Reply to the comments on WALLABY's early-science processing requirements

Version 1.0 — 20/07/2015

T. Westmeier, B. S. Koribalski, L. Staveley-Smith, P. Serra

1 Introduction

This document contains WALLABY's response to the comments on version 1 of the "*WALLABY Data Storage and Processing Requirements for Early Science Observations with ASKAP*" (WALLABY memo 19). We also refer to WALLABY memo 4 ("*WALLABY Response to ASKAP Science Processing*") for additional information about our data processing requirements.

2 ASKAP Project Scientist

General remarks

> ***Comment:*** *Several teams will need to take into account recent T_{sys}/η measurements for the Mk II PAFs and re-assess integration times per field and preferred survey strategies. Early science teams will need to complete their data requirements estimates (in particular the post-processing requirements) before the end of August, in time for us to submit a proposal to the Pawsey Centre in October this year.*

Reply: In our reassessment of WALLABY's storage and processing requirements we have assumed the revised value of $T_{\rm sys}/\eta \approx 90$ K. WALLABY's general approach to a decreased sensitivity of ASKAP–12 would be to reduce the total number of fields while keeping the sensitivity at the nominal WALLABY value. For $T_{\rm sys}/\eta \approx 90$ K this would translate into observing 5 fields with about 120 h of integration time per field, or 600 h in total.

3 ASKAP Computing

Response to individual points

Integrated cubes

> ***Comment:*** *A fundamental question for both programs is that of how the integrated cubes are created. The ASKAPsoft pipeline will produce cubes for each ~8 h observation of a field. These individual observations will need to be combined in some fashion. It is not clear that imaging all 7 datasets at once is possible within the currently-envisaged ASKAPsoft pipeline, which points to image-plane addition of the image cubes. No facility for this currently exists within ASKAPsoft, although doing this operation should not present too many problems. The integrated cubes may be regarded, however, as Level 7 products if they are not produced as part of the standard pipeline.*

Reply: We note that the *data volumes* produced by ASKAP–12 will be significantly smaller than for ASKAP–36. The number of baselines will be an order of magnitude smaller (66 versus 630), and the suggested time averaging of visibilities from 5 to 30 s will reduce WALLABY's data volume by another factor of 6. Hence, we would have expected ASKAPsoft to be capable of jointly imaging data from approximately 60×8 h on ASKAP–12. Computational requirements could be further reduced by processing the data in smaller chunks of frequency rather than all at once.

While joint imaging of visibilities would be desirable, combination of image cubes in the image domain would be acceptable for WALLABY if for some fundamental reason joint imaging would

not be possible. We note, however, that *additional deconvolution* of the resulting combined cube would be required due to decreased noise. This might pose a problem depending on how and when image deconvolution will be carried out in ASKAPsoft and in which form the PSF information will be stored. If multiple visibility sets could not be imaged together, then our storage requirement for single-beam image and PSF cubes would grow by an unreasonably large factor, as we discuss in the following section on PSF images.

PSF images

> ***Comment:*** *Both programs request PSF image cubes, produced for each beam, that have twice the linear size as the image cubes. This is currently not how the PSF images are produced by the ASKAPsoft imager – it produces PSF images on the same pixel grid as the images. Such functionality is unlikely to be present in time for Early Science, and would need a good justification to be included in our development. It is likely, however, that the PSF images will be produced on a per-beam basis.*

Reply: In order to assess our requirements for the storage of PSF information, we request further information on how imaging and beam information is expected to be made available. Will we be provided with mosaicked cubes or individual cubes of each of the 36 beams from ASKAP–12? How would deconvolution be carried out in practice in either of the two cases? If individual image cubes for each ASKAP beam were provided, what would their spatial size (in pixels) be? We note that, should additional image deconvolution be required (e.g. as the result of combining individual 8-hour observations into a single, deep cube as part of our Level 7 data processing), we will require a beam cube twice the size of the area on the sky within which we can expect to detect emission requiring deconvolution.

We also note that providing individual image and beam cubes for each ASKAP beam would result in unreasonably large data volumes to be stored. Our calculation indicates that in this case the total size of the imaging data would exceed that of the visibility data by a significant factor, unless individual cubes were combined on a daily basis and then immediately discarded.

Continuum maps

> ***Comment:*** *Both programs request continuum maps in each Stokes parameter (I, Q, U, V). These are presumably made over the 300 MHz bandwidth. While fine for Stokes I (and V), it is not clear how much the Faraday rotation across the band will degrade the Q and U images. Are these indeed necessary?*

Reply: Continuum information will be important in the interpretation of WALLABY's data, and we therefore require images of all four Stokes parameters. In order to alleviate the effect of image degradation by Faraday rotation, we changed our request to Stokes I, Q, U and V cubes with 304 spectral channels of 1 MHz width.

Measurement sets

> ***Comment:*** *The measurement set size calculation assumes an integration time of 30 s. While this is feasible for the correlator to produce, the fact that it is different from the nominal 5 s integration time means some additional time may be required to commission it. Section 3.1 gives our estimates of the practical size of datasets expected for these surveys (which are slightly different to those quoted in the documents provided).*

Reply: We carried out additional simulations and calculations which confirm that 30 s time resolution will be sufficient for visibility data storage in light of the 2 km maximum baselines and the dynamic range anticipated for HI sources in the WALLABY early-science data.

We note that the measurement set data sizes specified in section 3.1 of "ASKAP Computing's Response to Early Science Requirements" are smaller than our calculation of the expected raw visibility data volume,

$$(66\ \mathrm{BL} + 12\ \mathrm{AC}) \times 36\ \mathrm{beams} \times 4\ \mathrm{pol.} \times 16416\ \mathrm{ch.} \times 72\ \mathrm{bit} \times (1\ \mathrm{h}/30\ \mathrm{s}) = 185.5\ \mathrm{GB/h},$$

compared to 167.5 GB/h in section 3.1. Our calculation includes autocorrelations and is based on the specifications in version 1.0 of the "ASKAP Science Processing" document (the resulting values are consistent with the data volumes listed in that document).

Targeted – Level 5/6 data products

> ***Comment:*** *The detailed requirements document requests image cubes to be made at 4 different resolutions (30″, 90″, and 270″ as essential requirements, with 20″ desirable). Assuming these are obtained through their own imaging pipelines (that is, with their own weighting schemes, and not just smoothing the 30″, say, to coarser resolution), this implies a large amount of processing. It is hard to comment on the feasibility of this given the current developmental state of the ASKAPsoft imager, and our uncertainties about the run-time for spectral-line imaging at ASKAP–12 scale. We are certainly committed to providing at least the basic 30″ resolution spectral-line products.*
>
> *The production of 20″-resolution images is actually beyond the scope of the ASKAPsoft pipeline – our requirements are for full spectral-line imaging at 30″ resolution. While the data sizes are smaller than for full ASKAP, the increase in resolution does entail larger convolution functions and consequently larger memory footprint. While best efforts can be made at pushing the resolution limit, that would remain a research endeavour and cannot come at the expense of delivering the basic pipeline capabilities.*

Reply: Creation of 20″ data cubes will not be crucial for WALLABY's early-science goals, and we have removed all 20″ data products from the updated version of our document. We note that the final choice of configuration for ASKAP–12 will be optimised to provide maximal sensitivity at an angular resolution that is likely to be different from the 30″ resolution envisaged for ASKAP–12 spectral-line imaging.

90″ and 270″ image cubes would ideally be generated independently through re-weighting and tapering of the visibility data in order to maximise sensitivity. However, should this be impossible due to processing constraints, spatial smoothing of the 30″ data cubes will be an acceptable alternative (and identical to Gaussian tapering of the 30″ cubes in the uv domain). We also note that, irrespective of which method is used, data cubes at 90″ and 270″ resolution will only be required for the local universe (across approximately 50 MHz bandwidth, equivalent to $z \lesssim 0.04$ or $cz \lesssim 10\,000$ km/s) where extended emission can be expected. This would significantly reduce processing requirements.

Targeted – Continuum subtraction

> ***Comment:*** *This is not mentioned anywhere in the documentation. Some specifications about whether this is needed and to what accuracy would be useful.*

Reply: We implicitly assumed that all spectral-line image cubes would be gain-calibrated, bandpass-calibrated, flux-calibrated, and continuum-subtracted to ensure that they are ready for immediate scientific analysis.

The ideal process for continuum subtraction would require (1) subtraction of a sky model of continuum emission from the visibility data and (2) polynomial fitting and subtraction of any residual continuum emission from the imaging data (see WALLABY memo 4 for further details). Continuum subtraction to *well below the noise level of 1.6 mJy* will be required for WALLABY.

Targeted – Post-processing requirements

> ***Comment:*** *In assessing the requirements for the kinematics pipelines and the SoFiA processing, it would be good to get some idea of how suitable the software is for distributed, high-performance computing. How well does the code perform on large numbers of processors? Can it*

make use of all cores available on a single galaxy node? (If a single job is run on a galaxy node, it will be allocated all 20 of that node's cores, so for efficiency's sake it helps if the software can make use of them).

In the meeting on June 16, it was indicated that SoFiA could process large cubes by splitting them up into sub-cubes and running multiple processes. But are these individual processes single-threaded?

Reply: SoFiA is not currently set up to automatically distribute the processing of sub-cubes across multiple cores. Implementing multi-core parallelisation in SoFiA is still work in progress and will be the ultimate goal of the SoFiA development team.

Comment: *It would also be good to understand the contingency factor of 5 included in the calculations. Will it be necessary to re-run the entire dataset 5 times for optimisation, or could this be done with a small subset of the data before arriving at the ideal parameterisation?*

Reply: It is correct that parameter settings can be tested on a small subset of the data. The contingency factor of 5 has been included as a protection against potential failures of individual runs (e.g. due to technical problems) and to be able to run the source finding and parameterisation pipelines multiple times with different parameter settings, the results of which could then be combined into a single source catalogue with improved completeness. Experience from previous projects has shown that a factor of 5 is rather optimistic and will leave little room for failure.

4 CASDA

General remarks

Image cut-outs

Comment: *CASDA will be including a VO Simple Image Access Protocol (SIAP) service that will include cut-outs for images and image cubes. This is planned for the second production release in February 2016.*

For projects requesting image cut-outs around detected sources, further discussion may be needed on whether these should be provided as Level 5/6 data products and stored in the archive, generated as Level 7 data products, or some combination of both.

Reply: Noted. For the time being, we have moved all cut-outs to our Level 7 storage requirements table, as these will be created as part of our source finding pipeline anyway.

CASDA support for Level 7 data products

Comment: *As discussed in the User Requirements document, CASDA provides archive support for science catalogues produced as Level 7 data products. Procedures for depositing have been implemented and will be available for the first production release in November 2015.*

At present CASDA does not archive other Level 7 products including measurement sets, images or spectra. Science teams will need to make their own arrangements for archiving these data products.

However, I note that additional support for Level 7 image and spectral data products is on a 'wish list' for future CASDA project proposals. Funding for significant future CASDA upgrades including this, is not yet determined.

Reply: We removed all data products produced by the source finding pipeline—with the exception of source catalogues—from our Level 5/6 requirements table. These are now listed in the Level 7 data products table only.

Spectral line measurement sets

> ***Comment:*** *CASDA does not archive the spectral line measurement sets from ASKAP data. If access to these is critical this should be discussed with the science data processing group.*

Reply: We understand that there are currently no plans to archive spectral-line measurement sets from ASKAP–12 in CASDA. This is a major concern for WALLABY, as early-science observations with ASKAP–12 will be crucial in testing and debugging the ASKAP data reduction and imaging pipeline. We find it difficult to believe that the first iteration of data reduction will provide us with satisfactory imaging results; multiple iterations will likely be required to identify and rectify any potential problems with the pipeline. Another concern is the current uncertainty about the joint imaging capabilities for multi-epoch data in ASKAPsoft.

Longer-term storage of raw visibility data from ASKAP–12 will therefore be crucial in our opinion. We note that this requirement differs from the full ASKAP array for which visibility data will not need to be stored due to the expected technical maturity of the data reduction pipeline. Should spectral-line measurement sets not be archived in CASDA, an alternative storage location for measurement sets from ASKAP–12 will need to be found such that these sets could be reprocessed with ASKAPsoft if required.

In principle, the amount of storage required for visibility data from WALLABY early-science observations could be drastically reduced in a number of ways, including binning in time by a factor of 6 (from 5 s to 30 s) and reduction of the frequency range to be stored to a redshift range of $z \lesssim 0.1$ (within which most detections are expected and beyond which the RFI situation will deteriorate).

Response to individual points

> ***Comment:*** *From CASDA perspective the required Level 5/6 data products are feasible. However, producing spectral line image cubes at up to three different resolutions looks very demanding on science data processing pipelines and I expect this will need some discussion.*

Reply: Noted. As stated above, 90″ and 270″ cubes can potentially be generated through spatial smoothing of the 30″ cubes without the need to re-run the entire imaging pipeline. If that is not possible, they will be created as Level 7 products by the WALLABY team. Hence, we changed their status from “required” to “desirable” in our requirements document.

> ***Comment:*** *A minor question: Are gain cubes needed separately to sensitivity cubes or are these the same thing?*

Reply: Sensitivity cubes store the variation in rms noise level across the image cube, whereas gain cubes store gain variations across the image cube (relative to a unit point source). Sensitivity cubes will be required to allow noise normalisation prior to source finding. Whether gain cubes are needed depends on the mosaicking approach for ASKAP–12 observations in ASKAPsoft and whether that would lead to gain variations across the cube, in particular near the edge of the cube.

> ***Comment:*** *As above, note that CASDA is not archiving measurement sets for spectral line data.*

Reply: See our response above.

> ***Comment:*** *Also see the general notes on cut-outs and archiving of level 7 data products.*

Reply: See our response above.