SDP System-level Data Model View

Document number…… SKA-TEL-SDP-0000013
Document Type……..... REP
Revision…… 05
Author……....K Kirkham, F. Graser, P. Wortmann, P. Alexander, V. Allan, C. Laker
Release Date…… 2018-04-18
Document Classification…… Unrestricted
Status……..... Released

Author	**Designation**	**Affiliation**
Kechil Kirkham	SDP SE Team	SKA-SA
Signature & Date:	Kechil Kirkham Kechil Kirkham (Apr 19, 2018)	

Released by	**Designation**	**Affiliation**
Paul Alexander	SDP Project Lead	University of Cambridge
Signature & Date:	Paul Alexander Paul Alexander (Apr 19, 2018)	

Table of Contents

List of Abbreviations

FOV	Field Of View
GSM	Global Sky Model
LSM	Local Sky Model
PSF	Point Spread Function
SB	Scheduling Block
SDM	Science Data Model
SRC	SKA Regional Centre
TelMod	Telescope Model
TM	Telescope Manager

1. Primary Representation

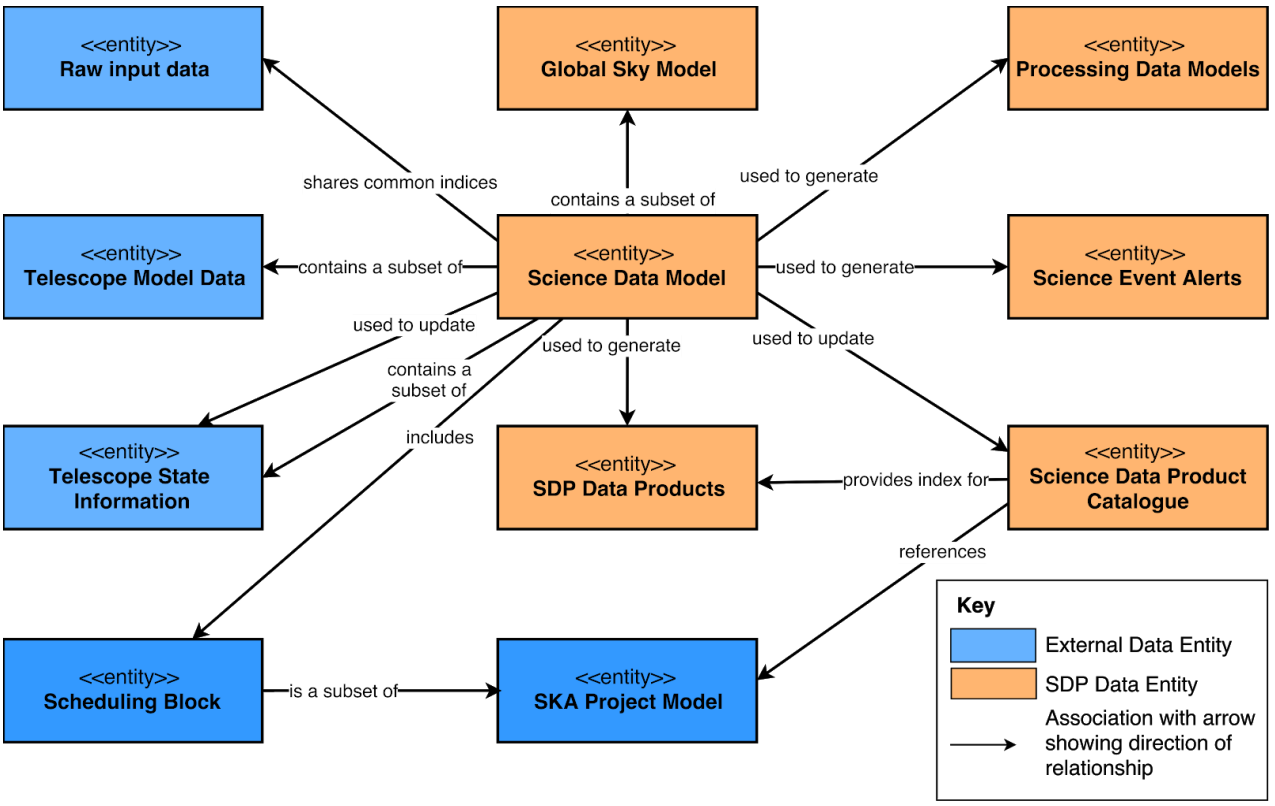

Figure 1: *SDP Conceptual Data Model diagram showing relationships between entities.* [1]

This document describes the various data models to be found in the Science Data Processor (SDP), the relationships between the SDP data models and relationships to external data models. This document should act as the central reference for SDP data models and other related documents (workflow, pipelines, etc.) should be consistent with it.

What we aim to show here is the decomposition of SDP data models, their relationships, and architectural considerations concerning these data models. Therefore each data model is described briefly and further related architectural views ('View Packets') will support each data model, and provide greater detail. This document does not describe data flows but rather logical data models. These are conceptual entities, and their relationship to data structures and data items is not detailed here. There is no common data structure associated with the SDP, unlike the ALMA science data model.

[1] *The arrows and their associated words should be read in the following way. Create a sentence in which the subject is the data model in question, followed by the verb (in the direction of the arrow), with the object being the data model on the other end of the arrow, e.g. the Science Data Model is used to generate Processing Data Models.*

During processing, different elements of the data models will be treated differently by the SDP. For example, gain tables are handled by queues, other elements may go into a DB, and there may be multiple representations of some elements. What is common is that the SDP will create data products which contain elements of these data models, and these data products will be stored and delivered to SRCs.

Note that this architectural view excludes data items used to configure and manage SDP internal components (e.g. configuration database), because they don't intersect with the data models described here.

2. Element Catalogue

2.1. Elements and Their Properties

Note that these are conceptual data models and therefore certain properties needed for implementation such as primary keys are not mentioned in this document.

2.1.1. Global Sky Model

This is a catalogue of sky components stored as a searchable database. It accommodates the following:

- Point sources with full Stokes parameters
- Time variability
- Resolved sources using compact basis set representations
- Complex spectral structures
- Information on planetary and solar radiation sources

Multiple ways to represent the sky model position are supported and interpreted by an API. Note that the Global Sky Model is not complete without a module to interpret the data.

2.1.2. Science Data Model

The Science Data Model consists of the following high level data entities:

- Local Sky Model (LSM)
- Subset of the Telescope Model (refer to Telescope Model section)
- Subset of the Telescope State Information (refer to Telescope State section)
- Gain Tables (Calibration Solutions)
- SDP QA metrics
- Processing Block
- Processing Logs
- Other required ObsCore [RD7] Data Model elements

Relationships to other data entities:

- Shares common indices with Raw Input Data
- Is used to generate Science Event Alerts
- Is used to update the Science Data Product Catalogue
- Is used to generate Science Data products
- Includes the Scheduling Block
- Contains a subset of the Telescope State Information

- Updates the Telescope State Information
- Contains a subset of the GSM
- Contains a subset of Telescope Model Data
- Is used to generate Processing Data Model

The SDM does not include the Raw Input Data, Processing Data Models or SDP Data Products, but does include references to these data entities.

At least one instance of the SDM is associated with each Scheduling Block. The working assumption is that there is one instance of the SDM associated with each Processing Block.

Note that there is no common data structure for the SDP, unlike the ALMA SDM.

2.1.2.1. Local Sky Model

The Local Sky Model (LSM) is a data structure that is initially populated by a subset of the GSM. The structure of the LSM is identical to the GSM. It contains the description of all celestial sources used during the data processing. Initially, the LSM will be populated with the sources from the GSM in the field of view (FOV) and bright sources outside the FOV affecting the visibilities. During data processing, newly found sources will be added to the LSM and existing sources in the LSM may be updated.

2.1.2.2. Calibration Solutions (a.k.a. Gain Tables)

The measured data from the telescope is corrupted by various effects. As a result an image produced from the measured data is limited in its quality. To improve the quality of the image, a model of the instrument and its environment is fitted to the measured data and used to correct the measured data for all corrupting effects. The resulting corrected image will have an improved quality.

Calibration Solutions are the fitted parameter values for the model of the instrument and its environment. These solutions are created in the Real-time Calibration pipeline and in the Calibration and Imaging pipelines that run on the Buffer. These solutions are created for each observation, but there may also be dedicated calibration observations that will use dedicated calibration pipelines.

Calibration Solutions are applied throughout the SKA telescope. Therefore, these solutions are fed back to TM. TM makes sure the solutions are distributed to the rest of the SKA System. SDP uses the Calibration Solutions internally, both in its real-time pipelines as well as in its pipelines that run on the Buffer. CSP uses the solutions to calibrate the Tied-Array Beamformer. The solutions may also be fed back to LFAA.

Calibrations Solutions are 2x2 matrices of complex values (so called Jones matrices). There is one matrix solution per Dish / AA-Station, per Direction on the Sky, per time-frequency domain (i.e. its validity domain), and per Primary Beam (only AA-Station will have more than one Primary Beam).

A Calibration Solution matrix may be a general, complex 2x2 matrix which describes all corrupting effects combined, or it may be constructed from a set of special 2x2 matrices that each describe a special effect. In the latter case, the total corrupting effect is described by multiplying all special matrices into one complex 2x2 matrix. Effects to be calibrated for are:

- Delay (K);
- Instrumental Gain (G);
- Bandpass (B);
- Cross-hand phase (XYf);
- Polarization Leakage (D);
- Atmospheric / Ionospheric Effects (Z);
- Faraday Rotation (F).

Each of these effects will have their own time- and frequency scale on which they vary.

2.1.2.3. SDP Quality Assessment metrics

A set of metrics to determine the quality of Science Data Products.

A metric is given by these parameters [RD6]:

- Name
- Value
- Date/time
- Origin (observation-id, pipeline component, …)

It should always be possible to produce a new Quality Assessment metric. There is no definitive list of Quality Assessment metrics at this moment.

2.1.2.4. Processing Block

A Processing Block is an atomic unit of processing from the viewpoint of scheduling. A Processing Block is a complete description of all the parameters necessary to run a workflow. A given Scheduling Block may contain multiple Processing Blocks. Processing Blocks have no independent existence beyond SDP.

Types of Processing Blocks:

- Real-time processing blocks
    - Receive visibilities
    - Real-time calibration
    - Imaging transient detection
    - Pulsar timing receive
    - Pulsar search receive
    - Single pulse transient detection
    - Transient Buffer receive
- Offline (batch) processing blocks
    - Calibration and Imaging
    - Pulsar timing processing
    - Pulsar search processing

This is the current list of Processing Block types but note for future reference

that this is extensible.

2.1.2.5. Processing logs

Processing logs are defined by SKA1-SYS_REQ-2336 [AD1] as “a log detailing the processing configuration” and are essential for interpretation of Science Data Products. Processing logs may also contain other information necessary for the interpretation of Science Data Products.

2.1.2.6. Other required ObsCore Data Model elements

Work still needs to be done to identify these elements.

2.1.3. Processing Data Models

Note: Previously called Intermediate Data Models / products. Processing Data Models is a placeholder for all intermediate data that a workflow may create. As such it must be extensible. The Processing Data Models contain data items used during the processing of data.

List of data items:

- Facet
- Visibility Set
    - Block Visibility
    - Coalesced Visibility
- UV Grid
- Intermediate imaging products[2]
    - Residual Image
    - Sky Components / shapelets
    - Dirty Image
    - PSF Image
- Imaging Kernels
    - W-kernel
    - A-kernel
    - Anti-aliasing kernel
    - Over-sampled kernels

2.1.4. SDP Data Products

This section provides a list and description of the data products which the SDP will deliver. SDP Data Products are independent of each other and are not derived from a common object type.

Science Alert Catalogue

Calatogue of Alerts produced during science data processing.

[2] Note that the intermediate imaging products can be used both in processing and can comprise the final preserved science data products.

Transient Source Catalogue

Time ordered catalogue of candidate transient objects pertaining to each detection alert from the realtime Fast Imaging.

Science Product Catalogue Data Product

A database relating to all Science Products which have been 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.

Image Products 1: Image Cubes

1. Imaging data for Continuum, as cleaned restored Taylor term images (n.b. no image products for Slow Transients detection have been specified – maps are made, searched and discarded).
2. Residual image (i.e. residuals after applying CLEAN) in Continuum.
3. Clean component image (or a table, which could be smaller).
4. Spectral line cubes:
    a. Spectral line cube with continuum emission remaining
    b. Spectral line cube after continuum emission subtracted
5. Residual spectral line image (i.e. residuals after CLEAN applied).
6. Representative Point Spread Function for observations (cutout, small in size compared to the field of view (FOV)).
7. Sensitivity Cubes (as per SDP_REQ-397).

Image Products 2: UVgrids

1. Calibrated visibilities, gridded at the spatial and frequency resolution required by the experiment. One grid per facet (so this grid is the FFT of the dirty map of each facet).
2. Accumulated Weights for each uv cell in each grid (without additional weighting applied).

Calibrated Visibilities

Calibrated visibility data (for example for EoR experiments) and direction dependent calibration information, with time and frequency averaging performed as requested to reduce the data volume.

Sieved Pulsar and Transient Candidates

A data cube which will be folded and de-dispersed at the best Dispersion Measure (DM), period and period derivative determined from the search; a single ranked list of non-imaging transient candidates from each scheduling block; for those transients deemed of sufficient interest, the associated “filterbank” data will also be archived; a set of diagnostics/heuristics that will include metadata associated with the scheduling block and observation.

Pulsar Timing Solutions

For each of the observed pulsars the output data from the pulsar timing section will include the original input data as well as averaged versions of these data products (either averaged in polarisation, frequency or time) in PSRFITs format; the arrival time of the pulse; the residuals from the current best fit model for the pulsar; an updated model of the arrival times.

Dynamic spectrum data

1. RFI Cleaned PSRFITS file (for Dynamic spectra mode)
2. Calibrated PSRFITS file (for Dynamic spectra mode)

Transient Buffer Data

Voltage data passed through from the CSP when the transient buffer is triggered.

Science Data Model

Each instance of the Science Data Model (refer to section 2.1.2) is stored as a data product when the processing finishes.

2.1.5. Science Event Alerts

Alerts produced by SDP following the detection of astronomical events.

Types of alerts:

- Single Pulse (non-imaging transient) Detection
- Pulsar Detection
- Imaging Transient Detection

The alerts themselves are formatted as IVOA alerts and sent to TM. These are recorded in the Science Alert Catalogue which provides a searchable and retrievable record of past alerts.

2.1.6. Science Data Product Catalogue

Catalogue of SDP Data Products. This is in itself a data product which is distributed to regional centres.

The architecture for this catalogue is of a flat data structure - there are no use cases at the present time which would require a more complex data structure. Each element within the catalogue will index all the elements of the *collection of data items* (see below) that form a SDP Data Product. The full metadata associated with a data product will itself be one of these data items. Each entry in the catalogue will however contain a subset of the metadata against which searches can be made. The one relational link is to the SKA project model (SKAPM) to associate each data item with an entry in the SKAPM,

Collection of data items: Physical data files/objects when considered together constitute a single SDP Data Product. Example: A spectral line image cube where *n* multiple files each contain a subset of the channels together with associated metadata file(s).

2.1.7. Raw Input Data

Raw input data:

- Visibility Data
- Pulsar & Transient Search Data
  - Pulsar Search Candidates
  - Single Pulse Candidates (fast transients)
- Pulsar Timing Data
  - Pulsar timing mode data (folded pulse profile data)
  - Dynamic Spectra Data
  - Flow Through mode data (raw beam-formed voltage data)
- Transient Buffer Data

2.1.8. Telescope Model Data

SDP requires a subset of the Telescope Model Data items for processing.

Minimal set of data items which SDP needs to subset:

- Beam Model:
  - Antenna Voltage Beams (Table defining beam parameters)
    - MID: Dish voltage beam
    - LOW: Embedded element pattern (Antenna dipole voltage beam) & station voltage beam
  - Antenna Positions
    - MID: Dish position
    - LOW: Station nominal centre position
- RFI (flagging) Mask. This requires a database of known RFI sources whose contents will be: Source name, Location, Strength and occupancy as a function of date/time and frequency.
- Calibration parameters (per antenna, frequency channel and polarisation)
- Measure data: Table defining leap seconds and all kinds of corrections for celestial movements.

The interface to get a subset of the Telescope Model Data needs to provide the ability to subset other items as required in the future.

2.1.9. Telescope State Information

Telescope State Information contains all information about the telescope's state, environment and behaviour. SDP requires a subset of this information for processing, and updates Telescope State Information.

Subset of Telescope State Information needed by SDP:

- Parallactic angle, or (preferred):
  - longitude of telescope
  - latitude of telescope

    - RA of phase centre
    - Declination of phase centre
- *Flags on antennas (missing etc) - if not already applied
- Time (in UTC?) or MJD reference
- Actual Time increment/step
- *Antenna pointing (actual pointing)
    - Elevation (antenna based)
    - Azimuth (antenna based)
- *Antenna on/off source (noise source)
- Current values of each of the receptor gains
- GPS ionospheric measurements (TEC values) for each station
- Weather

* Telescope State Information data needed in real-time (a.k.a fast telescope state).

SDP updates to Telescope State Information:

- Real-time calibration solutions (Jones Matrices)
- Pointing Solutions
- Antenna/station location and delay calibration solutions

2.1.10. Scheduling Block

According to [AD1], Scheduling Blocks are the indivisible executable units of a project and contain all information necessary to execute a single observation. A scheduling block may be stopped and cancelled but not paused and resumed.

All Scheduling Block data items are included in the Science Data Model. Scheduling Block data items are needed for processing or for the Science Data Product Catalogue.

Scheduling Block data items are concerned with configuration of the observation, and contain the ‘recipe’ for the observation.

The architecturally important features of the Scheduling Block are that it should be extensible, and that one Scheduling Block contains one or more Processing Blocks. The SDP cannot receive a Processing Block except as part of a Scheduling Block.

The details of what data items are required for a Scheduling Block are dealt with in a separate document called ‘Scheduling Block Data Items’.

The data items in the Telescope Model/Telescope State Information may have the same or similar name as those in the Scheduling Block. The Scheduling Block contains observation configuration parameters whereas the Telescope Model or Telescope State Information contain parameters that describe the state of the telescope, e.g. commanded pointing (in Scheduling Block) vs. actual pointing (in Telescope State Information).

2.1.11. SKA Project Model

List of data items that SDP needs to reference (via Scheduling Block ID):

- All data related to a particular project (e.g. Project ID, PI)

Note: There is an SKA system level architectural question about whether SDP should subset required data items in the Science Data Product Catalogue or whether it would be done via a database relation to the SKA Project Model database.

2.2. Relations and Their Properties

The primary representation shows all major relationships between entities. The table below lists these relationships. These are not exceptions to the primary representation diagram, but simply list the relationships depicted in this diagram. Note that the relationships do not represent data flow or interfaces between SDP and other elements. It shows the context of logical data model entities.

Relationship	**Description**
shares common indices with	This means that Item A shares common indices with item B. In this case, the Science Data Model shares common indices with raw input data. These common indices are provided by TM, and are used by SDP in processing. For example, SDP will only process raw data according to the indices provided by TM.
contains a subset of	Item A contains a subset of the data model in item B. For example, the Science Data Model contains a subset of the Telescope State Information, pertaining to the observation. It does not contain the entirety of Telescope State Information for the telescope.
is a subset of	Item A is a subset of item B, for example the Scheduling Block is a subset of the SKA Project Model. It contains information pertinent to the upcoming observation and not the entire SKA Project Model data set.
used to generate	Item A is used to create item B. For example, information in the Science Data Model is used to create Processing Data Models. This was once called intermittent data, and is used for pipeline processing.
used to update	Item A is used to update item B. For example the Science Data Model updates the Science Data Product Catalogue. Metadata concerning a processed observation is stored in the Science Data Product Catalogue, and this will have come from the Science Data Model. This metadata data itself, in the Science Data Model, will have originated from several different locations.
provides index for	Item A provides an index for item B. The Science Data

	Product Catalogue provides an index for the SDP Data Products stored in long term preservation so that they may be queried and retrieved.
references	Item A references item B. The Science Data Product Catalogue includes information from the SKA Project Model, for example data about the observation proposal, the PI etc..
includes	Item A includes item B. In this model, the Science Data Model (for a given observation) includes all Scheduling Block information relating to that observation.

Table 1: Relationship explanations, relating to the Primary Representation

2.3. Element Interfaces

Interfaces are described in the relevant ICDs.

CSP: The interface to CSP is described in the documents 300-000000-002_04_MID_SDP-CSP-ICD and 100-000000-002_04_LOW_SDP-CSP-ICD (Mid and Low respectively).

LFAA: The interface to LFAA is described in 100-000000-033_01_LOW_SDP-LFAA-ICD.

TM: The interface to TM is described in the 300-000000-029_04_SDP_to_TM_MID_ICD and 100-000000-029_04_SDP_to_TM_LOW_ICD.

The Delivery-centric Component and Connector View Document SKA-TEL-SDP-0000013 describes the relationship to the SRCs and the Observatory.

3. Context Diagram

The SDP Conceptual Data Model diagram shown in Figure 1 is a context diagram and therefore no additional context diagram is necessary.

4. Variability Guide

There is no variability at this level.

5. Rationale

5.1. Constructability

Requirements: SDP_REQ-828 (Constructability)

We needed to introduce a single high-level data entity called the **Science Data Model**, which encapsulates all information needed to process a particular grouping of data. The

grouping of data models in the Science Data Model (and its relationships) allows for the ability to pull together information from a number of sources and provides flexibility in implementation.

We have chosen not to have a unified structure (like ALMA), because of cost and implementation considerations.

5.2. Primary Functionality

Functionality required by L1 requirements is addressed by the following data models.

The **Global Sky Model** is a long-lived data entity which evolves over the course of the project. It is an explicit requirement of the observatory and serves other elements of the telescope.

Processing Data Models are required because the processing is organised as a directed acyclic graph and therefore intermediate data products must be managed explicitly as parts of the graph.

Science Data Products, Science Event Alerts and **Science Data Product Catalogue** are required by L1 requirements.

6. Related Views

Not applicable.

7. References

7.1. 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 PHASE 1 SYSTEM REQUIREMENTS SPECIFICATION SKA-TEL-SKO-0000008 Rev 10+

7.2. 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]	ALMA SDM Tables Short Description. COMP-70.75.00.00-00?-A-DSN. June 12 2015. Draft. The ALMA Export Data Format, ALMA-70.00.00.00-004-A-SPE: http://docplayer.net/48193909-Alma-export-data-format.html
[RD2]	Data Models for the SDP Pipeline Components, Ger van Diepen et al, 2/12/2016 Rev B. (unclassified).

[RD3] SDP Memo 033: Sky Model Considerations. DRAFT. Ian Heywood.

[RD4] Anna Scaife's Confluence Page on Intermediate Data Products
https://confluence.ska-sdp.org/display/SD/Intermediate+Data+Products

[RD5] Data Model Merged: Offline processing
https://drive.google.com/open?id=0B1RMQ-UciHvaN2taZVB4eWFNams

[RD6] Workflow_(pipelines)_data_model_view
https://drive.google.com/open?id=1c94AqgbIaQvplXEqyYIkCEMI68CXStVF48K3zKyPFyo

[RD7] IVOA ObsCore
http://www.ivoa.net/documents/ObsCore/20170509/REC-ObsCore-v1.1-20170509.pdf

[RD8] TMC Software Architecture Document, Rev 01, Subhrojyoti Roy Chaudhuri. SKA-TEL-TM-0000242

[RD9] OSO Software Architecture Document, Rev 01, Alan Bridger. TM number T4000-0000-AR-002. Document number 601-000000-002.

[RD10] SKA-TEL-SKO-0000893, Telescope Model Project Report, Rev 01s

8. Version History

Version	Date of Issue	Prepared by	Comments
04	2017-12-15	K. Kirkham	Submitted for SDP M19 deliverable. This document forms part of a pack implementing the following ECPs: ECP-150007 ECP-160012 ECP-160040 ECP-160048 ECP-160056 ECP-170031