eScience

Report for Working Group 2.3
Australian Astronomy Decadal Plan 2016-2025

Executive Members: Darren Croton (Chair), Jessica Chapman, Ray Norris, Greg Poole, Andreas Wicenec, Christian Wolf

1. Executive Summary

- eScience, better described as *data intensive research*, is fast becoming a new branch of modern astronomy. It complements and extends the traditional areas of observational and theoretical astronomy.

- Large future observing projects will deliver their data through online *software telescopes,* hosted at *data hubs* by dedicated high performance computing. Similarly for simulated data and models. Such virtual telescopes will become the observatory of the future for an increasing, even dominant, fraction of our community.

- *Data federation* will bring together data of different types and wavelengths by connecting the independent data hubs. The cost of such data federation must be budgeted for when projects are being planned. A first step towards data federation has come in the form of the All-Sky Virtual Observatory (covering theory and optical survey science).

- Astronomy data is most efficiently exploited when it is exposed to the largest number of astronomers, which can be effectively achieved through online federated data hubs. From pure arguments of economy, this both maximises the investment of taxpayer's money and the opportunity for scientific discovery, and hence should be a community and institutional priority.

- Such data hubs are especially important for astronomical surveys, in which Australia has and continues to be a world leader. It is therefore essential that Australia complement its next-generation survey telescopes by prioritising the data facilities necessary to extract science from them.

- Astronomers can no longer do it all. Building and running modern high performance computing, efficient online databases, and large data hubs will require non-astronomy professionals. These may include hardware experts, software engineers and data project managers. The existing career structures within our astronomy centres will need to adapt to provide a clear career path if a critical mass of such essential personnel is to be reached.

- Given the importance of non-astronomy professionals to astronomy eScience, the connections between academia and industry should be strengthened in the coming decade. This will facilitate the exchange of skills and resources needed for such projects to be successful, and provide new opportunities for our young astronomers looking for an alternative career path.

- It is imperative that astronomy education not only prepares our students with a deep understanding of the fundamental physics, but also equips them with the modern tools required to successfully compete in both academia and industry (where many will end up working). Such tools both broaden understanding and maximise opportunity, and should not be left to happenstance.

KEY RECOMMENDATION 1

The Decadal Plan should identify the All-Sky Virtual Observatory (ASVO) as a key infrastructure to be supported and developed over the course of the decade. The ASVO can and should serve, through nodes hosted at appropriate National Facilities and Institutions, as a federated repository for datasets of national significance, allowing the Australian community to maximise the scientific exploitation of existing and anticipated major astrophysical datasets.

KEY RECOMMENDATION 2

The Decadal Plan should identify the establishment of a Centre of Excellence for Data intensive Astronomy Research (CEDAR), or similar, as a priority for the coming decade. This Centre would be tasked with (1) addressing the unique big data challenges relevant for maximising the science return from our large and complex datasets, (2) federating our data hubs across wavelength and data type, (3) be responsible for employing a critical mass of data intensive research astronomers and professionals and locating them at the nodes, (4) training new astronomers in data intensive research skills, and (5) act as a bridge between the astronomy community and industry partners. A multi-disciplinary model spanning areas of science with similar big data challenges could also be considered.

2. Science – the case for prioritising data and the resources to exploit it

Big data science and the unexpected

Big data has and will continue to have a profound impact on the culture and conduct of science and society. Such datasets can be characterised as being abundantly rich in information yet challenging in size and complexity. They enable science that is simply not possible with small datasets, such as data stacking and data mining, or the opportunity to explore the consequences of different physical phenomena through vast suites of theoretical simulations.

Sometimes numbers alone provide the advantage. For example, the EMU survey being performed on the ASKAP telescope will be a superb cosmology machine, not because the telescope parameters are particularly well suited to cosmological tests, but because the sheer numbers of galaxies enables one to beat down the errors. When N is enormous, even $\sqrt{N}$ is big!

And often data richness opens new possibilities not previously envisaged. Discoveries that end up characterising a project are frequently absent in the original science case. For example, HST was driven by three key projects, but only one resulted in science significant enough to make the HST "top ten" discovery list: measuring the Hubble Constant from Cepheids. Furthermore, half of the top ten was not planned by anybody for HST (e.g. dark energy).

Data philanthropy

How will such massive datasets be exploited? It is usually true that the more costly a dataset was to obtain the richer it is and the more science can be derived from it. As the richness of modern astronomy data continues to increase there will be only one way to maximise its value, and that is to make the data public.

Allowing others to mine your data yields many benefits, specifically in the areas of exposure (your name and research is repeatedly featured by those who are using the data, for example at conferences), citation (by being credited on all subsequent work by others), and future funding opportunities (funding agencies

are always looking to maximise data exploitation; those who give their data away ultimately facilitate far more science from the original investment than those who keep their data close).

At an institutional level, arguments about value per dollar invested can be made by looking at publication rates for facilities or projects, regardless of who ultimately wrote the papers. For example, four times as many papers result from the HST archive than from the proposal-driven observations themselves. In this sense, institutions will be increasingly looking to maximise the number of astronomers working on their data for free. This can only happen if the data is publicly *and* easily available.

Data reach

Depending on the *reach* desired by the data owner, different levels of preparation are required. A data owner might be happy with a few individual groups using their work, for which it would be sufficient to simply zip the data into smaller files and put them online through a file sharing service or the like.

However, for very massive datasets this is unfeasible. Maximum reach for large and/or complex data can only be achieved by building a *data hub*, where the data is stored in some form of database and placed on a server exposed to the world (with adequate levels of security), supported by the appropriate high performance computing. To enable a wide range of astronomers to access the data in the way that suits them best, a user interface and set of data tools must be developed.

It is a mark of our times that citizens, scientists and educators alike now communicate by sharing data; not only the raw data but simulations, processed data and visualisations. Large projects like the Zooniverse (https://www.zooniverse.org/) and offshoots like Radio Galaxy Zoo (http://radio.galaxyzoo.org/) translate such reach into new highly successful modes of data analysis through citizen science.

Building data hubs

There are significant challenges when building a data hub.

The first involves data processing and "cleaning", including validation and data integrity. This step may also require some form of data reduction if the rate at which the data is generated exceeds the rate at which it can be moved to the host high performance computing infrastructure, or if the total size of the dataset is larger than the available storage intended to house it.

The second challenge involves efficient storage of the data, typically through the smart use of hardware and the construction of an optimised backend system. Long-term storage can often have its own special needs, where the hardware and software may go through several cycles of updating with time.

The third and most important is the challenge of community access, as discussed above.

To be successful, the construction and ongoing support of data hubs requires individuals with expertise. Like most large projects, these individuals must have the skills to manage the hub and develop and maintain the hardware and software. Beyond funding, the most significant threat to the success of a data hub is the hiring and retention of the people with the skills to run it.

Connecting data hubs

The real value in astronomy data hubs comes through their connection, or *federation*. For example, both observations and simulations can play a complementary role in facilitating the science objectives of each individual community, by providing theorists with the latest observations with which they can refine their models, and by providing observers with the latest simulations with which they can interpret their results. Similarly, observational datasets stored in individual data hubs taken across many wavelengths (say by the

optical, infrared and radio communities) provides a multi-dimensional picture of the universe that is not accessible from any single hub alone.

Implementing such connectivity is non-trivial. The data must be matched in such a way that the any incompleteness, bias, or calibration issues can be adequately managed in the subsequent analysis.

For many future datasets, moving data for cross-analysis will be prohibitive due to their size. Thus, some data hubs will need to be multi-wavelength by design, and not constructed for a single telescope or single project. Many science applications, like stacking and cross-correlation, will be challenging even within a single data hub. They will require experts with experience in statistics, computer science and software engineering to build.

The cost to the community

A question the community should consider (and for which there is no easy answer) is: what is the optimum ratio of dollars spent on data facilities to dollars spent on telescopes? For the SDSS this ratio was approximately unity (Quinn et al. 2014), where computing was about half the cost of the total project. For most next generation survey facilities we suspect the ratio of expenditure will be similar. That such amounts are often not included in the original budgets of new projects and facilities should be considered a significant threat to the ultimate success of the project.

3. People

Moving forward, it is clear that data-intensive projects will be a mainstay feature to many (if not most) Australian astronomy researchers. For Australia to be competitive in this research space the use of trained high performance computing and data science professionals will be absolutely essential, including hardware experts, software engineers, data science project managers and others.

However, despite the fact that Australia has been entrenched in very large data science projects for quite some time (with the likes of HIPASS, GAMA, SkyMapper and MWA, simulation projects such as TAO and GiggleZ, and emerging astronomy frontiers like ASKAP/SKA and LIGO), the ability to employ and retain such expertise has been consistently difficult.

This can be traced - at least in part - to a circular problem involving the career model (or lack thereof) for astronomy eScience professionals in Australia. Historically, the principal funding pathway for such people has been through the Australian Research Council or serendipitous University hiring that generally requires a demonstration of publishing excellence. This is not an asset that naturally results from the training or career progressions of these individuals. The work of the most talented eScience professionals rarely focuses on publication, and as enablers of science they often are not the authors of publications utilising or made possible by their efforts. As a result, they usually cannot drive their own initiatives and are generally relegated to support roles.

Such personnel, as highly skilled professionals, have a distinct set of cultural norms for demonstrating excellence. As such, forcing them to adapt to the career models of astronomers is not an optimal approach, and the Australian astronomy community will inevitably fail to hire and retain the best eScience professionals if it fails to accommodate this reality. In part, a Centre of Excellence for Data intensive Astronomy Research would address this issue by establishing a natural home for both technical and scientific staff.

GENERAL RECOMMENDATIONS

We have identified several steps that can be taken to improve the uptake and retention of the best eScience personnel for astronomy research in Australia:

- A broad survey of the composition, contributions, funding strategies, and career paths of eScience professionals is required to enable informed planning and the development of healthy career models.
- Metrics for evaluating the contributions, value and impact of eScience researchers (e.g. code citation statistics) need to be identified and suitable norms calibrated. This will be essential for hiring, retaining and promoting the most talented individuals.
- Funding streams specific to data intensive research (e.g. fellowships or industry connections) need to be identified or established to enable the most talented and valuable eScience researchers to drive their own research programs, and to further legitimise the efforts and cultural norms of this sector in our community.
- An emphasis should be placed on longer-term stability to enable career planning and funding for facilities. For short-term contracts, a positive step would be the pursuit of five-year appointments where the current norm is three.
- More generally, a cultural shift is required throughout the astronomy community whereby the cultural norms for demonstrating excellence amongst eScience professionals are learned and respected.

4. Infrastructure

Astronomy in the coming decade will be heavily project-driven, for which effective infrastructure is critically important. Appendix A1 describes current eScience-related infrastructure, focusing on high performance computing, data services, data archives and eResearch virtual laboratories. Looking to the future, data intensive science hardware, software and interconnectivity must be considered part of the project design processes from the beginning if such projects are to have the greatest chance of success.

GENERAL RECOMMENDATIONS

Four aspects of eScience infrastructure must be developed if Australian astronomy is to be prepared for our online digital future. For long-term success in data intensive research and data management there needs to be:

Infrastructure sustainability

- Continuity in funding for existing infrastructure, including required upgrades and/or replacement of equipment.
- Infrastructure funding should mandate planning for the long-term operational costs of facilities. It is not enough to build a facility and then assume it will continue without ongoing funds.
- Investment in regional data hubs that provide integrated data services, including long-term data storage, high performance computing facilities and support for data intensive research activities. Such hubs should aim to achieve an economy of scale and to provide a consistent high level of quality and expertise. They could form the core of a Centre of Excellence for Data intensive Astronomy Research or similar.
- Long-term secure archiving of data sets of national importance. Access to them should be given equal priority as the National Library. NCRIS or similar initiatives should facilitate sustainable

pathways for the long-term retention, management and use of data assets. These could come under the umbrella of the national All-Sky Virtual Observatory as a distributed online virtual laboratory.
- A sustainable infrastructure for mid-term and short-term archiving, linked to the data hubs, should be provided. This is required to support smaller teams and data of importance but not so large in scope.

Knowledge and expertise

- Investment in a sustained group of high performance computing, algorithm and data experts, which can be consulted by astronomers and science teams. Such experts could be located at the data hubs.
- Our favoured structure is a Centre of Excellence in Data Intensive Astronomy Research, or similar. This would also allow the establishment of new career pathways inside astronomy.

Reliable networks

- High speed and robust data networks that connect data hubs to each other, to Australian Universities and Commonwealth agencies, and connect data centres in Australia and overseas. The AARNet model has worked well and we recommend that a similar approach should continue to provide support to the academic and research community.
- Consolidated planning of interconnectivity, so that data can be easily moved between data centres. Researchers should be able to apply for time on any of the facilities without having to worry about data transfer and network connectivity issues.

Integrated processes

- A more homogenous user experience and broader access to facilities. As far as possible, infrastructure facilities funded as part of the national infrastructure should be available for applications from all Australian researchers.

5. Industry

Modern astronomy has strong overlapping interests with industry. Astronomical research is limited in its capabilities to the best technology of the day, be it in high-performance computing, software and algorithms, or telescope instrumentation and design. Hence, astronomy tends to work at the forefront of such technologies with the technology providers themselves, and astronomical research makes a natural test-bed for the high-tech industry to undertake their latest developments.

However, despite this the connections between astronomy and industry have remained relatively weak over the last decade, both at the University and R&D level. This need not be, particularly true now given the growing importance of big data eScience in astronomy.

Outside of academia big data and data mining are big business. This can be seen through the emergence and dominance of online social networks like Facebook and Internet search giants like Google. Data intensive research provides a new opportunity for the astronomy community to engage more closely with industry.

Such engagement will provide a number of benefits. Additional funding opportunities become available for research that has a connection with industry partners (e.g. LIEF). Astronomers can gain access to technology not yet commercially available (e.g. prototype GPU cards when partnering with a company like NVIDIA). Furthermore, strong ties to industry enable new career pathways for astronomers, in

particular PhD students, as a good fraction of them will end up following a career in the high-tech industry. Finally, the new data intensive landscape of astronomy will require the involvement of non-astronomy professionals if it is to meet the data challenges of the next decade, from hardware experts to software engineers to data project managers. Many of these people work in industry, and industry engagement provides a bridge to them.

The benefits go both ways. As astronomy data typically has a low commercial value, industry partners have the opportunity to test new technology "in the wild" with little risk. Involvement with astronomy research often has a positive public relations effect and companies can use this in their marketing and public engagement. A company may also attract additional government funding when partnering with academia. Finally, the industry partner gets to work with PhD astronomers, giving them the opportunity to recruit those who they see as the best and the brightest.

GENERAL RECOMMENDATIONS

- eScience-focused astronomers should seek out industry partners with which to build long-term relationships. Where appropriate, their PhD students should be exposed to industry methods and culture through joint projects relevant to their thesis.
- Astronomy departments should consider industry internships or similar, where PhDs and Postdocs spend time working on non-astronomy industry-focused problems. This will add diversity to the skill-set of early career researchers and provide invaluable real world experience and connections.
- Astronomy data hubs, high performance computing centres, and projects with critical eScience components should collaborate with industry to their mutual benefit. Astronomy Centres should seek out interested industry professionals and recruit them into their projects. This could be coordinated through a Centre of Excellence for Data intensive Astronomy Research.

6. Education

The job market is changing rapidly due to the emergence of new technologies, both in academic research and in the private sector. Graduate students desire and deserve to receive training that prepares them for careers both within and beyond astronomy, and which allows them to make stronger contributions.

The landscape of education itself is evolving rapidly as it comes under pressure from e.g. online learning like Massive Open Online Courses (MOOCS) and distance education programs, software boot camps like the popular Software Carpentry (http://software-carpentry.org/), and even changes to the structure of the education system (e.g. the currently proposed deregulation of the Australian university sector, which will surely have an impact on how students learn, for better or worse). Ultimately what matters is the ability for students to meet the challenges required of them as they transition from undergraduate studies into a research environment, and beyond.

For incoming PhDs, an education that lacks research specificity disadvantages them and increases the time they must spend during a PhD "getting up to speed" (although this is inevitable at some level). In a perfect world students should be maximally equipped to do world-leading research as soon as their PhD project begins. An important priority is thus to maximise the development of basic research-focused skills while at the Honours and Masters level, and not later when the PhD research funding clock is ticking.

Our core expectation for the next decade is that astronomy students will need to graduate with better skills in database technologies, data mining and machine learning, and the general practises of software engineering, irrespective of whether they pursue a career in academia or in industry. For example, a 2011

survey by Insight Data Science[1] has identified a rapidly increasing skills shortage in machine learning that will affect all developed countries and become a bottleneck for economic growth.

We expect that the following two key skills trends will change the astronomical research world of the future:

1. Survey databases will be the new observatories for most astronomers, the so-called *software telescope*, while much of the actual observing will be robotic or done in service mode. As instrumentation and observing become a more specialised field within astronomy, most astronomers will require more eScience-related skills than observing skills to undertake their science.

2. Astronomical projects are growing in size and complexity and need to cultivate a greater reuse of software components. Accordingly, software engineering skills will be required, which keep flexible reusability in mind and anticipate future workflows to minimise wasted efforts.

GENERAL RECOMMENDATIONS

- PhD program organisers should coordinate with senior-level undergraduate course creators to offer research-specific course-trees that focus on the identified skills required in a PhD. In particular, coursework in data intensive research should be considered a core part of any astronomy education. Additionally, this should be done with consideration to the needs of industry outside of astronomy, as this is where many graduates will end up.
- There is great value in expanding what is traditionally considered an astronomy PhD, into research areas that encompass astronomy-related eScience tools and technologies, software engineering and computer science, beyond the usual branches of physics. The fusion of traditional physics-based research with such emerging technologies will be a defining feature of the new research landscape.
- Astronomy departments should consider joint appointments with other research areas, such as mathematics and statistics, computer science, and information and communication technology, to foster greater cross-disciplinary education and interaction with astronomy.

[1] http://insightdatascience.com/Insight_White_Paper_2014.pdf

A1. Current Infrastructure

NATIONAL HIGH PERFORMANCE COMPUTING

National Computing Infrastructure (NCI)

NCI provides comprehensive and integrated high-end computing services to Australia's research communities. This includes Climate and Earth System Science, Earth Observation and the Geosciences, and Astronomy. These complement NCIs traditional emphasis in Computational Science.

The NCI infrastructure base is funded by the Australian Government through its NCRIS and Super Science programs. Its expert support team and the totality of the operational costs (amounting to $11+M p.a.) is provided through a formal Collaboration Agreement of Australia's major research organisations and research-intensive universities.

NCIs highest performance systems include:

- A 1.2 petaflop Fujitsu Cluster (Raijin).
- A high-performance cloud consisting Dell-Intel CPUs, IB backplane, 160 TB of SSDs for data-intensive services.
- A Lustre parallel file system (150 Gbit/sec on the supercomputer, and 45 Gbit/sec in persistent storage).

NCI is planning for the next generation of infrastructure and is seeking to learn from national research communities, such as astronomy, of their requirements so that this may shape the future specification. Between now and the major technology refresh, which it is hoped will be in place by 2016-17, there will be minor increments to the capacity of the facility, particularly in storage which today comprises in excess of 10 PB disk and 20 PB tape. NCI further advocates substantial investments in the development of expertise in computational and data-intensive science, comparable to those provided in data, to drive Australia's engagement with international initiatives and to increase the impact that may be leveraged from the considerable investments in infrastructure.

Pawsey Centre

The Pawsey Centre is a world-class super-computing centre located at the Western Australian Technology Park in Kensington, Perth. Its primary aim is to host petascale supercomputing facilities and expertise to support Square Kilometre Array pathfinder research, geosciences and other high-end science.

The Pawsey Project - to construct the Pawsey Centre and to commission the initial equipment - was funded in 2009 as a Commonwealth Super Science infrastructure project, with $80 million allocated through the Federal Education Investment Fund (EIF). The facility is owned by CSIRO and managed by iVEC, a joint venture between CSIRO and the four public universities in Western Australia. The Pawsey Centre is operated through funds from the Western Australian State Government and the iVEC Partners. Much of the hardware in the Pawsey Centre was delivered in mid-2013 and will be due for replacement in 2016.

The Pawsey Centre primarily hosts remotely accessed and real-time processing infrastructure. Twenty-five per cent of Pawsey Centre facilities are allocated to radio astronomy, with a further twenty-five per cent allocated to Geoscience. The radio astronomy share is predominantly for the operational requirements of major facilities such as ASKAP and the MWA. The CSIRO ASKAP Science Data Archive (CASDA) will provide tools to search, access and transfer the ASKAP data products as well as

the long-term data curation. The ASKAP data volumes are now ramping up to a total of around five petabytes per year by 2016.

The Pawsey Centre facilities include:

- A CPU-based cluster called Epic, hosted at Murdoch University. Epic is a HP Proliant system with 9600 cores, 500TB of network storage, and an Infiniband network. Epic entered service in mid 2011 as a supercomputing pathfinder and is due for retirement by 2015.
- A GPU-based cluster called Fornax, hosted at the University of Western Australia. Fornax is a SGI system with 96 nodes comprising a NVIDIA Tesla C2050 GPU with 72GB of memory. The system has 500TB of network storage and uses an Infiniband network.
- A CPU-based Cray supercomputer called Galaxy, devoted to the real-time operational requirements of radio astronomy.
- A CPU-based Cray supercomputer called Magnus, used for data intensive science.
- A SGI hierarchical storage system backed by SpectraLogic tape libraries with capacity for up to 100 Petabytes of tape storage.
- SGI data analysis engines, supporting various post- processing activities on data from the supercomputers and data repository.

In addition, the Pawsey Centre has been supporting members of the Australian Consortium for Interferometric Gravitational Astronomy (ACIGA) since the inception of the Pawsey Center. This includes substantial access to the GPU-enabled supercomputing facility, Fornax, mass data storage, and supercomputing expertise.

Gravitational Wave Signal Processing and Multi-Messenger Astronomy

The Laser Interferometer Gravitational Wave Observatory (LIGO) is on track to deliver the first direct gravitational wave (GW) detections by 2017. The first science run is on schedule for 2015. LIGO signal processing is both data and CPU intensive. More than 1 TB of raw data per day is expected from a single detector. The scientific outputs are currently bounded by the ability to process LIGO data.

The iVEC supercomputer facility has been supporting the Australian Consortium for Interferometric Gravitational Astronomy (ACIGA) members for GW signal processing. This involves substantial access to the GPU-enabled supercomputing facility, Fornax, mass data storage, and supercomputing expertise. GW signal processing with iVEC is currently funded by the Australian Research Council. This includes the successful 2013 LIEF bid “Equipment for Gravitational Wave Research” by ACIGA, in partnership with iVEC and AARnet. Efforts are underway to establish a LIGO data analysis centre in Australia, which will allow low latency GW data to stream from the US to Australia, with a large fraction to be analysed in real time by Australian as well as international researchers. This will ensure Australians play a major part in the first detection of GWs. In addition, around 50 astronomical telescopes around the world have signed up to take triggers from LIGO to conduct follow up observations. ASKAP and MWA, and UWA’s Zadko Telescope at Gingin are among them. This will help grow the synergy between gravitational wave, radio and optical astronomy in Australia, and ensure Australia's frontier role in multi-messenger astronomy.

gSTAR and swinSTAR

The gSTAR supercomputer is a GPU-based facility for performing world-class simulations and to enable rapid processing of telescope data. Funding for gSTAR is provided by an EIF grant obtained in co-operation with (and administered by) Astronomy Australia Limited (AAL). It is hosted at Swinburne and operated as a national facility. Swinburne University funds an additional set of computing nodes known as swinSTAR.

The gSTAR hardware includes:

- 50 standard SGI C3108-TY11 nodes that each contain two six-core Westmere processors at 2.66 GHz (each processor is 64-bit Intel Xeon 5650), 48 GB RAM, and 2 NVIDIA Tesla C2070 GPUs (each with 6 GB RAM).
- Three high-density GPU nodes that have the same CPU capabilities as the standard nodes but each contain seven NVIDIA Tesla M2090 GPUs. All GPUs perform at greater than 1 Tflop/s (single precision).

The national astronomy community has access to a minimum 40% of the total compute capability of the facility (all nodes combined) and a minimum 200TB of storage. Up to half of the astronomy time is allocated through a merit based proposal scheme judged by the Astronomy Supercomputer Time Allocation Committee (ASTAC), a committee of AAL. The remaining astronomy time is available through a general access job queue.

MASSIVE

The Multi-modal Australian Sciences Imaging and Visualisation Environment (MASSIVE) is a GPU-based facility with a primary focus on performing imaging and visualisation. Located in Clayton, Victoria, the main system operates at over 30 teraflops with its CPU component, and over 120 teraflops using the GPU co-processors. MASSIVE consists of 2,224 cores, including 244 GPUs and a number of nodes with large RAM (e.g., 10 nodes with 12 cores each @ 192 GB RAM per node, 20 nodes with 16 cores each @ 128 GB RAM per node, etc; for details see https://www.massive.org.au/high-performance-computing/resources). Significant quantities of these resources are put towards theoretical astrophysics research, which has included extensive studies of neutron star magnetohydrodynamics and smooth particle hydrodynamics. Time allocation for the MASSIVE facility is granted through partner shares, as well as through the National Computational Merit Allocation System.

GOVERNMENT INFRASTRUCTURE PROJECTS

Research Data Storage Infrastructure (RDSI)

The Research Data Storage Infrastructure (RDSI) Project is an initiative of the National Collaborative Research Infrastructure Strategy (NCRIS) funded by the EIF. RDSI is a $50m federally funded project that provides storage for nationally significant data collections that are made accessible to researchers. Priority science areas for RDSI storage include Climate Science, Geoscience and Astronomy.

The RDSI Project funds six primary data 'nodes' in Brisbane, Sydney, Canberra, Melbourne, Adelaide and Perth, with two additional nodes in Townsville and Hobart. The nodes are supported by over 50 Australian organisations including almost all (37/39) Australian Universities and Commonwealth Government Agencies. A fast network provided by AARNet connects these nodes to each other. Merit Allocation Committees for the nodes consider proposals for RDSI facilities taking into account consistent assessment criteria and the particular research interests of the node's stakeholders. The RDSI Project also provides support for Collections Development Storage. This can be used for the creation and development of data collections of national value prior to making these accessible.

Funding for the RDSI nodes was approved in 2012. The current RDSI program ceases on 31 December 2014. Although current operators may continue to provide access to storage, this will be on a node-by-node basis. Whilst discussion is ongoing, no further infrastructure funding has so far been confirmed. This situation puts the long-term curation and storage for nationally significant data collections at risk.

National eResearch Collaboration Tools and Resources (NeCTAR)

NeCTAR is an Australian Government project conducted as part of the Super Science initiative and financed by the EIF. This funding agreement ends on 30 June 2015 and arrangements beyond this are uncertain.

NeCTAR aims to enhance research collaboration and research outcomes by providing Information and Communication Technology (ICT) infrastructure that:

- Creates new information centric research capabilities;
- Significantly simplifies the combining of instruments, data, computing, and analysis applications; and
- Enables the development of research workflows based on access to multiple resources.

NeCTAR funding for astronomy has provided financial support for the ASVO (see below), for the ANDS Project, and for RDSI. NeCTAR initiatives also include a Federated Research Cloud Project and a National Servers Program. The Federated Research Cloud has a network of around eight partner nodes. This aims to provide self-service abilities for users to publish research data, share knowledge and rapidly deploy and access software applications without the burden of operating their own computer servers. The National Servers program will provide a network of virtual servers and platforms to support eResearch activities.

DATA NETWORK CONNECTIVITY (AARNet)

Fast data connectivity between major data hubs within Australia and between Australia and overseas is an essential requirement for scientific research over the coming decade. For astronomy there are requirements to transfer data in 'real' time. For example, approximately 75 Terabytes of data taken with ASKAP at the Murchison Radio Observatory will be transferred each day to Perth, in real time, using four 10 Gbps (Gigabits per second) links. For Very Long Baseline Interferometry (VLBI) experiments, Terabytes of data recorded at observatories around Australia are transferred in real time to Perth for data correlation. Astronomers also require access to data that have been collated and are made accessible through data collections managed by data hubs, CSIRO or Universities.

As noted elsewhere in this report, astronomy is a highly international discipline. Almost all science teams with Australian astronomers also include astronomers from overseas and this is reflected in the authorship of publications in the major journals. There is already a need to transfer high data volumes to other countries. As examples, petascale data from the Murchison Widefield Array is routinely transferred to the US, whilst the dominant group of users who access data from the pulsar data archives is in China. When ASKAP starts full-time operation, we expect a sustained data flow from the Pawsey Centre and NCI to users, observatories, and data centres around the world. The need for international connectivity will grow strongly over the next decade with the emergence of global projects such as the Square Kilometre Array.

Australia's Academic and Research Network (AARNet) has provided data networks for nearly twenty five years. AARNet is a not for profit company that operates Australia's National Research and Education network (NREN). The AARNet shareholders are CSIRO and 38 Australian universities. AARNet provides high-capacity internet and advanced communication services to academic (schools, universities, cultural institutions), health and research sectors. It serves over one million end-users.

Within Australia, the AARNet backbone network connects to 'Points of Presence' at the major cities (Perth, Adelaide, Melbourne, Canberra, Sydney, Brisbane, Darwin and Hobart), with additional links to regional centres. AARNet is now rolling out the AARNet4 network. Network links on this network will be

able to support up to 80 channels at 100 Gbps, corresponding to a maximum data transfer rate of 8 Terabytes per second.

Internationally, AARNet provides connections across the world via Points of Presence in Singapore, Honolulu and the US. The organisation is working with other national research networks to develop a shared global fabric.

eSCIENCE PROJECTS

All Sky Virtual Observatory (ASVO)

The All-Sky Virtual Observatory (ASVO) Project is funded by NeCTAR, with additional funds from NCRIS and Astronomy Australia Limited (AAL). The support of AAL in particular has been instrumental to the success of the ASVO.

The ASVO provides a modular and scalable infrastructure for storing, curating, and serving datasets of national significance that enhance the scientific resources of the broader community. The long-term goal is to establish a National Data Federation, allowing interoperability across all datasets of national significance, seamless access to distributed datasets, and consequently to enable a new class of scientific investigation of a scale and mode not otherwise possible.

The ASVO already has two nodes deployed, the SkyMapper node at ANU and the Theoretical Astrophysical Observatory (TAO) node at Swinburne. SkyMapper provides an integrated and comprehensive environment for the hosting, analysis, and exploration of the SkyMapper Southern-Sky Survey. TAO provides access to several cosmological simulations and galaxy formation models and is hosted on Swinburne University's gSTAR supercomputer. Together, these provide a direct and vital link between the theoretical and observational aspects of data collection and analysis.

Two new nodes are planned, to serve datasets from the MWA and AAT. These new nodes will be developed over the 2015-2016 timeframe. The ASVO facility provides a natural and scalable infrastructure that can serve to underpin all the major national facilities Australia operates, and to add value to the international facilities in which we are partners or to which we have access.

Pulsar Data Archives

The pulsar surveys and timing data taken with the Parkes radio telescope have value to the community long after the initial processing. The CSIRO Data Access Portal (DAP), initially released in 2011, provides access to pulsar data from the Parkes telescope to astronomers worldwide. The DAP Project was established with funding from the Australian National Data Services (ANDS) with additional support for construction and operations provided by CSIRO. Efforts are underway to directly link the pulsar data collection to high performance computing facilities.

Swinburne University of Technology provides access to more than 1 Petabyte of data from the High Time Resolution Universe Survey so that scientists from around the world can reprocess the search data and use timing points to augment their own data to gain more precise ephemerides, measure dispersion measures or study pulse profiles. This project was also supported by ANDS.

Australia Telescope Online Archive (ATOA)

The CSIRO Australia Telescope Online Archive (ATOA) provides access to unprocessed data taken with the Australia Telescope National Facility over nearly 25 years. The ATOA currently holds around 100 Terabytes of unprocessed data obtained with the Australia Telescope, Compact Array, Mopra radio

telescope and Parkes radio telescope. Following an embargo period of 18 months, data held in the ATOA are made openly available astronomers worldwide. Processed image data cubes, from large-scale surveys of molecules in our Galaxy, taken with the Mopra radio telescope, are also provided. The ATOA is funded as part of the ongoing operations of CSIRO Astronomy and Space Science.

The Zooniverse, Galaxy Zoo and Radio Galaxy Zoo

The Zooniverse (http://www.zooniverse.org/) is an online environment where public education and outreach contributes towards new science discoveries and publications. Currently, the Zooniverse has an army of over 1 million volunteers spread across 22 projects. According to ADS, as of early 2014, 43 refereed papers (with over 1300 citations) alone have resulted from the Zooniverse Galaxy Zoo project alone, which uses public data from the Sloan Digital Sky Survey (SDSS). It should be noted that the Galaxy Zoo project was only launched in 2007 (based on the SDSS which started in 2000) after the main slew of survey science papers were published by the SDSS team.

We can expect 70 million sources from the upcoming EMU survey (Norris et al. 2013). Unfortunately, current algorithms will fail to match these radio sources with their host galaxies for approximately 10% of the sources. We are currently testing the feasibility of using the Zooniverse citizen science methods to identify host galaxies via the Radio Galaxy Zoo project (http://radio.galaxyzoo.org/). The Radio Galaxy Zoo project currently uses archival datasets from the FIRST (Becker et al 1995) and ATLAS (Middelberg et al 2008) surveys.

A2. Working Group Members

- James Allen, Sydney University
- Matthew Bailes, Swinburne University
- Lindsay Botten, NCI
- Jessica Chapman, CSIRO (Executive)
- Darren Croton, Swinburne University (Chair, Executive)
- Maria Cunningham, UNSW
- Michael Drinkwater, UQ
- Alan Duffy, Swinburne University
- Yeshe Fenner, AAL
- Christopher Fluke, Swinburne University
- Andy Green, AAO
- Amr Hassan, Swinburne University
- Alexander Heger, Monash University
- Andrew Hopkins, AAO
- Ben Humphreys, CSIRO
- Jarrod Hurley, Swinburne University
- Arna Karick, Swinburne University
- Slava Kitaeff, UWA
- Iraklis Konstantopoulos, AAO
- Baerbel Koribalski, CSIRO
- Paul Lasky, University of Melbourne
- Nuria Lorente, AAO
- Andrew Melatos, University of Melbourne
- Ray Norris, CSIRO (Executive)
- Quentin Parker, Macquarie University
- Greg Poole, Melbourne University (Executive)
- Peter Quinn, ICRAR
- Andreas Wicenec, ICRAR (Executive)
- Linqing Wen, UWA
- Matthew Whiting, CSIRO
- Christian Wolf, ANU (Executive)
- Ivy Wong, UWA

A3. Glossary

AAL	Australia Astronomy Limited
AAO	Australian Astronomical Observatory
AARC	Australian Resources Research Centre
AARNet	Australia's Academic and Research Network
ANDS	Australian National Data Service
ANITA	Australian National Institute for Theoretical Astrophysics
ANU	Australian National University
ARDC	Australian Research Data Commons
ARRC	Australian Resources Research Centre
ASTAC	Astronomy Supercomputer Time Allocation Committee (run by AAL)
ASVO	All Sky Virtual Observatory
ATOA	Australia Telescope Online Archive
ATNF	Australia Telescope National Facility
ADE	ASKAP Design Enhancements
ASKAP	Australian Square Kilometre Array Pathfinder
BETA	Boolardy Engineering Test Array
BoM	Bureau of Meteorology
CASDA	CSIRO ASKAP Science Data Archive
CASS	CSIRO Astronomy & Space Science
CDC	Canberra Data Centre (CSIRO)
COTS	Commercial-Off-The-Shelf, i.e. generally in reference to a ready-to-use software product purchased from a vendor
CPU	Central processing Unit
DAE	Data Analysis Engine
DAP	Data Access Portal (CSIRO)
DDP	Dynamic Disk Pool
DMF	Data Management Framework
EIF	Education Investment Fund
EMU	Evolutionary Map of the Universe
DAE	Data Analysis Engine
DIRP	Data Intensive Research Pathfinder
FITS	Flexible Image Transport System
FLOPS	Floating Point Operations per Second
GA	Geoscience Australia
gSTAR	GPU Supercomputer for Astrophysical Research
GPU	Graphical Processing Unit
HPC	High Performance Computing
HSM	Hierarchical Storage Management
IB	Inifiniband. The network connectivity between the storage, the DMF servers and PDMos.
iCAD	iVEC Committee for Allocation of Data
ICRAR	International Centre for Radio Astronomy Research
IM&T	CSIRO Information Management & Technology

Intersect	Intersect Australia Ltd
iVEC	An unincorporated 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
LCM	Logical Collection Manager
LUN	Logical Unit Number
MAID	Massive Array of Idle Disks
MRO	Murchison Radio Observatory
MWA	Murchison Widefield Array
NCI	National Computing Infrastruture
NCMAS	National Computational Merit Allocation System
NCRIS	National Collaborative Research Infrastructure Strategy
NeCTAR	National eResearch Collaboration Tools and Resources
OPAL	Online Proposal Applications and Links, a CASS system used to assist with management of astronomy operations.
OSM	Optimised Storage Manager
PAF	Phased Array Feed
PDMo	Parallel Data Movers
PDSI	Petascale Data Storage Institute (Switzerland)
PRAC	Pawsey Radio Astronomy Committee
PSF	Point Spread Function
RDS	Research Data Service (Used in CSIRO)
RDSI	Research Data Storage Infrastructure
ReDS	Research Data Services (Use by RDSI)
RFI	Radio Frequency Interference
RTC	Real Time Computer (in Pawsey Centre)
QCIF	Queensland Cyber Infrastructure Foundation
SESKA	Sustainable energy for the SKA
SIAP	Simple Image Access Protocol
SKA	Square Kilometre Array
SLA	Service Level Agreement
SOA	Service Oriented Architecture
SOC	Science Operations Centre
SSP	Survey Science Project
SST	Survey Science Team
STAP	Supercomputing Technology and Application Program
SUT	Swinburne University of Technology
SwinSTAR	Swinburne Supercomputer for Theoretical Astrophysical Research
TAO	Theoretical Astrophysical Observatory
TAP	Table Access Protocol
VLBI	Very Long Baseline Interferometry
VO	Virtual Observatory