LOFAR Data Format ICD
File Naming Conventions

Document ID: LOFAR-USG-ICD-005

Version 2.00.11

SVN Repository Revision: 9374

A. Alexov, K. Anderson, L. Bähren, J.-M. Grießmeier,
A. Gunst, H. Holties, M. Wise, G.A. Renting

SVN Date: 2012-01-10

Contents

Change record

Version	Date	Sections	Description of changes
0.1	2009-09-23	all	Document creation following format discussion
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	2010-09-21	3	Optional Description table matches the current BF naming convention of Sub-Array Pointing and Beam; added Stokes to the same table. Text updated to reflect the need to pad numbers with zero's.
2.00.02	2010-11-23	3	Added `.raw` and `.incoherentstokes` all possible extension names.
2.00.03	2010-12-06	3	Added "p" for "part" as a new file option descriptor.
2.00.04	2011-03-10	all	Maintain list of references through BibLaTeX database.
2.00.05	2011-03-28	all	Added text to the first few sections about motivation/scope/etc, since they had been blank.
2.00.06	2011-04-01	4	Added suggested naming convention for non standard filenames.
2.00.07	2011-04-21	B	Added glossary. Expanded list of authors.
2.00.08	2011-04-25	3, 4, 5	Introduction of *prefix* as part of the filename; adding section with examples.
2.00.09	2011-09-28	3, 5	Added new fields which were requested by the TBB group; added order preference listing to the optional descriptors.
2.00.10	2011-11-18	3	Added and updated naming conventions for SkyModel and InstrumentModel.
2.00.11	2012-01-10	3	Added Subband Group.

1. Introduction

1.1. Purpose and Scope

This document sets forth a formal data interface specification for **LOFAR file naming conventions**. The specification applies to data structures produced by various LOFAR online and offline processing pipelines that will be called LOFAR data products.

This document is intended to be the formal interface control agreement between the LOFAR project, observers/users of LOFAR data products, and the LOFAR long-term archive (LTA) facility.

1.2. Context and Motivation

This document is not an Interface Control Document (ICD), per say, but describes the names of the LOFAR data products, and is therefore is considered to be part of the group of ICDs.

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.

Reference	Title	Description
ICD-001 [6]	TBB Time-Series Data	Digitized voltage output, as received by the individual LOFAR dipoles.
ICD-002 [5]	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 [1]	Beam-Formed Data	Hosting structure for LOFAR Beam-Formed data.
ICD-004 [7]	Radio Sky Image Cubes	Primary data product of the imaging pipeline.
ICD-005 [2]	File Naming Conventions	Conventions for the naming scheme applied to LOFAR standard data products.
ICD-006 [13]	Dynamic Spectrum Data	Hosting structure for dynamic spectrum data, i.e. intensity as function of time and frequency.
ICD-007 [4]	Visibility Data	Hosting structure for LOFAR UV Visibility data, primary output of interferometer operations.
ICD-008 [3]	RM Synthesis Cubes	Hosting structure for LOFAR Rotation Measure Synthesis Cubes output data.

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

2. Overview

This document is structured as follows: Section 3 will describe the convention to follow for required and optional naming tags for LOFAR file names. Section 4 gives suggestions to follow for file names for files that can not be covered, or not entirely by this convention.

3. Filename Convention

LOFAR data products [6, 1, 7, 13, 4, 3] will have file names of the form,

<Prefix><Observation ID>_<Optional Descriptors>_<Filetype>.<Extension>

In this the individual fields are defined as follows:

1. **Prefix**, for a standard LOFAR dataset, is a fixed single character, 'L'.
2. **Observation ID** – unique identifier of the observation as part of which (or based on which) this data product was created.
3. **Optional Descriptors** are used to further indicate the nature of the data stored within the file. It is best to keep the file name as short as possible; optional descriptors should only be used if necessary to better uniquely identify a file. Underscores should be used between multiple optional descriptors. Please note that the optional descriptors listed below are in their preferential order in which they should be used within the file name:

Description	Format	Example
Sub-Array Pointing (SAP)	3 digits	SAPxxx
Subband (SB)	3 digits	SBxxx
Subband Group (SBG)	3 digits	SBGxxx
Beam (B)	3 digits	Bxxx
Part (P)	3 digits	Pxxx
Stokes (S)	1 digit	Sx
Station Name (CS, RS, DE, etc)	3 digit	CSxxx, RSxxx, DExxx
Date (D)	8 digits	D<yyyymmdd>
Date (D) & Time	8+6.3 digits	D<yyyymmdd>T<hhmmss.sss>Z
Readout (R)	3 digits	Rxxx

   While the descriptors can be used to e.g. indicate a specific sub-band, beam or date it is not to be used for ranges. Digits should be padded with zero's. Please note that the number of digits can increase in the future; software should be made flexible to take this into account.

   The "P" (Part) parameter holds the number of parts/slits the Beam-Formed coherentstokes data files will be split by frequency range. It is anticipated that the typical 240 subbands will be evenly split into 8 sections (to match the number of cores for ease of processing). Each of these 8 sections will have a unique part number.

   The "SBG" (Subband Group) parameter is used when data from more than one subband is combined into one output dataproduct. This is then enumerated using the Subband Group index.

4. **Filetype** is a marker for the contents of the file. There will be several different kinds of data produced by LOFAR, see table below. Importantly, filetype signifies the kind of LOFAR data that comprise the particular data file, and therefore, will also signal the appropriate interface control document for further reference, should users find that necessary. The options for the file type along with their abbreviations are listed in the table below.
5. **Extension**

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

Table 2: Overview of standard LOFAR data products and the corresponding file type attribute value.

Extension	Type of data
.MS	CASA/casacore MeasurementSet
.h5	HDF5 file
.fits	FITS file
.log	Logfile
.parset	A parset file
.LSM	Local sky model
.IM	CASA/casacore image file (PagedImage)
.INST	Instrument model (ParmDB) file generated by BBS
.vds	Dataset description file
.gds	Dataset description file
.conf	Configuration file (mostly local to station)
.raw	Raw Beam-Formed (non-Incoherentstokes) file written from the Blue Gene/P
.incoherentstokes	Raw Beam-Formed incoherent Stokes file written from the Blue Gene/P

Files generated by CASA/casacore will continue the currently existing conventions using upper-case suffixes.

4. Non standard file names

4.1. Introduction

Section 3 only covers dataproducts generated by the automatic systems in the Central Processing (CEP) facility of LOFAR. It is useful to also have file naming conventions for files that fall outside that regime. This is of particular importance if the data might need to be archived in the LOFAR Long Term Archive (LTA).

The biggest difference with section 3 is that for data generated by processes not under SAS control, there is no Observation ID. Two cases can be distinguished: where the data are derived from CEP generated data, and where the process is completely autonomous.

4.2. Derived from a SAS process

For data that is basically generated by further processing of a standard data product it is advised to keep the identifiers of the standard product and add extra descriptors as needed just before the Extension separated

by one or more dots.

```
<Prefix><Observation ID>_<Optional Descriptors>_<Filetype>.<Other descriptors>.<Extension>
```

In this the individual fields are defined as follows:

1. **Prefix**, for a standard LOFAR dataset, is a fixed single character, 'L'.
2. **Observation ID** is the SAS identifier for the process that generated the data product that this data product is derived from.
3. **Optional Descriptors** are the same as defined under the previous section 3, if applicable.
4. **Filetype** is the same as defined under the previous section 3, if applicable.
5. **Other descriptors** is relatively free form and could for example be things like `flagged`, `manual`, `pub_version3`, `Nature.page45.image`. Basically anything that can't be caught by the standard descriptors from the previous section. Note if needed these should preferably be separated by dots, not underscores, to separate them from the more standardized descriptors.
6. **Extension** should be last, so it is easy to identify the type of data product and what tools are needed to read it.

4.3. Not derived from a SAS process

For data that not derived from a dataproduct directly related to a SAS Observation ID, a slightly different naming convention is advised.

```
<Prefix><Non numerical identifier><Custom ID>_<Optional Descriptors>_<Filetype>
                        .<Other descriptors>.<Extension>
```

In this the individual fields are defined as follows:

1. **Prefix**, for a standard LOFAR dataset, is a fixed single character, 'L'.
2. **Non numerical identifier** is needed to make it clear that what follows is not a SAS Observation ID. This could be just one character, or the name of a project or instrument.
3. **Custom ID** is probably needed to identify dataproducts within your project or instrument.
4. **Optional Descriptors** are the same as defined under the previous section 3, if applicable.
5. **Filetype** is the same as defined under the previous section 3, if applicable.
6. **Other descriptors** is relatively free form and could for example be things like `flagged`, `manual`, `pub_version3`, `Nature.page45.image`. Basically anything that can't be caught by the standard descriptors from the previous section. Note if needed these should preferably be separated by dots, not underscores, to separate them from the more standardized descriptors.
7. **Extension** should be last, so it is easy to identify the type of data product and what tools are needed to read it. The 123th dataset of a project called DCI, could then for example be

   `LDCI123_B012_P345_sky.something.else.h5`

5. Examples

In order to illustrate how to construct file names based on the before-mentioned rules:

1. Beam-formed data
   a) Files containing data for a single sub-array pointing:

      ```
      L123456789_SAP000_bf.h5
      L123456789_SAP001_bf.h5
      L123456789_SAP002_bf.h5
      L123456789_SAP003_bf.h5
      ```

   b) Files containing data for a single beam within a sub-array pointing:

      ```
      L123456789_SAP000_B000_bf.h5
      L123456789_SAP000_B001_bf.h5
      L123456789_SAP000_B002_bf.h5
      L123456789_SAP000_B003_bf.h5
      ```

   c) File containing data for a single Stokes parameter of a beam within a sub-array pointing:

      ```
      L123456789_SAP000_B001_S0_bf.h5
      L123456789_SAP000_B001_S1_bf.h5
      L123456789_SAP000_B001_S2_bf.h5
      L123456789_SAP000_B001_S3_bf.h5
      ```

2. TBB time-series data

   ```
   L123456789_tbb.h5
   L123456789_CS011_D20110719T110541.036Z_R002_tbb.h5
   ```

A. Discussion & open questions

—/—

B. The GROUPTYPE attribute

The first attribute in every group/dataset must be the attribute GROUPTYPE. Since the Common LOFAR Attributes are in the root header, the value in the Common LOFAR Attributes for (GROUPTYPE) = 'Root'. The options for the group type are listed in the Group Type table below, grouped by category.

Group/Dataset	Value	Use in ICD...	Comments
File root group	'Root'	001, 003, 004, 006, 007, 008	Root group of HDF5 file.
SYS_LOG	'SysLog'	001, 003, 004, 006, 007, 008	System log files, parsets.
PROCESS_HISTORY	'ProcessHistory'	003, 004, 006, 007, 008	Processing history logs.
COORDINATES	'Coordinates'	001, 002, 003, 004, 006, 007, 008	Coordinates group.
Direction coordinate	'DirectionCoord'	002, 004	Coordinate describing a direction towards a position on the sky/celestial sphere.
Linear coordinate	'LinearCoord'	002, 004	Coordinate with a linear set of values along the world axis.
Tabular coordinate	'TabularCoord'	002, 004	Coordinate with a tabulated set of values along the world axis.
Polarization coordinate	'PolarizationCoord'	002, 004, 006	Representation of a polarization coordinate.
IMAGE_NNN	'Image'	004	Container for the various components associated with an image.
DATA	'Data'	004	Dataset containing the actual pixel data array.
SOURCE_TABLE	'SourceTable'	004	Table collecting a list of sources and their associated parameters.

As can be seen from the above table, some of GROUPTYPE attribute are common to all data format ICDs, whereas other are more specific to a given data product, depending on its hierarchical structure.

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. [9]

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. [10]

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 [15]. 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. [16, 17]

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. [14]

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 [11, 8, 12].

References

[1] A. Alexov, K. Anderson, L. Bähren, J.-M. Grießmeier, J.W.T. Hessels, J.S. Masters, and B.W. Stappers. Beam-Formed Data. LOFAR Data Format ICD LOFAR-USG-ICD-003, ASTRON, 2011.

[2] A. Alexov, K. Anderson, L. Bähren, A. Gunst, H. Holties, and M. Wise. File naming conventions. LOFAR Data Format ICD LOFAR-USG-ICD-005, ASTRON, 2011.

[3] J. Anderson, L. Bähren, M. Bell, T. Riller, and K. Anderson. Rotation measure synthesis cubes. LOFAR Data Format ICD LOFAR-USG-ICD-008, ASTRON, 2011.

[4] K. Anderson. Visibility data. LOFAR Data Format ICD LOFAR-USG-ICD-007, ASTRON, 2011.

[5] L. Bähren, A. Alexov, K. Anderson, and J.-M. Grießmeier. Representations of World Coordinates. LOFAR Data Format ICD LOFAR-USG-ICD-002, ASTRON, 2011.

[6] L. Bähren, K. Anderson, A. Corstanje, A. Horneffer, and J. Masters. TBB Time-Series Data. LOFAR Data Format ICD LOFAR-USG-ICD-001, ASTRON, 2011.

[7] L. Bähren, K. Anderson, and J. Masters. Radio sky image cubes. LOFAR Data Format ICD LOFAR-USG-ICD-004, ASTRON, 2011.

[8] M. R. Calabretta and E. W. Greisen. Representations of celestial coordinates in FITS. *A&A*, 395:1077–1122, 2002.

[9] Casa website. http://casa.nrao.edu/.

[10] Casacore website. http://code.google.com/p/casacore/.

[11] E. W. Greisen and M. R. Calabretta. Representations of world coordinates in FITS. *A&A*, 395:1061–1075, 2002.

[12] E. W. Greisen, M. R. Calabretta, F. G. Valdes, and S. L. Allen. Representations of spectral coordinates in FITS. *A&A*, 446:747–771, 2006.

[13] J.-M. Grießmeier, A. Alexov, K. Anderson, and L. Bähren. Dynamic spectrum data. LOFAR Data Format ICD LOFAR-USG-ICD-006, ASTRON, 2011.

[14] A. J. Kemball and M. H. Wieringa. Measurementset definition version 2.0. AIPS++ Note 229, NRAO/ATNF, January 2000.

[15] The HDF Group. *Hierarchical data format version 5*, 2000-2010.

[16] The HDF Group. *HDF5 Reference Manual*, November 2009.

[17] The HDF Group. *HDF5 User's Guide*, November 2009.