LOFAR & HDF5

Toward a New Radio Data Standard

WWW.LOFAR.ORG

For decades now, scientific data volumes have experienced relentless, exponential growth. As a result, legacy astronomical data formats are straining under a burden not conceived when these formats were first introduced. With future astronomical projects ensuring this trend, ASTRON and the LOFAR project is exploring the use of the Hierarchical Data Format, version 5 (HDF5), for LOFAR radio data encapsulation. Most of LOFAR's standard data products will be stored natively using the HDF5 format. In addition, HDF5 analogues for traditional radio data structures such as visibility data and spectral image cubes are also being developed. The HDF5 libraries allow for the construction of potentially distributed, entirely unbounded files. The nature of the HDF5 format further provides the ability to custom design a data encapsulation framework. The LOFAR project has designed several data formats that will accommodate and house all LOFAR data products, the primary styles and kinds of which are presented in this poster. With proper development and support, it is hoped that these data formats will be adopted by other astronomical projects as they, too, attempt to grapple with a future filled with mountains of data.

The LOFAR Radio Telescope

- The "LOw Frequency ARray" (LOFAR)
- Currently, the "largest radio telescope in the world"
- 48 stations (40 NL, 8 International); complete by end of 2011
- Baselines from 1 - 1000 km; ultimately achieve sub-arcsecond resolution over much of the band
- Low Band Antenna (LBA) bandpass: 30-80 MHz
- High Band Antenna (HBA) bandpass: 120-240 MHz
- Total Bandwidth: 48MHz
- Data Correlation: IBM Blue Gene/P supercomputer, Groningen, NL *(see photo on right)*
- Offline processing cluster has 100 nodes, each with: 24 cores, 64GB RAM, 21TB
- Long Term Archive (LTA) has: 2.2PB disk, 5PB tape
- Access to 22,600 cores via BigGrid and JUROPA

The LOFAR "Superterp", Exloo, NL

The Blue Gene/P supercomputer at Groningen

LOFAR Data: Variety, Complexity, Volume

Datasets produced by LOFAR observations will vary tremendously in type, size and complexity *(see tables below)*. Radio Sky Images, Beam-formed (BF) data, Transient Buffer board (TBB) time-series data and Dynamic Spectra data are just some of the datasets which are expected to produce files sizes in the several tens of terabytes.

- LOFAR Cosmic rays observation

Mode	Data source(s)	File size
UHEP	Time-series from BF data and TBBs	3.8 GB/event
VHECR	Time-series from TBBs (station)	25 MB/event
HECR	Time-series from TBBs (station)	25 MB/event
TS	Time-series from TBBs (full aray)	3.0 TB/event

- LOFAR Pulsar observation (BF data, 244 sub-bands, 5 stations)

Exposure Time	File size (Known)	File size (Search)
1 min	11.2GB	56GB
10 min	112GB	560GB
30 min	336GB	1.7TB
1 hr	672GB	3.4TB
2 hr	1.3TB	6.7TB
12 hr	8.0TB	40.3TB

Data product	Quantity	Array shape
TBB time-series	$s_j(t)$	1-dim
TBB spectral-data	$\widetilde{s}_j(\nu)$	1-dim
Beam-formed data	$\widetilde{S}(p,\nu,\mathrm{Dec},\mathrm{RA})$	3-dim
All-sky dynamic spectrum	$I(p,\nu,t)$	3-dim
Visibility data	$V(p,\nu,t,B_{12})$	4-dim
Image hypercube	$I(p,t,\nu,\vec{\rho})$	6-dim
Radio sky image	$I(p,\nu,Dec,RA)$	4-dim
CR image cube	$I(p,t,\nu,r,\mathrm{El},\mathrm{Az})$	6-dim
CR image cube	$I(p,t,\nu,\xi_3,\xi_2,\xi_1)$	6-dim
RM Synthesis cube	$\mathrm{DF}(p,\mathrm{Dec},\mathrm{RA},\phi)$	4-dim
RM Synthesis map	$\mathrm{RM}(\mathrm{Dec},\mathrm{RA})$	2-dim

- Data rates up to 8GB/sec
- File sizes, 10's to 100's TB
- Datasets ranging from 1 up to 6 dimensions
- Single dataset can be spread over multiple disks
- Data writing must be multi-threaded and parallelizable
- Data are Hierarchical, multidimensional
- Not a job for most astronomical data containers such as CASA nor FITS!

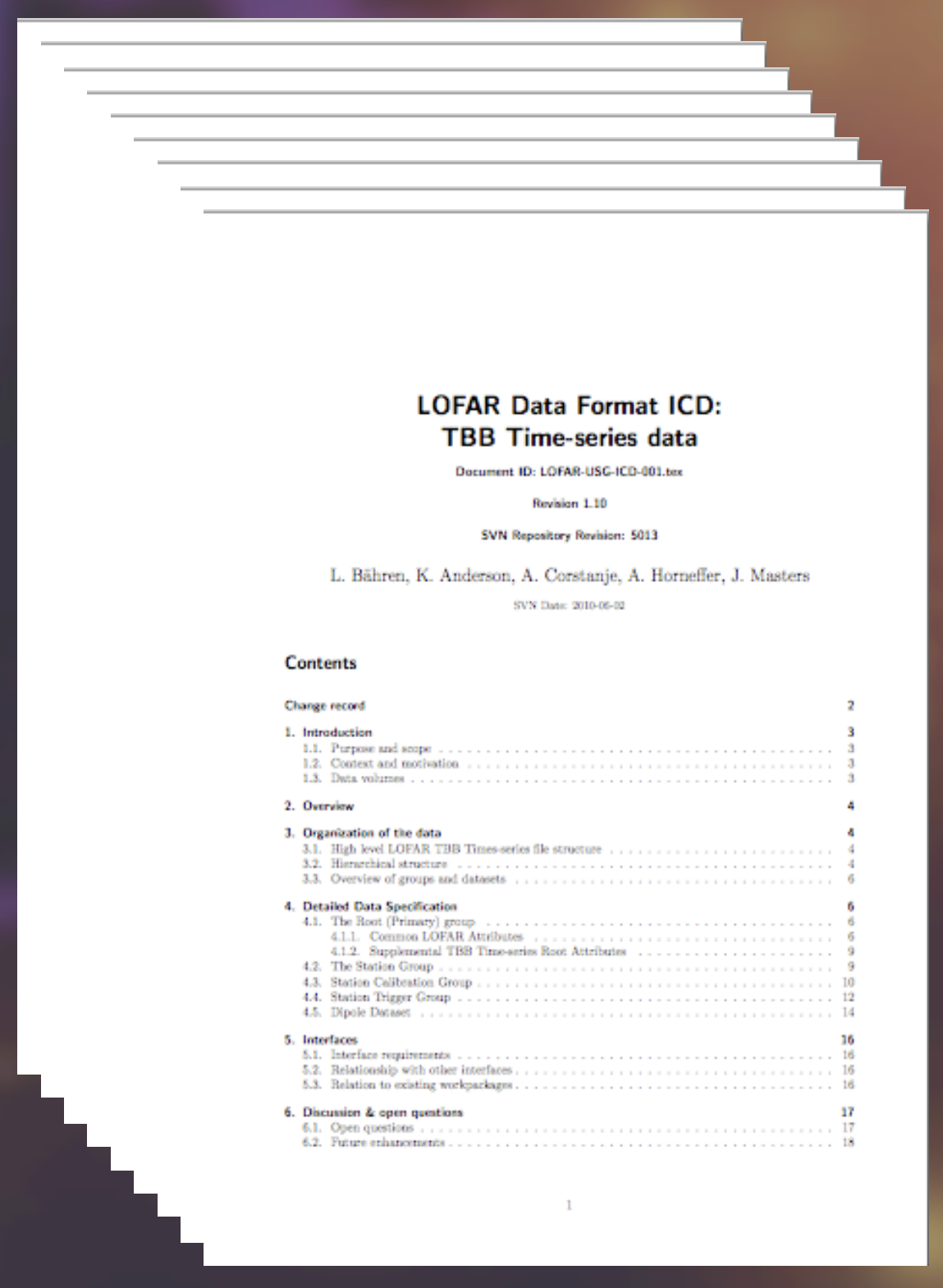
LOFAR Data Format Specifications -- HDF5

HDF5 was chosen as a viable solution for potentially massive LOFAR data products. HDF5 provides a framework allowing users to essentially design their own files to appropriately accommodate a vast variety of data. For over two years the LOFAR project has been engaged in designing a complete set of specifications for all LOFAR observational data, with a certain structural parallelism maintained across all file designs.

These **Interface Control Documents[3] (ICDs)** provide detailed descriptions of all expected LOFAR Data Products:

- **Radio Sky Image Cubes**
- **Dynamic Spectrum Data**
- **UV Visibility Data**
- **Beam-Formed (BF) Data**
- **Transient Buffer Board (TBB) Time-Series**
- **Rotation Measure (RM) Synthesis Cubes[4]**
- **Near-field Images[4]**

For the LOFAR project, a data product is defined within the context of HDF5. 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, logs, etc.. As an "all-in-one" wrapper, the HDF5 format simplifies the management complex datasets.



LOFAR Data Format ICD:
TBB Time-series data

Document ID: LOFAR-USG-ICD-001.tex

Revision 1.10

SVN Repository Revision: 5013

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

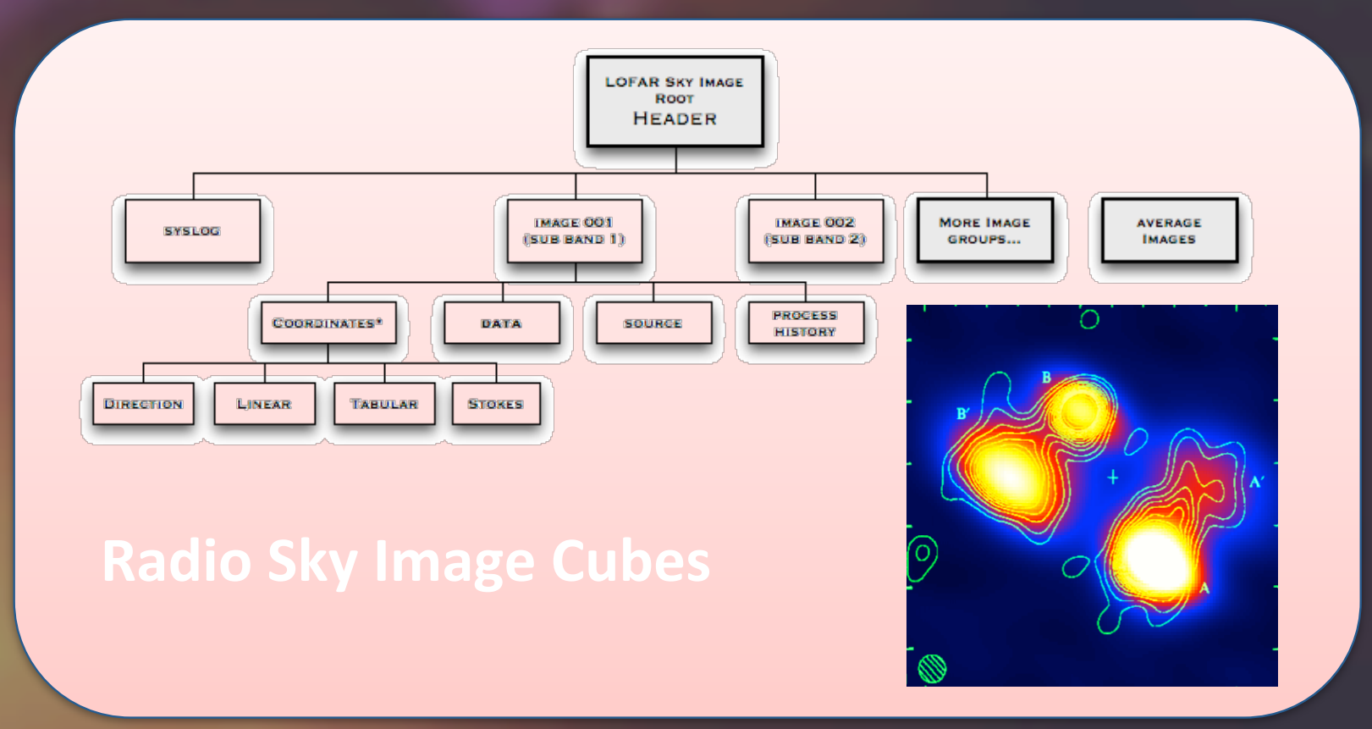
SVN Date: 2010-06-02

Contents

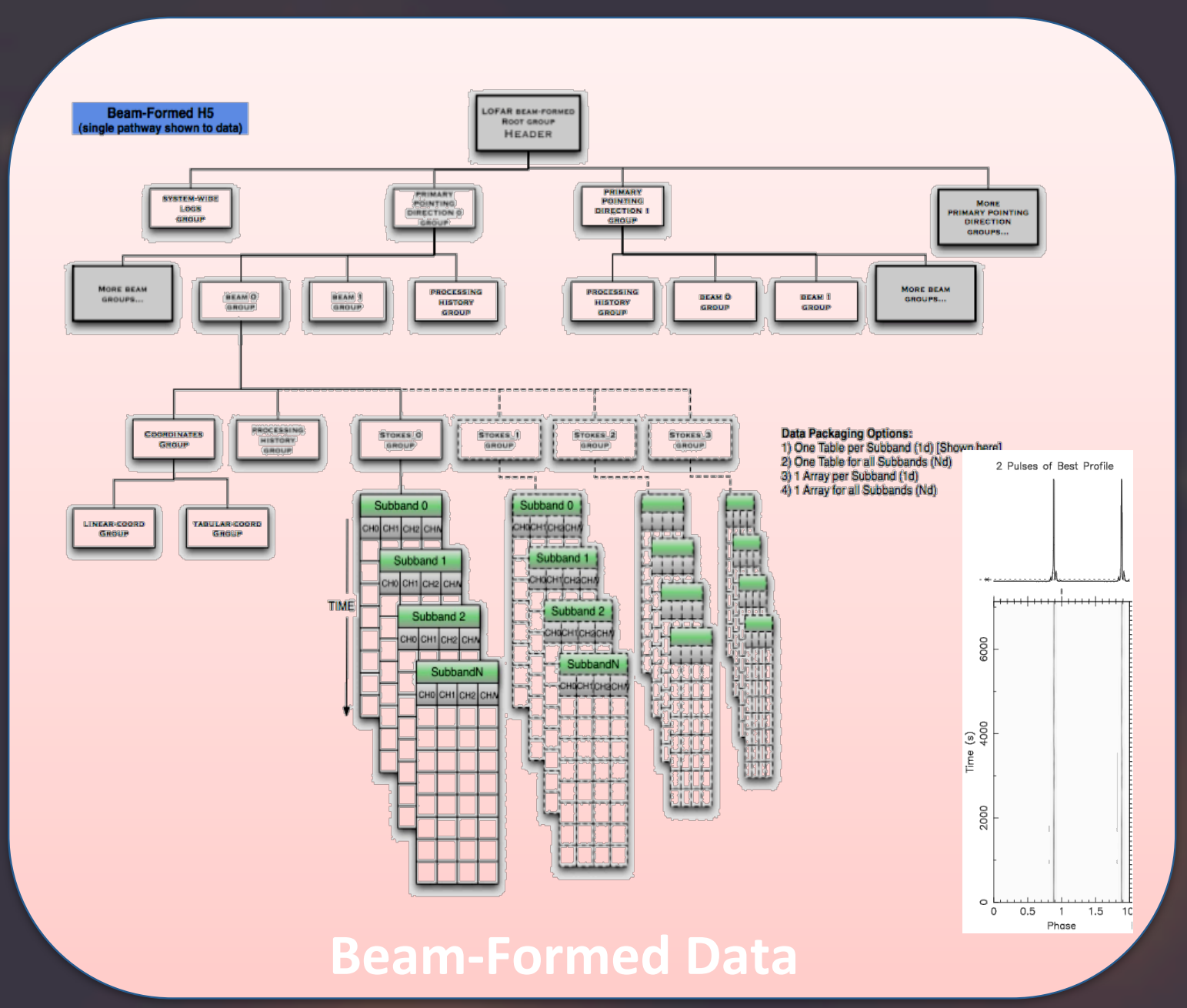
Change record 2
1. Introduction 3
1.1. Purpose and scope 3
1.2. Context and motivation 3
1.3. Data volumes 3
2. Overview 4
3. Organization of the data 4
3.1. High level LOFAR TBB Times-series file structure 4
3.2. Hierarchical structure 4
3.3. Overview of groups and datasets 6
4. Detailed Data Specification 6
4.1. The Root (Primary) group 6
4.1.1. Common LOFAR Attributes 6
4.1.2. Supplemental TBB Time-series Root Attributes 9
4.2. The Station Group 9
4.3. Station Calibration Group 10
4.4. Station Trigger Group 12
4.5. Dipole Dataset 14
5. Interfaces 16
5.1. Interface requirements 16
5.2. Relationship with other interfaces 16
5.3. Relation to existing workpackages 16
6. Discussion & open questions 17
6.1. Open questions 17
6.2. Future enhancements 18

1

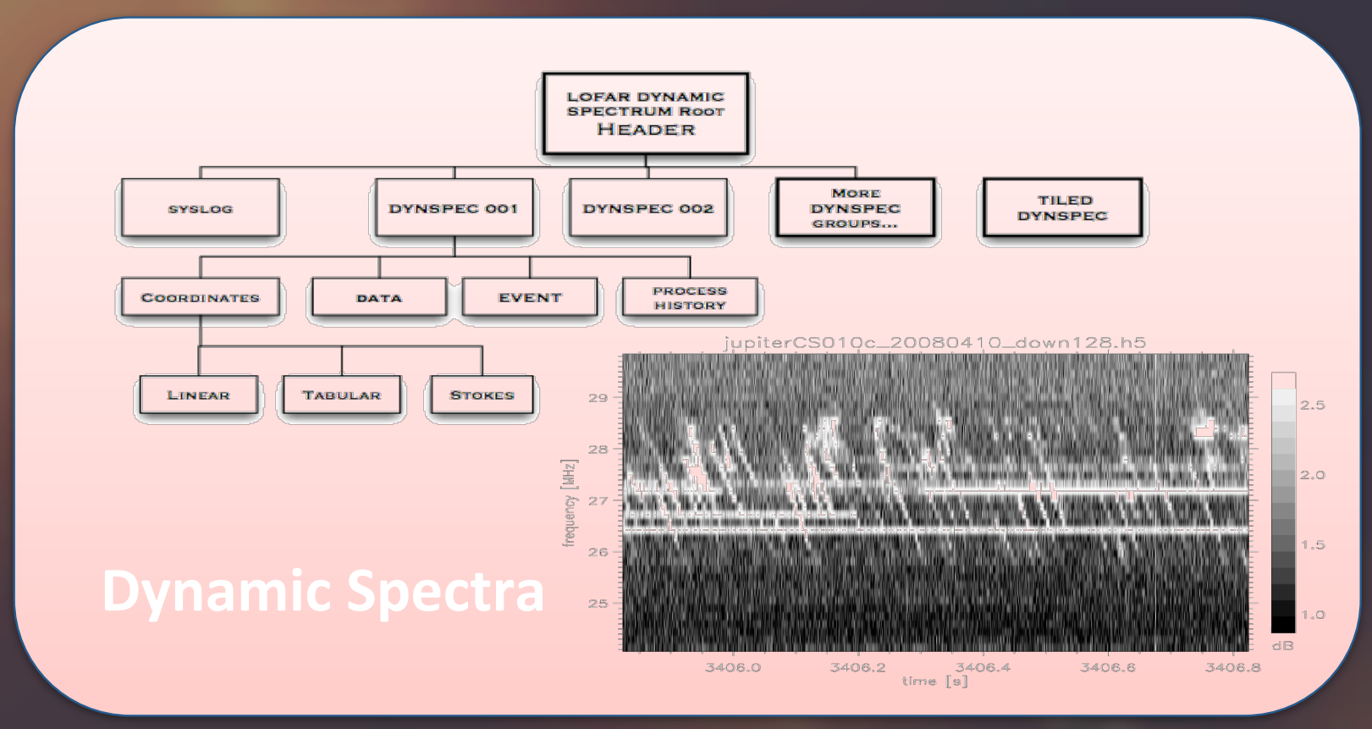
HDF5 Home Page:
www.**hdfgroup.org**



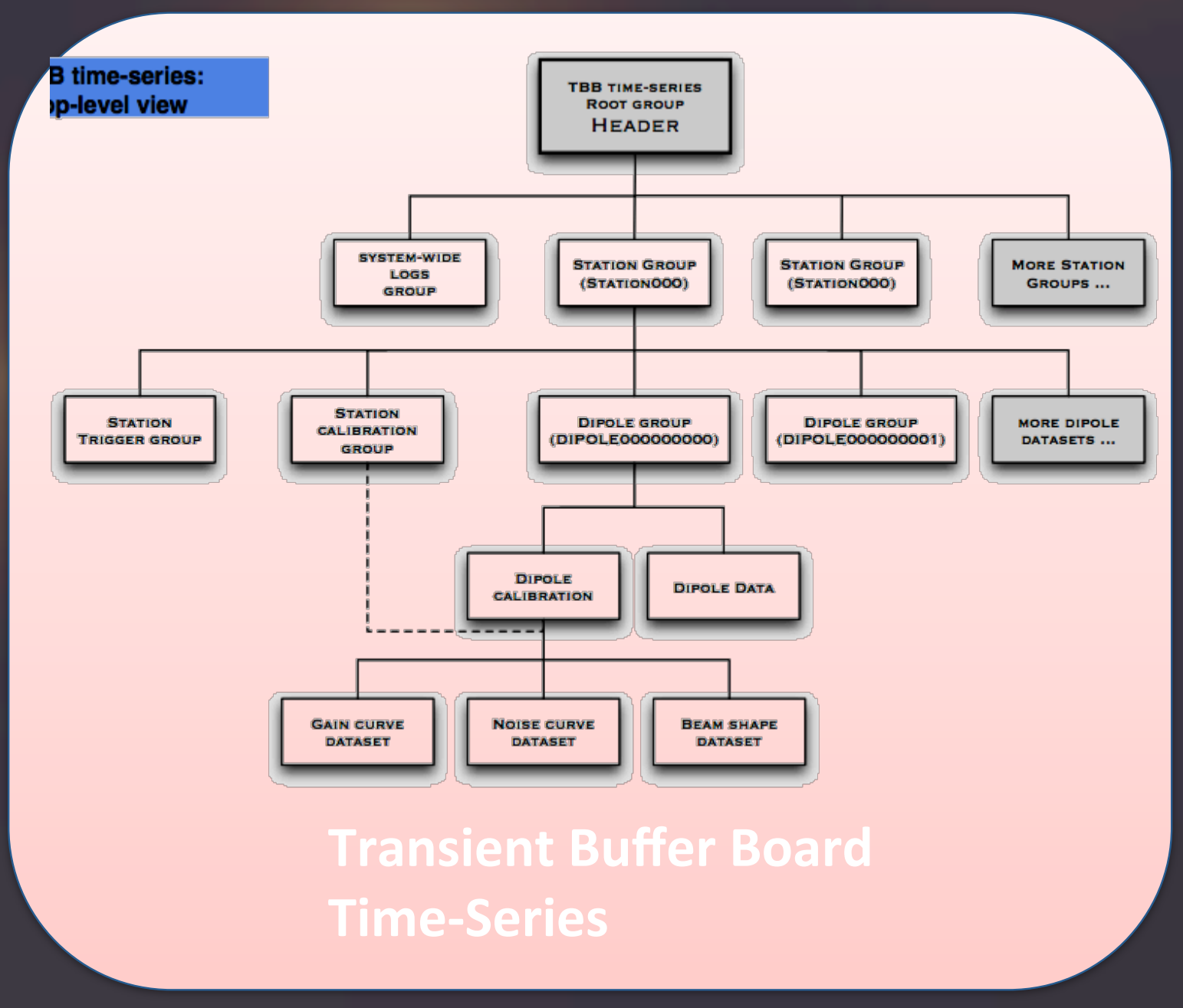
Radio Sky Image Cubes


Beam-Formed H5 (single pathway shown to data)
LOFAR BEAM-FORMED ROOT GROUP HEADER
SYSTEM-WIDE LOGS GROUP
PRIMARY POINTING DIRECTION 0 GROUP
PRIMARY POINTING DIRECTION 1 GROUP
MORE PRIMARY POINTING DIRECTION GROUPS...
MORE BEAM GROUPS...
BEAM 0 GROUP
BEAM 1 GROUP
PROCESSING HISTORY GROUP
COORDINATES GROUP
STOKES 0 GROUP
STOKES 1 GROUP
STOKES 2 GROUP
STOKES 3 GROUP
LINEAR-COORD GROUP
TABULAR-COORD GROUP
Subband 0
Subband 1
Subband 2
SubbandN
CH0 CH1 CH2 CHN
TIME
Data Packaging Options:
1) One Table per Subband (1d) [Shown here]
2) One Table for all Subbands (Nd)
3) 1 Array per Subband (1d)
4) 1 Array for all Subbands (Nd)
2 Pulses of Best Profile
Time (s)
Phase

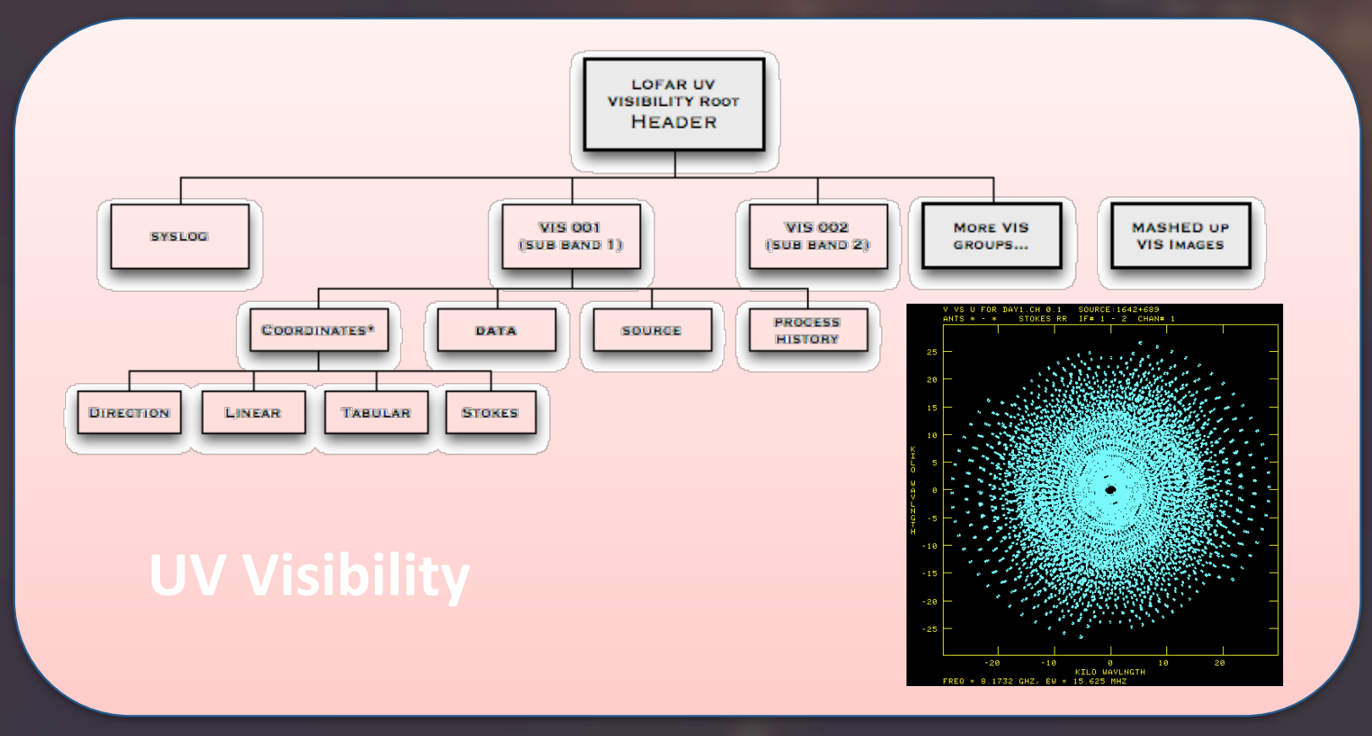
Beam-Formed Data



Dynamic Spectra


B time-series: p-level view
TBB TIME-SERIES ROOT GROUP HEADER
SYSTEM-WIDE LOGS GROUP
STATION GROUP (STATION000)
STATION GROUP (STATION000)
MORE STATION GROUPS ...
STATION TRIGGER GROUP
STATION CALIBRATION GROUP
DIPOLE GROUP (DIPOLE000000000)
DIPOLE GROUP (DIPOLE000000001)
MORE DIPOLE DATASETS ...
DIPOLE CALIBRATION
DIPOLE DATA
GAIN CURVE DATASET
NOISE CURVE DATASET
BEAM SHAPE DATASET

Transient Buffer Board Time-Series



UV Visibility

[3] *Documents available from the* ***LOFAR Wiki ("Data Products" link), http://usg.lofar.org/wiki***

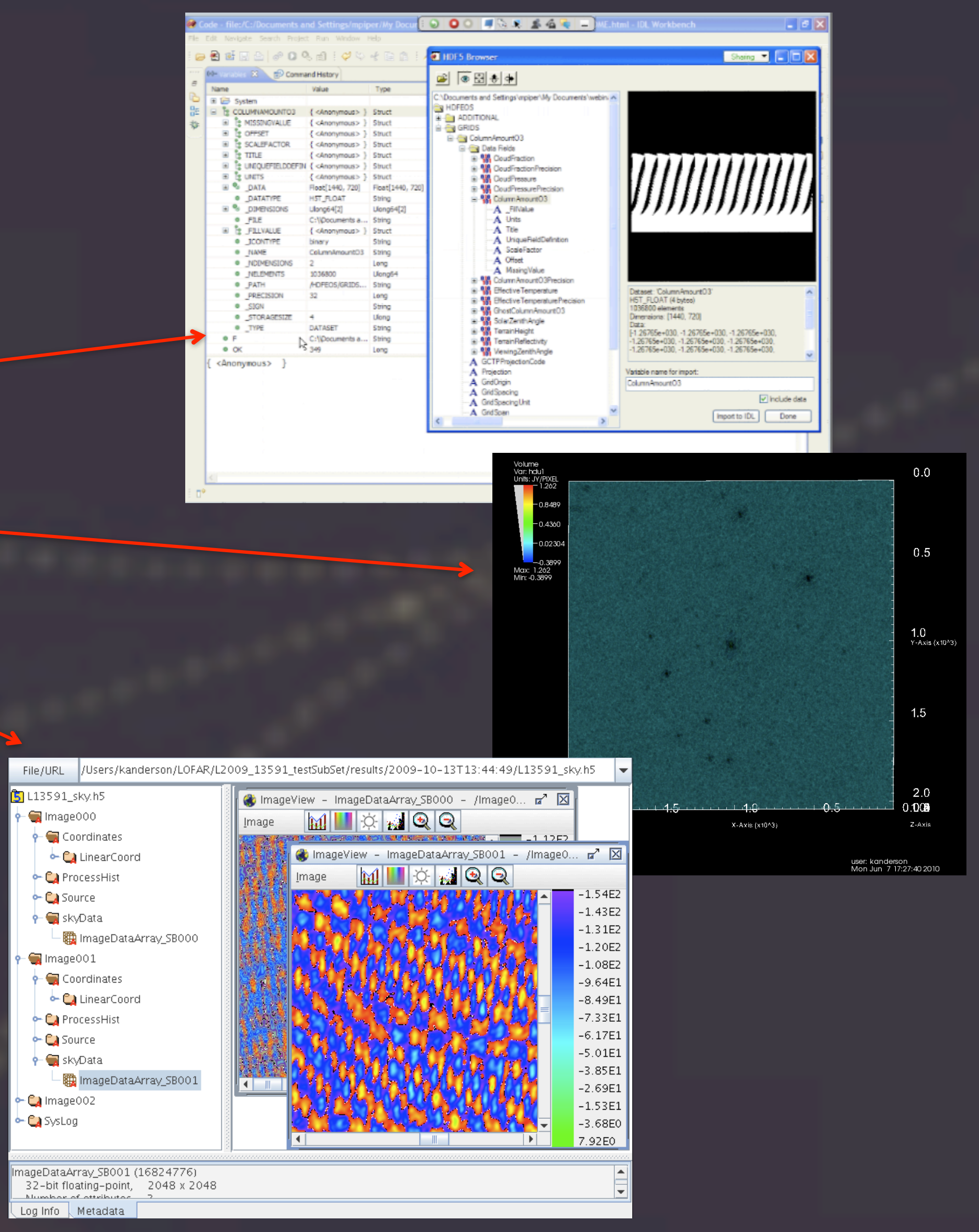
[4] *RM Synthesis and Near Field Imaging are under development*

HDF5-friendly Toolsets, Libraries and Packages

In addition to the HDF5 library itself, the body of libraries and tools available to work with HDF5 files has grown substantially:

- **IDL**
- **VisIt**
- **HDFView**
- **DAL** (LOFAR)
- **PyDAL** (LOFAR)
- **h5py**
- **PyTables**

The LOFAR project is developing the C++ **Data Access Library (DAL)**, which will provide full scope constructors for creating and accessing LOFAR data products.



Summary And Future Considerations

The LOFAR project would like to form collaborations to help develop a true set of standards for radio data which can also meet the demands of future projects such as LSST and the SKA. For this, we have set up a moderated majordomo email list called nextgen-astrodata@astron.nl. To sign up, send mail to majordomo@astron.nl with subject "subscribed nextgen-astrodata".

We are in active development of the C++ **Data Access Library (DAL)**, which ultimately will provide access to LOFAR data. It is a translation of the LOFAR ICDs into code. It abstracts the underlying file format (FITS, the Casa/AIPS++ Measurement Sets and HDF5) from the user. A python interface to the DAL, **PyDAL**, will also be forth coming. The DAL is an open source project on github:

https://github.com/nextgen-astrodata/DAL

Future work also involves developing an interface in **DS9** for HDF5 LOFAR data. Additioanlly, we hope to have a plugin for VisIt to understand HDF5 and astronomical WCS.

The time is ripe to solve the issue of large and/or complex data across wavelengths and projects!

LOFAR · The HDF Group · National Radio Astronomy Observatory, A facility of the National Science Foundation · NRAO · NSF · University of Amsterdam · NWO · ASTRON