SKAO SCIENCE DATA PRODUCTS: A SUMMARY

Document Number SKA-TEL-SKO-0001818
Document Type ..... NOT
Revision 01
Author ..... Shari Breen, Rosie Bolton, Antonio Chrysostomou
Date 2021-05-15
Document Classification ..... UNRESTRICTED
Status ..... Released

Name	Designation	Affiliation	Signature	
Authored by:				
Shari Breen	Interim Head of Science Operations	SKA Office	Shari Breen	
			Date:	2021-05-18
Owned by:				
Antonio Chrysostomou	Deputy Director of Operations	SKA Office	Antonio Chrysostomou	
			Date:	2021-05-18
Approved by:				
Lewis Ball	Director of Operations	SKA Office	Lewis T Ball.	
			Date:	2021-05-17
Released by:				
Lewis Ball	Director of Operations	SKA Office	Lewis T Ball.	
			Date:	2021-05-17

DOCUMENT HISTORY

Revision	Date Of Issue	Engineering Change Number	Comments
A	2021-04-07	-	First draft release for internal review
B	2021-04-28		Revisions including feedback from Operations staff and members of the SRCSC WGs
C	2021-05-11		Final comments from all signatories
01	2021-05-15		1st Release

DOCUMENT SOFTWARE

	Package	Version	Filename
Word processor	MS Word	Word 16.27	SKA-TEL-SKO-0001818-01_DataProdSummary.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 FIGURES

LIST OF ABBREVIATIONS

ADP..... Advanced Data Product
Co-I..... Co-Investigator
CSP.....Central Signal Processor
FFT.....Fast Fourier Transform
FOV.....Field of View
GSM.....Global Sky Model
IVOA..... International Virtual Observatory Alliance
KSP.....Key Science Project
LFAA.....Low-Frequency Aperture Array
LSM..... Local Sky Model
LTP..... Long Term Preservation
ODP..... Observatory Data Product
OLDP.....Observation-Level Data Product
PI..... Principal Investigator
PLDP.....Project-Level Data Product
PSS.....Pulsar Search
PST.....Pulsar Timing
QA..... Quality Assessment
SB..... Scheduling Block
SDP..... Science Data Processor
SKA..... Square Kilometre Array
SKAO.....SKA Observatory
SRC..... SKA Regional Centre
SRCSC .....SKA Regional Centre Steering Committee
SWG.....Science Working Group
ToO..... Target of Opportunity
WG..... Working Group

1 Introduction

1.1 Purpose of the document

This document provides a summary of the data products that SKA users can expect, as well as the processes through which they will be delivered. The information presented here is largely derived from the "Observatory Establishment and Delivery Plan" [AD1] but with some additional details or emphasis appropriate for the provision of this brief reference document for the SKA Regional Centre Steering Committee and Science Working Groups.

We further hope that the information provided here will guide the expectations of the community as users adapt to the SKAO workflow. The SKAO will provide users with science data products, processed according to the parameters and pipelines selected by the users prior to observations being executed. Following the submission of these processing details there will be no further user interaction with the Science Data Processor (SDP).

This is, however, not to say that users will not be able to interact with, and potentially amend the chosen SDP pipelines or parameters as the observations for a large project begin. The SKAO recognises that there will sometimes be a need to observe a small fraction (i.e., a few hours to a few percent of the project depending on the individual requirements) of a large project and deliver the derived data products together with the calibrated visibility data to the SRC Network for consideration by the project PIs. Based on these data, the PIs may wish to test and fine tune the selected SDP pipelines (which will be accessible within the SRC Network) and ultimately amend the requested SDP workflow for the full project. Following this process, it is expected that the remainder of the project will be processed by the SDP without further interaction with the project PIs. This is a mutually beneficial process to ensure that the completed project satisfies the science goals.

While a similar process might be available for standard PI projects until an observing mode or capability has been fully verified, PIs of these projects should not expect a similar SDP feedback stage to be available to them following these initial stages of Operations.

Some users might be inclined to request visibility data as the final SKA data product, and we emphasise that while this is a recognised SKA data product, the visibilities are only expected to be delivered in exceptional circumstances (perhaps less than 1% of projects in an observing cycle). A request for visibility data will need to be accompanied by a detailed plan for the generation of appropriate science data products.

1.2 Scope of the document

This document includes pertinent information about the provision and delivery of SKA data products to science users, as well as the advanced analysis that users might conduct within the SKA Regional Centre (SRC) Network. As a subset of the information provided in [AD1], we refer the reader to that document for further information. While the information provided here and in [AD1] provides the current plan for SKAO science data products, we recognise that the science community finds some of the additional detail provided in "Science Data Processor anticipated data products: A quick guide for SWG members" [RD1] to be useful. Until we endeavour to replicate the kind of detail, [RD1] may remain a reference document within the community, as a supplement to this document. We caution that in the event of conflict between [RD1] and this document, this document shall take precedence.

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] SKA-TEL-SKO-0001722, SKAO Establishment & Delivery Plan, Rev01

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] Science Data Processor anticipated data products: A quick guide for SWG members

3 Data products

In general, the SKAO defines three types of SKA science data products split between two categories:

Observatory Data Products (ODPs): Observation-level data products (OLDPs) are calibrated data products generated by SDP workflows and are based on data obtained from a single execution of a scheduling block (SB).

Project-level data products (PLDPs) are calibrated data products generated by combining several, related, observation-level data products, delivering the requirements of the PI as outlined in their original proposal.

Software pipelines to generate both OLDPs and PLDPs will be specified in advance of the SBs being scheduled. The Observatory is responsible for the generation of both types of ODPs, providing the workflows, software, ensuring quality assessment (QA), reproducibility and that both product types are appropriately stored and made available to users (i.e. archived). OLDPs will be generated by the SDP, but the generation of PLDPs will often require the combination of data taken at different epochs (e.g. for a deep integration or large mosaic) and will therefore require SRC Network resources. Generation of ODPs of either type (OLDP or PLDP) will not require any interaction by the science users. In the case of large projects, it is acknowledged that it may be necessary to allow PIs to check the output of the selected SDP pipeline on some test data (with a feedback loop to adjust SDP pipeline parameters before going “full steam ahead”), or to provide multiple data products from the SDP (e.g. using a family of different clean boxes, or a set of different imaging parameters etc.) whilst the visibility data are still available to establish the best parameters for future processing.

Advanced Data Products (ADPs): These are the user-generated products, produced through the detailed and rigorous analysis and modelling of Observatory data products (either at the observation or project level). The generation of ADPs will usually require some level of interactive visualisation and examination of

data, as well as comparison to data from other SKA observations or other facilities.

Science users are responsible for the generation of ADPs.

3.1 Pipelines and Data Products

OLDPs will be generated by the SDP using workflows and pipelines specified within the SBs. There is no user interaction with the SDP, but users will be able to define workflow parameters in their project SBs which contain definitions of “processing blocks”, including technical details such as required spatial and spectral resolutions as well as continuum image bandwidth intervals. These will be set in advance of the observations being scheduled using the Observation Design Tool. A processing block is an atomic unit of data processing for the purposes of the SDP’s internal scheduler. Each processing block will reference a processing workflow and each SB will indicate one or more processing blocks to be used (specifying, e.g., ingest, self-calibration, Data Product preparation).

The complete list of OLDPs that SDP will be capable of generating is reproduced below, from [RD1]. Each of these will have associated data processing and QA information included in a log file.

Image Products 1: Image Cubes	• Imaging data for continuum, as cleaned restored Taylor term images (N.B. no custom image products for slow transient detection ("fast imaging") have been specified – maps are made, searched, and discarded). • Residual image (i.e., residuals after applying CLEAN) in continuum. • Clean component image (or a table, which could be smaller). • Spectral line cube after (optional) continuum subtraction. • Residual spectral line image (i.e., residuals after applying CLEAN). • Representative point spread function for observations (cut-out, small in size compared to the field of view (FOV)).
Image Products 2: uv Grids	• Calibrated visibilities gridded at the spatial and frequency resolution required by the experiment. One grid per facet (the FFT of the dirty map of each facet). • Accumulated weights for each uv cell in each grid (without additional weighting applied).
Calibrated Visibilities	Calibrated visibility data and direction-dependent calibration information, with time and frequency averaging.[1]
LSM Catalogue	Catalogue of a subset of the Global Sky Model (GSM) containing the sources relevant for the data being processed. These are the sources in the FOV, as well as, potentially, strong sources outside of the current FOV. Initially, the Local Sky Model (LSM) is filled from the GSM. During data processing, the sources found in the images are added to the LSM. The resultant LSM might be superior to the GSM (not just to account for variability, but possibly also because of the addition of longer baselines, for example) and can be used as a starting LSM for future observations of the same field.

[1] a null calibration table with zero averaging could be applied to allow access to raw visibilities in exceptional circumstances.

Imaging Transient Source Catalogue	Time-ordered catalogue of candidate transient objects pertaining to each detection alert from the real-time, fast imaging pipeline.
Pulsar Timing Solutions	For each detected pulsar, the output data from the pulsar timing section will include the original input data (channelised time-series of complex voltages) as well as averaged versions of these data products (either averaged in polarisation, frequency, or time) in PSRFITS format: arrival time of the pulse; residuals from the current best-fit timing model for the pulsar.
Transient Buffer Data	Voltage data passed through from the CSP when the transient buffer is triggered.
Sieved Pulsar and Transient Candidates	A data cube which will be folded and de-dispersed at the best dispersion measure, period and period derivative determined from the search; A single ranked list of non-imaging transient or pulsars (as appropriate) candidates from each SB. For those transients or pulsars deemed of sufficient interest based on a set of parameters, the associated “filterbank” data will also be archived. A set of diagnostics/heuristics will include metadata associated with the scheduling block and observation. Discovery of sufficiently interesting pulsars will generate an alert as well as being recorded in a log.
Science Alerts Catalogue	Catalogue of (IVOA formatted) science alerts produced and communicated by the SDP. This catalogue provides a searchable and retrievable record of past alerts.
Science Product Catalogue	A database relating to all science data products processed by the SDP. It includes associated scientific metadata that can be queried and searched and includes all information so that the result of a query can lead to the delivery of data.

Multiple different data products can be produced from the same observation, limited only by scientific justification and resource availability. No combination of data products should be considered innately mutually exclusive. However, consideration of the overall SDP processing load required for the generation of all products associated with a particular SB will be needed at the time of proposal assessment (as part of technical feasibility), but also at the project planning stage and as the SB is scheduled.

Delivery of raw visibility data as a data product (with or without averaging and/or calibration) is technically possible and is likely to be necessary for limited cases while the development of robust calibration pipelines continues in early Operations. However, in steady-state Operations, the SKAO is responsible for the delivery of calibrated data products and proposals requesting raw visibility data are expected to be very much the exception and will require a detailed plan for calibration and the generation of data products.

3.1.1 Project-level Data products

For many projects it will be necessary to combine multiple related OLDPs to fulfil the science goals outlined in the observing proposal. Given the limited capacity of the SDP and the time between the execution of the respective SBs for a project, PLDPs will be generated within the SRC Network but will remain the responsibility of the SKAO. Like OLDPs, these will be created using SKAO workflows and will have associated data processing and QA information included in a log file. As the combination of OLDPs, PLDPs will necessarily be drawn from the same list of possible SDP data products given above.

Users will have access to each of the OLDPs that have been used to create their PLDPs and will therefore retain the ability to create their own versions within the SRC Network. This gives users the flexibility to, for example, exclude an epoch in a deep combined integration if scientifically advantageous (since each constituent OLDP will have passed QA it shouldn't be technically necessary).

3.2 Uniformly reduced Science archive data products

Given that, in the majority of instances, visibilities will not be retained once final user-requested data products have been generated, the SKAO will consider options for producing a uniform set of data products using standard processing parameters, in addition to those requested by the PIs. This could increase the legacy value of SKA observations, but the feasibility of the additional processing and archiving load is yet to be fully considered. Whatever the case, any additional data products generated will not be released beyond the original project team until the proprietary period has concluded.

3.3 Quality Assessment

All ODPs will have an associated QA log from the entire SKA processing chain, stored in the project log, ensuring that it is accessible to the Operations staff at the Observatory and to science users. These QA logs will chiefly contain information generated by the SDP (on astrometry, photometry, radiometry, polarimetry and spectrometry), but will also contain other relevant information from the CSP (specifically in the PSS and PST) and the SKA-Low Monitoring, Control and Calibration System. Logs will be linked to the relevant ODP and stored in the project log.

3.4 Lifecycle of Data Products

OLDPs generated within each SDP will be delivered to SRC Network where users will access them. Once all planned SDP pipelines needing a particular set of raw data have been completed, those data can be deleted to free space in the SDP's buffer. This applies to both the hot buffer, where data are stored only whilst batch processing is being run and from which an almost immediate deletion of raw data is anticipated once a batch pipeline completes, and also to the cold buffer.

In addition to being delivered to SRC Network, all OLDPs will be stored in a long-term preservation (LTP) system for both telescopes. Once delivered to the SRC Network, users will be able to access these data products if they have appropriate permissions. As products move out of their proprietary periods, user access will be opened up to the general public. Some products may be public from the outset.

The timescales for data to make their way through the SDP and into the SRC Network are relevant to mention here. Execution of an SKA project may take anything from minutes to months, depending on the length of each scheduling block instance, the number of different scheduling blocks required and

any special weather or observational conditions they may need. For each scheduling block instance actually executed, the SDP will ingest data and work on initial pipelines in real-time, but thereafter data is placed into a "cold" buffer for storage until the bulk SDP processing, generating the OLDPs can occur - this will depend on the loading of the SDP. For example, the most challenging processing, resulting from a long (e.g., 10 hour) SB, may take several times longer than the SB duration to be processed by the SDP, thus data processing will sit in a queue, with the next-most-pressing processing selected once the SDP batch processing system finishes its current work. Based on our current estimates, this means that it could take a couple of weeks after SB execution for OLDPs to be generated by the SDP. Once OLDPs are created by the SDP they will be queued for delivery to the SRC Network. Again, this process is a bottleneck that depends on the available bandwidth, so further delays in delivery are possible and should be expected if high data volume projects are undertaken. (The metric to bear in mind here is that the anticipated connectivity of each SDP into the SRC Network is 100 Gbit/s, or ~1 petabyte per day. Capacity on this link is a resource that will need to be planned for alongside storage and compute capabilities.) The delivery of OLDPs to the SRC Network, is an area that will need to be prototyped extensively, and timescales analysed, during the construction and science verification periods.

As each OLDP is generated and made available in the SRC Network, PIs and Co-Is will have access to them in that form, but the full PLDPs (if applicable to a project) which are generated by the SKAO using SRC resources can only be created after all necessary OLDPs are in place. As the generation of the PLDPs will be the last stage before the PIs can access their final requested data products, it will be important that there is minimal delay in their creation (subject to compute availability within the SRC Network), and that user expectations on these timescales are well managed.

3.5 Roles and Responsibilities

In Figure 1, a 'swimlane' diagram is used to show where the responsibilities of the SKAO, the SRC Network, and the SKA user (whether as a PI/Co-I, or an archive user) lie with respect to delivering the SKA science programme. It shows two phases: the project execution phase and the science extraction phase. The SKAO is responsible for the project execution phase, which includes the generation, calibration, and delivery of OLDPs and PLDPs into the SRC Network. The SRC Network is then responsible for supporting the SKA community of users in extracting the science from the ODPs delivered to them for publication and dissemination.

The science extraction phase will generally result in the generation of further, more advanced data products (ADPs) as a consequence of the advanced analysis and modelling that will be employed by the science community. Those ADPs that will appear in publications, or that will be made public, will be added to the SKA science archive and made available to all users (while respecting the appropriate proprietary access periods). This will raise efficiency in the SRC Network by avoiding repetition of the processing to generate those products.

3.6 Summary of SKAO, SRC and User responsibilities in the generation of data products

The SKAO, SRC Network and user responsibilities in the generation of the different types of data products are shown in Figure 1. The boundary between each reflects the need to strike a balance between centralisation and the desire for software quality and data traceability, and the ability to declare that the Observatory's responsibilities to a specific PI or KSP team have been achieved, against the somewhat competing need to encourage scientific freedom and innovation.

It should be noted that while this figure highlights the ultimate responsibility for each of the data products, it doesn't fully capture all of the important roles played in their generation. Specifically, the generation of PLDPs are listed as the responsibility of the SKAO since we will provide and maintain the software pipelines to generate them from OLDPs, but the actual processing to generate them will be undertaken within the SRC Network.

The generation of the OLDPs and ADPs are somewhat more simply presented in Figure 1; the generation of OLDPs are the sole responsibility of the SKAO and the generation of ADPs are the joint responsibility of the users and SRC Network.

3.7 Data Management Model for the SKAO and the SRC Network

Once generated, ODPs will be delivered to the SRC Network but copies of the ODPs will remain at SDP sites in the long-term preservation (LTP) system. The LTP will be a high-latency data storage system that will store copies of the data products so that they can be re-delivered to the SRC Network if all copies in the SRC Network are lost. In other words, the LTP system is a back-up of last resort and not an actively managed storage element in the network of SRCs. The SKAO has a responsibility to ensure that data products are preserved, forever, in the LTP which is independent of SRC activity – so for example, it will *not* be possible to delete items from the LTP on the grounds that there are copies in place in one or more SRCs.

Within the SRC Network the management of data products can be more flexible. By agreeing to share the burden of data storage and access to users, SRCs can provide coverage of the whole SKA archive of Observatory and advanced data products without requiring that each SRC must keep a full (and fully backed-up) copy of each data product. Instead, there can be a global data management service that applies rules to data products or collections of data products and manages data transfer between SRCs in order to maintain adherence to the rules. The service can tag "spare" copies of data products for deletion (without necessarily performing the deletion) so that individual SRCs can clear space when resources become limited.

Using a global data management service to perform this function is essential to avoid confusion – SKAO's role in this will be to provide coordination and the necessary software to run the data management service as well as the hardware (e.g., servers) to maintain the catalogue of data product locations.

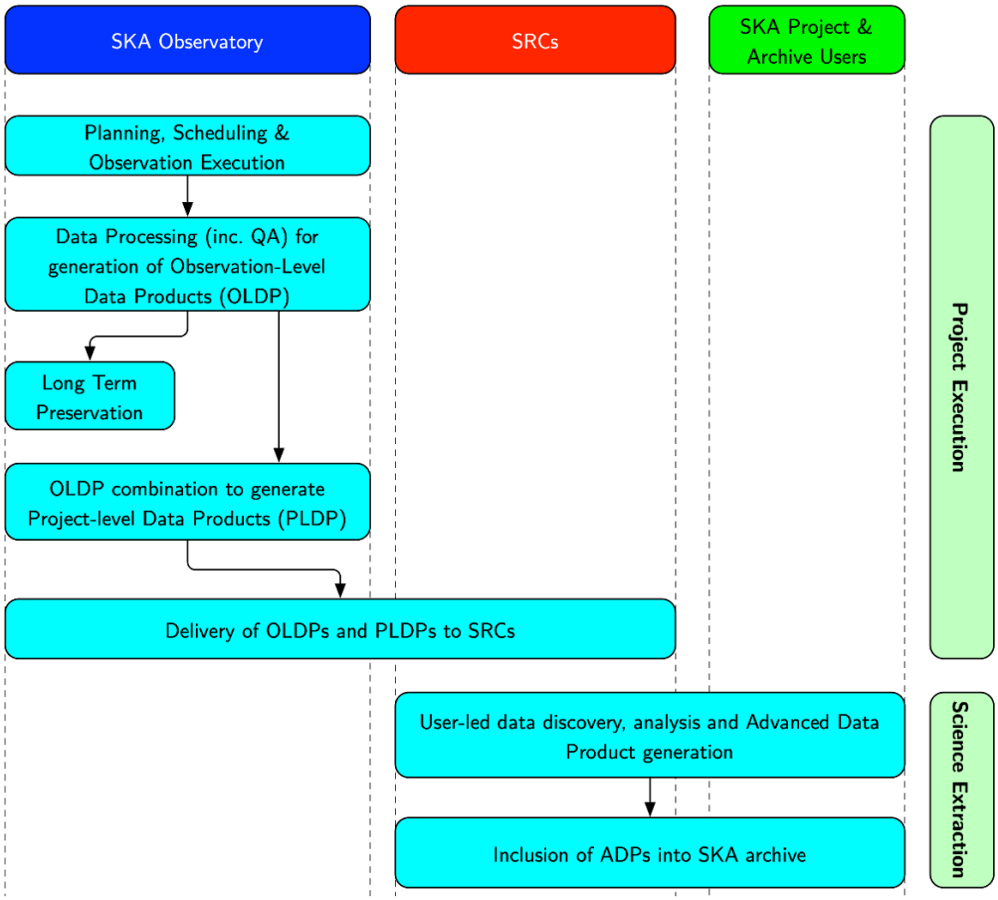

Figure 1. Swimlane diagram showing the responsibilities for the generation of science data products during the project execution and science extraction phases of a science project. The left lane depicts the Observatory's responsibilities, the middle lane those of the SRC Network, and the right lane for the PIs and Co-Is of SKA projects and general archive users. Observation-Level and project-level data products are generated by the Observatory, while advanced data products are produced by users within the SRC Network.