

# SciDB Radio astronomy science case

Ger van Diepen  
ASTRON, Dwingeloo, Netherlands  
20-Aug-2009

## Summary

Interferometers are used to obtain the radio-astronomical data in the so-called visibility domain. Antennae receive the (polarized) data and a correlator combines each pair of antennae to form a baseline. The resulting data are stored and processed in offline mode, although for some instruments the visibility data are not stored but processed in streaming mode and only the resulting images are stored.

After doing flagging and calibration, the data are gridded and FFT-ed to an image (which is comparable to an optical image). Thereafter sources are extracted from the image (and deconvolved). Furthermore rotation measure analysis can be done on the image data.

Note that optical astronomy is also starting to use interferometric techniques.

## Data sets

The visibility data are stored in a so-called MeasurementSet, which is described in

<http://www.astron.nl/casacore/trunk/casacore/doc/notes/229.html>. Basically the visibility data consist of 2-dim arrays (with axes frequency and polarization) per baseline (antenna pair) and time stamp. Newer instruments (like LOFAR and ASKAP) can observe multiple beams (parts of the sky) simultaneously which adds another axis. More axes (like frequency band) are also possible.

So the visibility data can be seen as an N-dim array of complex data. The array can be irregular. For instance, the sampling time can be dependent on baseline (shorter times for longer baselines).

The new instruments can have data rates of 2 GBytes/sec or more. Large MeasurementSets can consist of a few million rows, each row containing large data arrays.

Typical numbers for an 8 hour ASKAP spectral line observation are nbeam=32, nfreq=16384, npol=4, nbaseline=666, ntime=5760.

The data rate of SKA will be much higher.

The image format is quite simple; it is a rectangular 4-dimensional array with coordinates right ascension, declination, frequency, and Stokes (polarization). It can have one or more masks attached to it, where a mask is a boolean array with the same shape.

Images can be created as a time-series to detect transients. This adds another axis.

Image cubes can be quite large; for ASKAP an 8 hour observation can result in a spectral image cube of 4-byte float pixels with shape [12188,12188,16384,4].

Furthermore there are auxiliary data sets. The main ones are:

- Calibration tables containing instrumental (e.g. receiver gain) and sky (e.g. ionospheric distortion) calibration parameters that are valid for a given time and frequency interval.
- The sky model describing the astronomical sources in the sky.
- Observation catalogue telling the observations done and where the data are located in the archive.

## Operations on visibility data

- Flagging looks for RFI and other disturbances in the visibility data. This is often done by means of medians in a running time/freq window on the visibility data. However, simpler techniques like thresholding are also used. This is an area in development, because until now astronomers did it by looking at the data, but with the (future) high data rates it has to be automated.
- Data can be averaged in time and/or frequency to reduce the amount of data up (typically 10-100 times).
- Calibration solves for instrumental and sky parameters. It does it by comparing a chunk of visibility data (usually all baselines, part in time and freq) with a model and solving (some) parameters.

A parameter can be a simple scalar or some function in time/freq in which the coefficients of the functions are solved.

The model is non-linear, so multiple iterations are needed (using e.g. Levenberg-Marquard).

Examples of parameters are:

- gain and phase per antenna (station in LOFAR terminology).
- Direction-dependent gain/phase per antenna.
- Beam parameters.
- Source parameters (position, flux, ...)
- Ionosphere.
- Known sources are modeled and subtracted from the data. Calibration parameters are applied to the data.
- Imaging grids the visibility data to a regular grid and performs an FFT to form the image. An image is formed from all visibility data for a frequency channel; sometimes multiple frequency channels are combined into a single image channel.

The gridding is a complex operation, especially for the newer telescopes with a large field of view. Each visibility data point has to be gridded to a box of  $n \times n$  grid points using a convolution function. Also direction-dependent corrections (e.g. ionosphere, beam) need to be applied. Basically there are two approaches:

- W-projection can handle a large field (see <http://www.aoc.nrao.edu/evla/geninfo/memoseries/evlamemo67.pdf>).

For LOFAR or ASKAP  $n \times n$  will typically be around  $50 \times 50$ .

- Faceted imaging divides the field of view into smaller pieces. All data have to be phase-shifted to the center of each facet, gridded and FFT-ed.  $N^2$  can be much smaller depending on the size of the facets. The advantage of faceting is that the grid to fill is much smaller and might fit in the cache, but it requires more calculations.
- A combination of both methods is possible (when using large facets).

Besides these fundamental operations, the data are accessed for inspection. For example, making plots of the data (possibly averaged).

### **Operations on image data**

- Finding sources by searching for peaks in each frequency plane and fitting source parameters. The source parameters can be refined in the calibration process (see above). The sources found have to be matched against the overall sky catalogue.
- Cleaning the image (involves deconvolution). See for example <http://www.cv.nrao.edu/~abridle/.../node10.html>
- Image analysis (moments, statistics, combining images)
- Image display.
- Image subsets and regridding (for Virtual Observatory).
- Rotation measure synthesis to reconstruct the intrinsic polarization properties along a line of sight, using a Fourier relationship between the observed polarization products and a function describing the intrinsic polarization (the Faraday dispersion function).

### **Currently used software**

To store their data most new radio telescopes use the casacore Table System which is a simple form of an RDBMS, but with array extensions. Arrays can only contain basic types (bool, integers, reals, complex, string). Each column can have its own storage manager of which there are several types.

In this way data are stored in a columnar way to get faster queries (we do not use indices). There is an SQL-like query language TaQL (see <http://www.astron.nl/casacore/trunk/casacore/doc/notes/199.html>) to do selection, sort, update, etc.

Image arithmetic is done using a language called LEL (see <http://www.astron.nl/casacore/trunk/casacore/doc/notes/229.html>)

Pyrap is the python interface to casacore and is used heavily for ad-hoc access to the data (with pylab for plotting, etc.).

### **Distributed processing**

The possibly terabyte sized data sets for LOFAR and ASKAP are stored in a distributed way by partitioning the data along the frequency axis. Often each partition can be processed fully independently, but sometimes communication

is needed. It is tried to limit the amount of data to be exchanged as much as possible. For instance, getting a joint calibration solution is done by sending the normal equations to a central solver process instead of sending the much larger visibility data arrays. Similarly, imaging sends the smaller images instead of the visibility data.

## Data access patterns

1. Visibility data
  - Flagging accesses the data in time order. For some flagging algorithms a running time window is needed.
  - Calibration also accesses the data in time order. Often a time window is read to get sufficient signal-to-noise. Sometimes the (unflagged) data are averaged to reduce the amount of data. This will also change the weights of the data.
  - Imaging makes an image per frequency channel, thus accesses the data in frequency order.
  - Ad-hoc queries can be in any order.
2. Image data
  - Display is usually done of RA-DEC planes, but plotting of frequency profiles is also done.
  - RM-synthesis accesses images in frequency order.
  - Transient detection needs to compare RA-DEC images in time order.

## SciDB wishlist

- Guaranteed timely insertion into the data base (in binary form) to ensure the real-time system can store its data.
- Array axes in row major order (as in C and numpy).
- 0-relative indices.
- Data retrieval and updates in binary form because converting to/from ASCII is (too) expensive.
- Handling of masked arrays (i.e., ignore flagged data points)
- Tiled (chunked) data storage for fast access along all axes and automatic adaptation of chunk cache size to access patterns.

## Glossary

ASKAP	Australian SKA Pathfinder ( <a href="http://www.atnf.csiro.au/projects/askap">www.atnf.csiro.au/projects/askap</a> )
CASA	Common Astronomical Software Applications
LEL	Lattice Expression Language
LOFAR	Low Frequency Array (see <a href="http://www.lofar.org">www.lofar.org</a> )
RFI	Radio Frequency Interference
SKA	Square Kilometer Array
TaQL	Table Query Language
casacore	core of CASA package (see <a href="http://casacore.googlecode.com">casacore.googlecode.com</a> )
pyrap	Python Radio-Astronomical Package python interface to casacore (see <a href="http://pyrap.googlecode.com">pyrap.googlecode.com</a> )