antenna is CS001_HBA)
  incl. antenna field orientation per antenna field, ...

See ms2_description_for_lofar_2.07.01.pdf for details.

This is how the system works:
BG outputs contains polarization, but not yet the polarization
calibration. This is currently done offline, but it may be moved to BG.
If this happens, beamshapes etc. are not required in ICD_003/006.
Depending on whether this implementation changes, the tables from the MS
description  will be added to ICD_003/006... or not.

Health status will be added to SYS_LOG.

TODO:
- A2 - discuss this with Aris N.


(email thread)

Hi Ger,

There are a couple of questions which have arisen in the ICD group,
which we're hoping you could shed some light on or point me to someone
who could answer these.  We want to know the source of some of the
MeasurementSet metadata information.  Some of the same information
will be needed in other LOFAR data products for non-interferometric
observations.

1) Some of the ICDs will require tracking information of the
pointings, at some regular time internal, to be stored in tabular
form.  This is similar to the "field" and "pointing" table in the
MeasurementSet.  Where is the pointing information stored within the
system (say for cases when it's not written to a MS file)?  Is it
stored at a particular time interval already per OBSID?  How could one
best retrieve this information in order to pack it into the output
product other than a MeasurementSet?

2) There are several Antenna field-related tables in the
MeasurementSet, such as the HBA_JOINED table.  These contain the
health and status of the antennas observation for calibration.  Some
of the other data formats (Dynamic Spectra and Beam-Formed) need
similar information about the health and status of the antennas per
OBSID.  Where can we pull this information from the LOFAR system?  Is
is available at certain time intervals or just start/end of
observations?

Thanks much in advance.
```

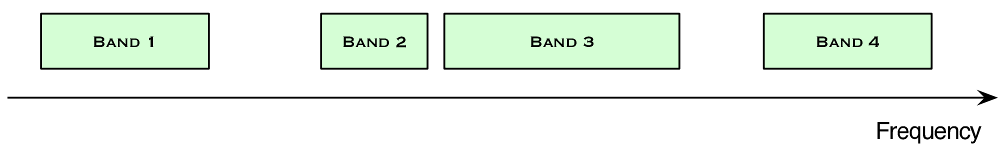
```
best,
A2

Hi Anastasia,

The pointing table is not filled in for the LOFAR MS.
The only pointing info filled is in the FIELD table which
contains (in J2000) the DELAY_DIR and PHASE_DIR, thus the
delay center and phase center. Usually they are the same,
but not necessarily. The info is coming from the parset
file and written by Chris' data writer.
Calculations of AzEl or apparent coordinates are done on
the fly (using casacore's Measures classes). It is possible
to add a virtual column to the MS main table giving AzEL
or HA. There are also TaQL functions doing it.

The info needed for the beam calibration is filled in by the
data writer using the BeamTables class (in LOFAR/LCS/MSLofar).
It gets its info from several sources:
- The station config files (text files) tell the layout of a
station (one file per station). Such a file is read by the
AntField class.
- The iHBADeltas.conf text files tell the dipole
layout of HBA tiles for each station.
- The AntennaSets text file defines for each configuration
(like LBA_INNER) which tiles/dipoles are used.
- The SAS DB contains info about broken dipoles/tiles. Ruud
has written a query to extract that info and Sven is working
on a program addbeaminfo to store that info in the MS.

I hope this helps.
I assume you can see our source files, but I can send some if you want to.

Cheers,
Ger
```

B. Coordinates Group Examples

B.1. The Time & Spectral Coordinates Representation

The 'Time' Coordinate can be represented linearly, and therefore will use the coordinate type 'Linear'. Time gaps are also possible, but are filled; therefore, the time axis will be linear. The padding of the gaps will be filled in with zero values for each missed time sample.

Given the nature and organization of the data, the sequence of frequency values along the respective data axis cannot be represented properly by a simple linear coordinate. As a consequence of this, the spectral/frequency axis spanning multiple frequency-bands is represented using a tabulated coordinate.



The above figure shows gaps in the frequency axis for an observation with 4 bandpasses which are not similar in frequency range.

A coordinate group to represent such a BF Coordinate Group would look like:

```
.
`-- COORDINATES
     |- RADEC_SYS                  'FK5'
     |- EQUINOX                    'J2000'
     |- NOF_COORDINATES            2
     |- NOF_AXES                   2
     |- COORDINATE_TYPES           ['Time', 'Spectral']
     |- COORDINATE_0
     |   |- COORDINATE_TYPE        'Time'
     |   |- STORAGE_TYPE           'Linear'
     |   |- NOF_AXES               1
     |   |- AXIS_NAMES             ['Time']
     |   |- AXIS_UNITS             ['us']
     |   |- REFERENCE_VALUE        [1.0]
     |   |- REFERENCE_PIXEL        [0.0]
     |   |- INCREMENT              [1.0]
     |   `- PC                     [1.0]
     `- COORDINATE_1
         |- COORDINATE_TYPE        'Spectral'
         |- STORAGE_TYPE           'Tabular'
         |- NOF_AXES               1
         |- AXIS_NAMES             ['Spectral']
         |- AXIS_UNITS             ['MHz']
         |- AXIS_LENGTH            1024
         |- AXIS_VALUES_PIXEL      [0, 1, 2, ..., 512, 513, ...]
         `- AXIS_VALUES_WORLD      [140, 140.1953125, 140.390625, ...,
                                    150, 150.1953125, ...]
```

Glossary of terms

Az Azimuth.

AIPS++ The AIPS++ project was a project from the nineties supposed to replace the original Astronomical Information Processing System or classical AIPS. The ++ comes from it being mainly developed in C++. It's also known as AIPS 2. It evolved into CASA, casacore and casarest (see those entries).

BBS BlackBoard Selfcal, pipeline used for LOFAR imaging data.

Beam A beam is formed by combining all the SubArrayPointing, one for each station, which are looking in a particular direction. There may be more than one beam for each SubArrayPointing, and different types of beams are available.

BF Beam-Formed data (time series structure).

CASA The Common Astronomy Software Applications package. User software for radioastronomy developed out of the old AIPS++ project. The project is led by NRAO with contributions from ESO, CSIRO/ATNF, NAOJ and ASTRON. [?]

casacore The set of C++ libraries that form the basis of CASA and several other astronomical packages. It contains classes for storing and handling visibility and image data, RDBMS-like table system and handling coordinates. Mainly maintained by ASTRON and CSIRO/ATNF. [?]

casarest The libraries and tools from the old AIPS++ project that are not part of casacore or CASA but still in use.

CEP Central Processing facility.

Channel The subband data of a LOFAR observation may be passed through a second polyphase filter to obtain a large number of channels (i.e. to increase the spectral resolution).

CLA Common LOFAR attributes. Set of root-level attributes that are used and required as attributes in all LOFAR science data products. If a value is not available for an Attribute, `'NULL'` maybe used.

Co-I Co-investigators on an observation project under the leadership of the PI.

Data Interface Set of definitions that describe the contents and structure of data files.

Data Access Layer (DAL) A C++ library with Python bindings providing read/write functionality for HDF5 format files, as well as access to Measurement Sets.

Dec Declination.

DPPP Default Pre-Processing Pipeline, pipeline used for LOFAR imaging data.

EAS Extensive Air-Shower.

El Elevation.

FITS FITS (Flexible Image Transport System) is a digital file format used to store, transmit, and manipulate scientific and other images. FITS commonly used in astronomy.

HBA High Band Antenna.

HDFView Hierarchical Data Format Viewer; a Java software tool for viewing the HDF5 structure and data. [http://www.hdfgroup.org/hdf-java-html/hdfview/]

HDF5 Hierarchical Data Format, 5 [?]. A file format capable of accommodating large datasets that comprises two (2) primary types of objects: groups and datasets. Implements self-organisation and hierarchical structures within the file format itself, facilitating self-contained data administration. [?, ?]

HDF5 group A grouping structure containing zero or more HDF5 objects, together with supporting meta-data.

HDF5 dataset A multidimensional array of data elements, together with supporting meta-data.

HDU Header-Data Unit Though typically used for FITS data descriptions, the term "HDU" can also be used more generically when discussing any data group that contains both data and a descriptive header.

Hypercube The hypercube is a generalization of a 3-cube to n dimensions, also called an n-cube or measure polytope. In data modelling a hypercube is a cube-like logical model in which all measurements are organized into a multidimensional space.

ICD Interface Control Document.

IVOA International Virtual Observatory Alliance.

KSP Key Science Project. One of several major observational and research projects defined by the LOFAR organization. These Key Science Projects are,

- Cosmic Magnetism in the Nearby Universe
- High Energy Cosmic Rays
- Epoch of Re-ionization
- Extragalactic Sky Surveys
- Transients - Pulsars, Jet Sources, Planets, Flare stars
- Solar Physics and Space Weather

LBA Low Band Antenna.

LOFAR The LOw Frequency ARray. LOFAR is a multipurpose sensor array; its main application is astronomy at low radio frequencies, but it also has geophysical and agricultural applications. [http://www.lofar.org/]

LOFAR Sky Image Standard LOFAR Image Cube. A LOFAR data product encompassing science data, associated meta-data, and associated calibration information, including a Local Sky Model (LSM) , and other ancillary meta groups that are defined in this document.

LSM/GSM The Local Sky Model/Global Sky Model. Sky Models are essentially catalogues of known real radio sources in the sky. A Local Sky Model for an observation is merely a subset of a Global Sky Model catalogue pertaining to that observation's relevant region of the sky.

LTA The Long Term Archive for LOFAR.

MJD Modified Julian Day. Derived from Julian Date (JD) by MJD = JD - 2400000.5. Starts from midnight rather than noon.

MS Measurement Set, a self-described, structured set of casacore tables comprising the data and meta-data of an observation. [**?**]

PI A Principal Investigator is the lead scientist resopnsible for a particular observation project.

RA Right Ascension.

RFI Radio Frequency Interference.

RM Rotation Measure.

RMSC The Rotation Measure synthesis cube is a data product which contains the output of LOFAR RM synthesis routines, namely the polarized emission as a function of Faraday depth. As with the Sky Image data files, all associated information is stored within an RMSC file.

RSP Remote Station Processing Board.

SIP Standard Imaging Pipeline or Submission Information Package within the context of the LTA.

Station Group of antennae separated from other groups. In it's current cofiguration, LOFAR has 48 stations.

SubArrayPointing This corresponds to the beam formed by the sum of all of the elements of a station. For any given observation there may be more than one SubArrayPointing, and they can be pointed at different locations.

Subband At the station level, LOFAR data are passed through a polyphase filter, producing subbands of either 156.250 kHz or 195.3125 kHz (depending on system settings).

TAI International Atomic Time (Temps Atomique International), atomic coordinate time standard.

TBB Transient Buffer Board.

TRAP Transients Pipeline.

USG LOFAR User Software Group.

UTC Coordinated Universal Time (UTC) is a time standard based on International Atomic Time (TAI) with leap seconds added at irregular intervals to compensate for the Earth's slowing rotation.

UV-Coverage A spatial frequency domain area that must be covered completely by observation in order to assure an optimal target image (Full UV- Coverage). During observation, the radio telescope turns with respect to its target, due to the earth rotation. A certain -instrument geometry dependent-rotation angle has to be covered in order to accomplish full coverage.

VHECR Very high-energy cosmic ray.

WCS World Coordinate Information (WCS). The FITS ”World Coordinate System” (WCS) convention defines keywords and usage that provide for the description of astronomical coordinate systems in a FITS image header [**?**, **?**, **?**].