







## SKA DATA CHALLENGES - OVERVIEW

Document number .....SKA-TEL-SKO-00001016  
 Document Type ..... REP  
 Revision ..... 01  
 Author..... SKAO Science Team  
 Date ..... 2019-03-05  
 Document Classification..... UNRESTRICTED  
 Status.....Released

Name	Designation	Affiliation	Signature	
Authored by:				
SKA Data Challenge Coordination Team p.p. Robert Braun	Science Director	SKAO		
			Date:	2019-03-06
Owned by:				
Robert Braun	Science Director	SKAO		
			Date:	2019-03-06
Approved by:				
Robert Braun	Science Director	SKAO		
			Date:	2019-03-06
Released by:				
Philip Diamond	Director General	SKAO		
			Date:	2019-03-06

## DOCUMENT HISTORY

Revision	Date Of Issue	Engineering Change Number	Comments
A	2013-08-16	-	First draft release for internal review
01	2019-03-05		First Release

## DOCUMENT SOFTWARE

	Package	Version	Filename
Word processor	MS Word	Word 2007	SKA-TEL-SKO-0001016-01_SKA Data Challenges V3.3
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	<a href="http://www.skatelescope.org">www.skatelescope.org</a>

## TABLE OF CONTENTS

<b>1</b>	<b>INTRODUCTION.....</b>	<b>5</b>
1.1	Purpose of the document .....	5
1.2	Scope of the document.....	5
<b>2</b>	<b>REFERENCES .....</b>	<b>6</b>
2.1	Applicable documents.....	6
2.2	Reference documents.....	6
<b>3</b>	<b>INTRODUCTION.....</b>	<b>7</b>
<b>4</b>	<b>SCIENCE DATA PROCESSOR CHALLENGES.....</b>	<b>7</b>
<b>5</b>	<b>SKA REGIONAL CENTRE DATA CHALLENGES.....</b>	<b>9</b>
5.1	Further detail about the ESCAPE project.....	9
<b>6</b>	<b>SCIENCE/KEY SCIENCE PROJECT CHALLENGES .....</b>	<b>10</b>
<b>7</b>	<b>RESOURCING AND FORWARD LOOK .....</b>	<b>12</b>

## LIST OF TABLES

Table 1. Data Challenges in the SDP area .....	8
Table 2. Data Challenges in the Science area .....	11

## LIST OF ABBREVIATIONS

AGN.....	Active Galactic Nucleus
DL.....	Data Layer
FITS.....	Flexible Image Transport System
FoV.....	Field of View
FWHM.....	Full Width Half Maximum
LAS.....	Largest Angular Size
PSF.....	Point Spread Function
SDP.....	Science Data Processor
SFG.....	Star-Forming Galaxy
SRC.....	SKA Regional Centre
SKA .....	Square Kilometre Array
SKAO .....	SKA Organisation

# **1 Introduction**

## **1.1 Purpose of the document**

The purpose of this document is to provide an overview of the SKA Data Challenges.

## **1.2 Scope of the document**

In this document, we briefly introduce the concept of SKA Data Challenges and subsequently provide an outline of how the various activities covered by this designation have been organised and are now being addressed.

## 2 References

### 2.1 Applicable documents

The following documents are applicable to the extent stated herein. In the event of conflict between the contents of the applicable documents and this document, **the applicable documents** shall take precedence.

[AD1] Applicable Document 1

[AD2] Applicable Document 2

### 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] Bonaldi, A, et al. 2018, "SKA Science Data Challenge 1: Data Description", SKA-TEL-SKO-00001001.R1 [https://astronomers.skatelescope.org/wp-content/uploads/2018/11/SKA-TEL-SKO-0001001-SKA\\_DataChallengesDataDescription-signed.pdf](https://astronomers.skatelescope.org/wp-content/uploads/2018/11/SKA-TEL-SKO-0001001-SKA_DataChallengesDataDescription-signed.pdf)

[RD2]

### 3 Introduction

SKA Data Challenges will be regularly issued to the community as part of the preparatory activities leading up to a fully operational SKA Observatory and Regional Centre network. Development of the Data Challenges is being coordinated by the SKA Data Challenge Coordination Group, which has representation of the SKAO, the SKA Regional Centres Coordination Group (SRCCG) as well as the Science Data Processor design team. Since there are such a wide range of activities within the Data Challenge realm, each with their own technical requirements, it has been found useful to further organise these efforts into three more specific areas of responsibility,

1. Science Data Processor Challenges
2. SKA Regional Centre Challenges
3. Science/Key Science Project Challenges

Examples of the types of activity foreseen within each of these areas are outlined in the following sections, together with other details pertaining to each.

### 4 Science Data Processor Challenges

Activities within this area are coordinated by Feng Wang (SKAO/Guangzhou Univ.).

Some of the key focus areas are:

- Algorithms for Calibration and Imaging, particularly where there are competing approaches (hence suitable for challenge)
  - Example: DD Calibration: A-Projection versus DD-Faceting
- Computational efficiency of pipelines running on the SDP platform, working in conjunction first with Bridging activities in this area, followed by construction.
- Algorithm development community/SDP engagement

It is foreseen that these challenges will sequentially target specific groups. There is an immediate need for simulations of realistic errors and their calibration/correction to inform error budgets and requirements for the System CDR. A specific example is: Pointing self-calibration.

Some of the complications in addressing these issues satisfactorily in a timely fashion are:

- Limited Resources. Current SDP Data Challenges mainly focus on the calibration and imaging rather than computing technology. The problem is that calibration and imaging are too specialised to find enough community volunteers. A web search for “radio interferometer [calibration, imaging]” in Google Scholar results in about 12000 relevant literature references. However, this is in contrast to attendance of the Radio Interferometer Calibration and Imaging (CALIM) group meetings over the past decade, where only about 35 – 50 people from 8 – 15 affiliations have taken part. This implies there are very few work teams in this area.
- Staffing Requirements. Reaching the objectives of SDP Data Challenges is not an easy task. Two types of people are urgently needed: radio astronomers and computer HPC/Cloud Computing specialists. In the latter case the project should look to Bridging (and eventually construction) resources to lead and provide sufficient support to the data challenges effort. However, such individuals lack in-depth knowledge and first-

hand experience of radio astronomy, especially interferometry, so this domain expertise has to be recruited as well.

- Furthermore, the data challenges need support from member countries in terms of the systems required. Timely access to suitable e-Infrastructures – HPC clusters, community clouds etc. – is clearly crucial. This support is also required as part of Bridging and then the main construction effort. The Board are requested to help facilitate such access through contacts within the relevant agencies for e-Infrastructures in the member countries.

Some more specific suggestions for SDP Challenges are outlined in the Table below. SDP data challenges will be fall into two main categories: (i) scaling of current state-of-the-art algorithms on the SDP platform as it evolves through Bridging and construction; (ii) performance of new algorithms as they emerge (having been initially tested and verified through the science data challenges).

NO.	Category	Data Challenges	Staffing Requirements	For HPC Community	Timeline
1	Algorithm Trade-offs	1. DD-Cal: A-Projection vs Faceting 2. W-Term 3. Grid/De-grid 4. Weighting methods		Partial	1 and 2 will be performed before July 2019. Partial work will be included in SDP Bridging.
2	Algorithm Optimisation		HPC/Cloud	Yes, but is very specific to the computing environment (CPU/GPU, Spark/StarPU/DALiuGE)	First challenge task will possibly be released in March 2019
3	High Performance Algorithm Development		Astronomer + HPC	Partially, but need to determine technical approach, computing platform, development language and interface.	
4	Engineering	System Simulations of realistic errors	Astronomer + Engineer	No	Already started in Jan 2019.

**Table 1.** Data Challenges in the SDP area



## 5 SKA Regional Centre Data Challenges

Activities within this area are coordinated by Rosie Bolton (SKAO).

Some of the key focus areas are:

- Data movement
- Data formats
- Protocols
- Security/data federation
- Databases

These activities relate to the development of SRCs and they will involve the global SRC community. As the Board will be aware, the SKA Regional Centre Coordination Group (SRCCG) is wrapping up its activities in Q1 2019, and a new body, the SRC Steering Committee will take on the role of developing the SRC models. Members of the SRCSC will have access to resources that can be used to further the design and implementation of SRCs.

A very likely important role of the SRC Steering Committee - most likely via a dedicated sub-group - will be to collaborate with the office on technical SRC prototyping tests that are needed and to offer resources (hardware, network link capacity, expertise and actual FTE effort) to help this effort. These may include data challenges that will not necessarily require direct input from the SKA science community, unlike some other Data Challenges that will be developed. Further detail on plans for these tests will follow after the SRCSC meets and determines its first subgroups and objectives.

Within continental Europe, a significant amount of SRC prototyping and development work will be carried out as part of the H2020-funded “ESCAPE” project. SKAO is a beneficiary of this project and will appoint three new employees. In addition to working within the ESCAPE project these new staff will have sufficient time available to ensure that this work can be made globally relevant in collaboration with the SRCSC.

### 5.1 Further detail about the ESCAPE project

ESCAPE is a large consortium project, running from Feb 2019 to July 2022, answering an EU Horizon 2020 call to develop the European Open Science Cloud vision for the astronomy and particle physics EOSC cluster. ESCAPE will involve extension of existing software technologies to incorporate use cases for the ESFRI landmark projects involved.

For SKA this means interest in three key areas:

1. the development and deployment of an SKA data lake - this is a shared data management system (e.g. CERN’s rucio software) running across several international sites, using representative data sets and with quality of service requirements set by the various needs of different SKA science objectives. This represents a very important opportunity, fostering good collaboration with the other

ESFRI projects, most especially with CERN. This work will also address networking protocols as applied to SKA's large data products.

2. The extension of the virtual observatory to include radio astronomy data.
3. The development (and assessment) of a prototype Science Platform - this must include the systems by which global teams of scientists can collaborate on workflows and then deploy these to run efficiently on distributed computing and with the distributed storage. This will involve colleagues at ASTRON primarily, but also at LSST (via the UK LSST Data Access Centre).

The SKAO has been awarded £700k funding from this project, and our focus will be to roll-out the European-focussed ESCAPE prototyping work to make it globally relevant for SKA. The ESCAPE project runs until mid 2022, and the prototyping work runs continuously through the project with definition of SKA relevant work by September 2019. This will then lead into an implementation of a prototype data lake for SKA starting immediately after, and substantially mature by Q2 2020.

SKA's work within ESCAPE will feed into, and be influenced by, the SKA Regional Centre Steering Committee - most likely though a dedicated sub-group. We foresee using this SRCSC subgroup to engage with regional data centres (e.g. national HPC infrastructures, prototype SRC sites).

The distributed data management as applied to SKA-scale data, and the continued work on network transfer protocols are data challenges in the sense that the data activities are challenging to the software and hardware being tested, but they will not require open public calls for participation. When complete however, we should be able to use the systems developed under the ESCAPE umbrella as the platform to support future astronomer- (or algorithm developer-) led data challenges, and to support new sites wishing to join by providing a set of assessment criteria and tests they will need to meet if they are to perform well in the prototype SRC network.

## 6 Science/Key Science Project Challenges

Activities within this area are coordinated by Anna Bonaldi (SKAO).

Some of the key focus areas are:

- Engagement with the scientific community by demonstrating SKA capabilities
- Preparing the scientific community for SKA data formats and sizes
- Converting standard SDP products
- Defining useful added value data products
- Algorithm development for SRCs and other centres
- Feedback on specific issues e.g. transient buffers, SDP and SRC design

Some more specific suggestions for Science Challenges are outlined in the Table below.

NO.	Category	Data Challenges	Staffing Requirements	HPC Requirements
1	Source Finding and Characterisation in Image Data Products with/without instrumental systematics	<ul style="list-style-type: none"> <li>- Continuum</li> <li>- Polarimetric</li> <li>- HI Galaxy</li> <li>- Slow Transients</li> <li>- HI Intensity Mapping</li> <li>- EoR/CD Mapping</li> <li>- Solar/Heliospheric</li> </ul>	External community input + dedicated resource at SKAO	Variable from laptop/desktop to supercomputer
2	Source Finding and Characterisation in Non-image Data Products	<ul style="list-style-type: none"> <li>- Fast Transients</li> <li>- VLBI</li> </ul>	External community input + dedicated resource at SKAO	Variable from laptop/desktop to supercomputer
3	Science extraction from data products with/without noise, foreground emission or instrumental systematics	<ul style="list-style-type: none"> <li>-EoR/CD signal detection</li> <li>-BAO signal detection with/without foregrounds</li> <li>-analysis of large-scale structure signatures</li> <li>-detection of cosmic shear through weak lensing</li> </ul>	External community input + dedicated resource at SKAO	Variable from laptop/desktop to supercomputer
4	Efficient mining and science extraction from catalogues/archival data		External community input + dedicated resource at SKAO	Variable from laptop/desktop to supercomputer

**Table 2.** Data Challenges in the Science area

The science data challenges mainly target the SKA scientific community, which can be consulted and engaged through the network of the Science Working Groups (SWGs). However, actual participation of SWG members in the challenges depends on how well each specific challenge aligns with their own scientific interest and research priorities, and on their level of commitment to other projects. For example, some are heavily involved in the analysis of precursor/pathfinder data, which may prevent them from participating in our challenge exercises. Careful planning and design of each challenge is necessary for them to be useful and interesting enough for the community, while fulfilling SKAO needs. Putting in place some form of reward, such as publishing the challenge results in a refereed journal authored by the participants, will be undertaken.

The Science Data Challenge team has released the first challenge in November 2018, with a provisional deadline of 15<sup>th</sup> March to submit results (further details are given in [RD1]). The challenge was advertised through our SWGs, as well as through media and a white paper on arXiv to reach beyond the SKA community. There are currently 13 participating teams both within and outside the SKA SWGs, distributed across Europe, Canada, Australia, India, China, South Africa and Chile. The outcome of this challenge in terms of number of participants, time to completion, feedback of the participants and quality and usefulness of the results, will inform the future strategy. Additional Science Data Challenges will be released at intervals of about 9 months, depending on complexity of both challenge preparation and submission processing.

## **7 Resourcing and Forward Look**

The activities outlined in the previous sections will require significant resourcing to address effectively. Coordination with relevant external and joint initiatives, such as the EU-funded AENEAS, ASTERICS and ESCAPE programmes, is ongoing and this is vital to ensure there is no duplication of efforts. Close coordination with the SKA Science Working Groups is also ongoing to ensure SKAO activities are complementary to the excellent challenges already being developed by these Groups. Staff resourcing is being provided by a number of sources, including existing SKAO staff, secondments and a dedicated post-doctoral researcher who will begin in Q3 2019 to support the Science Data Challenges in particular. Computational resourcing will be sought from a variety of sources both internal and external to the SKAO. Early challenges will be constrained by the capabilities of accessible systems.

Priorities for addressing the SKA Data Challenges are defined by the three individual Challenge coordinators and their teams, reflecting the assessed urgency of the work. Overall progress and priorities are regularly reviewed by the SKA Data Challenge Coordination Group. It is foreseen that the three streams will progress in parallel in the first instance, with later challenges involving a more nearly end-to-end approximation of the actual data flow and astronomy user interaction. Over time, increasingly realistic data rates and processing complexities will be simulated.