SKA REGIONAL CENTRE REQUIREMENTS

Document Number..... SKA-TEL-SKO-0000735
Document Type ..... RSP
Revision 03
Author ..... R. C. Bolton and the SRCCG
Date 2019-01-23
Document Classification..... UNRESTRICTED
Status.....Released

Name	Designation	Affiliation	Signature	
Authored by:				
Rosie Bolton and the SRCCG	SRC Project Scientist	SKAO	R Bolton	
			Date	2019-01-23
Owned by:				
Antonio Chrysostomou	Head Science Operations Planning	SKAO	Antonio Chrysostomou	
			Date	2019-01-23
Approved by:				
Gary Davis	Director of Operations Planning	SKAO	GRD	
			Date	2019-01-23
Released by:				
Phil Diamond	Director General	SKAO	Phil Diamond	
			Date	2019-02-12

DOCUMENT HISTORY

Revision	Date Of Issue	Engineering Change Number	Comments
01	2017-07-04	-	First version of SRC requirements
02	2018-11-12		Revision 2, updated requirements on processing load and storage. Minor changes to text.
03	2019-01-23		Corrected mistake in Req 14 which had incorrect online data volumes

DOCUMENT SOFTWARE

	Package	Version	Filename
Word Processor	MsWord	Word 2007	SKA-TEL-SKO-0000735-03_Regional Centre Requirements.docx
Block diagrams			
Other			

ORGANISATION DETAILS

Name	SKA Organisation
Registered Address	Jodrell Bank Observatory Lower Withington Macclesfield Cheshire SK11 9DL United Kingdom Registered in England & Wales Company Number: 07881918
Fax.	+44 (0)161 306 9600
Website	www.skatelescope.org

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

ADP .....Advanced Data Product

FACT .....Fairness, Accuracy, Confidentiality, Transparency

FAIR .....Findability, Accessibility, Interoperability, Reusability

IVOA .....International Virtual Observatory Alliance

MoU .....Memorandum of Understanding

ODP .....Observatory Data Product

SKA .....Square Kilometre Array

SKAO .....SKA Observatory

SRC .....SKA Regional Centre

SRCCG.....SKA Regional Centre Coordination Group

TBC .....To Be Confirmed

TBD.....To Be Determined

W3C.....World Wide Web Consortium

GLOSSARY

In this document we use some words in specific ways:

1. Near-line storage: Data storage that may involve some latency performing processing jobs on data products held in near-line storage.

2. Observatory: The SKA Observatory.

3. Online storage: Instantaneously accessible storage – users do not perceive significant latency when performing processing on data products held in online storage.

4. Provenance: the description of how a data product was generated – including input and output data, software used and configuration parameters.

5. Reproducibility: the ability for the results of an experiment to be confirmed by an external researcher, using the same configuration, input data and methods.

6. Science Data Products: This is the general term encompassing two categories of data products:
    a. Observatory Data Products: those generated within the Observatory and delivered to the SRCs,
    b. Advanced Data Products: those generated within the SRCs.

7. Scratch: temporary working area of the computing facility, associated with a user's execution of a particular work flow or project, which is not part of the archive.

8. User: A user of the SRC

1 Introduction

1.1 Purpose of the document

This document describes and presents requirements and goals that the SKA Regional Centres (SRCs) will meet, individually or collectively, with explanations given to introduce each item and place it into context.

1.2 Scope of the document

To develop preliminary requirements for the SRCs the SKA Regional Centres Coordination Group (SRCCG) started from the "SKA Regional Centres: Background and Framework" document [RD1] and considered the role the SRCs will need to take in delivering and making accessible SKA Observatory data products to the user, and in performing further processing tasks on the SKA data products to produce Advanced Data Products with enhanced scientific qualities. Our primary focus was on the needs of the Observatory – how can the SRCs maximise the scientific output of the SKA? In addition, we have also considered the user's perspective – what do the SRCs need to provide to each user to make their experiences with SKA data useful and productive? The former perspective gives rise to a series of requirements and goals around the bulk capability of the SKA Regional Centres taken as a collective whole, their ability to receive SKA Observatory Data Products and their total processing capability. The latter perspective gives us requirements around the "Accessibility and Software tools".

In 2018 Q4 we updated these requirements after conducting a high-level assessment of the possible data rates out of the SKA telescopes, and making clear and simplistic assumptions about the way that SRCs will use these data products and about the potential use made of the telescopes themselves. The focus has been on the early years of the SKA – including considering the role of SRCs whilst the SKA telescopes are under construction. In the current roll out plan, array capability emerges around 2022 and the arrays grow to full scale by year-end 2027. During commissioning, the scale of the SRCs ramps up to a full scale compute system C2024 and with continuously growing data storage. The role of the SRCs will evolve during this period, initially, as the SKA capabilities grow, focussing on training users with SKA-scale data and data challenges (possibly involving pre-cursor and pathfinder data sets) and on enabling world experts to contribute to commissioning challenges, to (c.2028) reaching the steady-state model incorporating user support and science-generation as outlined in [RD1].

The detailed explanation of the assumptions that have gone into the updated numerical requirements is presented in [RD3].

We have separated out requirements (absolutely essential items, or bare-minimum capabilities) from goals. For performance figures, our requirements correspond to our current best estimates of compute capability or storage figures whilst the goals include additional contingency based on the uncertainties in the numerical analysis. Goals are desirable as bringing additional benefits or reducing risk compared to bare-minimum requirements.

All dates in this document are subject to the Project schedule. We refer dates in years relative to T0, the date at which approval of the Construction proposal is granted by the SKA Council and funds are released[1].

[1] At time of writing T0 is anticipated to be mid 2020, and full-scale scientific operations with Key Science Projects are anticipated to start c. 2028.

2 References

2.1 Applicable documents

There are no applicable documents that take precedence over this document.

2.2 Reference documents

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

[RD1] SKA-TEL-SKO-0000706 SKA Regional Centres: Background and Framework

[RD2] The FAIR Guiding Principles for scientific data management and stewardship: https://www.nature.com/articles/sdata201618

[RD3] SKA Regional Centre Scaling Background document (in prep)

3 SKA Regional Centres – Context and Boundary Conditions

A high-level overview of the SKA Regional Centre context and motivation is given in [RD1], and we refer the reader there for a more complete description than we have space for here. In brief summary, SRCs will provide a mechanism through which a given community or communities will have access to resources and data products necessary to do science with the SKA. They are necessary for the astronomy community to produce scientific results with the SKA Observatory (SKAO). Their existence, however, does not alter the fact that the responsibilities for observing programme adjudication and execution, and the quality assessment of Observatory Data Products, remain fully with the SKAO.

Not all member nations will necessarily host an SRC, however, SKA users within each member nation must have access to their SKA data products. This means that the SRCs must be useable remotely, ideally in a coherent way, irrespective of the geographical location of the user or the data products. Users will also require significant computing resource (software tools and processing power) within, or facilitated by, the Regional Centres in order to manipulate, analyse and visualise their data products, in order for the Observatory to be as productive as possible.

Not all SRCs will be the same. Some may wish to take up specific roles, perhaps driven by proximity to SKA infrastructure, available national or regional expertise, scientific interests or nationally driven focus on technology developments. These choices will be dependent on available resources, funding opportunities, and community obligations or priorities.

In order to be accredited as an SRC, those organisations or group of organisations seeking such a designation will need to meet a number of requirements and commit to providing some minimum resources to the SRC pool. These cover six broad areas:

- “Governance”: How the SRCs will function together with the SKAO and be managed to deliver the needs of the user community;
- “Science Archive” and “Storage Capacity”: Provision of an archive for the storage and curation of, and access to, SKA Science Data Products;
- “Accessibility and Software Tools”: How users will interact with the SRCs;
- “Data Processing”: Generation and visualisation of Science Data Products;
- “Network connectivity”: Transfer of Science Data Products from the Observatory into SRCs and between SRCs.

4 Draft SRC Requirements and Goals

4.1 Governance

At the time of writing, a governance model for the SKA Regional Centres is not defined. Whatever model is adopted, it is expected that for any SRC to retain its designation as such it will be required to abide by MoUs for the provision of services and resources, and be compliant with SRC requirements.

Over the lifetime of the observatory, membership of the SKA is likely to change. An individual SRC may seek to terminate its SRC status, but must do so according to some notice period and under the conditions specified in its MoU. Any such SRC will be required to ensure that all SKA Science Data Products (both observatory and advanced products) it hosts are available to copy out to other SRCs until they have all been safely saved to alternative locations in other SRCs.

The observatory will be delivering data products to multiple SRCs, so each SRC must be compliant with the interfaces and data format policies of the SKAO (TBD). SRCs must also be able to share data between each other.

Table 1: Requirements and Goals Derived from Consideration of Governance

ID	Type	Name	Description
1	REQ	SRC designation	SRC designation will be awarded if individual prospective SRCs meet and maintain all the criteria set out in appropriate MoUs and Accreditation criteria. The ability of each SRC to meet its criteria, and the criteria themselves will be reviewed annually (TBC).
2	REQ	Graceful exit of SRC	On seeking to terminate SRC status, an SRC shall ensure that all data products and software tools held by it are available elsewhere in another SRC.
3	REQ	Graceful exit, data saving	Collectively the SRCs shall manage the redistribution of data products hosted by any individual SRC if and when that SRC seeks to terminate its SRC designation.
4	REQ	SRC to SKAO interface	Interfaces between each SRC and the SKAO will be compliant with policies set out by the SKAO.
5	REQ	SRC to SRC interfaces	Interfaces between SRCs will be compliant with policies set out by the global network of SRCs.

4.2 Science Archive

The SKA will generate Observatory Data Products that will be delivered to the SRCs for (authorised) access by the scientific community. Users will interact with these Observatory Data Products, to visualise them, assess results, and to use them to generate further data products for more detailed analyses. We envisage that once the user deems their data products ready for use and/or publication, they can label them as "Advanced Data Products" (ADPs). Once a data product has been tagged in this way, it is archived and becomes visible to *authorised* users (as per the data access policies) across all SRCs, just as Observatory Data Products are.

It is natural to expect that one copy of the all Observatory Data Products from the same project will be located at the same Regional Centre, since generation of any Advanced Data Product will entail combining these data sets (e.g. to achieve integration times longer than 12 hours). Likewise, as part of the archiving of the Advanced Data Products, it is expected that one copy of the Advanced Data Products will be located in the same Regional Centre that generated them.

SKA will be a world-leading facility and as such should expect to follow, if not to lead, best practice in scientific integrity. This means that the SRCs, individually and as a whole, must enable adoption of best-practice as it evolves[2], and we should encourage users to participate in this. Currently we have identified the areas of Open Access (providing public links to data products to go alongside publications), and of *Provenance* and *Reproducibility*. Provenance is the description of how an ADP was generated – including input and output data, software used and configuration parameters. Provenance should be stored in the ADPs but also ingested into a standard data model for querying and viewing (e.g. IVOA, W3C). Reproducibility in science is a more abstract concept that refers to the idea that the results of an experiment can be confirmed by an external researcher, using the same configuration, input data and methods. To achieve reproducibility, it is necessary to preserve and make discoverable and accessible all the elements involved in the experiment (input and output data, software implementing the scientific method, annotations to understand the experiment, etc.). Recently, the concept of a Research Object has arisen as a digital solution to preserve scientific experiments.

Therefore, preserving the workflows used to generate each one of the ADPs, and their provenance, will provide a step towards SKA science reproducibility.

The SRCs should also meet the needs of general users who wish to work with the public data products. To enable this, the Science Data Products (ODPs and ADPs) would need to be publicly accessible and mechanisms would need to be provided to allow searches for relevant data products according to some criteria (e.g. sky position, source name, frequency band, etc).

[2] For example, see the FAIR[RD2] and FACT principles.

Table 2: Requirements and Goals Derived from Consideration of the Science Archive Functionality

ID	Type	Name	Description
6	REQ	SRC Data policies	Each SRC will preserve and make available to users, the SKA Science Data Products, in adherence to SKAO data access policies and data security standards.
7	REQ	SRC Data Sharing	Each SRC will, when required, distribute the SKA Science Data Products to other SRCs.
8	GOAL	Minimise data transfer between SRCs	Data products will be located within the network of SRCs such that any transfers between individual SRCs are minimised.
9	REQ	Open Access	The SRCs will enable users to provide public links to SKA Science Data Products in their research publications. Published and non-proprietary data must be publicly available.
10	REQ	Reproducibility: Provenance and workflow preservation	Each SRC must be capable of saving the complete workflow and provenance associated with any ADP, in such a way that they can be queried, viewed and the associated workflows can be re-used to create new ADPs.
11	GOAL	Advanced Data Product re-generation	Each SRC must be able to save the software environment associated with the provenance and workflow of an ADP that is required to re-execute the workflow in order to regenerate it.
12	REQ	Data product index	Collectively, the SRCs will maintain and provide access to an index of all Science Data Products (including Observatory Data Products and Advanced Data Products), capable of showing the location(s) of each one.

4.3 Storage Capacity

The Regional Centres will be required to provide at least two types of storage capacity. The first is the storage required for the Observatory Data Products and the Advanced Data Products (with appropriate back-up). The second is to support a variety of different capabilities for users, including scratch spaces and project collaboration spaces. Different applications will be suitable for different hardware architectures and we imagine that a mixture of different options will make up the SRCs, quite probably with heterogeneity even within a single SRC.

A set of delivered capacity pledges for each Regional Centre will be agreed and maintained, in order to provide a planning road-map for future pledges. The analysis in [RD3] does not extend beyond 2028 in detail – it is important to understand that the storage requirements for the SRCs will continue to grow as time goes by, but that because these are sensitive to the ways the SKA telescopes will be used we cannot accurately predict them now – instead we must ensure that the relevant parties develop and share a rolling road map detailing the next few years' needs.

Here we include here a bare-minimum requirement for storage based on the output of [RD3] and a goal for storage capacity which is large enough to cover for the uncertainties in that study.

The sizing work in [RD3] does not consider the storage needed to support incorporation of pre-cursor and pathfinder data into SRCs to facilitate community engagement and training. To allow for this we include a goal of having 5 PBytes storage available by approximately 2022.

We anticipate that this storage capacity would be divided up between online and near-line storage with the understanding that online storage is desirable as it provides synchronous data access to support processing and to support users for retrieval or interactive use. The previous version of requirement 16 was ambiguous – it stated that data products must be available no later than 12 hours after being requested, but "availability" was not defined and the time to copy a data product from near-line into online storage will depend on the size of the product at the local area network. A new version of this requirement will be developed, ensuring that it protects users from unacceptably long waits accessing data products.

Table 3: Requirements and Goals Derived from Consideration of Storage Capacity

ID	Type	Name	Description
13	GOAL	Overall archive storage capacity of the SRCs	In aggregate, the SRCs will have a net storage capacity of at least these totals per year: T0+2: 5 PBytes (TBC) in total, all to be online T0+8: 2.5 Exabytes (TBC) by year 2028, increasing at an annual rate of around 1 Exabytes (TBC).
14	REQ	Bare-minimum storage capacity of	In aggregate, the SRCs shall provide the following annual storage capability:

		the SRCs at start of operations	T0+3: At least 24 PBytes, of which at least 22 PBytes to be online T0+4: At least 150 PBytes, of which at least 125 PBytes to be online T0+5: At least 365 PBytes, of which at least 240 PBytes to be online T0+6: At least 895 PBytes, of which at least 530 PBytes to be online T0+7: At least 1400 PBytes, of which at least 700 PBytes to be online T0+8: At least 1.7 ExaBytes, of which at least 700 PBytes to be online
15	REQ	Data security	Collectively, the SRCs will ensure that data are secure (with file loss not more than 1 per year per (TBD) files) and monitored.
16	REQ	SRC Data Availability	PLACEHOLDER (TBD)

4.4 Accessibility and Software Tools

Efficient performance of the global collective of SRCs will rely on the principle that users do not need to know within which SRC their data are located. This is so that data can be placed at the most cost-effective location, considering the size and storage costs of each object but also taking into account the anticipated processing needs of different experiments. For users to be unaware of this they must all interact with their data and the SRCs via a single platform, which we here call a "Science Gateway", common across all SRCs. The SRCs must therefore function as a federated cloud, within which software can be shared across elements interoperably.

We must also consider the various needs of the end users – many will not be trained interferometry radio astronomers and these users will need access to software that enables them to interact with and run processing pipelines on the data products. At the other end of the spectrum, there will be many users (or collaborations of users) who are world experts in their data processing areas and who need to be able to develop and run their own algorithms within the SRC environment, and to share these with other users.

Table 4: Requirements and Goals Derived from Consideration of the Accessibility and Software Tools

ID	Type	Name	Description
17	REQ	Common Environment	Each SRC shall support use of a common environment across all SRCs.
18	REQ	Common Software Tools	Each SRC shall maintain, at a minimum, a common set of software tools.
19	REQ	Science Gateway	The SRCs will host a single Science Gateway used by all SRC users, compliant with SKAO policies on User access interface.
20	REQ	External software	The SRCs will enable users to develop and run software in their sites.

4.5 Data Processing Capacity

The Regional Centres will be required to perform a variety of different tasks for users, including combining Observatory Data Products from within the same project (to achieve deep integration on the sky), performing simulations to compare results to models, and visualising (archived and scratch) products in some way. Different applications will be suitable for different hardware architectures and we imagine that a mixture of different options will be present across different SRCs, quite probably with heterogeneity even within a single SRC.

Resource pledges for each Regional Centre will be collected, and a planning road-map for future pledges will be maintained, ensuring that this is both justified and feasible. We do not yet know how much compute capability will be needed across the SRCs overall. Indeed, it is likely to grow with time in response to changes both in the SKA and in the cost of processing. The actual desire is, of course, to have sufficient capacity to meet the scientific demands of the community. In [RD3] we identified a need to scale up to a 35 PFlops capability from 2024, based on the data output of SKA itself. During the years prior to that the processing scaled from SKA's data rates is expected to be small. However, provision of significant (10PFlops scale) compute resource to the SRC prior to 2024 would enable users to run data challenges, perform simulations and in general gain confidence with the SRC framework, so this has been included as a goal.

This processing capability might be divided up between batch processing and interactive sessions to ensure that interactive work is able to be suitably responsive.

Table 5: Requirements and Goals Derived from Consideration of Data Processing Capacity

ID	Type	Name	Description
21	REQ	Overall processing capability of the SRCs	In aggregate, the SRCs will provide an annual average of 35 (TBC) PFlops (peak) by T0+4.
22A	GOAL	Early processing capability of SRCs	The SRCs will collectively provide annual average processing capability of at least 10 (TBC) PFlops (peak), by T0+2.
22B	GOAL	Enhanced overall processing capability of the SRCs.	The SRCs will collectively provide an annual average of at least 80 (TBC) PFlops (peak) by T0+4.

4.6 Network Connectivity

For the SRCs to be maximally useful, every SKA Science Data Product must be stored and made available. The Observatory will therefore need to monitor the achieved network speeds of the links into the SRCs to enable delivery of data products to be scheduled. When data products are transferred into or between SRCs, this must be done without loss of integrity.

Table 6: Requirements and Goals Derived from Consideration of Network Connectivity

ID	Type	Name	Description
23	REQ	Network monitoring	The SRCs will provide a system to regularly monitor the end-to-end performance of all network links.
24	REQ	Observatory data product ingest rate	Across the SRCs, the rate of ingest of SKA Observatory Data Products must match the rate at which they are dispatched. This is expected to be up to 100 Gbit/s per telescope site by 2025.
25	REQ	Data integrity	Each SRC will use data transfer protocols that ensure data integrity during data replication into the SRCs from the SKAO and between SRC sites.