CSIRO ASKAP Science Data Archive: Overview, Requirements and Use Cases

ASKAP-SW-0017

Version 1.0 21 May 2014:

Status: Open Release

Authors: Jessica Chapman (CASS), Ben Humphreys (CASS), Matthew Whiting (CASS), Dan Miller (CSIRO IM&T), Ray Norris (CASS)

Keywords: ASKAP, Data, Archives

Enquiries should be addressed to: Jessica.Chapman@csiro.au

Document distributions

VERSION	DATE	AUTHORS	DESCRIPTION OF CHANGE
0.1	01 Jul 2008	Ray Norris	Initial Version
0.2	19 Sep 2008	Ray Norris	Updated draft version
0.5	19 March 2013	Jessica Chapman Ben Humphreys Matthew Whiting Dan Miller Ray Norris	Substantial revision to the original document. Limited distribution of draft document to participants of a data meeting held in March 2013.
0.8	8 Nov 2013	Jessica Chapman Ben Humphreys Matthew Whiting Dan Miller Ray Norris	Public release of draft version to seek input from the science community.
1.0	21 May 2014	Jessica Chapman Ben Humphreys Matthew Whiting Dan Miller Ray Norris	Public release. Document updated following consultation with the science community, CASDA Science Reference Group and many other individuals.

Contents

1. Introduction 5
1.1 Summary 5
1.2 Scope 5
1.3 Acknowledgments 6
Abbreviations 7
2. ASKAP Overview 9
2.1 ASKAP specification 9
2.2 Locations 10
2.3 ASKAP timeline 10
2.4 Telescope Operating System 12
2.5 Central Processor 13
2.5.1 Data conditioning and calibration 13
2.5.2 Imaging pipelines 14
2.5.3 Source detections 16
2.5.4 Simultaneous pipelines 18
2.6 Data levels 18
2.6.1 Data Validation 19
3. ASKAP Operations and science 20
3.1 ASKAP science observations 20
3.2 Early Science 21
3.3 Survey Science Projects 22
3.3.1 EMU 23
3.3.2 POSSUM 24
3.3.3 WALLABY 25
3.3.4 DINGO 25
3.3.5 FLASH 26
3.3.6 GASKAP 27
3.3.7 VAST 27
3.3.8 COAST 28
3.3.9 CRAFT 29
3.3.10 VLBI 30
3.4 Guest Science Projects 31
3.5 Target of Opportunity observations 31
4. The science archive 31
4.1 Overview 31
4.2 Pawsey Centre Infrastructure 33
4.3 Primary data products 35
4.4 Virtual Observatory protocols 36
4.5 Data volumes 36
5. Archive search and access 39
5.1 Archive searches 39
5.1.1 Automated queries 39
5.1.2 Graphical interface 39
5.2 Data access 40
5.2.1 Data transfer rates 40
5.3 Applying for supercomputing facilities 42
5.3.1 Pawsey Centre facilities 42

5.3.2 Other supercomputing facilities 43
6. Requirements and use cases 44
6.1 High-level requirements 44
6.2 Survey Science Projects use cases 46
6.3 Use cases for science users 63
6.4 Other use cases 65
Appendices 67
Appendix A: Data volumes 67
Appendix B: CASDA data products 69
Appendix C: Survey parameters 72
References 83

1. INTRODUCTION

1.1 Summary

The CSIRO ASKAP Science Data Archive will provide the long term storage for ASKAP data products and the hardware and software facilities that enable astronomers to make use of these.

ASKAP is, in many ways, a data driven facility where the data rates are extremely high. The ASKAP data rates arriving at the Pawsey Centre are approximately 2.5 Gbytes per second, equivalent to 75 Petabytes (PB) per year. This is beyond the current ability to archive data and so raw visibility data and calibrated spectral line visibility data will not be archived. Such high data rates require instead that ASKAP data processing is carried out in quasi real time using automated pipelines to produce data products and associated metadata that are stored and made available through the science archive. The archive can be thought of as the end stage of the full system.

The CSIRO ASKAP Science Data Archive (hereafter CASDA) will include calibrated visibilities for continuum data, and image cubes for both spectral line and continuum data. Source detection algorithms will be used to search image cubes for radio sources and source-related information will be captured in catalogues. Calibration and scheduling information related to the observations will also be stored. The total volume of archive data is expected to reach 5 PB per year.

1.2 Scope

This document discusses the user requirements and use cases for CASDA as needed to support scientific observations with the ASKAP array located at the Murchison Radio Observatory (MRO). CASDA will provide the archive support from the start of Early Science onwards. Early Science will begin following the installation, commissioning and verification of the first 12 MkII phased array feeds (PAFs) on the antennas.

The document is written for a broad audience that includes ASKAP Survey Science Teams, the general astronomy community and groups from CASS, CSIRO IM&T, ICRAR and iVEC who are working on the radio astronomy archives at the Pawsey Centre. In particular, it is intended to provide the high level requirements and use cases to the CASDA development team as input for the more detailed design and architecture specifications, and is intended as a reference source for the Science Survey teams and general astronomy community to facilitate discussions towards verifying user requirements and use cases.

Some readers may not be familiar with ASKAP specifications, or with radio astronomy techniques. To help provide context, sections 2 and 3 provide an overview of the ASKAP system and operations. The science archive, requirements and use cases are discussed in sections 4 and 5.

In addition to CASDA, a separate commissioning archive will be used for the data collected from BETA – the initial array of six ASKAP antennas equipped with MkI PAFs. This archive will store and provide access to commissioning data as MkII PAFs are installed and tested on

the antennas, and will include data from science demonstrations. This archive is the responsibility of the CASS Science Data Processing group. The requirements of this commissioning archive are NOT discussed further in this document.

This document provides only minimal information on the user support for CASDA. Further information will be provided through the ATNF website (www.atnf.csiro.au).

1.3 Acknowledgments

This document draws on previous ASKAP documents. In particular it builds on and replaces the earlier document *ASKAP Science Data Archive: Draft Requirements Document* (2009, Norris and Johnston [6]) and has made extensive use of *ASKAP Science Processing* (2011, Cornwell et al. [2]).

Many individuals have contributed input towards this document. We thank in particular members of the CASDA Science Reference Group for their substantial contributions.

Abbreviations

Acronym	Definition
ACES	ASKAP Commissioning and Early Science
ADQL	Astronomical Data Query Language
API	Application Programming Interface
ASKAP	Australian SKA Pathfinder
ATOA	Australia Telescope Online Archive
ATNF	Australia Telescope National Facility
BETA	Boolardy Engineering Test Array
CASA	Common Astronomy Software Applications
CASS	CSIRO Astronomy and Space Science
CASDA	CSIRO ASKAP Science Data Archive
CDS	Collections Development Storage
CPU	Central Processing Unit
COAST	Compact Objects with ASKAP: Surveys and Timing
CRAFT	Commensal Real-time ASKAP Fast Transients
DAP	Data Access Portal
DINGO	Deep Investigations of Neutral Gas Origins
DRAO	Dominion Radio Astrophysical Observatory
EMU	Evolutionary Map of the Universe
EVACAT	EMU Value Added Catalogue
FDF	Faraday Dispersion Function
FITS	Flexible Image Transport System
FLASH	The First Large Absorption Survey in HI
FLOPS	Floating Point Operations per Second
FWHM	Full width at half maximum
GAMA	Galaxy and Mass Assembly [survey]
GASKAP	The Galactic ASKAP Spectral Line Survey
Gb	Gigabit (10^9 bits)
Gbps	Gigabits per second
GB	Gigabyte (10^9 bytes)
GBps	Gigabytes per second
GSP	Guest Science Project
HPC	High Performance Computing
HSM	Hierarchical Storage Management System
IB	Infini-Band
ICRAR	International Centre for Radio Astronomy Research

IM&T	Information Management and Technology
iVEC	iVEC is a joint venture between CSIRO, Curtin University, Edith Cowan University, Murdoch University and the University of Western Australia
IVOA	International Virtual Observatory Alliance
LBA	Long Baseline Array
MAID	Massive Array of Idle Disks
MB	Megabyte (10^6 bytes)
MRO	Murchison Radio Observatory
MWA	Murchison Widefield Array
NCI	National Computing Infrastructure
NCMAS	National Computational Merit Allocation Scheme
NED	NASA/IPAC Extragalactic Database
OPAL	Online Proposal Applications and Links
PAF	Phased Array Feed
PB	Petabyte (10^{15} bytes)
POSSUM	Polarization Sky Survey of the Universe's Magnetism
PPC	POSSUM Polarisation Catalogue
PSF	Point Spread Function
PSRFITS	Pulsar FITS (data format)
PVACAT	POSSUM Value Added Catalogue
RDS	Research Data Services
RDSI	Research Data Storage Infrastructure
RFI	Radio Frequency Interference
RTC	Real Time Computer
SIAP	Simple Image Access Protocol
SIMBAD	Set of Identifications, Measurements, and Bibliography for Astronomical Data
SKA	Square Kilometre Array
SRG	Science Reference Group
SOC	Science Operations Centre
SSP	Survey Science Project
SST	Survey Science Team
STILTS	Starlink Tables Infrastructure Library Tool Set
TAP	Table Access Protocol
TB	Terabyte (10^{12} bytes)
VAST	Variables and Slow Transients
VLBI	Very Long Baseline Interferometry
VO	Virtual Observatory
WALLABY	Widefield ASKAP L-Band Legacy All-Sky Blind Survey

2. ASKAP OVERVIEW

2.1 ASKAP specification

This section gives an overview of the ASKAP system. This is largely extracted from previous ASKAP documents [2, 4, 5].

ASKAP is an array of 36 12-m diameter prime-focus parabolic dish antennas located at the Murchison Radio Observatory in Western Australia. The array is designed to be a fast survey instrument for centimetre-wavelength observations with high dynamic range and a wide field-of-view.

The ASKAP system specification is given in Table 1.

Table 1: ASKAP specification

Number of antennas	36	**Notes**
Dish diameter	12 m	Corresponds to a full-width half maximum primary beam of approximately one degree.
Maximum baseline	6 km	30 antennas are located within a region of 2 km in diameter. The remaining 6 extend the baselines to a maximum of 6 km.
Frequency range	700 – 1,800 MHz	Equivalent to approximately 42 cm (700 MHz) to 17 cm (1,800 MHz)
Field-of-view (area)	30 square degrees	
Processed bandwidth	300 MHz	
Number of channels	16,200	18.5 kHz per channel, plus zoom modes ~ 1 kHz per channel
Correlator integration time	5 s	Minimum time per visibility sample
Number of Phased Array Feed elements	188	The number of elements for Mk II PAFs
Digitisation levels	14 bits	
Dynamic range	50 dB	
Sensitivity (Ae/Tsys)	65 m^2 K^{-1}	
Survey speed	1.3 x 10^5 m^4 K^{-2} deg^2	

2.2 Locations

Physical locations for ASKAP sub-systems are:

- The antennas, beamformers and the correlator are located at the Murchison Radio Observatory (MRO).
- Operational engineering support is provided by CSIRO Astronomy and Space Science (CASS) staff located in Geraldton with some additional support provided from technical staff in Marsfield, Sydney.
- Data are transmitted over high-speed dedicated links to the Pawsey Centre in Perth.
- The Central Processor used for real-time data processing is located at the Pawsey Centre. The platform within the Pawsey Centre which hosts the Central Processor is known as the Real Time Computer.
- ASKAP visibility and image data files are archived by CASDA at the Pawsey Centre.
- CASDA user interfaces are provided through the CSIRO Data Access Portal in Canberra.
- CASDA backend databases, catalogue data products and indexes are stored in both Perth and Canberra.
- ASKAP observations will normally be carried out and monitored by CASS staff located at the Science Operations Centre in Marsfield, Sydney.
- First-level user support for the archive will be provided by CASS staff in Perth and/or Sydney.
- ASKAP will also provide data used for education and outreach programmes. The coordination of these programmes will be from the CASS Headquarters, in Sydney.

2.3 ASKAP timeline

As at May 2014:

- The site infrastructure including roads, a RFI-shielded control building, waste, water, initial power and fibre links are complete.
- The installation of the 36 ASKAP antennas is complete.
- MkI Phased Array Feeds (PAFs) are installed on the BETA array.
- BETA is now being used for commissioning purposes. The first BETA spectral line image using six antennas and 15 baselines was obtained in April 2014.
- System tests of the first full MkII PAF are in progress.
- Installation, commissioning and science verification of the MkII PAFs will continue throughout 2014.
- Early Science with ASKAP will begin following the commissioning and science verification of the first 12 MkII PAFs.

- Further MkII PAFs will be added to the array during 2015.
- Fibre links between the MRO and the Pawsey Centre provide data rates of 40 Gb/s.
- The Pawsey Centre building was completed in April 2013 and installation of a Cray supercomputer and storage facilities began soon after. Installation and acceptance tests are completed.
- The CASDA project began in August 2013. The first stage of the project to carry out a detailed analysis of user requirements, system design and architecture is essentially complete. A Preliminary Design Review was completed in April 2014.
- CASDA construction will begin in June 2014 as a partnership between CSIRO Astronomy and Space Science, CSIRO Information Management and Technology and iVEC.
- It is intended that CASDA will be available from the start of Early Science around mid-2015.

2.4 Telescope Operating System

Figure 1 summarises the ASKAP data flow.

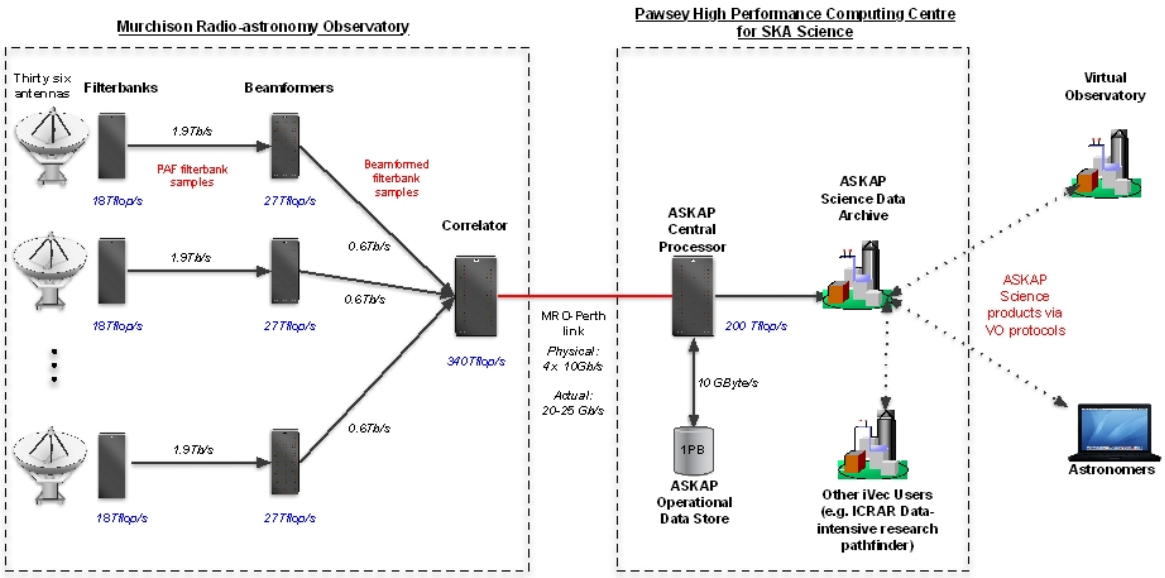

Figure 1: ASKAP data flow (adapted from [1])

The ASKAP computing architecture has three major components: the Telescope Operating System, the Central Processor and CASDA. The Telescope Operating System is responsible for the control and monitoring of the antennas. This includes the antennas, beamformers and correlator.

The ASKAP large field of view is achieved using phased array feeds with 188 detection elements at the focus of the antennas. For each antenna the voltages measured by these elements are amplified, digitised and filtered into 304 coarse channels of 1 MHz each.

The beamformer for an antenna constructs beams by summing and weighting the signals from the individual elements. ASKAP will be configured to give a total of 36 observing beams.

The samples for each beam are further filtered to high resolution. Each 1 MHz channel is split into 54 fine channels, giving 16,416 channels in total. Edge channels are later discarded and a total bandwidth of 300 MHz and 16,200 channels are used given a spectral resolution of 18.5 kHz per channel. For some spectral line projects, zoom modes providing a spectral resolution of ~ 1kHz per channel across a narrower frequency range will be used.

The signals from one antenna beam are correlated with the signals from the corresponding beams from the other antennas. In effect this allows ASKAP to operate in a way that is equivalent to a number of conventional radio arrays operating simultaneously. The correlator forms the cross-products between each pair of antennas. ASKAP antennas have two linear polarisation axes allowing four polarisation products (called XX, YY, XY and YX). For each integration period of 5 seconds, one cross-correlation (also called a 'visibility') is output from the correlator for each beam, *baseline*, channel and polarisation. The correlator also outputs one auto-correlation for each beam, *antenna*, channel and polarisation.

For 36 beams, 630 baselines, 36 auto-correlations, four polarisation products and 16,416 channels the correlator produces 1.6 billion distinct correlations and a total data volume, every 5 seconds, of 12.6 Gigabytes (GB). Thus the maximum data rate from the correlator, for a full array of 36 antennas is 2.5 Gigabytes per second (GBps). For a smaller number of antennas the data rate scales as the number of baselines. During the data processing stages the data volumes are reduced.

The correlation samples are sent over high speed links to the Pawsey Centre at the maximum data rate of 2.5 GBps. Four 10 Gigabits per second (Gbps) links are available for ASKAP.

2.5 Central Processor

The Central Processor is a hardware and software subsystem that is responsible for all of the stages of data processing from the correlator to the production of science data products such as image cubes and source catalogues. The processor as a system can be thought of as a sophisticated ‘backend’ to the array.

The processor includes a Cray supercomputer with 9,440 Central Processing Unit (CPU) cores, a total memory of 32 TB and a total compute power of 200 TFLOPS. This is supported by a 1.4 PB Lustre disk-based file system that is used to buffer the visibility data during data processing and to temporarily store the data products produced prior to sending these to the archive.

2.5.1 Data conditioning and calibration

Data processing is carried out using a set of pipelines. A schematic of the data conditioner pipeline (also known as the ingest pipeline) is shown in Figure 2.

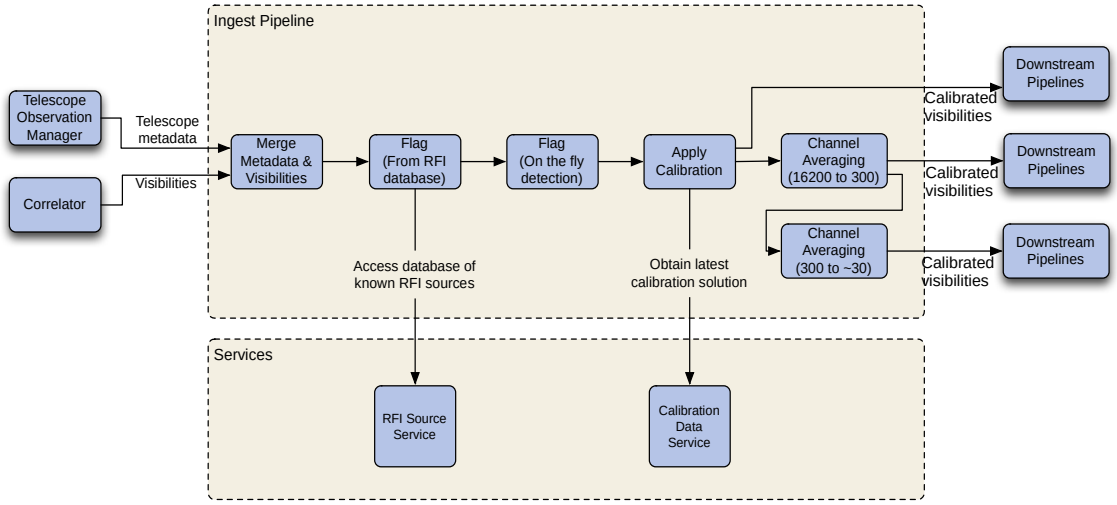

Figure 2: Data Conditioner Pipeline

Data arriving from the correlator are acquired through a set of 16 ingest nodes and merged with telescope-related metadata provided by the Telescope Operating System. The data are then

'conditioned' prior to being forwarded to the science processing pipelines. Conditioning steps include flagging the data for known sources of radio frequency interference (RFI). This is done using a database of known interference sources as well as the dynamic detection of new interference sources. Other bad data are also flagged.

After conditioning, the visibilities are calibrated to correct for atmospheric and instrumental visibility variations and for the instrumental bandpasses. The calibration of ASKAP data with many beams requires a novel approach to data calibration. The full ASKAP array will use a self calibration technique where a pre-determined global model of the sky, based on information derived from known bright sources, is used to correct the observed visibilities. This model will be updated and improved as ASKAP observations progress [2]. During commissioning and Early Science where a smaller number of antennas are used, alternative calibration methods may be applied. After calibration the data are averaged as needed and the calibrated visibilities are sent to imaging pipelines.

2.5.2 Imaging pipelines

A schematic diagram for the data processing pipelines is shown in Figure 3.

The imaging pipelines grid the visibility data and Fourier transform these to the 'image plane'. A single radio astronomy image is a map of the sky brightness across an observed region of sky (also known as a 'field'). An image cube is a set of images contained within a single file that covers a range of frequencies and is represented by three dimensions. For a standard image cube, the x and y-axes correspond to the plane of the sky whilst the third axis corresponds to the channel number or frequency.

For an ASKAP antenna the FWHM primary beam at a wavelength of 20 cm is approximately one square degree. To cover the field-of-view of 30 square degrees, the data from the 36 beams are 'mosaiced' together to produce a single image. To correct for edge effects some overlapping of adjacent beams is used.

The ASKAP specifications include three different imaging pipelines.These will be used for continuum observations, spectral line observations, and transient observations. For the purposes of this document the letters C, S and T are used to label the three types.

C: Continuum Imager

For continuum data processing the visibilities are averaged into 1 MHz bins. This reduces the total number of channels from 16,200 to 300 and thus substantially reduces the data processing load. Further averaging may be applied. Continuum imaging will generally use one of two modes:

All 300 channels are retained and the data products formed are continuum image cubes. Image cubes may be retained for all four polarisation products known as Stokes I (total intensity), Q (linear polarisation), U (linear polarisation), and V (circular polarisation).

For some continuum surveys, a 'multi-frequency synthesis' technique is used where the full set of frequency information is used to produce images for three 'Taylor terms'. These correspond to the source flux density at a given frequency, the spectral index and the spectral curvature.

(For a radio continuum source, the spectral index and curvature characterise how the flux density of a source varies with frequency).

For polarisation surveys , full-Stokes spectra are extracted at the positions of known sources, and a technique known as ‘rotation measure synthesis’ is used to obtain a value of the rotation measure of each source.

S: Spectral Line Imager

For spectral line imaging the visibility data from 16,200 spectral channels are processed to generate image cubes. Spectral line image processing normally includes removal of any radio continuum emission. However, where needed, continuum images or image cubes will also be retained so that the continuum levels can be measured.

Data volumes for image cubes are large. As an example, the data volume for an image cube with 3,600 x 3,600 pixels in the x- and y-directions and 16,200 spectral channels is 840 GB. Due to the high data volumes, calibrated visibility data for spectral line observations will not be archived. Spectral line processing will normally only be carried out for the Stokes I polarisation product. For some projects ‘cut-out’ image cubes will be stored in addition to the full cubes. Cut-out cubes correspond to smaller parts of a cube, centred on pre-determined positions.

Spectral line image cubes can be used to generate two-dimensional images known as ‘moment maps’. Moment maps are a way of summarising the information contained in a three-dimensional cube into a single image. Standard moment maps used are for integrated intensity (M0), velocity field (M1) and velocity dispersion (M2).

Limitations in computing power and memory impose some restrictions in spectral line data processing. For full-sized data cubes covering 30 square degrees, image processing is expected to be restricted to baselines shorter than 2 km, corresponding to an angular resolution of 30 arcsec at 1.4 GHz. For some science projects higher resolution ‘postage stamp’ image cubes may be produced for smaller regions around known sources. The imaging pipeline allows *either* full-size data cubes (angular resolution ~30 arcsec at 1.4 GHz) *or* postage stamp cubes (angular resolution ~10 arcsec at 1.4 GHz) but not both simultaneously. There are also some limitations on the total number of postage stamp cubes that can be produced in the pipeline [2].

During commissioning and Early Science experience will be gained to better determine the capabilities and limitations for spectral line processing.

T: Transient Imager

The transient imaging pipeline will produce one image cube every 5 seconds. This allows for searches of bright sources that vary over time or may be detected as a single ‘burst’ of emission. Information on bright sources derived from the transient data processing may be used to update the Global Sky Model.

The compute requirements for such rapid imaging are very high. To enable fast processing, the visibility data from transient observations are averaged over bins of ~ 10 MHz corresponding to 30 spectral channels.

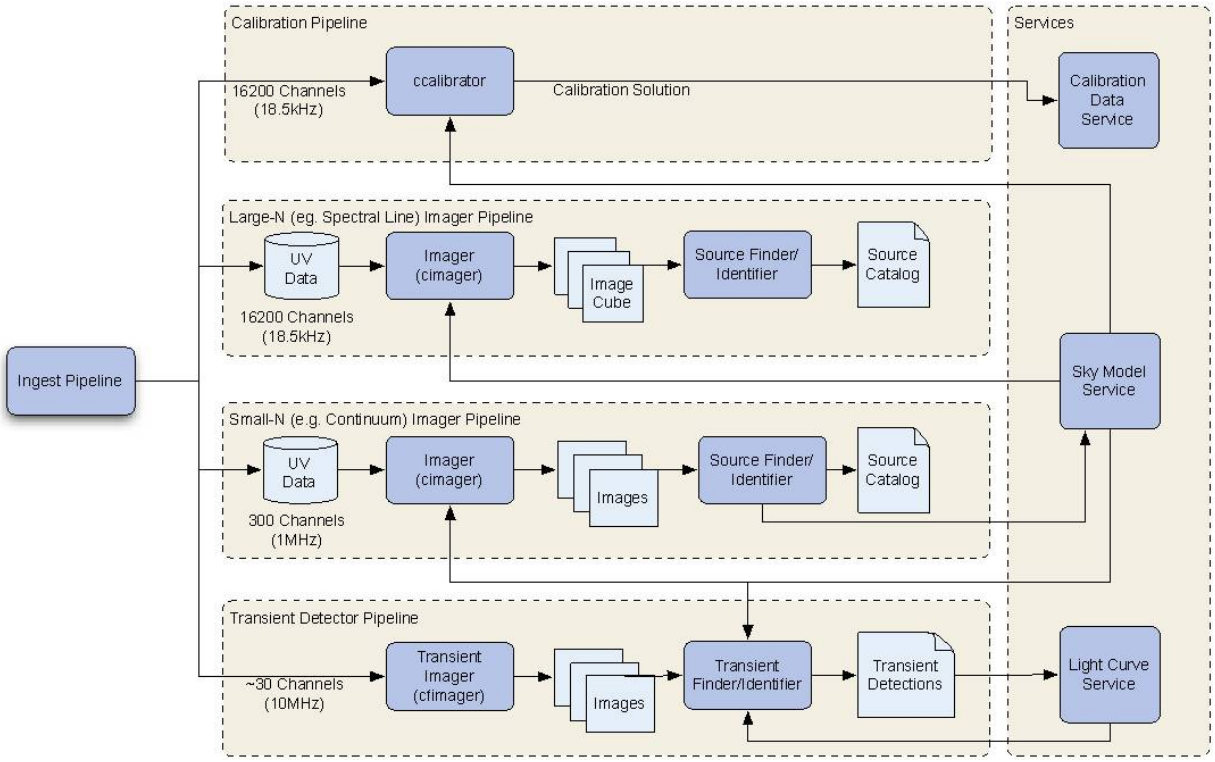

Figure 3: ASKAP data processing pipelines

2.5.3 Source detections

The data processing pipelines include automated searches for sources (or components of sources). The ASKAP source finder software builds on the package Duchamp [7, 8] and can be used to search for sources in both single-channel images and multi-channel image cubes. Groups of pixels or voxels (three-dimensional pixels) that lie above a specified flux or signal-to-noise threshold are identified, possibly following some pre-processing (through smoothing or multi-resolution reconstruction) to enhance the signal-to-noise of real sources. Parameters characterising the source detections, such as their position on the sky, size, position angle, strength and frequency are written into source catalogues, in effect with one source detection per catalogue row. (For CASDA, a catalogue is conceptually equivalent to a two-dimensional table where each row contains a set of attributes for an object. For example, for a source detection catalogue each row will include the right ascension, declination, size, measured brightness and other attributes for one source).

For transient observations, each image cube is searched and the results written into catalogues with a cadence of 5 s. The image cubes themselves are not retained. The transient catalogues allow for the construction of catalogues containing the time-dependent information needed to generate source light curves and to allow subsequent sampling or smoothing over longer time intervals. This capability will enable studies of sources that vary on timescales longer than 5 seconds.

Table 2 provides some examples to illustrate data volumes for data products produced by the science data processing pipelines. For detailed survey parameters and for a full list of data product types see Appendix B.

Table 2: Example data sizes

Survey Type	Product	Parameters used	Output size	Notes
C	Full polarisation continuum calibrated visibility data set	36 beams 300 channels 4 polarisations 666 baselines (includes auto-correlations) Time per sample 5s 12 hours integration	2.24 TB	Data volume calculated as 9 Bytes per sample = 8 Bytes per visibility + 1 Byte for weighting.
S	One spectral line image cube	3,600 x 3,600 pixels 16,200 channels Stokes I polarisation	839 GB	A image cubes of 3,600 x 3,600 pixels and pixel size of 7.5 arcsec corresponds to an array configuration with baselines below 2 km.
S	2000 postage stamp image cubes	256 x 256 pixels 512 channels Stokes I polarisation	0.27 TB	
C	Set of 11 continuum images generated using 'Taylor-term' images	10,800 x10,800 pixels 1 channel	5.2 GB	0.47 GB per image. Data are averaged to a single frequency channel. 11 images per field produced for multi-frequency synthesis.
C	Set of 4 polarisation continuum image cubes	10,800 x 10,800 pixels 300 channels 4 polarisations	560 GB	139 GB per polarisation
S	Source detections catalogue generated from one 12 hour spectral line	500 detections 300 Bytes per row	150 KB	Estimate only

	image cube			
T	Bright source detections from one 5s image cube	1000 detections 300 Bytes per row.	300 KB	Estimate only

2.5.4 Simultaneous pipelines

ASKAP has been designed so that the imaging pipelines can run concurrently. Data arriving from the correlator can be simultaneously passed through the three types of imager to produce spectral line, continuum and transient results. This provides a very powerful data processing capability.

Different science projects may make use of the same data stream from the MRO correlator. For example a spectral line survey of neutral hydrogen from galaxies, a continuum survey and transient observations for the observed regions of sky could all use the same data sets. For this scenario the transient imager runs constantly producing image cubes and bright source detections every five seconds. The continuum imager and spectral line imager start up following the end of a scheduled block of observations, with continuum data processed prior to spectral line data.

It is intended that ASKAP will be used to observe multiple programs wherever possible. In practice this may be complicated by other considerations such as the different regions of sky required by different surveys and/or different sensitivity requirements.

2.6 Data levels

Figure 4 shows the data flow and data processing stages for ASKAP as a set of increasingly higher levels. As discussed by Cornwell et al. [2], levels 5 and 6 represent the primary data products that are stored in CASDA. CASS is responsible for the generation of all data products up to and including level 5. For major surveys, the survey science teams will be responsible for validating the science data products prior to release for general use. Validated data products are classified as level 6. Note that data products themselves are *unchanged* between levels 5 and 6.

The science teams and/or astronomers from the general astronomy community may develop ‘enhanced’ data products and these are classified as level 7. The tools and processes for doing this are their responsibility. Examples of enhanced products are a final catalogue for a major survey, or a set of image cubes that have been processed by stacking together a larger set of cubes. Science teams should endeavour to carry out their data processing in a timely manner.

CASS staff will take responsibility for ensuring that the data are not released to users until they have been quality approved by the relevant science team.

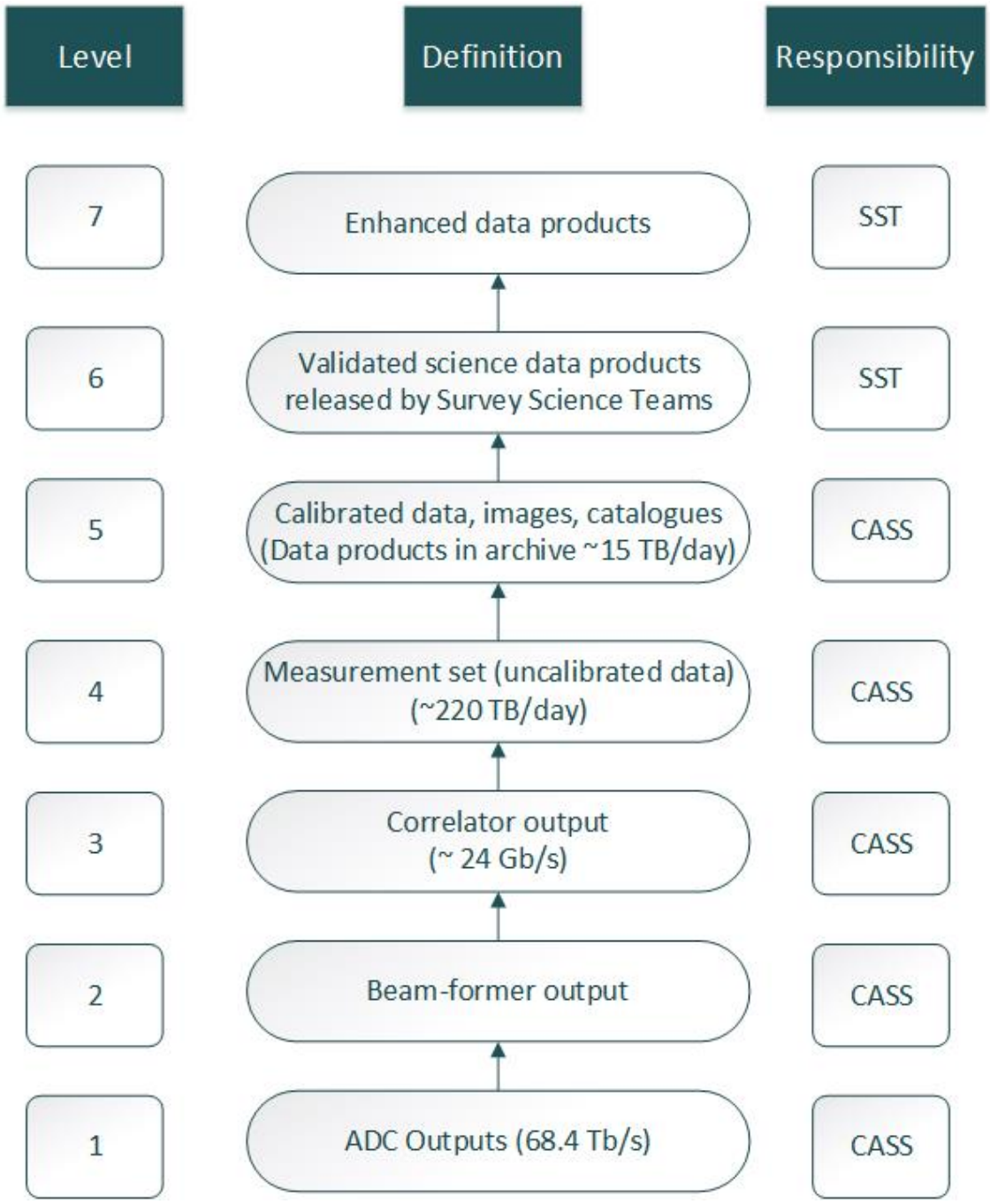

Figure 4: ASKAP data processing levels

2.6.1 Data Validation

CASS will retain the ultimate responsibility for the quality of all ASKAP data products but will delegate responsibility to the Survey Science Teams for validating the data quality for the large-scale science surveys.

The purpose of data validation is to determine whether the data products are 'science ready' to a state where they can meaningfully be used for scientific research. In Early Science the assessment of data quality will be carried out by an ASKAP Commissioning and Early Science (ACES) team (section 3.2). For full ASKAP operations, it will be the responsibility of the Survey Science Project (SSP) teams (section 3) to determine the specific validation criteria for their own projects and to carry out any data analysis required for validation. To reduce the effort involved, data validation should be automated as much as possible and should make use of reports generated in the science data pipelines to provide information on data quality and system performance. Such reports will be made available through CASDA. In some cases it may be necessary for science teams to retrieve visibility and or image data files from the archive for validation purposes. Where files are not archived special consideration for data access may need to be considered.

A CASDA tool will be used so that data validation metadata flags are set in the science data archive. Following assessment of the data quality the science team will enter information to assign values to data quality flags. Science teams will also provide information on specific problems encountered so that this can be shared with other users. After data quality flags are set CASS staff will 'publish' the data products for general use.

In general, observations with bad data will be repeated. However, bad data sets will not be removed from the archive as these could potentially be useful for engineering tests or other purposes.

Some administration and operations staff will also be able to set or potentially override the data validation flags. Changes to data quality flags will be tracked.

3. ASKAP OPERATIONS AND SCIENCE

3.1 ASKAP science observations

This section provides an overview of ASKAP observing and operations. For additional information see documents [1, 3, 6, 7, 8].

ASKAP will be operated by CSIRO Astronomy and Space Science (CASS) as part of the Australia Telescope National Facility (ATNF). The ATNF also includes the Australia Telescope Compact Array, the Parkes radio telescope, and the Mopra radio telescope. These facilities are used together for Very Long Baseline Interferometry (VLBI) observations with the Long Baseline Array. All data taken on ATNF facilities belong to CSIRO.

Due to the remoteness of the MRO, ASKAP science observations will be taken in a remote-observing mode, normally from the Science Operations Centre in Marsfield, Sydney. The control and monitoring of the antennas will be carried out by CASS Science Operations staff using facilities provided by the Telescope Operating System. The science teams will not be present for the observations. Instead they will interact with the data products and information provided in CASDA. User support for CASDA will be provided by CASS staff.

The scientific use of ASKAP will be open to astronomers from around the world, with telescope time allocated on the basis of scientific merit and technical feasibility. ASKAP

science users will include science teams who submit proposals and are allocated time for their projects, and the international general astronomical community who make use of ASKAP results through the science data archive but are not directly included on the project teams.

As a rough estimate, the number of users of ASKAP data is expected to be at least 1500 individuals. This includes approximately 350 individuals on the Survey Science Projects (section 3.3), 400 individuals on Guest Science Projects (section 3.4) and 750 individuals from the general astronomical community. The science users of ASKAP include about 30 research scientists working for CASS who participate as members of the science teams.

3.2 Early Science

CASDA will provide archive support to ASKAP science observations taken from the start of Early Science.

ASKAP Early Science is an observing program aimed at producing scientifically useful data. It will commence when an array of twelve ASKAP antennas fitted with CSIRO's Mk II phased array feeds has been commissioned and scientifically verified. This is known as ASKAP-12.

During Early Science, observations will be classified as 'shared risk' and the time available at the array will be shared between commissioning activities and science use. The observations will be carried out by the ACES team on behalf of the community. Data processing and data validation will be carried out by the ACES team working together with the Science Data Processing team. Members of the astronomy community are invited to join ACES to help with commissioning, science data processing and data validation activities.

ASKAP Early Science data are non-proprietary. Data will be publicly released when they are deemed to be of appropriate quality by the ASKAP Project Scientist.

The priorities for ASKAP early science are:

1. Demonstrating the unique capabilities of ASKAP

2. Providing data sets to the astronomy community to facilitate the development of analysis and interpretation techniques

3. Providing a mechanism for feedback to CASS on the performance and characteristics of the system and opportunities for improvement

4. Achieving high scientific impact

Early Science programs will begin with short 'pilot' observations. Science teams will have access to the data products from the pilot studies and will present results at a second ASKAP Early Science community workshop. Following this workshop and taking account of scientific advice, CASS will decide which of the pilot observations warrant the initiation of a full Early Science survey with ASKAP-12.

Following extensive consultation with the user community two science programs are considered to be of high potential for Early Science. These are:

1 MHz and 18.5 kHz survey in full Stokes, from 700-1800 MHz over a wide area of sky with 6-12 hours integration time per field

This survey will provide a unique broadband data set, significantly increasing the number of radio sources known and constraining the evolution of radio-loud active galactic nuclei.

It will also allow the study of magnetic fields, density and turbulence at a range of redshifts, with a high spectral resolution component to probe the environments of HI absorbing systems at a range on intermediate redshifts that has not yet been studied.

An 18.5 kHz spectral line survey, over 1150-1450 MHz and targeted toward a small number of fields, with 50-60 hours integration time per field

This data set will enable the study of galaxy evolution as a function of environment as well as the morphology of (and interactions between) HI clouds and filaments between nearby galaxies.

In addition to the above two programs, the feasibility of three further science programs is being investigated:

- A deep observation over two to four fields at 1000 – 1300 MHz
- Searches for variables and slow transients using multiple observations of the same fields
- Fast transient capture (not currently supported but firmware effort is being investigated).

For further information on the Early Science Program contact the ASKAP Project Scientist (lisa.harvey-smith@csiro.au).

3.3 Survey Science Projects

For the first five years of routine science operations with ASKAP, it is envisaged that at least 75 per cent of time will be allocated to Survey Science Projects (SSPs). These are defined as projects that require more than 1,500 hours of observing time.

Typically, observations for a SSP will be carried out over extended periods of some months with the same instrumental set up and data processing pipelines used from day-to-day. The data products from the SSPs will be released after data validation without any proprietary period. The science teams are responsible for checking and validating the primary data before they are released to the user community, and for working together with CASS to ensure that their science goals are achievable and met.

In September 2009, ten Survey Science Projects representing 363 investigators from 131 institutions in Australia and overseas were selected by an international panel. Ten ASKAP Survey Science Projects were approved:

- AS014: Evolutionary Map of the Universe (EMU)
- AS016: Widefield ASKAP L-Band Legacy All-Sky Blind Survey (WALLABY)
- AS002: The First Large Absorption Survey in HI (FLASH)
- AS004: An ASKAP Survey for Variables and Slow Transients (VAST)
- AS005: The Galactic ASKAP Spectral Line Survey (GASKAP)
- AS007: Polarization Sky Survey of the Universe's Magnetism (POSSUM)
- AS008: The Commensal Real-time ASKAP Fast Transients survey (CRAFT)
- AS012: Deep Investigations of Neutral Gas Origins (DINGO)
- AS015: Compact Objects with ASKAP: Surveys and Timing (COAST)
- AS003: The High Resolution Components of ASKAP: Meeting the Long Baseline Specifications for the SKA (VLBI)

Of the ten projects: EMU and WALLABY were assigned the highest ranking and will receive full CASS support. Six projects (DINGO, FLASH, GASKAP, POSSUM, VAST and CRAFT) were highly ranked. CASS will make all reasonable efforts to support these projects. Two projects (COAST and VLBI) were designated as strategic priorities. CASS will work to ensure that the capabilities defined by these are enabled to the extent possible.

The following notes briefly describe some of the science goals of the Survey Science Projects and some of the technical challenges associated with these projects. Further information on the Survey Science Projects is given in sections 4 and 5 and in Appendix C.

3.3.1 EMU

EMU is a deep radio continuum survey that will cover the southern sky, extending up to declination of +30 degrees. The total survey area of about 31,000 square degrees will require over 10,000 hours of telescope time and, with a full array of 36 antennas, will detect approximately 70 million galaxies. This will be by far the most extended sensitive survey of radio sources available.

The EMU science data processing will produce catalogues of source detections. These detections will form the basis for a range of science goals that include studies of the evolution of star forming galaxies and galaxies with active nuclei (AGN), and exploring the large-scale structure of the Universe at radio wavelengths.

EMU observations will also cover our own Galaxy and will provide a sensitive wide-field atlas showing the distribution of thermal and non-thermal radio continuum sources in the Galaxy.

The EMU data products produced in the data processing pipelines will include Stokes I images and image cubes and a source detection catalogue with information for around 70 million source detections.

In addition, the EMU science team will produce a value-added set of source catalogues (known as EVACAT) that include associations with catalogues of sources with results from major surveys at other wavelengths. These cross-identifications are critical in associating the detected sources with known objects and with identifying new types of sources. It is expected that several versions of the source catalogues will be produced over time.

Copies of the EMU source detection catalogues are likely to be made available through the NASA Extragalactic Database (NED) service as well as directly from CASDA. External data centres such as NED will be able to access CASDA catalogues or parts of catalogues for re-use in other systems using the access tools provided for all users.

3.3.2 POSSUM

The POSSUM project will study large-scale astrophysical magnetic fields. Magnetic fields are associated with many fundamental astrophysical processes. For example, magnetic fields accelerate cosmic rays, moderate the collapse of gas clouds into stars, provide pressure support to galactic disks, and are a fossil record of large-scale structure formation at earlier times. Such processes take place across many different scale sizes in our Galaxy as well as in other galaxies and the inter-galactic medium.

POSSUM aims to improve our understanding of magnetic fields in the Universe by studying broadband Faraday rotation and polarisation from a large sample of extragalactic radio sources. The POSSUM data products will enable studies of magnetic field studies of our Galaxy, other galaxies and galaxy clusters, and will provide a census of magnetic fields as a function of redshift, or distance in the Universe.

The POSSUM observing strategy for ASKAP is complementary to EMU. Observations for both projects will cover the same regions of sky and it is likely that these two projects will be carried out commensally. In effect, the continuum pipeline data processor will process a single stream of visibility data arriving from the correlator to produce calibrated visibilities. Separate image pipelines will then produce the image data products. Whilst the EMU survey will use total intensity (Stokes I) images, the POSSUM survey will use image cubes obtained for all Stokes parameters (Stokes I, Q, U and V). From these, Faraday rotation and polarisation properties will be obtained for detected sources.

The polarisation-related catalogues will include a POSSUM Polarisation Catalogue (PPC) with polarisation properties for compact (barely resolved) sources and a POSSUM Value Added Catalogue (PVACat) with information on more extended sources (which may not be spatially coincident in each Stokes parameter). The PPC will use as its starting point the EMU catalogue (data level 6) and will list the polarisation properties of each entry in this catalogue. The PPC will robustly classify the observed polarised spectrum of each source as Faraday ‘thick’ (complex) or Faraday ‘thin’ (simple). The PVACat will also start from the equivalent EMU catalogue (EVACAT) and will make use of the clustering properties and cross-identifications in defining sources and components. Several versions of PVACat will be produced, building on the initial version. New versions will potentially expand in scope to include a separate source-finding step conducted in Stokes Q, U or in polarised intensity spaces.

POSSUM data analysis will use rotation measure synthesis to measure magnetic field properties of the polarised sources. Further processing, including model fitting to Stokes Q and U spectra, may be carried out. These further steps will be used to classify the level 5/6 data products as Faraday thick or thin. Advanced analysis techniques are currently being investigated by the POSSUM team.

3.3.3 WALLABY

The WALLABY survey is a 'blind' survey of the southern sky to search for neutral hydrogen (HI) emission from galaxies. HI is the principal component of cool gas and this can be used to study how galaxies are formed and evolve over time and how they may merge or interact with other galaxies.

The survey aims to detect HI from around half a million galaxies with redshifts of $0 < z < 0.26$, corresponding to a look back time of 3 billion years. The observations will enable studies covering distances from High Velocity Clouds associated with our own Galaxy, to the Local Group of galaxies, and beyond to more distant clusters and super clusters.

The data volumes arising from ASKAP spectral line surveys are large and some compromises are required to make the data processing and storage manageable. The full WALLABY survey will generate around 96 PB of calibrated visibility data that are then processed to form image cubes. Given this extremely large data volume, the calibrated visibility data will not be archived.

For WALLABY it is likely that two types of data cubes will be produced:

- Low spatial resolution data cubes with full spectral coverage including all 16,200 channels. These will be restricted to using data from baselines below 2 km. (Using a maximum baseline of 2 km instead of 6 km reduces the cube data size by a factor of nine and degrades the spatial resolution by a factor of 3.)

- Cut-out image cubes centered on the positions and velocities of the detected sources. The data volume of the cut-out cubes is relatively small, allowing these cubes to be much more easily transferred to other locations for further use.

In addition, HI spectra and moment maps for detected sources will be generated as part of the imaging pipeline. Subject to the pipeline processing capability, some postage stamp cubes with higher spatial resolution may be generated for small regions around the positions of known sources.

3.3.4 DINGO

The DINGO survey will study the evolution of HI in the universe from the present time, back to a time when the universe was approximately half of its current age. The survey aims to detect HI spectral line emission from about 100,000 galaxies with redshifts of $0.02 < z < 0.43$. Unlike WALLABY which is a 'blind' survey of the sky, the DINGO fields will be selected from the GAMA (Galaxy and Mass Assembly) survey.

DINGO will study cosmological ‘distribution functions’ that describe how HI is distributed in galaxies and galaxy clusters. By combining the radio data with extensive information available from the GAMA and other surveys it will be possible to study the evolution and formation of distant galaxies, and the co-evolution of the stellar, gaseous and dark matter components of galaxies.

DINGO will obtain sensitive observations of a small number of survey fields with each field observed many times. Approximately 2,500 hours will be spent observing five regions of sky. In addition a deeper search will be obtained for two fields with 2,500 hours observing time on each field.

Following each scheduling block the science data processing pipeline will produce the individual data cubes for each survey field and these will be processed using the source finder with results written into source detection catalogues.

The survey team will use image stacking techniques to combine the data cubes so that a single final data cube is produced for each survey region. Each of the final stacked data cubes may contain up to 10,000 galaxies. Other advanced techniques such as spectral stacking across many galaxies may also be used.

Once the final data cubes are produced, these will be made available to the general community. Stacked image cubes and the science catalogues produced by the survey science team may be released at phased intervals prior to the full completion of the survey.

3.3.5 FLASH

The FLASH project will carry out a blind survey to search for extragalactic neutral hydrogen seen in absorption. In these absorbing systems cool hydrogen gas located in a galaxy or galaxy halo absorbs radio continuum emission from a more distant background source such as a radio galaxy or quasar. The absorbing system is located along the sight line from the observer to the background source. The survey expects to detect up to 1,000 extragalactic hydrogen absorbing systems with approximately one detection per survey field. These will be used for studies of the galaxy evolution and star formation in particular for galaxies in a redshift range of $0.5 < z < 1.0$.

The FLASH survey will target 850 survey fields and will identify 150,000 known continuum sources within these fields so that in effect each survey field will include around 150 to 200 sight lines to background sources. Prior to the start of the survey the Survey Science Team will generate a Target Source Catalogue that includes the positions of the continuum sources.

The data pipeline processing for FLASH will produce small postage stamp image cubes with full spectral coverage, centered on the positions of the known continuum sources. The source detection process is relatively straightforward: For each survey individual spectra are extracted at the positions of continuum sources and searched for HI absorption using a spectral line-finding technique developed by the Survey Science Team.

3.3.6 GASKAP

The GASKAP Survey Science team will carry out several surveys of gas in our Galaxy, the Magellanic Clouds, and the regions between the Clouds (the Magellanic Bridge) and between the Clouds and our Galaxy (the Magellanic Stream). These surveys will study spectral line emission and absorption from neutral hydrogen atoms (HI) at a wavelength of 21 cm and from hydroxyl (OH) molecules at a wavelength of 18 cm. The surveys will provide image cubes of extended gas emission with greater spatial and spectral resolution and coverage than has previously been achieved. They will also lead to the detections of thousands of compact sources, in most cases associated with either star formation regions or with evolved stars and supernovae.

In total GASKAP will observe around 360 fields with the observations taken over approximately 8,000 hours. Three different integration times will be used with 12.5, 50 and 200 hours per field allocated to different survey regions. The GASKAP surveys pose some particular ASKAP challenges. In particular:

- GASKAP will require the use of ASKAP zoom modes. Standard ASKAP observations use 16,200 channels across a bandwidth of 300 MHz corresponding to a frequency resolution of around 18.5 kHz. This is too coarse a resolution for Galactic spectral line studies where a resolution of around one kHz is typically needed. To achieve the required resolution, three zoom bands will be used to cover the HI and OH (1612 and1665/1667) transitions.

- To produce the final image cubes for the HI surveys, the ASKAP data cubes will be combined with data cubes already obtained from single dish observations. The addition of single dish data greatly improves the image quality for extended and complex structures. In principle several different techniques can be used for combining single dish and interferometric observations. Decisions on the approach to use are still to be made. At present it is not yet clear whether the HI data combination will be carried out as part of the pipeline data processing or will require post processing.

3.3.7 VAST

The VAST project will use the fast survey speed of ASKAP to investigate astrophysical objects that vary on timescales of 5 seconds or longer. Such sources span a huge range of scales, from Galactic to cosmological distances. They include flare stars, intermittent pulsars, X-ray binaries, magnetars, intra-day variables, supernovae and gamma ray bursts. Although the range of phenomena is very large, the underlying physics is generally associated with explosive events, propagation effects or by events linked to accretion and magnetism. VAST is likely to discover types of variable sources that so far are not known.

The VAST project observing strategy has two approaches:

- Where feasible, VAST will make use of ‘piggy-back’ observing where data taken for other projects is also analysed for transient sources.
- VAST will also make use of dedicated blocks of observing time. This will be used for repeated observations of target fields. A large-scale survey (VAST-wide) covering

approximately 5,000 square degrees is planned with the entire survey region observed daily using short integrations for each survey field. A deeper survey (VAST-deep) of the same survey region but with longer integration times, and a smaller survey of the Galactic Plane may also be undertaken.

The science data pipeline processing requirements required to meet the VAST requirement of variability timescales as short as 5 seconds are highly computing intensive and will pose many challenges. The full ASKAP data processing pipeline for transients is not expected to be in use for Early Science. However, Early Science is likely to include repeated observations of selected fields (section 3.2).

3.3.8 COAST

The COAST Survey Science Project will use the ASKAP array to study radio emission from pulsars. These are highly compact evolved stars that rotate and emit highly beamed radio emission as a series of radio pulses. Pulsars fall into two groups – 'standard' pulsars with periods of typically one second and 'millisecond' pulsars where the rotation rate is much faster. The time-related properties of pulsars can be measured to extremely high precision and this allows pulsars to be used as tools across a range of studies including tests of general relativity and gravitational wave studies. A key goal for pulsar astronomy is to detect gravitational waves, either from individual sources, or from a stochastic background. In addition pulsars are used to study the properties and evolution of neutron stars. Understanding their internal structures, emission mechanisms and magnetic fields remains highly challenging.

The COAST ASKAP pulsar observations will use the array in a special mode where subsets of antennas are used together in a tied-array mode. In effect each tied array acts as a single dish. The use of multiple tied arrays is anticipated as this would significantly improve the survey speed for ASKAP pulsar surveys.

The COAST planning includes two types of pulsar observations corresponding to timing and search modes:

a) Timing-mode observations of pulsars with known rotational periods. For this mode voltages at the antennas are sampled directly without using the correlator. The data are streamed off-line to another location where they are de-dispersed (to correct for dispersion and propagation effects in the interstellar medium) and 'folded' to the known pulsar period. By using multiple tied-array beams ASKAP will be able to observe 10s of pulsars at the same time giving it a multiplexing advantage when compared to a single dish such as Parkes. The main data products produced by timing observations are folded pulsar profiles and time series data.

b) Sensitive targeted search-mode observations will be carried out to look for pulsar emission from compact sources that are identified in other ASKAP surveys such as EMU. As for timing observations this mode takes the data stream from the MRO before it reaches the correlator. The search mode data volumes produced by timing and targeted search modes are comparable to those generated at Parkes. Data processing generates a list of pulsar candidates. These are then followed up with further observations to determine whether pulsars are present (Table 5).

In addition, a more complex search mode may be used where the data correlator is used to produce visibility files with an extremely high data rate of 2 milliseconds per sample. This 'fast dump visibility search' mode requires additional custom hardware and generates high data rates (Table 6). The feasibility of this is still under discussion.

Almost all pulsar data are retained using a standard PSRFITS file format. This is compatible with VO protocols.

COAST pulsar data will NOT be processed at the Pawsey Centre as part of the standard ASKAP science data processing pipelines. Instead these data will processed off-line by the science team using specialised pulsar data reduction software.

Pulsar data obtained with the Parkes radio telescope is provided to the community through the CSIRO Pulsar Data Archive and are made available using the CSIRO Data Access Portal (DAP). The pulsar data in this archive are stored in Canberra. The ASKAP pulsar data will be provided either through CASDA with data storage in the Pawsey Centre or, or alternatively, as an extension to the CSIRO Pulsar Data Archive, with data storage in Canberra. In either case, user access is through the DAP. This will be considered further as planning for pulsar observing with ASKAP progresses.

3.3.9 CRAFT

CRAFT is a project to search for and study fast transient sources that vary on timescales from approximately one millisecond to 5 seconds. The CRAFT project science goals are complementary to VAST and to some aspects of the COAST pulsar studies. CRAFT observations will be taken in a fully commensal mode. The study of fast transient sources is expected to open up new windows in astronomy that include sources that are so far unknown but represent extreme states of matter and very strong magnetic and/or gravitational fields. Such sources may include Galactic neutron stars that emit irregular or giant pulses in addition to sources of extragalactic origin.

As an example of fast transients, a single intense burst of radio emission lasting for a few milliseconds was found by Lorimer et al. (2007) [11] in archival Parkes data. The subsequent detection of several of these so-called Lorimer bursts by Thornton et al. [12] confirm their origin as extragalactic and suggest there may be 10,000 such bursts per day over the sky. The origin of these bursts remains unclear and they have not yet been identified in any other wave band. Detection rates for fast transients are very poorly known. An initial estimate of the possible detection rate for Lorimer bursts using ASKAP is one per day per 30 square degree field of view.

The large field of view of ASKAP together with the ability to use coherent data to determine a source position and to separate transient events from terrestrial interference provide very strong advantages for the study of transient sources. However, the data processing and data handling requirements are highly challenging and specialised hardware and software are required.

Instead of analysing visibility data from the ASKAP correlator, CRAFT will make use of the total power (auto-correlations) output from the beamformers at each antenna. The total power data streams will be sampled at a time resolution of about 1 millisecond and a frequency resolution of 1 MHz and can be summed together after accurate time alignment. For 36

ASKAP antennas and 36 beams per antenna each the data rate for a single polarisation output is ~ 6 Gpbs. This is sufficiently low that it is feasible to analyse the data stream in real time, de-disperse the data and search for fast transient sources. This use of incoherent data allows for the detection of transient sources but does not provide source positions.

In addition to using total power data streams, the ASKAP antenna beamformers will have data buffers to retain coherent baseband data over 'rolling time' intervals. The detection of a suspected transient event would trigger a readout of the buffers with the data sent off-site for follow-up processing. By using these coherent data it is possible to determine source positions to high accuracy. Data buffer sizes of around 8 GB (for each antenna, beam and polarisation) are sufficient to hold approximately 12 seconds of ASKAP data at the highest frequencies used.

At a later stage more sophisticated techniques may be employed including the use of tied-array data, larger data buffers and the use of very high time resolution data output from the correlator.

The data processing for CRAFT will not make use of the ASKAP science data processing pipelines. Some CRAFT level 7 data products may be uploaded to CASDA. These may include, for example, event related information, image data products associated with transient detections and possibly some coherent data from the data buffers. CASDA tools to support such transient data products are not expected to be available from the start of Early Science but may be considered later.

3.3.10 VLBI

Very Long Baseline Interferometry (VLBI) is a technique used in radio astronomy where radio astronomy signals are recorded at different, widely separated locations and then brought together for correlation. The Australian Long Baseline Array (LBA) includes radio telescopes at Parkes, Mopra, Narrabri, Hobart and Ceduna together with the recent inclusion of antennas at the MRO and in New Zealand. This array includes extremely long baselines of up to 5,500 km and this enables high resolution studies of compact objects.

The inclusion of VLBI as a Survey Science Project is primarily as a technical demonstrator that will trial and demonstrate many of the techniques that will be required for the SKA. These include high-speed data recording and data transport networks, innovative correlation facilities and the development of new approaches to VLBI. VLBI science observations taken with ASKAP have so far made use of a single antenna equipped with a single-pixel feed. This will later be extended to include all available antennas linked together as a 'tied array' whilst innovative techniques such as cluster-to-cluster observing may be tried.

There are currently no plans to provide user access to VLBI data through CASDA. The inclusion of ASKAP antennas for VLBI observations will be managed as part of standard CASS LBA operations. Currently, ASKAP VLBI data files are written locally to data disks at the MRO and transferred to Perth for correlation with data from the other radio telescopes used. The correlated data are stored using the iVEC PBStore facility and made available to users through ftp.

3.4 Guest Science Projects

The Guest Science Projects (GSPs) are observational programs that require less than 1,500 hours of observing time to complete. Proposals for Guest Science Projects will be submitted through the Online Proposal Applications system (OPAL) and will be assessed by the Time Assignment Committee. It is expected that proposals will be requested, as for other ATNF facilities, every six months. Up to 25% of the total time available will be scheduled for GSPs, corresponding to about 750 hours of telescope time every six months. A typical time for a project may be around 100 hours, so approximately 10 GSPs are likely to scheduled in a six-month semester.

In most cases, the GSP data and data products will be released publicly into the ASKAP Science Archive without any proprietary period. However, if reasonable grounds are established in the proposal, the Time Assignment Committee will have the discretion to allow a proprietary period of up to 12 months.

The data products for the GSPs will not be validated by the science teams as these groups cannot be expected to have the level of expertise required to do this. Some data quality flags will be provided by the ASKAP hardware and science processing pipelines and these will be available to users. However the survey science data flag will be unset.

3.5 Target of Opportunity observations

These are unexpected astronomical events of extraordinary scientific interest for which observations on a short time scale are justified. Such an event might for example be a supernova explosion or the need for radio data following the detection of an unidentified burst of emission with an X-ray telescope. Observing time for Target of Opportunity is granted by the ATNF Director (who is also the CASS Chief).

Target of Opportunity observations are of immediate interest to the astronomy community. The data products will be released without any proprietary period and with minimal data validation.

4. THE SCIENCE ARCHIVE

4.1 Overview

The CASDA software application will be responsible for taking the data products output from the Central Processor, archiving these to storage media and generating the databases and metadata needed for a science archive. CASDA will provide astronomers with the web and VO protocols needed to search and access data products and to further use these for scientific research. It will also manage administrative functions such as queue controls and monitoring system usage and performance.

The CASDA application and data services will be integrated with the CSIRO Data Access Portal (DAP). The CSIRO DAP is hosted in Canberra and acts as a central portal for many CSIRO data collections including radio astronomy pulsar data from the Parkes radio telescope. Access to the DAP is through the web (see https://data.csiro.au/dap/home).

Figure 5 shows a highly simplified overview of CASDA interactions. Users will interact with the archive through the CASDA application. The application will connect with the Real Time Computer (RTC) and with a 'middleware' layer that acts as an interface between the CASDA application and the Hierarchical Storage management (HSM) system.. The RTC is a compute platform provided by iVEC's Galaxy supercomputer. The middleware layer will manage the tracking of data files and their transfer and retrieval to/from storage media via the HSM.

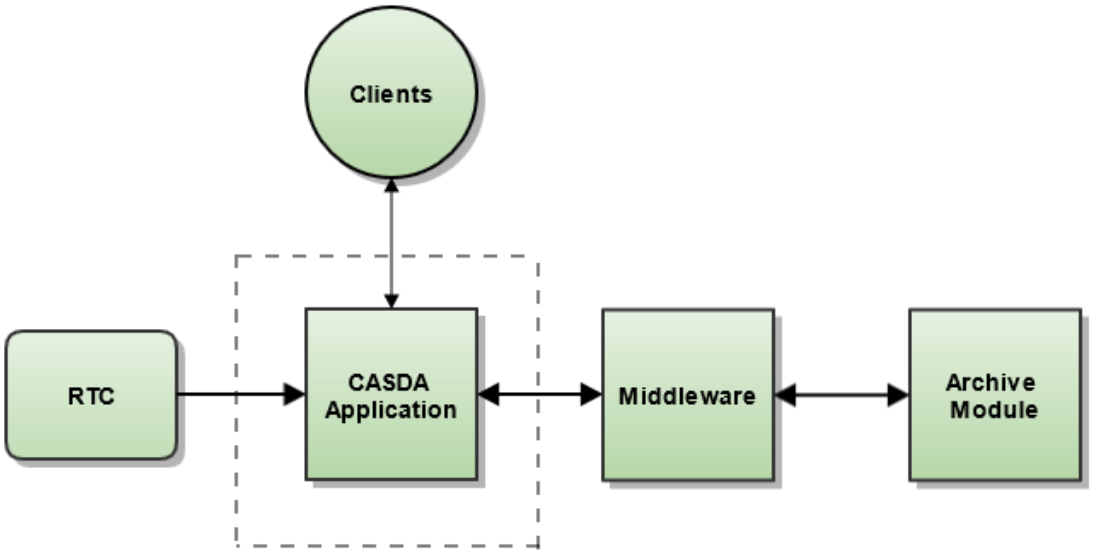

Image credit: James Dempsey

Figure 5: CASDA overview diagram. The RTC module includes the Lustre filesystem.

Figure 6 shows a context diagram for the CASDA application and the major activities of the different stakeholders. CASDA will ingest data products released from the Central Processor and will interact with the Survey Science Teams, Guest Science teams and the worldwide astronomy community with technical and operations support and administration provided by CSIRO and iVEC. Large data centres such as NED and SIMBAD will be able to access ASKAP catalogues using tools provided to the community and may make information from these available through other external services.

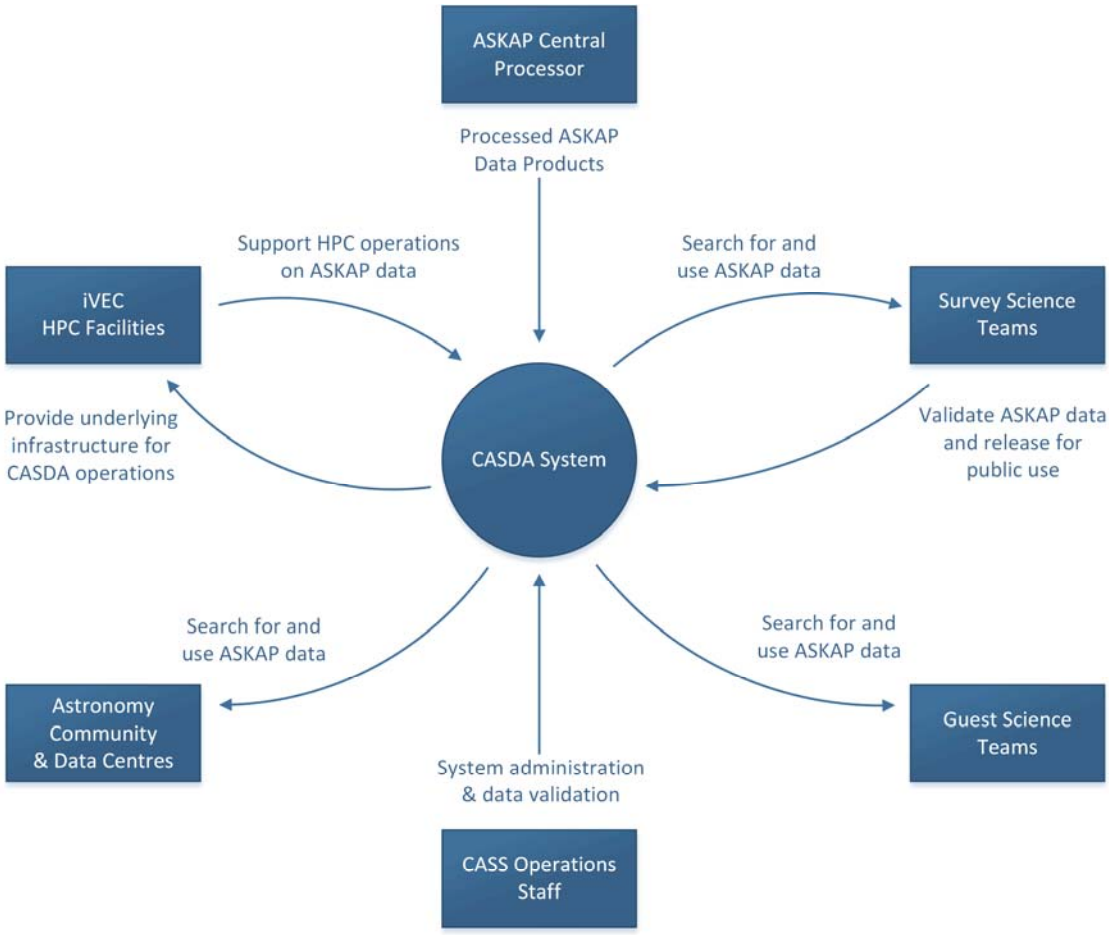

Figure 6: CASDA context diagram

4.2 Pawsey Centre Infrastructure

Figure 7 shows components of the physical infrastructure at the Pawsey Centre that are likely to be of relevance to the science archive. These include:

Processing platform

- RTC (Galaxy): This provides the platform for the ASKAP Central Processor sub-system.
- Lustre Filesystem: 1.3 PB of usable disk space attached to the RTC is provided as scratch space for the ASKAP Central Processor.

Data products produced by the RTC are written onto the Lustre file system and transferred from there for data storage using the HSM provided by SGI.

Storage platform

- Tape libraries: The HSM includes two Spectra Logic tape libraries, currently each has 20 PB of storage (shared between all users), with one library acting as a backup of the other. The tape libraries have capacity to increase to 2 x 50 PB;
- Massive Array of Idle Disks (MAID) array: This has 450 TB of additional HSM data storage provided on disks.
- Cached disk storage: Approximately five PB of disk storage distributed across a large disk pool. ASKAP has requested approximately one PB of this disk storage.
- The ASKAP disk storage on the HSM will be configured to handle visibility and image files. In addition to tape storage, files will be retained on disk for as long as is practically possible.
- Data is transferred between the RTC and data storage systems at the Pawsey Centre through edge servers (gateway nodes) and Infini-Band (IB) high performance connectors. Transfer rates across the IB networks are expected to be around 5 GB/s with similar read/write transfer rates for the Lustre filesystems.

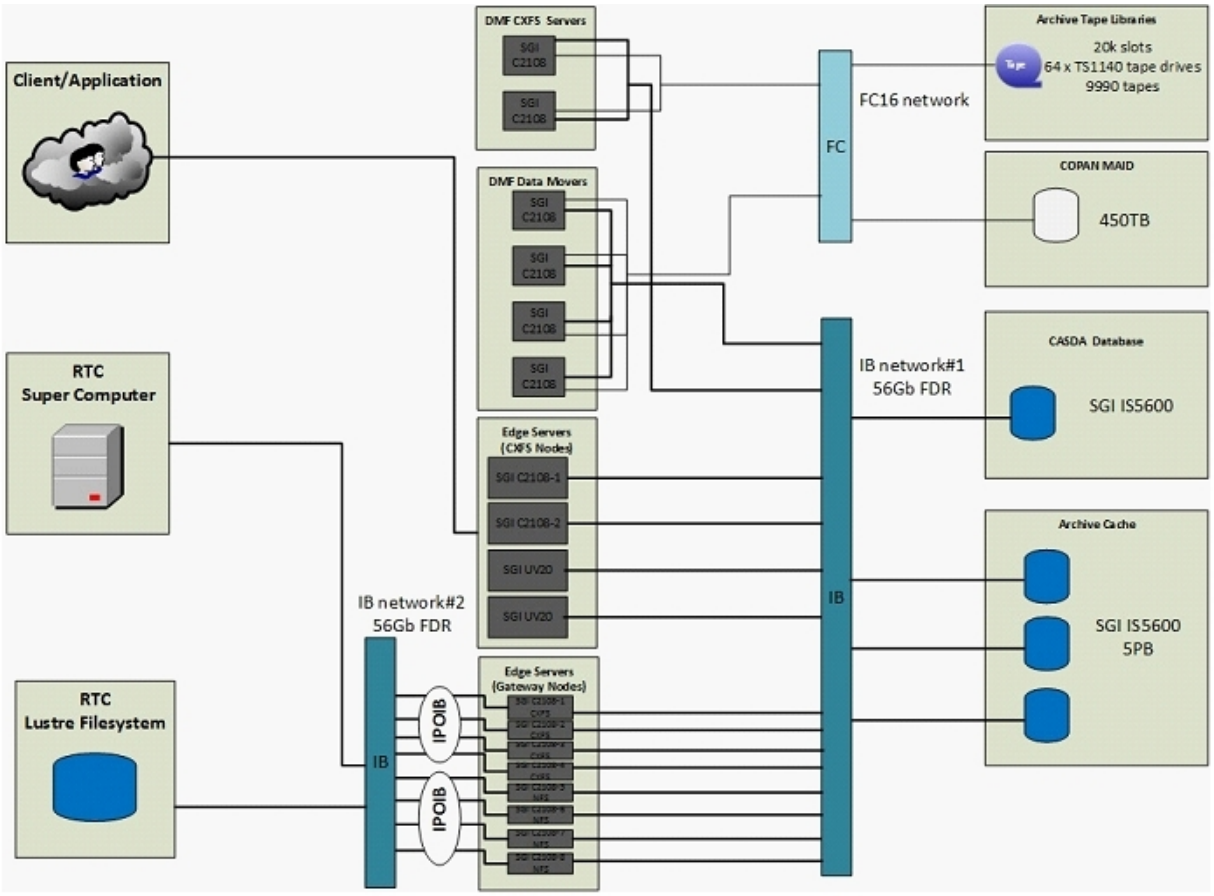

Image Credit: Dave Morrison

Figure 7: Components of the Pawsey Centre physical infrastructure relevant to CASDA

4.3 Primary data products

Table 3 lists the 'primary' data products that will be produced by the science data processing pipeline and made available to users through the science archive. This does not include specialised data products from the VLBI, COAST and CRAFT surveys. A more detailed breakdown of data products is given in Appendix B.

As an extension to previous requirement statements, CASDA will provide tools so that the Science Survey Project teams can load VO compatible science survey catalogues, with access provided to users through CASDA and metadata to establish the ownership and provenance of such data products. By making these available to the worldwide astronomical community, the association of these science catalogues with ASKAP (and CSIRO) will be much stronger than would otherwise be the case, whilst the science benefits obtained from using the archives will be significantly increased.

The FITS format used by ASKAP will be compatible with international FITS standard. (The FITS standard is available at http://fits.gsfc.nasa.gov/fits_standard.html.) Note that other data formats are being considered for use with radio astronomy and it is possible that archive support additional or alternative data formats may be required in the future.

Table 3: CASDA data products

Ref	Survey Types	Data Product	Format
P1	C	Calibrated continuum visibility data	CASA measurement set
P2	C	a) Full-size continuum images and image cubes b) Cut-out continuum images and image cubes for detected sources	FITS
P3	S	a) Full-size spectral line image cubes b) Cut-out spectral line image cubes for detected sources	FITS
P4	S	Spectral line postage stamp image cubes for detected sources	FITS
P5	S	Moment maps generated from cubes for detected sources	FITS
P6	S	Spectra of detected sources extracted from cubes for detected sources	FITS
P7	All	Calibration and system information	Catalogue
P8	All	Scheduling block information	Catalogue
P9	All	Global Sky Model	Catalogue
P10	All	Image quality reports	Catalogue
P11	C	Continuum source detection catalogues	Catalogue

P12	C	Polarisation-related catalogues	Catalogue
P13	S	Spectral line source detection catalogues	Catalogue
P14	T	Transient source detection catalogues	Catalogue
P15	All	Survey Science Teams: Level 7 source catalogues	Catalogue
P16	C	Polarisation-related spectra	FITS

4.4 Virtual Observatory protocols

Science users of the archive will be able to access images, image cubes and catalogues using Virtual Observatory (VO) protocols. VO protocols are unlikely to be used to provide direct access to visibility data files.

The International Virtual Observatory Alliance (IVOA) provides a set of standard and internationally recognised protocols that allows users to search and access data, including images and image cubes, catalogues and derived products. The current VO protocols include:

Cone Search Protocol: This is used to search a catalogue for information corresponding to a region of sky around a specified position. The protocol uses three arguments for right ascension, declination and search radius. In response, the server sends a VO table with the results (other formats can be requested). Cone searches are commonly used in astronomy and are likely to be the most frequent search mode carried out by general users.

Simple Image Access Protocol: This enables searches around a specified sky position and is used to find and access images and image cubes. The service can generate cut-out sections of images or cubes with user-specified dimensions. It returns a link to the requested outputs.

Simple Spectral Access Protocol: Allows for queries to return source spectra from an image or image cube.

Table Access Protocol (TAP): Allows for the construction of more complex queries of catalogues. Queries can be expressed in several ways: The result of a TAP query is generally returned as another VO table.

Note that additional VO protocols may be developed for radio astronomy, by CASS and/or other organisations and later considered for use with CASDA.

4.5 Data volumes

Table 4 summarises indicative data volumes for the seven Survey Science Projects where the data products are obtained from pipeline data processing at the Pawsey Centre. For each project the last three columns give the estimated total visibility, image-related and catalogue data volumes for the science archive based on indicative parameters. For assumptions used, detailed calculations and notes see Appendix C.

In Table 4, column 4 corresponds to the total time per field as given in the project proposal. Note that this may differ from the actual scheduled time as some adjustments are likely to allow for commensal observing. Data volumes given in this table are also NOT adjusted for commensal observing. Commensal observing will reduce the total data requirements through sharing of data between projects. For example, EMU, POSSUM and VAST may all use the same incoming set of visibility data, whilst several surveys will make use of the EMU Stokes-I images and catalogue information.

Table 4: Survey Science Projects indicative data volumes

Survey	**Type**	**N survey fields**	**Total time per field**	**Visibility data size per field**	**Image data size per field (all images or image cubes)**	**Total visibility data volume**	**Total image / image cube data volume**	**Total tables data volume**
Unit			**h**	**TB**	**TB**	**PB**	**PB**	**GB**
EMU	C	1200	12	2.4	0.004	2.8	0.004	25
POSSUM	C	1200	8	1.5	1.0	1.8	1.2	25
WALLABY	S	1200	8	not archived	1.8	not archived	2.1	<1
DINGO Deep	S	5	500	not archived	1.3	not archived	1.3	<1
DINGO Ultradeep	S	2	2500					
FLASH	S	850	2	not archived	0.48	not archived	0.4	<1
GASKAP	S	644	12.5	not archived	1.3	not archived	0.9	<1
VAST	T	1200	12	0.224	7.0	0.27	Most image cubes are not archived	See App B

Data volumes given in this table are indicative only and may vary considerably depending on actual scheduling and data processing parameters used. For example, the number of fields that can be observed per day for a given project will depend on a number of factors such as the rise

and set times of the regions to be observed, the location of the sun and whether projects are observed commensally with each other.

Table 5 provides examples of data volumes estimated for the three projects where the data are not processed through the data processing pipelines. For COAST values are given separately for different observing modes.

Table 5: Example data volumes for COAST, CRAFT and VLBI

Survey	Data volume per 12 hours (GB)	Total data volume for full project (TB)	Notes	CASDA archive required?
COAST (a)	12	2.0	Timing of millisecond pulsars. Data volume after de-dispersion and folding. Assumes single beam.	Possibly see 3.3.8
COAST (b)	20	1.0	Timing of non-millisecond pulsars. Data volume after de-dispersion and folding. Assumes 20 tied beams	Possibly see 3.3.8
COAST (c)	2,000	333	Search mode for targeted sources. Data volume for raw data	Possibly see 3.3.8
COAST (d)	85,000	88,000	Fast-dump visibility search mode observations. Data volume for raw data	No
CRAFT (e)	10,400 Per event	12, 500 @ one event per 12 hours	Corresponds to the data volume for coherent data from buffers	Possibly See 3.3.9
VLBI	50	15	Correlated visibility files are currently stored using PBStore, managed by iVEC	No

Notes:

a) Total data volume assumes 2,000 hours for timing of millisecond pulsar with observations taken over a five year period.

b) Total data volume assumes 250 hours for timing of non-millisecond pulsars.

c) Total data volume assumes 2,000 hours for targeted search mode observations.

d) Total data volume assumes 1,250 hours for fast dump search observations taken over five years. Raw data are likely to be discarded after processing for candidate detections.

e) Data volume corresponds to baseband data from the beam former data buffers with 36 antennas, 36 beams, 1 polarisation, 8 GB buffer size (per beam, antenna and polarisation).

Given the potentially very large amount of data it is unlikely that baseband data will be archived. A much smaller volume of data products associated with the transient events may be included in CASDA.

5. ARCHIVE SEARCH AND ACCESS

5.1 Archive searches

To search the CASDA archive, most queries will be undertaken using either web form interfaces or VO interfaces:

Web Form Interfaces: Customised web-based search forms will allow searches of all data product types held in the archive. Following a query set up through a web-based search, data may be made available through a number of mechanisms as discussed below.

VO services: CASDA will provide access to ASKAP catalogue and image–related data products using VO services and protocols (section 4.4). Links to the CASDA VO services will be available through the DAP and through external VO registries such as an IVOA registry.

To make use of the CASDA VO services, users will have several options including installing the freely available and widely used TOPCAT software, making use of web-based applications (such as SIMBAD or Aladin), or using astronomy software libraries. These applications provide tools that allow users to set up VO queries that search data archives and retrieve data. Users can connect to VO services provided by many different suppliers and this allows retrieval of data across multiple surveys.

5.1.1 Automated queries

CASDA will provide support for repeated automatic querying via the VO interfaces. The Astronomical Data Query Language (ADQL) will be supported whilst Python astronomy libraries, PyVO and astropy, and the Starlink Tables Infrastructure Libraries Tool Set (STILTS) may be used to interface with the VO Application Programming Interface (API). These will provide the means to do comprehensive catalogue and data product searches, retrieve image cut-outs and request file access via http.

5.1.2 Graphical interface

In addition to web-based and VO searches, CASDA may provide an interactive graphical interface to ASKAP data products. This tool is considered highly desirable but is of lower priority than getting core CASDA functions such as data deposit and retrieval in place. If resources permit, a graphical interface facility will be developed for the first release of CASDA. Otherwise this will be considered for a future upgrade.

5.2 Data access

From Early Science onwards, all access to the ASKAP validated data products (corresponding to data level 6) will be through the CASDA science archive.

CASDA provides tools to search and access archive data but does not directly provide access to data storage or high performance computing. Where working data storage or high performance computing facilities are needed to work with ASKAP data products accessed from the archive it will be the responsibility of the science teams to make suitable arrangements.

The use cases for CASDA cover a broad range of data access requirements from small data volumes below one GB to large data volumes of many TB. In general, it is expected that there will be a large number of users requesting relatively small volumes of data and a small number of users requiring high data volumes. The Survey Science Teams in particular are likely to require access to large data volumes for scientific research and analysis of their ASKAP data (see section 6.3).

For small to moderate data volumes, users will be able to access ASKAP data products by transferring data to other locations, either using standard web protocols (http) and/or other network transfer protocols.

Web transfers using http are generally efficient for relatively small volumes of data where the transfer can be completed in a short time. Other network transfers protocols (such as rsync or scp) allow transfer of higher volumes of data. Generally, such data transfers allow a transfer to continue following a network outage or other interruption. We note here that the use of different network transfer protocols with VO services, is being investigated.

The volume of data that can be transferred across a network depends critically on the available network speeds. In addition, data transfer rates can be significantly reduced by institutional firewalls and other factors. Actual data rates vary greatly between different organisations and locations.

5.2.1 Data transfer rates

To illustrate data transfer rates, Table 7 gives the minimum time needed to transfer a range of data volumes at given network speeds. In this table the blue cells indicate where a data transfer could reasonably be carried out as a *single* transfer transaction. For example, for a network speed of 10 MB/s it should be feasible to transfer 250 GB (over 8 hours), whereas for ADSL2 it would not be possible to transfer this data volume in a single transaction. Higher volumes could be achieved using multiple transactions.

In Table 6, column five refers to ‘fast’ network transfers between two data centres using dedicated paths and/or multiple data streams and appropriate configurations at both ends. Networks such as AARNET provide high data rates between participating nodes. The example below assumes a fairly conservative fast data transfer rate of 100 MB/s. Currently, data rates of at least several hundred MB/s, within Australia or between Australia and overseas, can be achieved with appropriate networks.

Table 6: Data transfer times for given network speeds

Transfer rate	**500 KB/s**	**3 MB/s**	**10 MB/s**	**100 MB/s**	**500 MB/s**
Example usage	**ADSL2**	**'Standard' network transfer**	**CSIRO network**	**'Fast' network transfer between data centres**	**Transfer within Pawsey Centre**
Transfer time					
1 GB	34 mins	6 mins	100 sec	10 sec	1 sec
10 GB	6 hours	1 hour	17 min	100 sec	10 sec
100 GB	2 days	10 hours	3 hours	17 min	100 sec
250 GB	6 days	25 hours	8 hours	45 min	4 min
1 TB	23 days	100 hours	28 hours	2.8 hours	17 min
10 TB	230 days	39 days	12 days	28 hours	2.8 hours
100 TB	6.3 years	390 days	116 days	12 days	28 hours

Table 7 shows a set of potential CASDA data tiers and indicates whether the different access methods are feasible for each tier. Note that the boundaries between tiers have not yet been fully established and are likely to evolve over time as technologies develop. However, CASDA will impose some restrictions on the volume of data that can be transferred at a time and the transfer mechanisms available for different data volumes.

Table 7: CASDA Data Tiers

		Data transfer feasibility			
Data Tier	**Data volume**	**Web access**	**Standard network access**	**Fast network access to another Data Centre[a]**	**Access within Pawsey Centre[b]**
A	0 – 10 GB	Yes	Yes	Yes	Yes
B	10 – 250 GB	No	Yes	Yes	Yes
C	250 GB – 5 TB	No	Possibly	Yes	Yes
D	5 TB – 50 TB	No	No	Yes	Yes
E	50 TB +	No	No	Yes	Yes

5.3 Applying for supercomputing facilities

This section provides some general information, as current in April 2014. Please note that this information is subject to change over time. For further advice please contact Jessica Chapman (Jessica.Chapman@csiro.au) in the first instance.

5.3.1 Pawsey Centre facilities

The Pawsey Centre facilities are managed by iVEC, a joint venture between CSIRO and the four public universities in Western Australia. The Pawsey facilities are allocated as follows:

- 25% for radio astronomy. The radio astronomy share is predominantly for the operational requirements of facilities such as ASKAP and the Murchison Widefield Array;
- 25% for Geoscience;
- 30% to the iVEC partners. These are CSIRO, University of Western Australia, Curtin University, Edith Cowan University and Murdoch University;
- 15% through the National Computational Merit Allocation Scheme (NCMAS) managed by the National Computing Infrastructure (NCI); and
- 5% Director's time.

Science teams with significant computing requirements may need to submit applications for the use of Pawsey Centre facilities, including access to disk space to work on the data, and to computing power. Currently applications for the use of Pawsey Centre facilities can be submitted through the Partner Share scheme, NCMAS or (for smaller requests) through a request for Director's time.

Both the NCMAS and Partner Share allocations can host large scale applications that have considerable HPC requirements. Application deadlines are currently around early October of each year. NCMAS is available to researchers at Australian higher education institutions and public sector research organisations in Australia. Applications to NCMAS may include a request for 'scratch' disk space. If required a separate application can be made to request archival data storage (see below). Where applications to NCMAS are not successful (or are only partly supported) they may be forwarded to iVEC (and/or Geosciences) for consideration through the Partner Share scheme. An application to the iVEC Partner Share scheme must be led by a member of one of the partner organisations. Applications may include individuals from Australia or overseas who are not members of the partner organisations.

In some cases ASKAP science teams may require access to data storage without requiring significant HPC facilities. In this case, teams may wish to consider applying to iVEC for appropriate storage, such as Pawsey storage or Research Data Services (ReDS) storage.

ReDS storage is provided by the Australian Research Data Storage Infrastructure (RDSI) and is intended for nationally significant data collections. This includes merit-based storage, which is mainly for collections that are complete, and Collection Development Storage (CDS), which is for collections that are still under development. CDS is applicable where data are being worked

on and/or where data are still being collected. There is no requirement for data to be made available to others.

The ReDS Merit-based Storage and CDS schemes are open all year round via an online, iVEC process. Generally a response is sent within six weeks. If approved, storage is allocated for a specific duration, taking into consideration iVEC's resource constraints and the requirements of each project.

For further information on iVEC data services, including the iVEC Data Storage and Management Policy, see the iVEC web pages including http://www.ivec.org/services/data-storage/ and links from there.

5.3.2 Other supercomputing facilities

ASKAP science teams may also wish to consider applications for use of facilities at other Data Centres in Australia or overseas. Within Australia, astronomers can apply for data storage and high-performance computing facilities provided by CSIRO, Swinburne University, iVEC, National Computer Infrastructure (NCI) and other RDSI nodes around Australia. Conditions apply for all facilities.

6. REQUIREMENTS AND USE CASES

6.1 High-level requirements

Table 8 summarises the high-level CASDA requirements. The requirements for the ASKAP data archive form part of the full set of ASKAP project requirements [10]. Column three provides cross references to [10].

Table 8: High-level CASDA requirements

Essential requirements	**Notes**	**Cross-reference to [10]**
	Data Access	
ASKAP data products are open access and made publically available as soon as possible.	Survey data products to be released to the public domain as soon as they are validated. Guest Science data products to be publicly released immediately following any required proprietary period.	4.1.1 4.4.1
CASDA will ensure authenticated user access for data downloads.	Access control will be based on a user authentication and registration system (such as OPAL). All users will have access to see what has been observed. Registration required for data downloads.	4.1.2
Survey Science teams will have access to data quality flags and will set these following data validation.	Survey Science data products restricted to science teams and administrators prior to validation. Changes to data flags tracked.	4.5.1 4.5.2
CASDA will provide user access to data products via web interfaces with appropriate search tools.	Searches will be made on metadata held in archive databases. Searches by name will use standard name resolvers.	4.1.3
CASDA will provide access to images, image cubes and catalogues using VO protocols.	VO protocols include: • VO cone searches. • VO Simple Image Access Protocol service • VO Table Access Protocol service. • VO Simple Spectral Access Protocol CASDA will register available services with appropriate registries and will develop VO capabilities that comply with IVOA protocols and standards.	4.1.4 4.1.5

Data ingestion		
CASDA will handle the ingestion of data products generated by ASKAP Survey Science Projects, Guest Science Projects and Target of Opportunity observations in a timely and efficient way.	ASKAP data products estimated at about 15 TB per day.	4.2.1 4.2.2 4.2.3 4.4.3
CASDA will provide a repository for Survey Science Teams to upload pre-defined and VO-compatible science catalogues and will provide search tools for such catalogues.	Level 7 catalogues owned by Science Teams. Metadata provided to identify ownership and provenance.	New item
Operations and user support		
Long term data storage will be provided at the Pawsey Centre.	Data products to be archived on an indefinite basis. Two copies of data sets will be stored at the Pawsey Centre.	4.3.1
The CASDA design will not restrict the potential future requirement for one or more copies of the archive to be stored at other locations.	Current plans do NOT include mirroring of the image and visibility data files to other locations. However, mirroring may be considered as a potential future requirement.	4.3.2 (modified)
The CASDA architecture and design will be flexible, extensible and scalable to allow for future growth and technology changes.	Such changes may include: • Additional catalogues • New data formats • New 'software instruments' with associated data processing pipelines • Increased and/or replaced physical infrastructure	4.3.3
CASDA will preserve the history of changes to the archive.	For example, earlier versions of catalogues will be retained.	4.3.5
Regular backups of the CASDA database, catalogues and metadata will be made with a copy stored at another location.	Off-site storage of the science databases and associated tables and metadata will reduce risks associated with on-site disasters. As above image and visibility data files are not mirrored.	New item
CASDA operations will provide appropriate levels of user support.	Specification will be included in a CASDA User Support Model.	4.3.7
The CASDA archive will normally be in operation 24 hours per day and seven days per week.	Some downtime will be required for maintenance and to resolve technical issues.	4.3.8
The CASDA support team will provide prompt responses to user requests for support.	Specification will be included in a CASDA User Support Model.	4.3.9

CASDA will provide appropriate archive administration tools.	See Table 11	4.3.10
CASDA will capture user feedback and levels of satisfaction.	Monitor levels of user satisfaction and capture comments and suggestions for improvements.	New item
CASDA will allow external data centres to access source catalogues and associated metadata for provision through their facilities	Survey Science Teams may wish to provide Catalogues to other data centres such as NED or SIMBAD.	4.1.9
Desirable requirements		
CASDA may use data visualisation tools to facilitate user interactions with the archive.	For example, Google-Map type interactive tools may be used to navigate the sky and obtain information on archive contents	New item
CASDA may provide a repository to enable science teams to load level 7 image cubes to the archive.	Subject to specific agreements on a case-by-case basis. May be required for stacked image cubes generated for DINGO.	New item
CASDA may provide facilities to load data from projects where the science data processing is not carried out using the ASKAP data processing pipelines in the Pawsey Centre.	May include pulsar data products from COAST. May include transient data products from CRAFT.	New item

6.2 Survey Science Projects use cases

Tables 9a to 9j provide a summary of data products and use cases for each of the Survey Science Projects. These tables are intended to facilitate discussions between CASDA and the Survey Science teams.

Each table lists the level 5/6 data products where these are generated in the ASKAP science data processing pipelines and level 7 data products that may be generated by the science teams. CASDA use cases and notes on validation are also given.

Note that other level 7 data products may also be produced in addition to those include in Table 9. CASDA will provide support for loading level 7 science catalogues produced (in appropriate formats) by the Survey Science Teams into the archive. The data management and storage of *all other* level 7 data products is the responsibility of the science teams (section 5.3).

In most cases science teams will need to consider access to high performance and/or data storage facilities for any analysis of data products accessed from the archive. Arrangements for doing this are also the responsibility of the science teams (section 5.3).

Table 9a: EMU

Science Team

PI: R. Norris (CASS)

Team includes ~ 100 individuals from 13 countries

Level 5/6 data products

- Full polarisation calibrated visibility data
- Images and image cubes are stored for Stokes I only. Images and cubes are produced for each integration block of 12 hours. Images correspond to the restored, residual and model images for each of three Taylor terms corresponding to the source intensity, spectral index and spectral curvature. For each Taylor term, images are also stored for sensitivity and the Point Spread Function (PSF).
- Source detection catalogue. The full survey is expected to result in approximately 70 million source detections.

Level 7 data products produced by Survey Science Team

- EMU Value Added Catalogue (EVACAT)
  Several versions of this catalogue will be produced as additional or updated information becomes available. The EMU catalogues will include associations with known sources from other major surveys as well as other derived scientific quantities, e.g. estimates of redshift, class of radio source etc.
- Stacked images
  The EMU team may investigate using image stacking techniques to increase the survey sensitivity.

CASDA Use cases

- Access summary information on observation blocks completed including information on system performance and RFI etc.
- Access information on status of tasks submitted for processing at Pawsey Centre
- Access help from archive support staff
- Query catalog(s) for a summary of visibility files archived and sky regions observed
- Access visibility files if needed (expected to be infrequent).
- Access catalogue information on image quality
- Set data validation flags following review of data validation reports.
- Access survey images or image cut outs for data validation or post processing purposes.
- Generate a large number of image cut-outs and transfer to another location.
- Access image cut outs for outreach purposes such as the ‘Radio Zoo’ collaboration with Galaxy Zoo.
- Access information from source detection catalogues
- Use complex filters to search catalogues – for example restrict a field to be within a range of values or define filter criteria based on a mathematical operation of several fields.
- Run a catalogue search using an input user catalogue as an input list of sources or positions for the search.
- Access full source detection catalogues for post-processing
- Load source catalogues (level 7) into archive and allow general user access
- Add new versions of source catalogues to archive, and retain previous versions
- Search all source catalogues

- Provide catalogues or parts of catalogues to external data centres – using the catalogue access tools. (No special tools will be provided but external data centres will be able to transfer catalogues or parts of catalogues and make these available through their services.)

Data validation notes

The science team expect to largely use automated statistical reports provided through the archive. However it may be necessary to access some visibility and/or image data products for data validation purposes

Other notes

The calibrated visibility data for EMU are likely to be shared with POSSUM. EMU will make use of Stokes I data only whilst POSSUM will use full polarisation.

The science team is likely to require access to working data storage and high performance computing for data validation and post-processing activities.

EMU source detections will be linked to corresponding POSSUM and FLASH continuum data.

Tables 9b: WALLABY

Science teams

PIs: B. Koribalski (CASS), L. Staveley-Smith (ICRAR)
Team includes ~ 80 individuals from 12 countries

Level 5/6 data products

- Spectral line data cubes. Up to four full-sized data cubes per field corresponding to total intensity (cleaned), total intensity (uncleaned), point spread function and sensitivity.
- Cut-out cubes with one total intensity (cleaned) cut-out cube for each detected source.
- Moment maps for detected sources corresponding to integrated intensity, mean velocity field and velocity dispersion. Moment maps will be produced for both full cubes and for cut-out cubes.
- Full resolution spectra for target positions (FITS and possibly PNG formats)
- Source detection catalogues with information derived from the individual data cubes
- Some postage stamp image cubes centred on positions of detected sources may be generated. Parameters and computing capability for postage stamp images to be further discussed.

Level 7 data products produced by Survey Science Team

- Clean image cubes produced with improved cleaning.
- Catalogues, cubes, masks, spectra and moment maps generated following additional source detection searches.
- Value added catalogues including multi-wavelength source properties obtained from non-ASKAP data and/or catalogues. `

CASDA use cases

- Fast transfer of high-volume data within the
- Pawsey Centre for data validation and post-processing purposes
- Access summary information on observation blocks completed including information on system performance and RFI etc.
- Access information on status of tasks submitted for processing at Pawsey Centre
- Access help from archive support staff
- Generate a summary of sky regions observed for project
- Access catalogue information on image quality
- Set data validation flags
- Access all or parts of data cubes
- Access moment maps
- Access HI spectra for set of sky positions
- Access information from EMU source detection catalogues
- Access source catalogues for post-processing
- Generate cut-out cubes for areas of interest from data cubes in archive
- Load value-added source catalogues (level 7) into archive and allow general user access
- Update science catalogues
- Search all source catalogues

Data validation notes

Calibrated visibility data files for spectral line data are not archived. During Early Science access to visibility data for a period of some days for validation purposes will be required. It will likely also be necessary to access some image data products for closer analysis.

Use of automated reports to facilitate validation is expected as system matures.

Other notes

The science team is likely to require access to working data storage and high performance computing for data validation and post-processing activities.

Table 9c: POSSUM

Science team

PIs: B. Gaensler (University of Sydney), T. Landecker (DRAO, Canada), R. Taylor (University of Calgary, Canada)

Team includes ~ 70 individuals from 15 countries

CASDA level 5/6 data products

- Full polarisation calibrated visibility data
- Restored, residual and model images cubes for Stokes I, Q, U and V with 300 spectral channels per cube, plus at least two image cubes for Stokes I point spread function and sensitivity.
- POSSUM Polarisation Catalogue
- Stokes I, Q, U and V spectra extracted from compact sources detected in the EMU catalogue (level 6).
- Noise spectra in Stokes I, Q, U and V measured around the detected sources.
- Model Stokes I spectrum (such as a low-order polynomial fit) for use in rotation measure synthesis algorithms (details still to be decided)
- 'Dirty' Faraday Dispersion Function created using the rotation measure synthesis algorithm. Output as a spectrum.
- Rotation Measure Spread Function calculated from the frequency sampling. Output as a spectrum.
- Cut-out cubes in Stokes Q, U and V for each detected source (details still to be decided).

Level 7 data products produced by Survey Science Team

- POSSUM Value Added Catalogue
- Stacked images for Stokes I, Q, U, V
- Additional products such as Faraday Depth cubes obtained using advanced polarisation analysis techniques
- Collation of additional polarisation information from other surveys
- Bayesian probability analysis of spatial matches to sources detected in other surveys.

CASDA use cases

- Access summary information on observation blocks completed including information on system performance and RFI etc.
- Access information on status of tasks submitted for processing at Pawsey Centre
- Access help from archive support staff
- Generate a summary of visibility files archived and sky regions observed
- Regular access to visibility data files (high data volumes may be required)
- Access catalogue information on image quality
- Set data validation flags following review of data validation report.
- Access all or parts of survey image cubes
- Access to EMU image data products
- Access information from EMU source detection catalogues

- Access information from POSSUM catalogues
- Access full source catalogues for post-processing
- Load value-added source catalogues (level 7) into archive and allow general user access
- Add new versions of source catalogues to archive, and retain previous versions
- Search all source catalogues

Data validation notes

The science team will largely use automated statistical reports provided through the archive. However it may be necessary to access some data products for closer analysis.

Other notes

The level 5/6 data products that will be provided through the pipeline data processing are still under discussion. Additional data cubes for PSF and/or sensitivity may be included.

The calibrated visibility data for EMU are likely to be shared with POSSUM. EMU will make use of Stokes I data only whilst POSSUM will use full polarisation.

The science team is likely to require access to working data storage and high performance computing for data validation and post-processing activities.

Tables 9d: DINGO

Science team

PI: M. Meyer (University of Western Australia)
Team includes ~ 40 individuals from 6 countries

Level 5/6 data products

- Spectral line data cubes. For each project there are three cubes per survey field corresponding to total intensity (cleaned), sensitivity and point spread function.
- Moment maps corresponding to integrated intensity, velocity field and velocity dispersion.
- Source detection catalogues with information derived from the individual data cubes
- Full resolution spectra for target positions (FITS and possibly PNG formats)

Level 7 data products produced by Survey Science Team

- Final Survey Science catalogues with full parameterisation of sources
- Stacked image cubes
- Cut-out image cubes
- Cross-identifications against other catalogues
- Additional data visualisation

CASDA use cases

- Fast transfer of high-volume data purposes within Pawsey Centre for data validation and post-processing purposes
- Access summary information on observation blocks completed including information on system performance and RFI etc.
- Access information on status of tasks submitted for processing at Pawsey Centre
- Access help from archive support staff
- Generate a summary of sky regions observed for project
- Access catalogue information on image quality
- Set data validation flags
- Access all or parts of data cubes
- Access moment maps
- Access HI spectra for set of sky positions
- Access information from EMU source detection catalogues
- Access source catalogues for post-processing
- Load stacked image data cubes into archive (desirable but not initially supported)
- Generate ‘cut-out’ cubes for detected sources from data cubes in archive
- Load value-added source catalogues (level 7) into archive and allow general user access
- Update science catalogues
- Search all source catalogues

Data validation notes

Data validation should largely use automated statistical reports provided through the archive. Calibrated visibility data files for spectral line data are not archived. However, some access to visibility data for a period of some days for validation purposes may be required. It may be necessary also to access some image data products for closer analysis.

Other notes

The science team is likely to require access to working data storage and high performance computing for data validation and post-processing activities.

The DINGO team will make final stacked images available to the community. Tools to provide SST images will not be available in initial releases of CASDA but may be considered for later releases.

Tables 9e: FLASH

Science teams

PI: E. Sadler (University of Sydney)
Team includes ~ 35 individuals from 6 countries

Level 5/6 data products

- Postage stamp image cubes corresponding to total intensity, sensitivity and point spread function, for maximum spatial resolution and spectral coverage of 16,200 channels.
- Either Taylor continuum images or continuum image cubes to enable reliable measurements of continuum flux densities.
- Two-dimensional moment maps corresponding to the velocity field and velocity dispersion.
- Source detection catalogues with information derived from the individual data cubes
- Full resolution spectra for target positions (FITS and possibly PNG formats)
- **Desirable**: Ability to retain full-sized data cubes.

Level 7 data products produced by Survey Science Team

- Final Survey Science catalogues with full parameterisation of sources. (The project will produce approximately 150,000 spectra at positions given in the Target Source Catalogue.)
- Target Source Catalogue
- Cross-identifications against other catalogues
- Additional data visualisation

CASDA use cases

- Fast transfer of high-volume data purposes within Pawsey Centre for data validation and post-processing purposes
- Access summary information on observation blocks completed including information on system performance and RFI etc.
- Access information on status of tasks submitted for processing at Pawsey Centre
- Access help from archive support staff
- Generate a summary of sky regions observed for project
- Access catalogue information on image quality
- Set data validation flags
- Access all or parts of data cubes
- Access moment maps
- Access selected set of HI spectra
- Access information from EMU source detection catalogues
- Access data files and catalogues for post-processing
- Generate ‘cut-out’ cubes
- Load value-added source catalogues (level 7) into archive and allow general user access
- Update science catalogues
- Search all source catalogues

Data validation notes

Data validation should largely use automated statistical reports provided through the archive. Calibrated visibility data files for spectral line data are not archived. However, some access to visibility data for a period of some days for validation purposes may be required. It may be necessary also to access some image data products for closer analysis.

Other notes

The FLASH project has a high level of computing power and memory requirements. Parameters for the image data products will be further refined and clarified as experience is gained during commissioning and Early Science. The potential and feasibility of postage stamp image cubes is not yet fully determined. As an alternative to retaining postage stamp images, an option may be to retain a single full-sized cube cube for each field, at the highest possible spatial resolution.

The science team is likely to require access to working data storage and high performance computing for data validation and post-processing activities.

Tables 9f: GASKAP

Science teams

PIs: J. Dickey (University of Tasmania), N. McClure-Griffith (CASS)

Team includes ~ 78 individuals from 11 countries

Level 5/6 data products

- Spectral line data cubes. Equivalent in data volume to three full-sized data cubes with 16,200 spectral channels per field corresponding to total intensity, point spread function and sensitivity. Each of the full sized cubes is (in effect) split into three smaller cubes with 50%, 25% and 25% of the channels corresponding to three zoom bands for HI, OH 1612 and OH1665/1667.
- Moment maps corresponding to velocity field and velocity dispersion
- Continuum images or cubes to enable reliable measurements of continuum flux densities in spectral line absorption data.
- Cut-out data cubes at positions of compact sources detected in HI or OH
- Full resolution spectra extracted at positions of detected compact sources
- Source detection catalogues for detected compact sources
- Desirable: Final data cubes with combined single dish plus ASKAP data (if feasible)

Level 7 data products produced by Survey Science Team

- Final Survey Science Catalogues for detected sources (may include Gaussian fits to spectral features)
- HI and OH absorption catalogues
- Final data cubes with combined single dish plus ASKAP data (if not produced in data processing pipeline)

CASDA use cases

- Fast transfer of high-volume data within the Pawsey Centre for data validation and post-processing purposes
- Access summary information on observation blocks completed including information on system performance and RFI etc.
- Access information on status of tasks submitted for processing at Pawsey Centre
- Obtain help from archive support staff
- Generate a summary of sky regions observed for project
- Access catalogue information on image quality
- Set data validation flags
- Access all or parts of data cubes
- Access moment maps
- Access HI spectra for set of sky positions
- Access information from EMU source detection catalogues
- Access source catalogues for post-processing
- Generate cut-out cubes from data cubes in archive
- Search all source catalogues
- Load value-added source catalogues (level 7) into archive and allow general user access

- Update level 7 science catalogues
- *Desirable:* Load final set of ‘combined’ data cubes to the archive – if these are not created during pipeline processing

Data validation notes

Data validation should largely use automated statistical reports provided through the archive. Visibility data are not archived. However, some access to visibility data for a period of some days for validation purposes may be required. It may be necessary also to access some image data products for closer analysis.

Data validation is likely to include analysis of dynamic range, bandpass stability and continuum subtraction.

Other notes

GASKAP science data processing requirements present strong challenges. To obtain the full scientific value from the survey it will be necessary to merge data obtained from other radio telescopes with the ASKAP data. The level 5/6 data products listed above will have limited scientific value for ASKAP-only data. If feasible the data merging will be carried out as part of the pipeline data processing. Otherwise this will be carried out as a level 7 activity.

The science team may assist with the pipeline data processing for the survey and will require access to working data storage and high performance computing for data validation and post-processing activities.

Table 9g: VAST

Science team

PI: T. Murphy (University of Sydney), S. Chatterjee (Cornell University, USA)

Team includes ~ 75 individuals from 10 countries

Level 5/6 data products

- Calibrated visibility data
- Source Detection Catalogue that includes a variability flag to indicate a detected source is variable on a timescale of 5s or longer
- Light Curve Catalogue
- Postage stamp or cut-out image cubes corresponding to transient detections may be retained.

Level 7 data products produced by Survey Science Team

- Transient Source Detection Catalogue with identifications and cross-associations
- Light curves extracted and analysed for transient/variable sources
- Stacked image cubes to search for weaker transients (to be confirmed)
- Polarisation images may be generated after detection of a transient (to be confirmed)

CASDA use cases

- Access summary information on observation blocks completed including information on system performance and RFI etc.
- Access information on status of tasks submitted for processing at Pawsey Centre
- Access help from archive support staff
- Generate a catalogue summary of visibility files archived and sky regions observed
- Access visibility files to generate polarisation images
- Access catalogue information on image quality
- Set data validation flags following review of data validation report.
- Access survey images or parts of images
- Access information from source detection catalogues
- Access information from light curve catalogues
- Record transient event information and communication activities following an event
- Record transient event information subsequently provided by the science team
- Load source catalogues (level 7) into archive and allow general user access
- Add new versions of source catalogues to archive, and retain previous versions
- Search all source catalogues

Data validation notes

Discussion is needed given to establish data validation procedures, given the stringent requirements to generate source detection and transient information on very short timescales.

Other notes

The science team will require access to working data storage and high performance computing for data validation and post-processing activities. The archive requirements to support VAST are not fully determined. The science data pipeline processing and archive specifications for transient observations will be developed further as experience is gained during ASKAP Early Science.

For commensal observations the visibility data used for VAST will be shared with other projects.

Table 9h: COAST

Note: The data archive requirements for COAST are considered here. However, the pulsar data may be managed together with pulsar data from other radio astronomy facilities. These data are provided through the DAP with primary data storage in Canberra.

COAST Team

PI: I Stairs (University of British Columbia)

Team includes ~ 35 individuals from seven countries.

Level 5/6 data products

None: No pipeline data processing provided by ASKAP project

Level 7 data products produced by Survey Science Team

- PSRFITS format timing data processed for de-dispersion and folding
- PSRFITS time series data derived from timing observations
- PSRFITS search mode data for targeted observations
- Candidate detection catalogues
- Fast dump visibility files

Pulsar data use cases

- Transfer and publish data files from timing mode observations to archive
- Access information on status of tasks submitted for processing
- Access help from archive support staff
- For timing data provide thumbnail images for folded pulse profiles
- Transfer and publish data files from targeted search mode observations to archive
- Transfer and publish pulsar candidate catalogues to the archive
- Search candidate catalogues to retrieve candidate lists for further observations
- Access and transfer pulsar files through web user interface and/or through VO services
- Set validation flags to withhold or release data as appropriate (no embargo period for ASKAP pulsar data but data validation applies)

Data validation

Standard pulsar techniques will be used to assess data quality

Other notes

Further discussion will be held to determine whether ASKAP pulsar data will be stored in the Pawsey Centre in Perth or using CSIRO facilities in Canberra.

The science team is likely to require access to working data storage and high performance computing for data validation and post-processing activities. This is not provided by CASDA. The CSIRO Bragg computer in Canberra is now being used to trial the use of HPC facilities with Parkes pulsar data with a fast connection between the data archive disks and super computer.

Table 9i: CRAFT

Note: CRAFT has challenging technical requirements (section 3.3.9). Data processing for CRAFT will be carried out by the science team instead of as part of the ASKAP Science Data Processing pipelines. Although the data processing will be 'offline', CASDA may provide archive facilities for some CRAFT data products. The archive requirements for this project at the Pawsey Centre are not yet well established and further advice from the science team will be sought.

Science Team

PI: P. Hall (ICRAR/Curtin University), J. P. Macquart (ICRAR/Curtin University)

Team includes ~ 40 individuals from four countries.

Level 5/6 data products

No pipeline data processing provided by ASKAP project

Level 7 data products produced by Survey Science Team (preliminary list only)

- Baseband data recorded in data buffers corresponding to times of potential transient detections (if feasible)
- Event trigger information (time, flux, duration, frequency etc)
- Images produced from buffered data for detected sources
- Source-detection information for detected sources
- Event handling information (community information, emails sent)
- Source identifications and associations (possibly in real time)
- Fast Transient Source Catalogue

CASDA use cases (preliminary list only)

- Access summary information on observation blocks completed including information on system performance and RFI etc.
- Access information on status of tasks submitted for processing at Pawsey Centre
- Access help from archive support staff
- Load baseband data corresponding to the times of potential transient detections
- Load catalogues generated by science team
- Access and retrieve information from catalogues
- Load images generated by science team
- Access and retrieve information from images
- Set data release flag in CASDA after off-line data validation

Data validation

Further discussion is needed given to establish data validation procedures, given the stringent requirements to generate transient detections on very short timescales.

Table 9j: VLBI

Science Team PI: Steven Tingay (Curtin University) Team includes ~ 20 individuals from 5 countries.
Level 5/6 data products No pipeline data processing provided by ASKAP project
Level 7 data products produced by Survey Science Team None relevant to CASDA
CASDA use cases None
Data validation As for standard VLBI data processing. No specific CASDA requirements.

6.3 Use cases for science users

There is considerable overlap between the data products and use cases for the different Survey Science projects. Table 10 provides a 'merged' summary of use cases for Survey Science teams, Guest Science teams and general science users. Columns 2 to 4 give an indication of the estimated average number of users per day where:

A) 0 = none, B) 1 – 5, C) 5 – 50, D) 50 or more

Table 10: Summary of use cases for science users

	Estimated average number of science users per day		
Use case	**General science users**	**Guest Science Teams**	**Survey Science Teams**
Login to archive using OPAL or Nexus accounts.	D	B	C
Access online information on how to use the archive.	C	C	B
Work through online demonstration tutorials.	B	B	B
Provide user satisfaction feedback and comments including suggestions for improvements using online form.	B	B	B
Send a request for support to the CASDA administrator.	B	B	B
Obtain report that shows the status of current activities at the Pawsey Centre and status of queued tasks.	A	B	C
Obtain report with information from the observing schedules and scheduling blocks.	A	B	C
Obtain report with information on system performance including RFI and calibration information.	A	B	C
Run a cone search to obtain a listing of all ASKAP observations taken for given region of sky.	C	C	C
Generate more general report(s) to summarise all ASKAP observations taken to date.	C	C	C
Set data validation flags at the data products level and enter information on data-related problems identified.	A	A	B
Provide a user-generated catalogue with a list of source positions and use this to obtain matches against ASKAP catalogues.	D	C	C
Carry out searches of source catalogues using complex queries. For example find all sources with $0 < z < 0.5$.	C	C	C
Obtain report with image quality information for a selected set of images or image cubes.	A	A	B

Obtain report with information from continuum source detection catalogues.	C	B	C
Obtain report with information from spectral line source detection catalogues.	B	B	C
Obtain report with information from transient source detection catalogues.	B	B	C
Obtain report with information from light curve catalogues.	B	B	C
Obtain report with information from polarisation catalogues.	B	B	C
Transfer complete catalogues, or parts of catalogues,for post-processing analysis.	A	A	B
View displays on browser of images, moment maps and spectra without data transfer	C	C	C
Transfer set of selected visibility data files	A	A	B
Transfer set of continuum images or image cubes for data validation purposes. Provide capability to transfer cut-outs for science or outreach purposes	A	A	B
Transfer set of spectral line images or image cubes for data validation purposes. Provide capability to transfer cut-outs.	A	A	B
Transfer set of continuum images or image cubes or cut-out cubes for post processing purposes.	A	B	B
Transfer set of spectral line images or image cubes or cut-out cubes for post processing purposes.	A	B	B
Transfer moment maps and or spectra from spectral line surveys.	B	B	B
Access facilities to enable data validation and/or post-processing purposes within the Pawsey Centre.	A	B	C
Load data products generated outside of the Pawsey Centre such as files associated with the detections of fast transients.	A	A	B
Deposit (upload) level 7 Survey Science Catalogues.	A	A	B
Deposit (upload) additional versions of level 7 Survey Science Catalogues.	A	A	B
Desirable use cases			
Use Google-Map type interactive tool to navigate the sky and obtain information on archive contents.	C	C	C
Access data, for a given set of positions, from other archive services such as NED or SIMBAD	C	C	C
GASKAP only Load final set of ‘combined’ high resolution data cubes to	A	A	B

the archive – *if these are not created during pipeline processing.*			
DINGO only Load final set of stacked data cubes .	A	A	B
CRAFT only Load data products associated with Fast Transient detections (full specification to be determined)	A	A	B

6.4 Other use cases

Table 11 lists other use cases associated with CASDA administration and support. This list is not intended to be complete. Broadly, the day-to-day archive administration and user support will be provided by CASS (Tier 1 support). The system software and user interfaces will be supported by CSIRO IM&T (Tier 2 support) whilst the Pawsey Centre infrastructure will be supported by iVEC (Tier 3 support).

Table 11: Other use cases

Task
Ingest from the ASKAP Central Processor and Lustre file system metadata describing the science data products to be archived. This includes an enumeration of the data products, metadata describing the data products, and metadata describing the configuration of the telescope at the time the observations were carried out.
Ingest from the ASKAP Central Processor data product files for images and image cubes, visibilities and tables as are described by the metadata
Manage user access conditions: User groups include: • Administrators and developers • Survey Science Team members • Guest Science Team members • General astronomy community
Set proprietary period for Guest Science Projects (default is zero) where a proprietary period for a project is approved by the Time Assignment Committee.
Manage archive queues. Assign priorities to tasks based on pre-determined conditions with goal of ensuring fair access.
Adjust queues when needed to ensure fair access to archive services

Monitor system performance. Measures may include: Data transfer speeds between different system areas Amount of downtime due to CASDA faults Amount of downtime due to other (infrastructure) faults Response time to restore system following faults Percentage of data 'lost' in given year.
Trigger alerts to archive administrators following any interruptions to normal performance.
Provide statistical information on archive usage including: Number of users User demographics (CSIRO, Australia, overseas preferably by country) Volume of data transferred to other locations Volume of data archived to tape Volume of data recovered from tape
Transfer data from tape to disk.
Regular backups of the science archive database and source catalogues
Re-ingest a set of files after they have been modified. Update existing metadata in the archive database and associated indexes.
Provide access to data (subject to approvals) within the Pawsey Centre

APPENDICES

Appendix A: Data volumes

Data volumes used in this document have been using the following formulae:

Correlated visibilities

The data volume for a set of correlated visibilities is given by:

NB x NP x NC x NBAS x NS x Vvis where

NB = number of beams

NP = number of polarisations

NC = number of channels

NBAS = number of baselines plus auto-correlations

NS = number of samples

Vvis = volume per complex visibility

For ASKAP data Vvis = 9 bytes (including 1 byte for weighting). NS = Total integration time/time per correlated sample. For ASKAP the time per sample is 5s.

Image cubes

The data volume for a 'standard' image cube is given by:

Nx x Ny x NC x Vvox where:

Nx = number of pixels in one direction on the sky (usually right ascension or longitude)

Ny = number of pixels in the perpendicular direction on sky (usually declination or latitude)

Vvox = volume per voxel = 4 bytes.

Voltage stream from phased array feeds

The total data volume for the voltage stream out from from the beamformers at the phased array feeds is given by:

NA x NB x NP x NS x Vsam where:

NA = number of antennas

Vsam = data volume per sample = 1 byte.

For a bandwidth of 300 MHz the sampling rate is 600 MHz. Thus, for two polarisations, 36 antennas, 36 beams the full data rate for the ASKAP voltage stream from all antennas and beams is ~ 12 Tbps.

Total power sampled from phased array feeds beam formers

The data volume for sampling the total power (auto-correlations) data streams from the phased array feed beams is given by:

NA x NB x NC x NP x NS x Vax where

Vax = volume per auto-correlation = 2 bytes

As an example, if the auto-correlation data streams are sampled at a time resolution of 1 ms then for the beam former spectral resolution of 1 MHz, the total power data rate for a single polarisation output is ~ 6 Gbps.

Appendix B: CASDA data products

Table B1: CASDA Data Products

	Visibility Data			
Ref	**Visibility Data**	**Sub-types**	**Stokes polarisation products archived**	**Notes**
P1	Calibrated continuum visibility data	Continuum only	I Q U V	Visibilities may be stored in either CASA or FITS format.
	Continuum frequency synthesis images			
	Image type	**Image sub-type (what the image measures)**	**Stokes polarisation products**	**Notes**
P2	Restored	Intensity at fixed frequency Spectral index Spectral curvature	I	Three types of Taylor term images. All image-related data products will be stored as FITS single-channel image files.
P2	Residual	Intensity at fixed frequency Spectral index Spectral curvature	I	See above
P2	Model	Intensity at fixed frequency Spectral index Spectral curvature	I	See above
P2	Point spread function	Instrumental response to Point spread	I	Generated for each Taylor term image
P2	Sensitivity	Sensitivity	I	Generated for each Taylor term image
P2	Mask	Pixels or voxels with emission above a threshold value	I	Masks are (sometimes) used to identify regions of interest within a cube or cut-out cubes. Other areas in the cube are 'blanked out'.
	Continuum image cubes with polarisation			
P2	Restored	Intensity	I, Q, U, V	Output as multi-channel FITS format image cubes
P2	Residual	Intensity	I, Q, U, V	

P2	Model	Intensity	I, Q, U, V	
P2	Point spread function	Instrumental response to Point spread	I, Q, U, V	
P2	Sensitivity	Sensitivity	I, Q, U, V	
P2	Mask	Determines pixels or voxels with emission above a threshold value	I, Q, U, V	See above
P16	Polarisation spectra	Polarisation-related parameters.	I, Q, U, V and other quantities	See Table 9c
	Spectral Line Image cubes and derived image products			
	Type	**Image sub-type**	**Polarisation products**	**Notes**
P3	Restored	Intensity	I	Cleaned cube
P3	Residual	Intensity	I	
P3	Uncleaned	Intensity (uncleaned)	I	Also known as ‘dirty’ cube
P3	Model	Intensity	I	
P3	Mask	Determines voxels with emission above a threshold value	I	
P4	Postage stamp cubes	Intensity	I	
P5	Moment maps	M0: averaged intensity M1: velocity M2: velocity dispersion	I	Moment maps are two-dimensional images. The three types of moment maps are often used for HI studies of galaxies.
P6	Spectra	Intensity	I	Output as (one dimensional FITS spectra) May also be retained as png or other image format for quick look purposes
	Level 5/6 catalogues			
P7	Calibration and system information			
P8	Scheduling information			
P9	Global Sky Model			The full Global Sky Model will be held on the RTC for use with the science data processing prior. The archive may store periodic ‘snapshots’ of the model (or parts of the model) to make this

		available to the community.
P10	Image quality reports	
P11	Continuum source detection catalogues	
P12	Polarisation properties catalogue	Includes rotation measures
P12	Polarisation atlas	Includes frequency-dependent information
P13	Spectral line source detection catalogues	
P14	Transient source catalogues	
	Level 7 data products supported in CASDA (to be confirmed)	
P15	Target source catalogues	To be confirmed
P15	Survey Science catalogues	Generated by the Survey Science Teams.

Appendix C: Survey parameters

The tables in this appendix give parameters for the Survey Science Projects that will be processed at the Pawsey Centre. These are intended to indicate the data sizes for data processing and for archiving purposes. Values given correspond to different stages of the data flow, from the Telescope Operating System, through the data ingest, calibration and imaging pipelines to the archive. In most cases the parameters given are taken from [2]. Note that the image sizes are based on nominal values for cell sizes and number of pixels. These may be slightly revised as the surveys become better defined.

Note that scheduling arrangements for ASKAP are not yet in place whilst the actual allocation of time and use of commensal observing are still to be determined. The tables correspond to indicative parameters for full ASKAP capabilities with 36 fully equipped antennas. In practice there will be an extended period between the start of Early Science and full array operation with data rates ramping up over time.

Table C1: Survey parameters for Emu and POSSUM

	EMU	**POSSUM**
Project Information		
Project Code	AS014	AS007
Rating	**1**	**2**
Survey type	Continuum	Continuum
Array Parameters		
Number of antennas	36	36
Maximum baseline (km)	6	6
Number of baselines (includes auto-correlations)	666	666
Number of beams	36	36
Frequency channels from the correlator	16,200	16,200
Number of polarisations	4	4
Bytes per complex sample (includes 1 for weight)	9	9
Data volume per visibility sample (GB)	13.98	13.98
Data sizes and rates		
Frequency channels after averaging	300	300
Number of polarisations in calibrated visibilities	4	4
Integration time (s)	5	5
Data rate (Gbits/s): Correlator to Real Time Computer prior to channel averaging	22.37	22.37
Averaged visibility frame (GB)	0.26	0.26
Averaged data rate (GB/s)	0.052	0.052
Averaged data rate (TB/h)	0.19	0.19

Observing time (h)	12	8
Averaged integration time (s)	5	5
Averaged visibility data set per field (TB)	**2.24**	**1.49**
Image sizes for processing		
Number of image polarisations archived	1	4
Number of channels for images or image cubes	1	300
Field of view (degrees)	7.5	7.5
Cellsize (arcsec)	2.5	2.5
Full image size (pixels)	10,800	10,800
Full image size (degrees)	7.5	7.5
Total size for single image (EMU) or image cube (POSSUM) (GB)	0.47	140.0
Number of images or image cubes per field	15	14
Image size per field (GB)	**7.0**	**1,960**
Catalogues [initial estimates]		
Number of rows for full survey	70 million	70 million
Data size per row (Bytes)	300	300
Image sizes for archiving		
Final 1-d image or cube size (pixels)	7,800	7,800
Single image/cubesize (GB)	0.24	**72.6**
Image size per field (GB)	3.6	**1,016**
Fields per survey	1,200	1,200
Survey total sizes		
Survey size: images (TB)	**4.4**	**1,230**
Survey size: visibilities (PB)	**2.7**	**1.8**
Survey size: catalogues (GB)	**21**	**small**

Notes:

a) EMU image sizes are for single-channel images.

b) Data volumes are given separately for EMU and POSSUM and are currently based on 12 hours per field for EMU and 8 hours per field for POSSUM. In reality these two

projects are likely to be observed commensally and to make use of the *same* calibrated visibilities. This reduces the total visibility data volume that needs to be archived.

Table C2: Survey parameters for WALLABY

	WALLABY 2-km array	**Cut-out cubes (example)**	**Postage stamps (example)**
Project Information			
Project Code	AS016		
Rating	1		
Survey type	SL		
Array Parameters			
Number of antennas	36		
Maximum baseline (km)	2	2	6
Number of baselines (includes auto-correlations)	666		
Number of beams	36		
Frequency channels from the correlator	16,200		
Number of polarisations	4		
Frequency channels after averaging	16,200		
Data sizes and rates			
Bytes per complex sample (includes 1 for weight)	9		
Data volume per visibility sample (all pols) (GB)	13.98		
Number of polarisations	4		
Integration time (s)	5		
Data rate (Gbits/s)	22.37		
Averaged data rate (GB/s)	2.8		
Averaged data rate (TB/h)	10.08		
Observation time per field (h)	8		
Averaged visibility data set per field (TB)	**80.54**		
Image sizes for processing			
Number of image polarisations	1	1	1
Number of channels for images or image cubes	16,200	128	512
Cellsize (arcsec)	7.5	7.5	2.5
Full image size (pixels)	3600	32	256
Full image size (degrees)	7.5	0.067	0.178
Total size for single image cube (GB)	839	0.00052	0.134
Number of image cubes per field	4	2000	2000
Image size per field (TB)	**3.4**	**0.001**	**0.268**
Catalogues [initial estimates]			

Number of catalogue rows per scheduling block	420		
Number of rows for full survey	500,000		
Data size per row (Bytes)	300		
Survey Sizes			
Final image cube size (1-d pixels)	2,600	32	256
Single image cube size (GB)	438	0.00052	0.134
Image size per field (GB)	1,752	1.05	268
Fields per survey	1,200	1,200	1,200
Survey size images (PB)	**2.10**	**0.0013**	**0.32**
Survey size images (PB)	**2.12 (excluding postage stamps)**		
Survey Size visibilities (PB)	**96.6 (not archived)**		
Survey size catalogues (GB)	**0.15**		

Notes:

a) Spectral line visibility data are not archived.

b) Image parameters in column 2 correspond to baselines smaller than 2 km. Cubes are made using full spectral coverage (16,200 channels) with a pixel size of 7.5 arcsec. Four cubes are made for each field corresponding to the total intensity (cleaned), total intensity (uncleaned), sensitivity and point spread function.

c) Column 3 corresponds to a set of four cut-out cubes for each detected source. For an assumed 500 detections per field this corresponds to 2000 cut-out cubes per field. The cube dimensions of 32 x 32 pixels x 128 channels (pixel size = 7.5 arcsec) are given as an example only. It is expected that cut-out cubes will be produced as part of the pipeline imaging with final parameters (including the number of cut-out cubes) still to be determined. For each detected source it is likely that, as a minimum, a Stokes I cut-out cube and a mask cube will be generated.

d) Column 4 is included as an example for postage stamp cubes. This again assumes a set of four cubes per detection with 500 detections per field and cubes sizes of 256 x 256 x 512 pixels (pixel size = 2.5 arcsec). Note that the inclusion of postage stamp image cubes as part of the pipeline processing is desirable but is likely to be limited by computing power (section 2.5.2).

e) Full resolution spectra and moment maps will also be archived and will add a small amount to the total data volume.

Table C3: Survey parameters for DINGO and FLASH

	DINGO	**FLASH**
Project Information		
Project Code	AS012	AS002
Rating	2	2
Survey type	SL	SL
Array Parameters		
Number of antennas	36	36
Maximum baseline (km)	2	6
Number of baselines (includes auto-correlations)	666	666
Number of beams	36	36
Frequency channels from the correlator	16,200	16,200
Number of polarisations	4	4
Frequency channels after averaging	16,200	16,200
Data sizes and rates		
Bytes per complex sample (includes 1 for weight)	9	9
Data volume per visibility sample (all pols) (GB)	13.98	13.98
Number of polarisations	4	4
Integration time (s)	5	5
Data rate (Gbits/s)	22.37	22.37
Visibility frame after any channel averaging (GB)	13.98	13.98
Averaged data rate (GB/s)	2.8	2.8
Averaged data rate (TB/h)	10.08	10.08
Observation time per field (h)	8	2
Averaged visibility data set per field (TB)	**80.54**	**20.14**
Image sizes for processing		
Number of image polarisations	1	1
Number of channels for images or image cubes	16,200	16,200
Cellsize (arcsec)	7.5	2.5
Full image size (pixels)	3600	128
Full image size (degrees)	7.5	0.09
Total size for single image cube (GB)	839	1.06
Number of images per field	3	450
Image size per field (TB)	**2.5**	**0.48**
Catalogues [initial estimates]		

Number of catalogue rows per scheduling block	tba	150
Number of rows for full survey	tba	128,000
Data size per row (Bytes)	300	300
Survey Sizes		
Final 1-d image or cube size (pixels)	2,600	128
Single image/cubesize (GB)	438	1.06
Image size per field (GB)	1,314	477.8
Fields per survey	**966**	**850**
Survey size images (TB)	**1,270**	**406**
Survey Size visibilities (PB)	**77.8**	**17.1**
Survey size catalogues (GB)	Tba	**0.038**

Notes:

a) Visibility data are not archived.

b) DINGO: The 966 survey fields is estimated as 68 repeats on five survey fields and 312 repeats on two survey fields.

c) FLASH: 850 survey fields. Parameters are given assuming that three sets of postage stamp data cubes are made for each survey field corresponding to total intensity, point spread function and sensitivity. Data cubes produced for 150 pointings within each survey field. Continuum images or image cubes will also be retained and will add a small additional volume of data.

Table C4: Survey parameters for GASKAP

	2-km array
Project Information	
Project Code	AS005
Rating	2
Survey type	SL
Array Parameters	
Number of antennas	36
Maximum baseline (km)	2
Number of baselines (includes auto-correlations)	666
Number of beams	36
Frequency channels from the correlator	16,200
Number of polarisations	4
Frequency channels after averaging	16,200
Data sizes and rates	
Bytes per complex sample (includes 1 for weight)	9
Data volume per visibility sample (all pols) (GB)	13.98
Number of polarisations	4
Integration time (s)	5
Data rate (Gbits/s)	22.37
Averaged data rate (GB/s)	2.8
Averaged data rate (TB/h)	10.08
Observation time per field (h)	12.5
Averaged visibility data set per field (TB)	125.8
Image sizes for processing	
Number of image polarisations	1
Number of channels for image cubes	16,200
1-d field of view (degrees)	7.5
Cellsize (arcsec)	7.5
Full image size (pixels)	3,600
Full image size (degrees)	7.5
Total image size (GB) single image	839
Number of images per field	3
Image size per field (GB)	2517
Catalogues [initial estimates]	

Number of catalogue rows per scheduling block	50
Number of rows for full survey	30000
Data size per row (Bytes)	300
Survey Sizes	
Final 1-d image or cube size (pixels)	2600
Single image/cubesize (GB)	**438**
Image size per field (GB)	1312
Scheduling blocks per survey	644
Fields per survey	481
Survey size image cubes (TB)	846
Survey size – all image cubes TB)	880
Survey Size visibilities (PB)	125.8
Survey size catalogues (MB)	9

Notes:

a) Parameters for GASKAP are adapted from [9]. A total observing time of 8050 hours is assumed to be taken over 644 blocks of 12.5 hours. To increase the sensitivity, some fields are observed more than once.

b) Observations will be taken using three zoom bands to cover the frequency ranges for HI, OH 1612 MHz and OH 1665/1667 MHz.

c) To estimate the data volumes, parameters are given for three spectral line cubes per field with 16,200 channels. These will be generated as three smaller cubes with 50%, 25% and 25% of the channels correspond to three zoom bands.GASKAP cubes may have fewer than the full 16,200 channels, with some channels used for continuum.

d) A total of 30,000 point-like source detections is assumed (approximately 15,000 for HI and 15,000 for OH). This corresponds to an average of about 50 detections per 12.5 h scheduling block.

e) Cut-out cubes for detected sources are likely to be retained. The total volume for these is small compared to the full-sized cubes.

f) Higher-resolution postage stamp image cubes may be desirable but are likely to be limited by available computing power (section 2.5.2).

Table C5: Survey Parameters for VAST

	VAST	Notes
Project Information		
Project Code	AS004	
Rating	2	
Survey type	transient	
Array Parameters		
Number of antennas	36	Full array
Maximum baseline (km)	6	
Number of baselines (includes auto-correlations)	666	
Number of beams	36	
Frequency channels from the correlator	16,200	
Number of polarisations	4	
Frequency channels after averaging	30	
Data sizes and rates		
Bytes per complex sample (includes 1 for weight)	9	
Data volume per visibility sample (all pols) (GB)	13.98	
Number of polarisations	4	Assumes full polarisation VAST polarisation analysis is still quite uncertain
Integration time (s)	5	
Data rate (Gbits/s): Correlator to Real Time Computer prior to channel averaging	22.37	
Averaged visibility frame (GB)	0.026	30 spectral channels
Averaged data rate (GB/s)	0.052	
Averaged data rate (TB/h)	0.019	
Observing time (h)	12	
Averaged visibility data set per field (TB)	**0.224**	
Image sizes for processing		
Number of image polarisations	1	Assumes only Stokes I is imaged
Number of channels for images or image cubes	30	
Field of view (degrees)	7.5	
Cellsize (arcsec)	7.5	
Full image size (pixels)	3,600	

Full image size (degrees)	7.5	
Total image size (GB) single image	1.56	
Number of images per field	8,640	For 12 hours observing
Total image volume for 12 hours (TB)	**13.4**	
Catalogues [initial estimates]		
Number of catalogue rows for full survey	5.2 billion	No compression
Data size per row (Bytes)	300	
Image sizes for archiving		
Final 1-d image or cube size (pixels)	2,600	
Single image/cubesize (GB)	0.81	
Image size per field (GB)	7,000	8640 image cubes in 12 hours**.**
Fields per survey	1,200	Piggy-back mode
Survey total sizes		
Survey size: images (PB)	**8.4**	Total image volume generated after 1200 blocks of 12 hours
Survey size: visibilities (PB)	**0.27**	
Survey size: catalogues (GB)	see note (c)	

Notes:

a) VAST observations may be taken using a wide range of parameters depending on the system set up for other scheduled projects. The parameters given here correspond to 12 hours of observations with 30 spectral channels and full polarisation visibility data retained with images formed for one polarisation only.

b) The full-sized VAST image cubes are unlikely to be retained. However, postage stamp or cut-out image cubes for potential transient detections may be archived.

c) The number of catalogue rows for VAST is potentially very large. The number of rows given here (five billion) is based on an estimate of 500 detections every five seconds with every detection retained as a separate row. With no data compression the approximate size of a catalogue would be about 1.5 TB. However, by splitting the information, into separate catalogues for source detections and light curves, it should be possible to greatly compress the data volume for the transient catalogues to manageable levels below ~50 GB.

REFERENCES

1. Bock, D., Chapman, J., Lensson, E., Edwards, P., (2012), ATNF Operations in the *ASKAP Era, version B*
2. Cornwell, T, Humphreys, B., Lenc, E., Voronkov, M., Whiting, M., (2011) *ASKAP Science Processing*, ASKAP-SW-0020
3. Feian, et al., (2009), *ASKAP User Policy*, http://www.atnf.csiro.au/projects/askap/UserPolicy_final.pdf
4. Humphreys, B., (2011) *ASKAP Central Processor Pawsey Centre Requirements Document*, ASKAP-SW-0021
5. Humphreys, B., Guzman, J.C., Marquarding, M., Cornwell, T., Voronkov, M., Brodrick, D., *ASKAP Computing Architecture*, ASKAP-SW-0003
6. Norris, R., Johnston, S., (2009), *ASKAP Science Data Archive: Draft Requirements Document*, ASKAP-SC-0001, version 1.0
7. Whiting, M., (2012), *Duchamp: a 3D source finder for spectral-line data*, MNRAS, 421, 3242
8. Whiting, M., Humphreys, B., (2012), *Source-finding for the Australian Square Kilometre Array Pathfinder*, PASA, 29, 371 – 381
9. Dickey, J. M., McClure-Griffiths, N. et al., (2013) *The Galactic ASKAP Survey*, PASA 30,3
10. ASKAP Prject, ASKAP Requirements, (2013), version 0.2, ASKAP-SEIC-0007
11. Lorimer, D., et al., (2007*), A Bright Millisecond Radio Burst of Extragalactic Origin*, Science, 318, 777
12. Thornton D., et al., (2013), *A Population of Fast Radio Bursts at Cosmological Distances*, Science, 341, 53
13. Macquart, J. P., et al., (2010), *The Commensal Real-Time ASKAP Fast-Transients (CRAFT) Survey*, PASA, 2010, 27, 272