Data Centres

V1.0, 10 Dec 2018, Minh Huynh

This document summarises the capabilities of Australian astronomical data centres for the Decadal Plan for Australian Astronomy 2015-2025 mid-term review, and highlights upcoming opportunities.

Capabilities

CSIRO Astronomy and Space Science (CASS) provides access to radio astronomy data through several data archives, the Australia Telescope Online Archive (ATOA), the Parkes Pulsar Data Archive and the CSIRO ASKAP Science Data Archive (CASDA):

- The ATOA is a mature Australia Telescope National Facility (ATNF) archive which provides access to unprocessed data from ATCA, Parkes and Mopra. The current size of the ATOA dataset is about 250 TB, with current growth at about 30 TB per year.
- The Parkes Pulsar Data Archive allows users to access fold-mode, search-mode and calibration data of single dish pulsar observations from Parkes. The size of the pulsar data in the archive now totals more than 1 PB (more than 1,000,000 observations) and growth is roughly 400 TB per year.
- CASDA was developed to store and serve science-ready data products produced by the ASKAPsoft pipeline (software custom built by CASS to process raw ASKAP data.) The data to be ingested by CASDA is expected to be about ~20 TB per day (~5 PB per year), at full ASKAP operations. CASDA has been in development since 2013, with the first production release in late 2015. Additional enhancements and functionality have been added to CASDA over the last few years, and more development planned as ASKAP moves into full operations.

Data Central, hosted by AAO Macquarie, is a data archive developed to serve optical astronomical survey data:

- It provides a next-generation user-focused interface to explore astronomical data of national significance (e.g., SAMI, GAMA, DEVILS, GALAH, OzDES, with more due this year), providing interoperability between all ingested data sets to allow for intuitive data exploration.
- An image cutout service allows astronomers to compare the same astronomical source at different wavelengths, side by side. The Query service makes the hundreds of catalogues hosted at Data Central queryable via an intuitive web interface, while the Archive service provides online access to historical and future data taken using Australian-based telescopes.

The Skymapper Archive contains optical data from the Southern Sky Survey, obtained using the SkyMapper telescope at the ANU’s Siding Spring Observatory. SkyMapper DR1, released in 2017, contains reduced images and photometric catalogues for objects in each image. A DR2 release is planned for late 2018, with approximately bi-annual data releases planned for the remainder of the survey. Future planned development of the Skymapper Archive includes work on interoperability with AAO Data Central.

The MWA Archive has served data to the MWA collaboration since mid-2013, when science operations began. In that time the telescope has collected and stored around 28 PB of raw visibility data with growth of 3-8 PB per year. International teams of scientists are now able to access the raw MWA data via a web portal. On-the-fly calibrated MWA data are expected to be available from Dec 2018. Future upgrades to the MWA under discussion would see data rates increase by at least a factor of four to between 12 and 30 PB per annum from 2020.

The Theoretical Astrophysical Observatory (TAO) archives queryable data from multiple cosmological simulations (e.g. Millennium, Bolshoi) and galaxy formation models (e.g. SAGE) which can be funneled through higher-level modules to build custom mock galaxy catalogues and images. TAO serves over 400 registered users and currently uses approximately 40TB of disk space on the OzStar supercomputer at Swinburne (this includes simulation/model data, user-generated data, and other supporting materials).

CASDA, AAO Data Central, Skymapper, MWA and TAO are all nodes of the All-Sky Virtual Observatory (ASVO), Australia's effort to provide researchers federated access to data from astronomical facilities using Virtual Observatory (VO) services. The VO technical standards already in place include protocols to generate scripted (or automatic) table searches, cone searches, image cutouts and spectra.

HPC facilities are required for (post-)processing of the astronomical data and for enabling simulations to model the data. CASDA and the MWA archives are hosted at the Pawsey Supercomputing Centre, Skymapper is hosted at NCI, and TAO on OzStar. Australian astronomers have merit-allocated access to Pawsey's Magnus (100M core hours/year) and NCI's Raijin (115M core hours/year) supercomputers for (post-)processing of data and simulations, but this time is shared across all scientific fields. Galaxy at Pawsey (roughly 1/4 compute power of Magnus, ~190 Tflops) is 100% allocated to radio astronomy (was ASKAP and MWA, now ASKAP only). OzSTAR at Swinburne has a minimum of 35% available for OzGrav and 20% for other astronomy usage, comprising access in excess of 18M cpu-hours and 1M GPU-hours per year (data processing and simulations). Magnus and Raijin at Pawsey and NCI can be considered Tier-1, or national scale, HPC facilities (~1 Pflop). Australian astronomers do not have access to a so-called Tier-0, or international facility, such as PRACE in Europe (~10 Pflop scale).

Opportunities (2021 to 2025)

Further work with ASVO nodes (CASDA, AAO Data Central, Skymapper, MWA and TAO) will allow Australian astronomers to more easily exploit the data and maximise the science from Australian facilities. The nodes will continue to develop functionality and add datasets over the next few years. As the individual nodes were developed independently there is scope for increasing the integration of the ASVO nodes and making the nodes more interoperable. Some funding is currently planned through AAL/ARDC and individual host institutions, but the amount is uncertain.

Astronomy Data and Computing Services (ADACS) was established in early 2017 by Astronomy Australia Limited (AAL) to empower the national astronomy community to maximize the scientific return from their data and eResearch infrastructure. ADACS is delivered through a partnership between Swinburne, Curtin and Pawsey - comprising Melbourne-based and Perth-based nodes - as well as a contribution from AAO Data Central that includes an ASVO coordination role. A key element of ADACS is to provide professional software development and data management services to astronomy researchers. This provides capacity (min. 2 FTE/yr) to enhance access to data, e.g. through development of data portals and user interfaces, and optimise data processing algorithms, e.g. GPU parallelisation. To date ADACS have worked on development projects involving gravitational wave, radio, optical and theoretical datasets. ADACS is currently funded to mid-2019 and several more years of funding is currently planned in the AAL roadmap. ADACS will optimise its services to address priorities outlined in the decadal plan (original and mid-term review).

There may be an opportunity to build strategic partnerships with Pawsey, NCI and OzSTAR to enable more HPC resources for astronomy. Pawsey and NCI have received approximately $70M each from the Australian Federal government to refresh HPC infrastructure. The new systems are expected to be in place in 2019 (NCI) and 2020 (Pawsey). The refreshed systems will provide much more compute for (post-)processing of astronomical data and simulations. For example, the Pawsey refresh should result in two multi-Pflop systems to replace Magnus and Galaxy. OzSTAR was refreshed in 2018 at a cost of $5M and will run to 2022 at least.

Construction of SKA1 is expected to begin in 2020/2021 and take about 5 years. Raw telescope data will be processed by subsystems of the SKA Observatory, leading to approximately 250-300 PB per year of science data products. The archiving/dissemination of these data products to the community and generation of advanced data products are not within scope of the current SKA observatory. A network of SKA Regional Centres has been proposed which will enable the community to exploit the SKA data for science. The functions of the SKA Regional Centres will include:

- Long-term persistent storage, management and curation of SKA data
- Provision of computational resources to support (post)-processing of SKA data
- Porting and maintenance of the necessary radio astronomy software for processing and analysis of SKA data
- Providing documentation, training and user support for SKA researchers

Planning for the Australian SKA Regional Centre is already underway, led by ICRAR and CSIRO.

With the advent of the ARC Centre of Excellence OzGrav, Australia is well-placed to pursue high impact science in gravitational wave research. An Australian Gravitational Wave Data Centre would store and process data from Advanced LIGO, facilitating the real-time detection of gravitational waves. This is expected to be funded by AAL (NCRIS) in partnership with OzGrav, building on OzSTAR infrastructure with co-investment from Swinburne.

Other international projects that may require data support for Australian scientists in the future include CTA, LSST and eROSITA.