Australia's National
Science Agency

ASKAP Pilot Survey Processing

10 August 2020

1 Overview

This document describes how the processing for the ASKAP Pilot Surveys is managed and scheduled, including the interactions with the science teams for defining pipeline configurations.

2 Running the processing

The processing of data is managed through the ASKAP pipeline scripts. These encapsulate well-understood processing workflows that use the ASKAPsoft calibration and imaging package. The pipelines run on the *Galaxy* supercomputer at Pawsey.

It is expected that the pilot survey processing will be the predominant activity on the CPU queue ("workq") on *Galaxy* for the foreseeable future. It will use the high-priority *askaprt* account, that gives its jobs higher priority in the queue than other tasks running in the regular *askap* account.

The processing is run on one scheduling block at a time. Most pilot observations were taken with one field per scheduling block, and the pipeline processing matches this arrangement. Observations for some projects have multiple interleaves for a given field and scheduling-block, and the pipeline will mosaic these interleaves together to form a single image product covering that field.

Observations that do not match this patterns are typically older observations done in rapid mode (to support RACS or VAST) – these are being processed under a specifically-developed setup, building on the rapid-mode of the pipeline and done with bespoke controlling scripts. Most of what is covered in this document refers to other, longer-integration observations.

The archiving of data (which can be triggered through the pipeline) also works primarily on a per-scheduling-block basis, and so it is expected that archiving will occur alongside processing, in an incremental fashion for each project.

3 Scheduling considerations

There is only a small number of people responsible for running the processing, and a limited amount of processing capacity. This limits the number of observations that can be processed at any given time to at most two or three when processed in full (depending on the level of interaction with them).

A guiding principle in determining the scheduling of processing is to ensure all science teams have access to some data. Each team, by now, has access to at least some pilot data for analysis and testing purposes (if not fully reduced and made available through CASDA). We will now endeavour to spread the processing effort evenly while ensuring we maintain a good throughput of data to get through the processing backlog.

There will be times at which we require feedback from the science teams regarding technical details of the processing (particular configuration settings, evaluation of new techniques, etc). It is likely that we will pause the processing for that project at that point and concentrate on other projects to avoid downtime.

The taking of Target-of-Opportunity (ToO) observations may necessitate a break to the planned schedule, to allow prompt processing and release of the ToO data. There are processes governing the allocation of such observations, and the processing implications will be considered as part of that allocation. The priority, however, will be to complete the processing of Pilot data.

4 Managing the schedule

The processing schedule is managed through a Trello board, at https://trello.com/b/PbrlbDRh/askap-science-processing. This board allows each observation or type of observation to be represented by its own card. The list of scheduling blocks associated with that card are provided, and configuration templates and other files can be attached to them.

The aim of this board is to provide a means of tracking which observations & surveys are in which state (planned, in progress, being archived). The intention is to record progress in each card, along with statements about the success or otherwise of the processing jobs.

The Trello board was not intended to be completely public, but we are willing to open it up to relevant science team members to allow monitoring of the progress of the pipeline processing. Please get in touch with Matthew Whiting to get an invitation to the board.

5 Timescales of processing

A key question is how long will the processing take? This varies from project to project, due to the different data sizes and the different processing strategies. The following table shows indicative times (drawn from experience), although note that there will be variability in the times according to processing considerations (such as number of cycles and depth of cleaning,

complexity of images, or even size of images), as well as the influence of the compute environment and state of the filesystems (heavy I/O load can slow down particular jobs).

Project	Typical number of fields/SB	Single beam overall	Single beam continuum imaging & self-calibration	Single beam continuum cube	Single beam spectral-line
EMU	1	10hr	9hr	N/A	N/A
POSSUM	1	11hr	9.25 hr	18 min	N/A
WALLABY	1	15hr	3.3hr	N/A	3.5hr
GASKAP – HI (no spectral imaging)	3	9hr	0.5hr	N/A	N/A
GASKAP – OH	3	24hr	1hr – 2.5hr	N/A	50 min
FLASH	1	24hr	8hr	N/A	4hr
DINGO	2	15hr – 24hr	3hr	N/A	Best: 25 min Can be 2.5hr
VAST	1	1.5hr	1.5hr	N/A	N/A

Table 1: Summary of processing times for different Pilot surveys. We give the times for a single beam, either the total time or just for particular types of imaging. Parallel processing will mean many (ideally all) beams can be processed simultaneously, assuming the *Galaxy* queue has room, so there is only a small additional time required for the full field.

We are able to complete the entire processing for a continuum field in about 15 hours, although this is highly dependent on the state of the system, how busy the queue is and, consequently, how many beams are able to be executed in parallel.

The rapid mode of the pipeline used for VAST (and for RACS) is notable for the very small overheads for the pre-imaging processing (bandpass calibration, flagging and averaging) – this all takes about a minute for a typical beam.

The spectral processing takes longer and is limited primarily by serial jobs for application of calibration and continuum-subtraction. The times will vary according to the distribution of work – there is a trade-off between number of cores and total time taken (the spread in times for DINGO indicates this, as we have used at least two different distribution schemes). Similarly, increasing the bandwidth increases the time almost proportionally, as can be seen with the FLASH processing (there is additional overhead in the I/O, particularly creating the larger files).

Mosaicking is an additional time impost – for the continuum this typically takes an hour or two, depending on the size, while for the spectral cubes it can take several hours at least. Work is underway to speed this up, particularly for the continuum case using multi-threading.

A large additional time burden at the moment is the effort of dealing with job failures. There can be a small number of jobs that fail due to slurm-related errors on the compute nodes – these need to be triaged and restarted, which means redoing the mosaicking (if it has already

been run). Having multiple fields (and, in the case of the GASKAP OH, multiple frequency bands) increases the likelihood of such failures occurring, due to the larger number of individual beams.

6 Interactions with science teams

We would prefer each science team to have a primary point-of-contact with the processing team. That person would be the one to refer questions about processing templates and other pipeline inputs (flagging information, for instance).

Before commencing with processing of a new type of observation, the processing team will clear the intended configuration file with the science team point-of-contact. In many cases, much of the configuration will be re-used from project to project, but there will inevitably be specific parameters that are required to meet particular science cases.

The pipeline configuration files will be kept in a git repository on bitbucket.csiro.au[1]. Access to this will be for the processing team and the designated science-team representatives (who will have write access to the repository). Template configuration files will be kept here, and anything used in the processing will be drawn from here.

Updates to the pipeline/ASKAPsoft functionality will be communicated with the science teams, particularly through the working group meetings, ACES, and the ASKAP forum, and some testing time will be allowed to validate any updates and demonstrate their effectiveness for various science cases.

7 Archiving and data access

The aim of the processing is to produce data products suitable for upload into CASDA. The pipeline scripts provide this upload mechanism, and the aim is to have this built into the natural processing workflow. For initial processing of observations for a survey (or for new types of observations), it is expected there will be some careful analysis of the data products prior to upload to CASDA, but further processing will likely go through to CASDA much more promptly.

The upload to CASDA involves the identification of a pre-defined set of data products and copying them to the CASDA storage. Upon completion of this process, the data will be available for validation. This is meant to be done by specific science team members – the Data Access Portal (DAP – http://data.csiro.au) provides mechanisms to manage the list of validators for a project. The science team point-of-contact will be notified as the data arrives in CASDA and will

[1] https://bitbucket.csiro.au/projects/ASKAPSDP/repos/askap-sst/browse

ensure that the relevant people within the project are aware of the need for validation. During the validation period, the data is accessible only to nominated people within the team. Once the validation is completed, the full set of data is released by observatory personnel and will be made publicly available.

In some cases, the science team validation will produce additional validation documents (plots etc). These can be uploaded to CASDA to be associated with the pipeline data products, and the science team point-of-contact should work with the processing team to facilitate this. This will be done prior to release.

8 Reprocessing and reobservation

There may be situations where the processing provided does not meet the science requirements. It is possible to request it be re-processed, although there are several considerations here.

A key restriction comes from CASDA, where there cannot be more than one version of a given file – filenames need to be unique. Subsequent versions of images can be stored, but must be named distinctly (we will typically do this via a version string, such as 'v2' or similar). The catalogues have more stringent requirements. When a dataset is released (following validation), the catalogue is merged into the global catalogue, and removing it if it needs replacing can be time-consuming. Leaving an unwanted catalogue out of the release is the preferred approach (assuming it has been replaced with a better version), as it will not then get merged with the global catalogue.

Re-observations can be requested[2] to replace observations that are not able to meet the requirements of the Pilot survey. The request should detail what went wrong and why the original observation was not suitable. Best efforts will be made to make up missing pilot observations, although other operational constraints (particularly testing and maintenance, as well as preparation for phase II of the Pilot surveys) may over-ride them.

In summary, it is best to ensure that questionable data products are rejected prior to release. However, this has policy implications in that unreleased data cannot be freely used and should not be incorporated into value-added products due to problems with traceability.

We expect to continuously improve the telescope and the processing software, so an alternative to re-processing is to re-observe. This has the advantage that the data will be indexed under a different scheduling block, removing the issues with having more than one version of the same file in CASDA. In general, re-observations within the same pilot phase will

[2] Observation requests can be submitted at https://tinyurl.com/askaprequest

only be considered if there are clear issues with the data that prevent use for the intended science case.

Commensal use of the data is a factor to be considered, since data may still be useful for a different science outcome. In cases where data exhibit slight issues but are still of science quality, the recommended approach is to release and re-observe in future pilot survey iterations. This maintains a record of improvements over time. Science teams should balance the desire for greater sky coverage with the desire to maximise data quality, given the finite amount of telescope and CPU time available.

9 Low-priority “filler” processing

The processing of the pilot survey observations is the highest priority at the moment. There will be times, however, where the pilot processing is not fully utilising the system. The processing team will attempt to make full use of the system by including lower-priority jobs, that might be test fields or non-pilot observations. Similar arrangements will be made to interact with the relevant science-team contacts, should this be necessary.

The low-priority *askap* account will continue to be available for use by ACES members and other ASKAP team members. This account has been the main way high-volume (but short-duration) jobs for RACS and VAST have been run. It also provides a mechanism for testing by ACES members and others of new processing techniques or parameterisations that can later be factored in to the operational processing.

We will continue to monitor the impact of this account, however, and may consider reducing the extent of the queue to which is has access, to limit its ability to block the queue with running jobs and prevent the pilot processing from going ahead. There will also be a greater need for coordination with the operational processing to manage the smooth operation of the queue.

As Australia's national science agency and innovation catalyst, CSIRO is solving the greatest challenges through innovative science and technology.

CSIRO. Unlocking a better future for everyone.

Contact us

1300 363 400
+61 3 9545 2176
csiroenquiries@csiro.au
csiro.au

For further information

CSIRO Astronomy & Space Science
Matthew Whiting
+61 2 9372 4683
matthew.whiting@csiro.au
csiro.au/cass