SDP Memo 033: Sky Model Considerations

Document number……………………………………………………………….……SDP Memo XX
Document Type…………………………………………………………………..……MEMO
Revision………………………………………………………………………………...DRAFT
Author…………………………………………………………………………...….Ian Heywood
Release Date………………………………………………………….YYYY-MM-DD
Document Classification…………………………………………………. Unrestricted

Lead Author	Designation	Affiliation
Ian Heywood		University of Oxford Rhodes University
Signature & Date:		13 Oct 2017

SDP Memo Disclaimer

Table of Contents

List of Tables

List of Abbreviations

AIPS: Astronomical Image Processing System
ASKAP: Australian Square Kilometre Array Pathfinder
ATCA: Australia Telescope Compact Array
CASA: Common Astronomy Software Applications
DFT: Direct Fourier Transform
EMU: Evolutionary Map of the Universe
FIRST: Faint Images of the Radio Sky at Twenty-cm
FFT: Fast Fourier Transform
GLEAM: Galactic and Extragalactic All-sky MWA Survey
GSM: Global Sky Model
LOFAR: Low Frequency Array
LoTSS: LOFAR Two-metre Sky Survey
LSM: Local Sky Model
MFS: Multi-frequency Synthesis
MSSS: Multi-frequency Snapshot Sky Survey
MWA: Murchison Widefield Array
NRAO: National Radio Astronomy Observatory
NVSS: NRAO VLA Sky Survey
SKA: Square Kilometre Array
SNR: Signal-to-Noise Ratio
SUMSS: Sydney University Molonglo Sky Survey
TIFR: Tata Institute of Fundamental Research
TGSS-ADR: TIFR/GMRT Sky Survey Alternate Data Release
VLA: Very Large Array
VLSS-r: VLA Low-frequency Sky Survey Redux

1. Introduction and context

Synthesis arrays have always relied on bright sources in the sky for instrumental calibration. The two main aspects of the instrument that must be calibrated are the bandpass (***B***, the system gain as a function of frequency, assumed to be slowly variable in time) and the complex gains for each antenna or station (***G***, time-variable due to electronic or atmospheric effects, and usually assumed to have a weak frequency dependence). The absolute flux-density scale and astrometric reference frame of the measurements are also set via the calibration process. Strong sources are used as they provide measurements with a high signal-to-noise ratio (SNR), minimising the time spent calibrating the machine, and providing a set of instrumental corrections with correspondingly high SNR. For a cm-wave instrument, with a parabolic antenna limiting the field of view, the assumption that a carefully-selected calibrator is an isolated point source at the phase centre is usually good enough.

For calibration to occur, the model of the sky must be inverted into a set of model visibilities, with time, frequency and baseline coordinates matching those of the actual measurements. Calibration is the process of constructing an instrumental model, and employing a numerical solver to determine the values of ***B*** and ***G*** that allow the measurements to best fit the model. These instrumental corrections are then applied to the observations until such time as the instrument drifts away from its current state and the calibration must be repeated. The special case of *self-calibration* involves an often-iterative refinement of an ad-hoc sky model generated from the observations themselves. After the reference calibration is applied to the visibilities they are imaged, and an interim LSM based on this image is used to further refine the (typically ***G***-only) telescope gains. The use of strong astronomical calibration beacons across the full field of view to mitigate direction-dependent effects during the self-calibration stage has also become more routine in recent years.

The sky model requirements of the SKA are in principle no different to those described above. However, particularly for LOW but also for MID, the field of view is wide enough, and the instantaneous sensitivity is high enough that there is effectively no such thing as an isolated point source that can be used for reliable calibration purposes.

This leads to a requirement that the Sky Model must contain a suitably accurate description of the sky brightness distribution as a function of position, for the entire sky visible to the SKA. It must describe the frequency dependence of this brightness distribution over a range suitable for both the LOW and MID components of the SKA, and be able to capture the temporal behaviour of sources known or found to be variable. We subsequently refer to this master database as the Global Sky Model (GSM).

For a given observation the GSM is queried and a suitable subset of sources spanning the field of view and frequency range of that observation are returned. This subset is referred to as the Local Sky Model (LSM). The content of the LSM describes the *intrinsic* properties of a region of the sky, however this will not match the sky as perceived by the telescope, as it is corrupted by the instrumental and atmospheric effects that must be mitigated via calibration. A particular example is the position-dependent sensitivity of the instrument imparted by the primary beam, which will cause sources away from the pointing direction to be attenuated. The LSM is passed to visibility prediction machinery for generation of an *apparent* model against which the instrument can be calibrated. How large the subset of sources described by the LSM is depends primarily on the field of view, and therefore the lowest frequency of the observing band. In practice, there is likely to be a requirement to include sources beyond the main lobe of the primary beam, and for LOW in particular the strongest sources in the sky (Cygnus-A, Cassiopeia-A, Taurus-A… usually referred to as the 'A-team') will likely have to be dealt with irrespective of where the telescope is pointing.

There should also be link between the end-stage of the data processing, linking the final derived science products to the GSM, whereby the GSM can be updated with an *intrinsic* description of the sky derived from the observations themselves. The reason for this is to steadily grow the database, particularly in terms of the number of faint sources it contains, as well as to keep track of sources that are known or found to be strongly variable. This leads to improved accuracy in subsequent calibration operations, as well as cheaper imaging for repeated observations of deep fields, whereby the best-fitting model visibilities may be subtracted as an initial step post-calibration, and imaging and deconvolution is only required on the residual visibilities.

2. Prior art

As mentioned above, the need for a sky model is nothing new, only the capabilities of the SKA make the scope of the required GSM somewhat unprecedented. Here we review some of the existing approaches, as well as existing and forthcoming databases and surveys that could potentially become the initial building blocks of the SKA's GSM.

2.1 Sky model formats in existing calibration packages

Every existing software package that can perform calibration of telescope data relies on predicting model visibilities from a model of the sky. These models typically take two forms, namely either an image or a catalogue. Image-based models usually contain clean components, i.e. a set of delta functions with a position, a brightness and sometimes a description of their spectral behaviour. Such groups of components are a natural by-product of deconvolution by numerous flavours of the clean algorithm. The advantage of them is that self-calibration can proceed using a natural intermediate product of the imaging process. The downside is that positional accuracies are quantized at the level of the pixel size in the spatial dimensions.

The AIPS **[1]** package handles clean components in database form, via the CL table in the FITS headers of data products, however CASA **[2]** generates images with the same dimensions as the sky image itself in order to store them. The latter approach balloons the data volume somewhat, as for most cases the overwhelming majority of pixels in the model are zero (although this leads to them being readily compressible). Spectral behaviour in CASA's imaging routines (MFS; **[3]**) is handled by modelling the frequency behaviour of each source as a Taylor polynomial pegged to the total intensity value at the reference frequency. Additional planes in the image domain are used to capture the higher-order Taylor terms. Stokes parameters to capture the polarization properties of a source are also represented by delta functions in images. A full-Stokes (IQUV) model of a field with polynomial order high enough to capture spectral curvature (3 terms) could therefore result in up to 12 full-field images being required. This approach is also adopted by the ASKAPsoft **[4]** suite of software.

The MFS approach is handled in slightly different ways by some other alternative imaging packages. For example, wsclean **[5]** and DDFacet **[6]** do not employ a Taylor expansion but perform deconvolution independently in an arbitrary number of sub-bands, with peaks being identified in the full-band image during the minor cycle. Polynomials of arbitrary order are fitted to the peaks in each sub-band, before being inverted into a visibility model for subtraction during the minor cycle. In these cases the frequency dimension is again represented by individual frequency plane images, the number of which is dependent on the number of sub-bands.

For the handful of ‘primary’ calibrators which are the sources of choice for things such as setting absolute flux scales for e.g. VLA observations, CASA comes bundled with image-based clean component models for each of the VLA’s observing bands. These are designed to capture both the spectral behaviour of the sources, but also their changing effective morphologies as the frequency changes. An example of the latter might be a source that has an extended radio jet visible at lower frequencies, whereas at higher frequencies it is effectively seen only as a flat-spectrum core.

Component-based models, represented by a table as opposed to an image, have the advantage that as a data product they occupy significantly less volume than images, but also that the position of the components can be specified with arbitrary precision. Visibility prediction from a component catalogue requires either a somewhat expensive DFT operation per component, or for the models to be gridded as delta functions prior to FFT inversion as per the image-based case. Software packages exist that are capable of taking a hybrid approach. CASA has a toolkit for handling component-based models, and the MeqTrees **[7]** package handles direction-dependent calibration by partitioning up the sky model, typically using DFT predictions for models towards problem sources, with a direction-independent component predicted from a so-called ‘brick’ of clean components. The flexibility to mix-and-match sky models depending on the calibration problem is desirable, and should be implemented for the SKA to cater for different calibration scenarios that may arise.

Alternatives to clean component representations for complex morphologies have been proposed over the years. One such example is the decomposition of a brightness distribution into a set of basis functions, the list of coefficients being an lightweight way to characterise complex morphological features. Shapelets **[8]** are one such example, as the basis functions are invariate between the image and Fourier domains, a property that is useful for synthesis imaging. However, none of these alternatives have gained traction enough for any mainstream software packages support them.

2.2 Large-area radio surveys

There are several large-area radio surveys that have been carried out with existing instruments, including SKA pathfinders and precursors, and several more that are forthcoming. These have led to radio images and subsequently-derived source catalogues at various frequencies, depths and angular resolutions. In some cases the survey covers essentially the entire sky visible to the observatory that conducted it. Most existing large-area sky surveys have been conducted in the northern hemisphere, although this is set to change as the SKA precursor instruments complete their initial batch of large-scale observing programmes.

The results of *any* significant radio survey that covers the sky visible to the SKA is of potential interest as a contributor to the initial GSM. The major surveys of interest are summarised in the table below.

Name	**Freq (MHz)**	**Ang.Res(``)**	**Depth**	**Nsrcs**	**Coverage**	**Year**	**Ref**
VLSS-r	74	75	100	92,965	>-30	2014	**[9]**
MSSS	30-160					2015	**[10]**
GLEAM	52-212	160	5	307,455	<+30	2017	**[11]**
LoTSS	120-168	5	3.5				**[12]**
TGSS-ADR	150	25	0.1	620,000	>-53	2017	**[13]**
SUMSS	843	45	1	107,765	<+30	2003	**[14]**
EMU	1100-1400	10	0.01		<+30		**[15]**
NVSS	1400	45	0.45	1,800,000	>-40	1998	**[16]**
FIRST	1400	5	0.15	946,432	NGC, SGC	2014	**[17]**

Table 1: *Radio surveys that could contribute towards populating an initial GSM database*

2.3 Existing calibrator databases

The large-area surveys listed above are certainly of immense use for the basis of a GSM for the SKA, challenges of cross-identification and quality control notwithstanding. However the databases of known calibrator sources maintained by both the VLA **[18]** and the ATCA **[19]**, each of which contains a couple of thousand entries distributed over the entire sky, should also be made use of. The advantage of these databases is that they have broadband spectral coverage, astrometric positions that are often measured using VLBI, and (particularly in the case of the ATCA database) feature several epochs of monitoring to determine source variability (usually more important at the higher frequencies). These databases are geared towards the assumption mentioned previously that the astrometric calibration source is effectively the only thing in the field of view, an assumption that certainly does not hold for LOW, and likely should not be made for MID. However the ATCA database in particular likely already forms the basis of a suitable astrometric reference frame for the SKA. Improved models of the surrounding fields can be initially constructed using the surveys listed in 2.2, and then latterly using observations from the SKA itself (or its precursors). Calibrator positions from **[18]** or **[19]** should likely be used to calibrate the astrometry for the entire GSM.

3. Proposed structure of the GSM database

Something approaching the ideal model from which to calibrate the SKA will likely not exist until many years into the operation of the SKA itself. The initial GSM will likely be populated from results already in-hand, such as those listed in Sections 2.2 and 2.3, and then refined with further SKA observations. Depending on the final observing mode, for the case of MID the GSM may be somewhat patchy across the sky. This may take the form of a suitably-spaced grid of well-known calibrator fields that are described in detail (such as those discussed in Section 2.3), that are visited occasionally, and relying on iterative self-calibration during general observing. The functionality to continually update the contents of the database must be present so refinement can be continual. The higher angular resolution observations of MID can be used to update the morphologies of sources to improve LOW calibration for example. The main thing to consider at this stage is that the database is designed to be as future-proof as possible, and will not need to be re-designed later. As such the following parametric model is proposed for the main GSM. Note that this includes uncertainties for all major parameters. These will generally not be used during prediction, however fully-Bayesian calibration approaches have been demonstrated **[20]** and if it becomes computationally feasible to deploy such things in future then these parameters may be required.

Parameter	Type	Unit	Description
SOURCE_ID	Integer		Unique identifier for each source in the database. Sources are formed from groupings of components.
COMPONENT_ID	Integer		Unique identified for each component in the database. Components belonging to a particular source all share the same SOURCE_ID
RA	Float	rad	Right Ascension of the component
DEC	Float	rad	Declination of the component
SIGMA_RA	Float	rad	Uncertainty in RA
SIGMA_DEC	Float	rad	Uncertainty in DEC
I	Float	Jy	Stokes I brightness at FREQ0
Q	Float	Jy	Stokes Q brightness at FREQ0
U	Float	Jy	Stokes U brightness at FREQ0
V	Float	Jy	Stokes V brightness FREQ0
SIGMA_I	Float	Jy	Uncertainty in I
SIGMA_Q	Float	Jy	Uncertainty in Q
SIGMA_U	Float	Jy	Uncertainty in U
SIGMA_V	Float	Jy	Uncertainty in V
RM	Float	rad / m2	Rotation measure
SIGMA_RM	Float	rad / m2	Uncertainty in RM

SPI	Array of floats		Polynomial coefficients describing the model spectrum of the component. Order of polynomial required is determined by length of this array.
SIGMA_SPI	Array of floats		Uncertainties in SPI.
FREQ0	Float	Hz	Reference frequency at which I, Q, U and V are specified, and about which any polynomial expansion described by SPI occurs.
MAJ	Float	rad	Major axis of the component, set to zero for point components.
MIN	Float	rad	Minor axis of the component, set to zero for point components.
PA	Float	rad	Position angle of the component, east of north, set to zero for point components.
SIGMA_MAJ	Float	rad	Uncertainty in MAJ.
SIGMA_MIN	Float	rad	Uncertainty in MIN.
SIGMA_PA	Float	rad	Uncertainty in PA.
TIME	Float	MJD	Date at which all other parameters were last measured. It is assumed that the prediction step will use the most recent values, however this can be used to track source brightness and spectral variability.
ORIGIN	String		Very brief description of the origin of the entry in the GSM. This could refer to an external survey (with version number for surveys with multiple data releases) or could contain a scheduling block ID number for sources derived from SKA observations.

Table 2: *Proposed fields in the GSM / LSM database*

In addition to the parametric model described above, the ability to store images in the database should also be provisioned. Such images would contain a data array as well as a header containing appropriate metadata, and be attached to a unique SOURCE_ID. The upstream processing machinery could be configured to ignore database entries belonging to SOURCE_ID and instead predict visibilities relevant to that direction from the image model. This would be particularly advantageous for morphologically complex sources which are not easily decomposed into sets of points and Gaussians. The issue of very large-scale diffuse emission that fills the primary beam (e.g. diffuse galactic synchrotron emission) poses a difficult problem in terms of modelling. The best way of mitigating this issue is to simply exclude the very shortest interferometer spacings that are most sensitive to it when solving for the instrumental parameters during calibration.

The image metadata should at a minimum replicate RA, DEC, FREQ0 and TIME, with image planes containing data relevant to I, Q, U, V, SPI and RM. Uncertainty maps could also optionally be provided. FITS-format images meet these requirements.

References

Reference Documents

Reference Number	Reference
1	http://www.aips.nrao.edu
2	http://casa.nrao.edu
3	Rau & Cornwell, A&A, 532, 71, 2011
4	https://www.atnf.csiro.au/computing/software/askapsoft/sdp/docs/current/index.html
5	Offringa et al., MNRAS, 444, 606, 2014
6	Tasse et al., *submitted*, 2017
7	Noordam & Smirnov, A&A, 524, 61, 2010
8	Refregier, MNRAS, 338, 35, 2003
9	Lane et al., MNRAS, 440, 327, 2014
10	Heald et al., A&A, 582, 123, 2015
11	Hurley-Walker et al., MNRAS, 464, 1146, 2017
12	Shimwell et al., A&A, 598, 104, 2017
13	Intema et al., A&A, 598, 171, 2017
14	Mauch et al., MNRAS, 342, 117, 2003
15	Norris et al., PASA, 28, 215, 2011
16	Condon et al., AJ, 15, 1693, 1998
17	Becker et al., ApJ, 450, 559, 1995
18	https://science.nrao.edu/facilities/vla/observing/callist
19	http://www.narrabri.atnf.csiro.au/calibrators/
20	Natarajan, I., et al., MNRAS, 464, 4306, 2017