Australian SKA Regional Centre

White paper on high level requirements

Authored: Slava Kitaeff et al.
ICRAR-CSIRO, AusSRC Project Engineer, ERIDANUS National Project Lead

Version: 1.0
Released: 22 May 2018

Table of contents

List of figures

List of abbreviations

AusSRC ……………….. Australian SKA Regional Centre
CSIRO ………………… The Commonwealth Scientific and Industrial Research Organisation
CSP …………………….. Central Signal Processor
ICRAR …………………. International Centre for Radio Astronomy Research
KSP…………………….. Key Science Projects
MWA ……………………. Murchison Widefield Array
RCA .... Regional Centre Alliance
SDP .... Science Data Processor
SKA .... Square Kilometre Array
SKAO .... SKA Observatory
SRC .... SKA Regional Centre
SRCCG .... SKA Regional Centres Coordination Group
VO .... Virtual Observatory

1. Introduction

The AusSRC White Paper initiates a discussion about the future Australian SKA Regional Centre (AusSRC) in the form of high-level requirements. The paper will inform the development of technical and organisational requirements for the AusSRC, in particular a series of projects under the ERIDANUS[1] umbrella.

Scope of document

The scope of this document is to provide the background context for AusSRC, and the role it will play in the future of Australian astronomy and the SKA. This document does not discuss details of technologies, their implementation or detailed requirements. These will follow in future documentation.

The paper is an evolving document that is regularly updated as details and plans become available and more concrete. It is anticipated that several versions of the document will be released over the next 2-3 years.

2. Background

SKA and Regional Centres

The first phase of the SKA Observatory will include two telescopes: SKA1-Low to be constructed in Australia (operating between 50-350 MHz), and SKA1-Mid in South Africa (operating between 0.35-14 GHz).

SKA1-Low will be comprised of 131,072 antennas distributed over 512 stations, while 197 15-m class dishes will comprise SKA1-Mid. By the time the SKA is in steady state operations it will generate a raw data rate close to 1 TB/s. Although the raw data will be processed by CSP and SDP subsystems of the SKA, the output of data products to be disseminated to science teams are estimated to be 250-300 PB per year for the two arrays combined.

The following key factors discussed in [RD5] that must be addressed by a model of Regional Centre Alliance (RCA), that is a collaboration of all SRCs, are:

1. Data products. The generation of advanced data products specific to the science goals is not within scope of the current SKA observatory.The combination of data products from across multiple observations must be undertaken outside of the observatory boundaries by the science teams.
2. Data volumes. The data volume that will be generated by the SKA will be very high. The data products need to be curated and served to the community while abiding by the SKA's data access policies. It is expected that SKA users will run several models for their analyses, and final data volumes associated with each experiment may also

[1] https://eridanus.net.au/

be large. In addition, future upgrades could increase the delivery from each site, or if additional “discovery” data products are generated. Given the overall volume of data products and the large potential size of each one, a traditional model where users download their data to a local machine to process or analyse is not viable. Instead, SKA users will need to “take the processing to the data”; submitting requests to run analysis or visualisation workflows on data held remotely.

3. Collaboration. In the first years of full SKA operations, the science programme will be largely defined by Key Science Projects (KSPs). These KSPs will be comprised of large teams, with membership drawn from across the whole SKA member nations, conducting ambitious science programmes that require thousands of hours of observing time on the SKA. To maximise the scientific productivity of the SKA will require new methods, algorithms and software. Such activities will necessarily be driven from the SKA scientific community, and will require an environment that enables innovation in research and successful collaboration.

In April 2016, the Data Flow Advisory Panel presented its recommendations to the SKA Board, which were broadly endorsed, with the SKA Organisation instructed to work with its member nations to establish a network of SKA Regional Centres that will enable the community to exploit SKA data for high impact science (see Figure 1) [RD5].

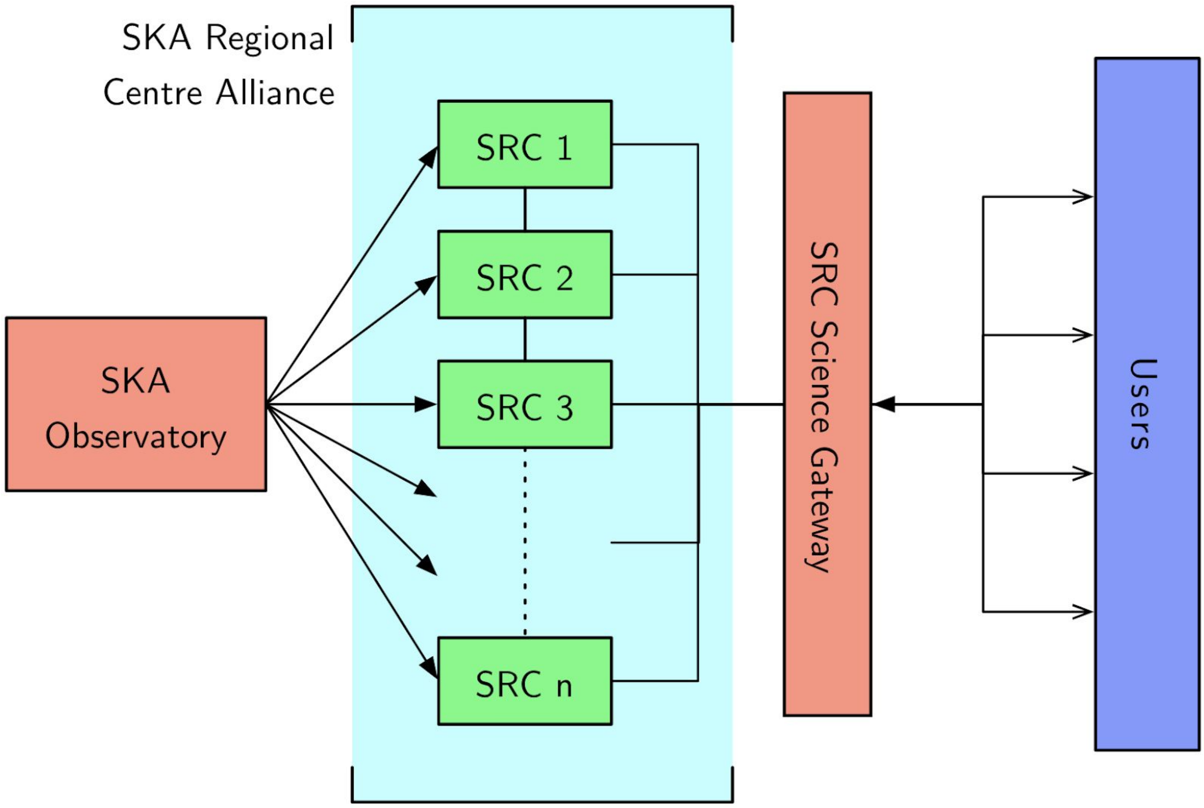

Figure 1: A network of SKA Regional Centres (the SKA Regional Centre Alliance – RCA) receives science data products from the SKA Observatory. Access to SKA science data products, as well as the tools and processing power necessary to fully exploit the science potential of those products, is provided via a Science Gateway. Access to science data products is irrespective of a SKA user’s geographical location, or whether their local region or country hosts an SRC (adopted from [RD5]).

The key objectives of SRCs will be to:

- Provide long-term, persistent storage capabilities.
- Provide sufficient computational resources to support processing and analysis of SKA data by the astronomical community at the appropriate scales and with reasonable latency.
- Provide long-term data management and curation including metadata allowing easy data discovery, examination of data provenance, and combination with other existing, relevant data sets.
- Provide security and data protocols capable of supporting a wide range of access paradigms from fully open access public datasets to proprietary data for individuals or consortia.
- Provide porting and maintenance of the necessary radio astronomy software stack to the cloud platform.
- Provide documentation, training, and user support for SKA researchers.
- Provide the environment that enables innovation in research and successful collaboration.

Australia has built two SKA pathfinders - ASKAP and MWA, and will host SKA1-Low. Developing an SRC in Australia is critical in order to support SKA, as well as SKA precursors science teams. Significant effort and investment has already been made that paves the way to the future Australian SRC:

- MWA and ASKAP have already produced over 25 PB of data stored in the Pawsey Supercomputing Centre. The volumes of data from these telescopes are rapidly increasing due to the recent upgrades.

- Pawsey, NCI and other facilities are used to store and process the data from the telescopes. AARNet provides the capability to transport these data across the country and internationally.

- Australia has built a significant scientific capability in astronomy, and in particular in radio astronomy. MWA science collaborations are conducting surveys such as GLEAM, EoR, Solar, FRB, spectral-line and many other observations. The ASKAP telescope will be used by at least 10 science teams to produce the surveys such as EMU, DINGO, WALLABY and others. All the surveys will collect or are already collecting a staggering amount or data that needs to be stored, managed, disseminated, processed, and analysed. Complex algorithms and software have been developing by many groups to achieve the planned science goals. This work already requires a significant capacity in data transport, storage and compute infrastructure in Australia, which currently cannot be considered as sufficient.

All the aspects of planning and executing the observations, collecting, managing and processing the data require highly specialised expertise that is not readily available to all the teams. Therefore, the collaborative aspects of planned development of AusSRC are very important.

Multi-wavelength or so called multi-messenger astronomy has become a powerful instrument of groundbreaking scientific discoveries, for example, the the first observations of neutron-star merger started from the detection of gravitational waves, followed by observations from radio waves to infrared to visible to X-rays to gamma rays. It is important that AusSRC enables such discoveries through sharing knowledge, expertise, data and algorithms reaching out to many cross-disciplinary areas, not just in astronomy but such disciplines as particle physics, cosmic rays and others.

The international collaborative effort required to develop AusSRC is important, especially in the South-Pacific region, where a regional alliance has many prerequisites to be formed and developed into a productive collaboration, specifically with China and New Zealand.

Australia participates in and leads a number of major SKA Science Working Groups. The work that Australia already does for the SKA precursor telescope and science with them is crucial for the success of the SKA globally, and must be used as a stepping stone in developing the AusSRC.

ERIDANUS Project

The ERIDANUS Project is a design study commenced in April 2017, aimed at deploying prototype data intensive research infrastructure and middleware, between and within Australia and China, capable of addressing SKA-class data and processing challenges. The project responds to the identified challenges, and works in a coordinated and collaborative manner with the European Aeneas (Advanced European Network for E-infrastructures for Astronomy with the SKA) project.
The ERIDANUS Project engages with:

- The astronomical communities within Australia, China and New Zealand,
- The providers of networking and computing research infrastructure (e.g. research network providers, supercomputer centres, commercial cloud providers),
- National and international committees (e.g. SRCCG, ASRCWG) coordinating efforts on SKA regional centre developments,
- SKA precursor/pathfinder projects – 21CMA, FAST, MWA, ASKAP, to enable large scale prototyping and the support of specific science programs.

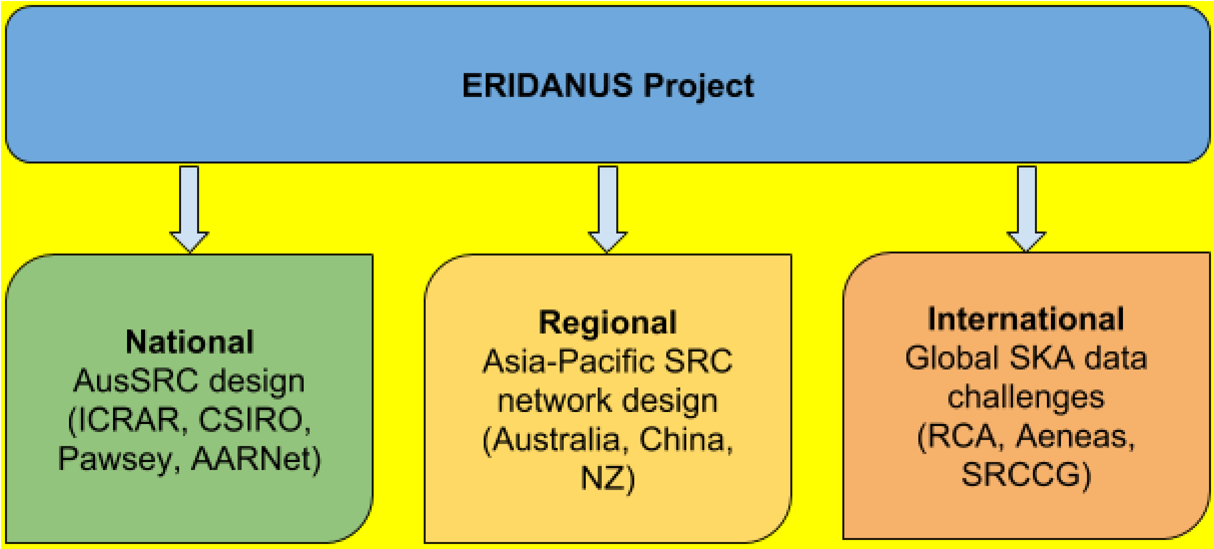

Figure 2: ERIDANUS Project is a design study aimed at deploying prototype data intensive research infrastructure and middleware capable of addressing SKA-class data and processing challenges, between and within Australia and China. The project addresses the challenges at three levels: national, regional (Asia-Pacific) and international.

Envisioned Process

AusSRC will require a significant investment into the infrastructure and development of capabilities and tools able to address SKA-class challenges. This requires a substantial preparatory work to be done before the SKA1 becomes operational.

Figure 3 depicts the steps necessary to be undertaken in order to develop the full Business Case that would include the project structure based on well understood System Architecture, and fully costed implementation plan.

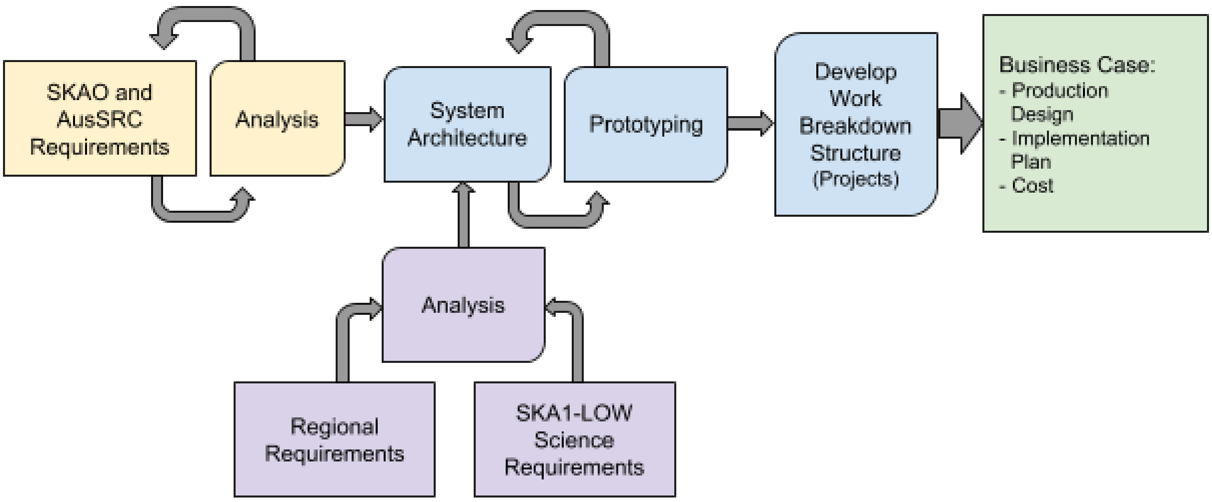

Figure 3: The envisioned process to develop an AusSRC business case that includes a well understood technical architecture and its fully costed implementation plan.

Firstly, community generated requirements need to be analysed. Such requirements will define an initial technical architecture of an AusSRC.

Then, the architecture needs to be refined through a series of prototypes focused on the needs of science teams involved with ASKAP and MWA, connecting to new system emerging at the Pawsey Supercomputing Centre and utilizing Australian developments produced through engagement in the SKA pre-construction effort. The architecture will need to be refined further with end-to-end tests of regional SRC capabilities with China and other regional partners responding to global data challenges to the architecture coming from the SRCCG. Australian and international SKA1-LOW science teams requirements will also need to be assessed for an AusSRC post 2022.

Finally, a product design along with its implementation plan and full costing will form a business case for an at-SKA-scale AusSRC.

3. AusSRC High Level Requirements (Level 1)

This chapter provides High Level Requirements (HLR) for the AusSRC based on the discussions that took place during the AusSRC workshop in November 2017 in Perth [RD2], SKA Regional Centre Requirements [RD1], and two Australia-China SKA Big Data workshops that took place in 2017 and 2018 [RD3], [RD4].

At the time of writing the levels of requirements have not been defined. The structure adopted in this document is as follows:

- Level 0 - RCA Requirements
- Level 1 - AusSRC High Level Requirements
- Level 2 - AusSRC Detailed Requirements

Breakdown

The Level 1 requirements are notionally grouped into the following categories:

- Governance
- SKAO and other SRCs
- Infrastructure
- Collaboration
- Software
- Science Archive and Post-processing
- User Support
- Compliance

Such a categorisation is relative as it is not always possible to assign a requirement to a single category due to existing overlaps. Tagging approach might be implemented in later versions of the document instead.

1. Governance

At the time of writing, a governance model for the AusSRC and the SKA Regional Centre Alliance (RCA) is unknown but might be defined towards the end of 2018. However, in any adopted governance model AusSRC will be a part of RCA.

During the AusSRC workshop [RD2] many discussion revolved around the necessity to retain and disseminate the critical expertise in Australia. AusSRC is seen as playing a significant role in that.

ID	Name	Description
1.1	AusSRC membership	[placeholder]
1.2	Multidisciplinary approach	AusSRC should adopt a multidisciplinary approach with the identified disciplines (Optical astronomy, IR astronomy, X-ray astronomy, Gravitational Waves, SETI, Dark Matter, Astrobiology, Ionosphere, Planetary Science, Cosmic Rays, Cherenkov Telescope Array, Neutrino Physics (KM3NET)). Not astronomy disciplines have been also identified: medical imaging, defence, space debris, bioinformatics, Earth observations, geological studies [RD2].
1.3	Retention policies for expert staff	AusSRC should have retention policies for expert staff in order to ensure the continuity of critical expertise in Australia. [RD2]
1.4	Dissemination of expertise	AusSRC should implement an effective policy that will help to share and disseminate expertise across Australia. Shared appointments have been noted as one effective way to achieve it. [RD2]
1.5	Resource Allocation Policy	AusSRC should have a policy that regulates the allocation of storage and compute resources.
1.6	Data policies	AusSRC should have policy/policies that regulate the archiving, sharing, ownership, access and any provenance of the data stored within AusSRC. Such a policy should inherit the appropriate rules of the SKAO and RCA data policies or be subject to those policies. [RD1][RD2]

2. SKAO and other SRCs

AusSRC will be Tier 1 SRC receiving SKA-Low data for post-processing and further sharing and dissemination to other Tier 1 and Tier 2 SRCs. AusSRC will be a part of the global

Alliance of SRCs - RCA, hence will be bounded by certain common rules, agreements and policies, and will have agreements with both, higher and lower-tier SRCs.

ID	Name	Description
2.1	SKAO and multi-tier SRCs	AusSRC should have agreements and policies that define the scope and responsibilities of AusSRC in relationship with SKAO and SRCs of other levels.
2.2	Planning support	AusSRC should provide support for the SKA observations, post-processing and archiving planning and proposal writing to Australian research teams and individuals [RD2].
2.3	Survey support	AusSRC should support the post-processing and archiving of the observations with the SKA to Australian research teams and individuals [RD2].

3. Infrastructure

One of the primary objectives of AusSRC is to develop and provide the infrastructure for receiving, storing, post-processing SKA1 data products, as well as providing a long term science archiving capabilities across Australia and further out to global astronomy community, most notionally Asia-Pacific region (China, New Zealand, India).

ID	Name	Description
3.1	Processing capacity	AusSRC should provide 25 PFlops peak performance at the start of SKA1 operation (2025) for post-processing SKA1-Low data products.
3.2	Storage capacity	AusSRC should provide 55 PBs per year storage capacity at the start of SKA1 operation (2025) to store SKA1-Low data products.
3.3	Network	AusSRC should provide 100 Gbps network connectivity between the tiers in Australia at the start of SKA1 operation (2025).
3.4	Monitoring	AusSRC should provide an ability to monitor, control, respond, and optimise the network, storage and processing infrastructure at all tiers and nodes. [RD1][RD3][RD4]

4. Collaboration

The SKA Key Science Projects (KSP) will drive the science program that needs to be supported by the AusSRC. KSP will be comprised of large teams from across the global scope of SKA and extended cross-frequency and cross-disciplinary collaborations. AusSRC will be a key facilitator for those collaboration to succeed.

ID	Name	Description
4.1	SKA collaboration	AusSRC should provide both, technical and non-technical means for global SKA collaboration [RD2].
4.2	Multi-messenger collaboration	AusSRC should provide both, technical and non-technical means for SKA KSP collaboration with other astronomy groups and instruments [RD2].
4.3	Interdisciplinary collaboration	AusSRC should provide both, technical and non-technical means for collaboration with adjacent disciplines such as Cosmic Rays and Particle Physics, and others [RD2].
4.4	Industry collaboration	AusSRC should collaborate with the industry partners to develop and deliver best technology and support [RD2].
4.5	Expert support	AusSRC should have expert resources available to support KSPs in planning and applying for the telescope time, planning and executing post-processing, archiving, and data analysis [RD2].
4.6	Physical space	AusSRC should provide a physical space to support collaborations in the form of busy weeks, workshops, hackathons etc [RD2].
4.7	Development tools	AusSRC should provide the frameworks and software tools for collaborative development [RD2].
4.8	Financial Support	AusSRC should provide financial means such as travel grants and meeting facilitation funds to support collaborative work at AusSRC [RD2].

5. Software

One of the current and significant issues of legacy radio astronomy software its highly limited usability on modern parallel platforms. Lack of optimisation, scalability, poor modularity translate into low efficiency, low performance and low maintainability of the software. These issues must be addressed in AusSRC in order to achieve the highest return on investment. This requires a significant cultural change in the way how the software is developed.

ID	Name	Description
5.1	Software Interoperability	Interoperability should be provided at the correspondent levels of post-processing between different tiers of AusSRC and to the global network of SRCs. [RD1] [RD2][RD4]
5.2	Software lifecycle management	AusSRC should have a common framework for development of software that supports a full lifecycle of software. [RD2]
5.3	Software Libraries and Repository of Algorithms and Repositories	AusSRC should have a repository of maintained software, algorithms and ML models. [RD2]
5.4	Software QC	AusSRC should have a quality control process and metrics for the software, algorithms and models in the repository. [RD2]
5.5	Development environment	AusSRC should provide an environment where organisational, community and individual development can co-exist. [RD2][RD4]
5.6	Software Publishing	AusSRC should have a policy and resources for software publishing. [RD2]
5.7	Common API	AusSRC should have a common API for the software. [RD2]
5.8	Collaboration environment	AusSRC should have a software environment that facilitates and enables collaboration and collaborative post-processing and data analysis. [RD2][RD4]
5.9	Common user environment	AusSRC should support use of a common environment across RCA. [RD1]

6. Science Archive and Post-processing

AusSRC will directly receive data products from SKA1-Low for long term storage and sharing. User will interact with the data products to visualise them, assess the results, and to use them to generate further data products for detailed analyses. The advanced data products (ADPs) will be put back into the AusSRC science archive, published, and accessed by other users.

SKA will be a world-leading facility and it is expected that the AusSRC will follow best practice in scientific integrity. To achieve reproducibility, it is necessary to preserve and make discoverable and accessible all the elements involved in the experiment. This includes input and output data, software, used algorithms, configurations, annotations, history of versions, etc.

AusSRC will need to provide at least two distinct storages. One to store the observatory data products and ADPs, and the other a variety of storages for the users, post-processing, project collaboration spaces, scratch space etc. The requirements for the technology, capacity and backup strategies are likely going to be heterogeneous and optimised to the use cases.

ID	Name	Description
6.1	System Access	AusSRC should implement or adopt a unified and secure way of user authentication. [RD2]
6.2	Data access	AusSRC should have data management tools providing transparent way to access the data across AusSRC and global network of SRCs [RD2].
6.3	Data provenance	AusSRC should implement SKA data provenance rules and policies within its data management tools [RD2].
6.4	Data stewardship	Any frameworks and software tools developed or deployed in AusSRC should support the FAIR Data Principles.
6.5	FAIR Principals in Policies	AusSRC data policies should follow FAIR Data Principles[2].
6.6	Data Sharing	AusSRC should have all the tools and infrastructure to enable global sharing and redistribution of data. [RD4]
6.7	Data transfer minimisation	AusSRC should support the RCA goal of minimisation of data transfer. [RD1][RD4]
6.8	Open Access	AusSRC should enable public sharing of data that is either published or non-proprietary. [RD2]
6.9	Data Security and Availability	AusSRC data policy and lower level requirements should include specifications on acceptable data losses and availability time. [RD2][RD1]
6.10	Data Formats	The data formats should provide interoperability at the correspondent level of post-processing

[2] https://www.go-fair.org/fair-principles/

		between different tiers of AusSRC and global network of SRCs. [RD2]
6.11	Data Publishing	AusSRC should have a policy, tools and resources for data publishing. [RD2]
6.12	Computational resource sharing	AusSRC should have a policy and agreements with other SRCs and RCA about sharing the computational resources. This requirement supports the requirement 9.4. [RD1]
6.13	Reproducibility	AusSRC should provide a way of saving the complete workflows, environments and all the provenance associated with any advanced data product, in such a way that it can be queried, viewed and the associated workflows can be reused. Using the saved workflows on the same data should reproduce the advance data product. [RD1]
6.14	Abstraction and reusability	AusSRC frameworks should provide the sufficient abstraction of layers of post-processing in such a way that all the knowledge of processing in encapsulated within the framework itself and the interface is sufficiently generic and abstract to provide a separation of concerns when a new pos-processing pipeline is developed. It should provide the highest level of reusability of previously developed modules and algorithms. [RD2]

7. User Support

There will be a broad range of activities that AusSRC will need to provide help to the users starting from preparing a proposal for observations, planning storage and computational resources, developing post-processing software etc. On going training for postgraduate student and established researchers will be critical in achieving the high impact outputs of the SKA.

ID	Name	Description
8.1	Training	AusSRC should provide an ongoing training for SKA and SRC tools and infrastructure, software and algorithms development. This should include the allocation of necessary for training resources. Postgraduate training as well as ongoing training of researchers should be developed. [RD2]
8.2	Experts integration	AusSRC should implement a management model that would enable integration of experts with the

		research teams. [RD2]
8.3	User support model	AusSRC should have a user support model that is continuously improved. [RD2]
8.4	Standard for Service Level	AusSRC shall adopt an industry standard (such as ITIL v4) for Service Level Agreements with commercial partners who provide services and compute resources.

8. Compliance

AusSRC will need to provide the highest level of interoperability to enable science across science archives and other telescopes globally. Compliance of used protocols, interfaces and data formats is critical in achieving interoperability.

ID	Name	Description
9.1	SKA compliance	AusSRC interfaces to the SKA should be compliant with the interfaces and protocols set by the SKAO. [RD1]
9.2	RCA compliance	AusSRC interfaces to other SRCs, including the SRCs of lower tiers should be compliant with the interfaces and protocols set by the RCA. [RD1]
9.3	IVOA compliance	AusSRC interfaces should be IVOA compliant whenever possible. If IVOA standard does not meet the requirements an effort should be made to contribute to improving IVOA interfaces and protocols. [RD2]

4. References

The following documents are referenced in this document. In the case of conflict between the contents of the referenced documents, or between the content of referenced documents and this document, this document shall take precedence.

[RD1] SKA-TEL-SKO-0000735 SKA Regional Centre Requirements
[RD2] Australian SKA Regional Centre Workshop: Summary of Discussions
[RD3] Australia-China SKA Big Data Workshop Report
[RD4] 2nd Australia-China SKA Big Data Workshop Report
[RD5] SKA-TEL-SKO-0000706 SKA Regional Centres: Background and Framework

The documents [RD2][RD2] and [RD4] can be found in https://eridanus.net.au/?page_id=296.