SKA REGIONAL CENTRE REQUIREMENTS

Document Number SKA-TEL-SKO-0000735
Document Type ..... RSP
Revision 01
Author ..... R. C. Bolton and the SRCCG
Date 2017-09-27
Document Classification ..... UNRESTRICTED
Status ..... Released

Name	Designation	Affiliation	Signature
Authored by:			
Rosie Bolton et al.	SRC Project Scientist	SKAO	R. Bolton
Owned by:			
Antonio Chrysostomou	Head Science Operations Planning	SKAO	Antonio Chrysostomou
Approved by:			
Gary Davis	Director of Operations Planning	SKAO	GRD
Released by:			
Phil Diamond	Director General	SKAO	

DOCUMENT HISTORY

Revision	Date Of Issue	Engineering Change Number	Comments
01	2017-07-04	-	First release of SRC requirements

DOCUMENT SOFTWARE

	Package	Version	Filename
Wordprocessor	MsWord	Word 2007	SKA-TEL-SKO-0000735_01-SKA_Regional_Centre_Requirements
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

FACT .....Fairness, Accuracy, Confidentiality, Transparency
FAIR .....Findability, Accessibility, Interoperability, Reusability
IVOA.....International Virtual Observatory Alliance
MoU.....Memorandum of Understanding
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. "Observatory": The SKA Observatory.

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

3. "Reproducibility" refers to the idea that the results of an experiment can be confirmed by an external researcher, using the same set-up, input data and methods.

4. Science data products fall into two categories:

    a. "Observatory data products": those generated within the Observatory and delivered to the SRCs
    b. "Advanced data products": those generated within the SRCs

5. "Scratch": Scratch refers to temporary working area of computing facility, associated with a users the execution of a particular work flow or project, which is not part of the archive.

6. "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 Centre Alliance (that is, the all the SRCs collectively), its ability to take in SKAO data products and its total processing capability. The latter perspective gives us requirements around the "Accessibility and Software tools".

We have separated out requirements (absolutely essential items, or bare-minimum capabilities) from goals. Goals are aspirations, often performance-related which are desirable, but for which inability to achieve would mean a reduction in performance.

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

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 though, SRCs will provide a mechanism through which a given community or communities 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 Alliance. These cover six broad areas:

- “Governance”: How the SRCs are made to function as an alliance and to work serving the needs of the SKAO and 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 Centre Alliance is not defined. Whatever model is adopted, it is expected that for any SRC that has attained the relevant accreditation to join the SRC Alliance, to remain a member 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 exit the SRC Alliance but must do so according to some notice period and under the conditions specified in its MoU. Any SRC terminating its membership 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 the SRC Alliance.

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 Alliance Membership	Membership of the SRC Alliance will be awarded if individual prospective SRCs meet and maintain all the criteria set out in appropriate MoUs. 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 exit an SRC shall ensure that all data products and software tools held by it are available elsewhere in the SRC Alliance.
3	REQ	Graceful exit, SRC Alliance	The SRC Alliance shall manage the redistribution of data products hosted by any individual SRC if and when that SRC might terminate its membership of the SRC Alliance.
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 SRC Alliance.

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 science data products ready for use 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 the SRC Alliance, just as observatory data products are.

It is natural to expect that one copy of the 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 and the Alliance as a whole must enable adoption of best-practice as it evolves[1], 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 set-up, 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 science data products 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). We await policy direction from the SKA Board before defining relevant requirements on the SRCs.

[1] 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 SRC Alliance such that any transfers between individual SRCs are minimised.
9	REQ	Open Access	The SRC Alliance 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	The SRC Alliance 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 SRC Alliance, quite probably with heterogeneity even within a single SRC.

The SRC Alliance will agree a set of delivered capacity pledges for each regional centre, and provide a planning road-map for future pledges. The storage capacity needed in the SRC Alliance overall is as yet unknown – we include here a bare-minimum requirement for 2025 of 600 PetaBytes (TBC) (roughly based on the anticipated minimum delivery rate from the Observatory, with a multiplier of 4^2), with a goal of 2 Exabytes (TBC). Indeed, the annual need is likely to grow with time in response to the increasing quantity of observatory data products and advanced data products, and as the SKA community grows. A goal for the capacity to increase at 1 Exabytes per year (TBC) is recommended, the aim being to have sufficient capacity to meet demands on the archive as they evolve.

Table 3: Requirements and goals derived from consideration of storage capacity.

ID	Type	Name	Description
13	GOAL	Overall archive storage capacity of the SRC Alliance	The SRC Alliance is expected to have a net storage capacity of at least 2 Exabytes (TBC) in year 2025, increasing at an annual rate of around 1 Exabytes (TBC).
14	REQ	Bare-minimum storage capacity of the SRC Alliance at start of operations	The SRC Alliance must provide a bare minimum of 600 PetaBytes (TBC) of storage at the start of SKA1 operations.
15	REQ	Data security	The SRC Alliance 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	The SRC Alliance will ensure that all data products will be made available within 12 hours (TBC) of being requested.

This storage capacity might be divided up between online and nearline 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.

[2] This number does not include the back-up of these products within the host SRC.

4.4 Accessibility and Software Tools

Efficient performance of the SRC Alliance 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 Science Gateway, common across all SRCs. The SRC Alliance 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 interferometric 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 the SRC Alliance.
18	REQ	Common Software Tools	Each SRC shall maintain, at a minimum, a complete set (TBD) of software tools.
19	REQ	Science Gateway	The SRC Alliance will host a single Science Gateway used by all SRC users, compliant with SKAO policies on User access interface.
20	REQ	External software	The SRC Alliance will enable users to develop and run software in the SRCs.

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 make up the SRC Alliance, quite probably with heterogeneity even within a single SRC.

The SRC Alliance will agree to provide processing via a set of resource pledges for each regional centre, and will maintain a planning road-map for future pledges, ensuring that this is both justified and feasible. We do not yet know how much compute capability will be needed in the SRC Alliance 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 goal is to have sufficient capacity to meet the scientific demands of the community, estimated here to be of the order of 500 PFlops in 2025, increasing over time to 1 ExaFlops by 2030.

Table 5: Requirements and goals derived from consideration of data processing capacity.

ID	Type	Name	Description
21	REQ	Overall processing capability of the SRC Alliance c. 2025	The SRC Alliance will provide an annual average of 500 (TBC) PFlops (peak) at the start of SKA1 operations (c. 2025)
22	GOAL	Growth in overall processing capability of the SRC Alliance with time.	The net processing capacity of the SRC Alliance will increase with time, so that it can provide an annual average of 1 (TBC) ExaFlops (peak) by c. 2030

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

4.6 Network Connectivity

For the SRC Alliance 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 SRC Alliance to enable delivery of data products to be scheduled. When data products are transferred into or across the SRC Alliance, 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 SRC Alliance will provide a system to regularly monitor the end-to-end performance of all network links. Each regional centre will respond to issues appropriately and in a defined manner (TBD).
24	REQ	Observatory data product ingest rate	Across the SRC Alliance, 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 in 2025.
25	REQ	Data integrity	Each SRC will use data transfer protocols that ensure data integrity during data replication into the SRC Alliance archive from the SKAO and between SRC sites.