SDP Memo 42: Data Model Summary for Pulsar/Transient Search & Timing

Document Number ..... SDP Memo 42
Document Type ..... MEMO
Revision C1
Author ..... R. J. Lyon, L. Levin, B. W. Stappers
Release Date 2018-03-26
Document Classification ..... Unrestricted
Status ..... Draft

Lead Author	Designation	Affiliation
R. J. Lyon	SDP.PIP.NIP Member	University of Manchester
Signature & Date:	(26/03/2018)	

SDP Memo Disclaimer

The SDP memos are designed to allow the quick recording of investigations and research done by members of the SDP. They are also designed to raise questions about parts of the SDP design or SDP process. The contents of a memo may be the opinion of the author, not the whole of the SDP.

Revisions

Revision	Date of issue	Prepared by	Comments
C	March 12th 2018	Robert Lyon	Initial version of the document.
C1	March 26th 2018	Robert Lyon	Document updated based on feedback from Andrea Possenti, Ingrid Stairs and Willem van Straten. Added text to the captions of all tables, to make it clear how parameter choices relate to / affect data product sizes. Figures 1 and 2 have been replaced by SEI diagrams. Some formulas have been reordered but the content unchanged. Clarification have been made in some places and there have been a couple of minor grammar corrections applied.

Table of Contents

Summary

This document summarises the time-domain data models, underpinning Square Kilometre Array (SKA) data processing. The models provide an abstracted representation of the information entering/exiting the SKA's science data processing system, the Science Data Processor (SDP). We solicit input from the pulsar science community to help refine them. Our ultimate goal is to ensure the SKA provides output data products which meet our collective science needs.

1 Introduction

As part of the Science Data Processor (SDP) design work, the Time-Domain Team (TDT) led by the SKA Group in Manchester, was tasked with developing data models describing a sub-set of SKA data products. Specifically, those time-domain data products created/modified when observing during,

1. pulsar search mode
2. transient search mode
3. pulsar timing mode
4. dynamic spectrum mode.

We summarise those data products here, to stimulate feedback from the research community. In particular we aim to improve our data models and clarify our metadata requirements to help ensure reproducible science is possible with the SKA. A preliminary list of metadata items is included in Table 6 for review (see Appendix A).

2 System Context

During an SKA scan[1], data is transferred from each individual antenna to the Correlator/Beamformer (CBF). The CBF yields beamformed data, which are subsequently sent to the Central Signal Processor (CSP). From here, the data are processed by internal CSP sub-systems.

- When in pulsar/transient search mode, data are sent to the Pulsar Search Sub-system (PSS). The PSS performs tasks including de-dispersion, interference removal (birdie zapping etc), acceleration searching, harmonic summing, sifting, candidate optimization and eventually folding. This PSS produces many pulsar/transient candidates which are sent to SDP for candidate selection and analysis.
- When in pulsar timing/dynamic spectrum mode, data are sent to the Pulsar Timing Sub-system (PST). The PST delivers a time-frequency-polarisation data product to the SDP for timing analysis.

In both cases data are output by the CSP sub-systems in real-time, or as close to real-time as possible, and sent to the SDP.

The SDP is responsible for performing the science analysis. This involves applying analytical methods to filter pulsar candidates (e.g. machine learning), updating timing models based on pulse arrival times, generating alerts where appropriate, and ultimately storing the data in regional centres or some other appropriate data storage/access facility. The practical realization of the data models summarised here describe i) what will be persisted by the SDP, and ii) the data researchers can access.

[1] For our purposes a scan is analogous to an observation.

3 System Constraints

The data models are constrained by fundamental SKA design requirements, and architectural decisions over which we have no control. These constraints are primarily driven by cost. This ultimately dictates the maximum possible data rate between SKA processing components, which in turn limits data product sizes. If the data models do not meet your needs[2] it is possible to modify the design. However, this may require an Engineering Change Proposal (ECP) subject to SKA office review. For example, enlarging a 1 MB pulsar candidate by 50% may seem like an insignificant change. However if 1.5×10^6 candidates are generated per scan, this increases the per scan data volume by 0.75 TB, greatly impacting data rates in practice. This would likely require an ECP. Changes that do not significantly impact data rates/volumes can made without an ECP[3]. Especially if useful for on-line/off-line science analysis.

4 Data Models

We present a condensed version of the conceptual data models that abstract away the engineering specifics. For those that wish to study the logical models, these can be found on-line (Lyon, 2017)[4]. Note these are described according to Software Engineering Institute (SEI) standards (see Merson, 2009; Clements et. al., 2010).

4.1 Pulsar Timing Use Case

An SKA sub-array comprising between 1-16 observing beams, is allocated for each pulsar timing mode scan. Given a timing target/group of targets, a single scan will observe for between 180-1800 seconds (user configurable), producing a time-frequency-polarisation data product.

4.1.1 Pulsar Timing Data Product

The timing data product consists of two logical components. This design is summarised in Fig. 1. The largest component is a d-dimensional data cube[5]. There is precisely one cube per data product. The cube describes a time and frequency averaged version of the timing observation. The size of the data cube in bits, is given by,

$$C_{size}^{time} = N_{chan} \times N_{bin} \times N_{sub} \times N_{pol} \times N_{bit} \ , \tag{1}$$

where N_{chan} is the number of frequency channels, N_{bin} the number of phase bins, N_{sub} the number temporal sub-integrations, N_{pol} the number of polarisations, and N_{bit} is the number of bits per sample in the cube. The values of these variables are user configurable. However, any combination of configuration choices must produce a data product whose transmission does not violate CSP to SDP data rate constraints. Some plausible configuration choices are listed in Table 1. Note that the full resolution data cube is also saved (via Dynamic Spectrum

[2] E.g., the time resolution is insufficient for a specific science case.

[3] Clearly this is context dependent.

[4] See `https://github.com/scienceguyrob/SKANIP_Data_Models`.

[5] In practice timing data is naturally divided in to time, frequency, phase and polarisation. In other words, a 4-d data cube. However we refer to d instead to stay consistent with SKA engineering documents.

Table 1: Anticipated pulsar timing parameter choices, which determine the size of the data products. These are described below via two typical observing scenarios, i.e. observations of individual pulsars versus observations of millisecond pulsars for pulsar timing arrays. The parameter values are not final, and will be user configurable. However they are bounded by maximum data rate considerations. The timing array figures provided (a worse case) yield a data cube ~ 12.08 GB in size per beam per scan. This size increases very marginally with metadata. When using all 16 pulsar timing beams, ~ 193.27 GB of data is produced per scan, at a rate of ~ 8.59 Gb/s (gigabits per second) when sent over 180 seconds.

Description	Pulsars	Timing Arrays
Channels (N_{chan})	512	4096
Profile bins (N_{bin})	2048	2048
Sub-int length (seconds)	1-10	1-10
Polarisations (N_{pol})	4	4
Bits per sample (N_{bit})	16	16
Sub-ints (N_{sub})	180-1800	180-1800
Scan length (seconds)	180-1800	180-1800
Number of beams	16	16

mode, See Section 4.3).

The timing data product also contains a set of metadata information. This is comprised of one or more metadata tuples[6]. The tuple format is simply *key:value:type*, which allows metadata to be self-describing. Metadata is either numerical or textual, and describes the observation/data product in some way. Metadata will include, for example, the pulsar observed, beam position, number of observing beams used, optimised dispersion measure (DM), and many more. The metadata will also describe the data product provenance[7]. The precise metadata to be used is yet to be decided. We provide a preliminary list in the Appendix, which borrows from the PSRFITS (Hotan et. al., 2004) and PSRCHIVE (van Straten et. al., 2012) header formats.

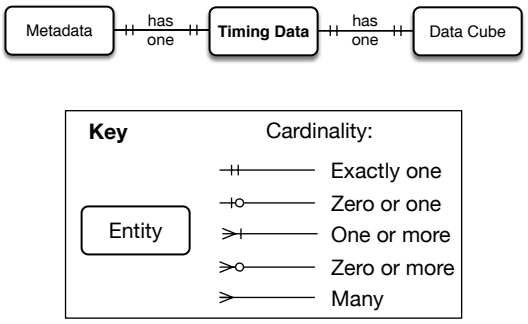

Figure 1: Conceptual overview of the generic timing data model. This is an SEI diagram (Merson, 2009) that describes the relationship between the data model entities. In reality the model can be realised, i.e. implemented, differently. It indicates that timing data has a single data cube entity, and a single metadata entity.

[6] An n-tuple is an ordered sequence of n elements.

[7] This information describes where the data product came from, the software was used to generate it, the software versions, the hardware resources used, and details of the observation (scan identifiers, noise background statistics).

4.1.2 Intermediate Data Products

Timing processing within SDP yields multiple intermediate data products. Each product is associated with metadata. This is used to ensure intermediate data are correctly linked to the 'parent' product in the SKA preservation system[8]. The intermediate products include,

1. **Averaged timing data products:** Much of the pulsar timing pipeline uses full-resolution data cubes during data processing. This is because interference mitigation and calibration greatly benefit from high resolution data. However some processing steps require higher signal-to-noise (S/N) values rather than high resolution. This requires partly averaged (also known as summed or scrunched) data products. One or more averaged intermediate data cubes will be sent to the archive for storage along with the original data cube. This information is saved for future post-processing. The quantity and size of the averaged cubes is scan specific, but will likely include cubes averaged over frequency ($1 \times N_{bin} \times N_{sub} \times N_{pol} \times N_{bit}$), time ($N_{chan} \times N_{bin} \times 1 \times N_{pol} \times N_{bit}$), frequency & time ($1 \times N_{bin} \times 1 \times N_{pol} \times N_{bit}$), or some combination.
2. **TOA List:** A Time-of-arrival (TOA) list describes pulse arrival times for a pulsar with respect to some time reference. There is 1 TOA recorded per sub-integration and frequency channel, as an intermediate data product in the pulsar timing pipeline. The TOA list is represented by a row vector of size $N_{chan} \times 1$ (or cube of $N_{chan} \times 1 \times N_{sub}$).
3. **Timing Residuals:** Timing residuals describe the difference [modulo the spin period] between the expected pulse arrival times, and the actual pulse arrival times. The residuals are represented by a row vector of size $N_{chan} \times 1$ (or cube of $N_{chan} \times 1 \times N_{sub}$).

4.2 Dynamic Spectrum Use Case

Dynamic spectrum mode (DSM) was introduced to allow the high resolution data product recorded during pulsar timing mode, to be preserved for off-line analysis. This is known as Channelized Detected Time Series data (CDTS) by the science community. DSM data is not processed within SDP, simply archived for later use by the science community (e.g. can be searched for pulsars etc.).

4.2.1 Dynamic Spectrum Data Product

The data product is structurally identical to that used for pulsar timing. There are some practical differences: i) DSM data cubes are much higher resolution than pulsar timing data products making them significantly larger in size, and ii) DSM data is not folded. As the dynamic spectrum product is not processed by SDP, there are no intermediate data products.

4.3 Pulsar Search Use Case

Pulsar search mode is used to undertake pulsar searches, via targeted observations (e.g. of the Galactic centre or Globular Clusters) or large-scale surveys. Anywhere from 1-1500 observing beams may be used when in pulsar search mode. The mode outputs a data product per beam, containing the strongest pulsar candidates found by CSP for analysis.

[8] We do not describe the metadata here for brevity.

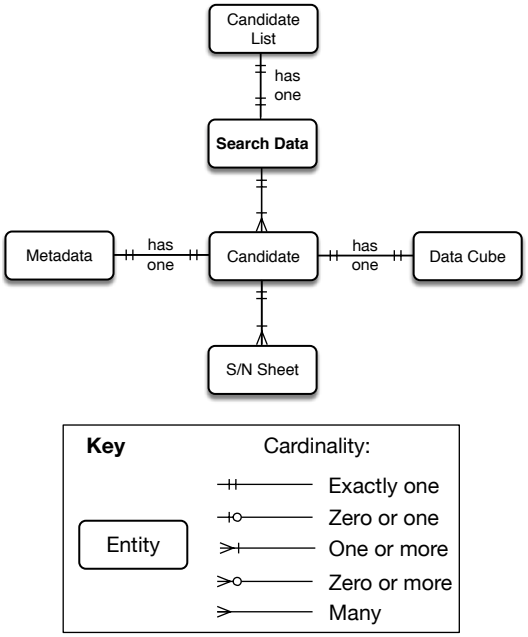

Figure 2: Conceptual overview of the generic search data model. This is an SEI diagram (Merson, 2009) that describes the relationship between the data model entities. In reality the model can be realised, i.e. implemented, differently. It indicates that search data has a single list entity, and one or more candidate entities. In turn each candidate has 1 data cube, 1 set of metadata information, and 1 or more sheet entities.

4.3.1 Pulsar Search Data Product

The search product consists of two logical components. These in turn possess subordinate 'child' components, which represent the individual candidates detected during a search. This design is summarised in Fig. 2. The search data entity is comprised of a single candidate list, and one or more candidates. The list is data structure which summarises all candidates detected during the CSP search prior to sifting. The list can have an unbounded length, though is limited in practice to describing only those detections rising above some user controlled significance level. For modelling purposes, we assume there are up to 10,000 list entries.

Each list entry describes a candidate with respect to sky location, S/N, DM, pulse width, acceleration, and period. The list will also contain flags, which indicate whether or not the candidate is considered a duplicate (i.e. identified as such during sifting). The list will be sorted, mostly likely according to S/N. The list is largely useful for speeding up on-line/off-line analyses. The current set of candidate list items are described in Table 2. We welcome suggestions for additional variables.

A search data product is primarily comprised of a collection of candidate entities. Each candidate entity is analogous to the traditional pulsar candidate. The 1,000 unique (i.e. after sifting) highest S/N candidates from the list, will be accompanied by their corresponding candidate entities. These describe a detection in detail. Whilst it is possible to store more candidates in the data product, 1,000 is practical limit given a maximum data rate constraint between CSP and SDP. The largest candidate component is a single d-dimensional data cube. It describes a time and frequency averaged version of the detection. The size of the

Table 2: The current set of candidate list attributes. This list is not final, and can be expanded according to our science needs.

Item	Description	Type
Candidate ID	The unique identifier for the candidate.	textual or numerical
Beam ID	The beam the candidate was found in.	textual or numerical
RAJ	right ascension (J2000) of the source.	textual or numerical
DECJ	declination (J2000) of the source.	textual or numerical
Period	period found during the search in milliseconds.	numerical
Pulse width	pulse width in milliseconds.	numerical
Acceleration	pulse acceleration	numerical
DM	DM found during the search	numerical
Spectral S/N	The spectral S/N found during the search	numerical
Folded S/N	The folded candidate S/N	numerical
Sifted	A flag indicating whether or not the candidate survived sifting	boolean
Duplicate	A flag indicating whether or not the candidate has harmonic duplicates.	boolean

data cube in bits, is given by,

$$C_{size}^{search} = N_{chan} \times N_{bin} \times N_{sub} \times N_{bit} \ , \tag{2}$$

where N_{chan}, N_{bin}, N_{sub}, and N_{bit} are defined as before. Each candidate is also associated with metadata, much like the pulsar timing product. However the search product also has one or more 2-d matrices which we call 'sheets'. Each candidate is expected to have three sheets computed by the CSP, which contain S/N values found during different optimizations of the folded data. The sheets used include,

1. A sheet describing the resulting S/N values for different values of period and period derivative.
2. A sheet describing the resulting S/N values for different values of period and DM.
3. A sheet describing the resulting S/N for different values of period-derivative and DM.

Sheet size is determined from the number of trial periods, period derivatives, and DMs. Note that the sheets are made available by the Folding and Data Optimization (FLDO) module, within CSP. The sheets have been included in our data models, as it is believed they may assist in helping filter out noise and RFI candidates downstream (i.e. via the application of machine learning tools). Finally the expected values for all pulsar search variables, are listed in Table 3.

4.4 Transient Search Use Case

Transient search mode is used to undertake single pulse searches, and will run in parallel with Pulsar Search.

Table 3: Anticipated pulsar search parameter choices, which determine the size of the search data products. These parameters are not strictly fixed, and will be user configurable in practice (though bounded by maximum data rate considerations). The figures provided yield a data cube ~ 1.07 MB in size. If producing 1000 candidates per beam per scan (including metadata), this equates to ~ 1.07 GB of data. When using all 1500 beams of SKA Mid, ~ 1.6 TB of data is produced per scan, at a rate of ~ 71.56 Gb/s (gigabits per second) when sent over 180 seconds.

Description	Value
Channels (N_{chan})	128
Profile bins (N_{bin})	128
Sub-integrations (N_{sub})	64
Bits per sample (N_{bit})	8
Trial periods ($Trial_p$)	256
Trial period derivatives ($Trial_{\dot{p}}$)	256
Trial Dispersion Measures ($Trial_{dm}$)	256
Scan length (seconds)	180-1800
Number of beams	1-1500
Max Candidates per beam	1000

Table 4: Anticipated transient search parameter choices, which determine the size of the data products. These parameters are not strictly fixed, and will be user configurable in practice (though bounded by maximum data rate considerations). The figures provided yield a data cube ~ 2.62 MB in size. Assuming an event rate of 1 single pulse event each second, over a 180 second scan, this yields ~ 471.93 MB of data (including metadata). When using all 1500 beams of SKA Mid, ~ 707.89 GB of data is produced per scan, at a rate of ~ 31.46 Gb/s (gigabits per second) when sent over 180 seconds.

Description	Value
Channels (N_{chan})	1024
Samples per event	640
Polarisations (N_{pol})	4
Bits per sample (N_{bit})	8
Trial Dispersion Measures ($Trial_{dm}$)	1024
Scan length (seconds)	180-1800
Number of beams	1-1500

4.4.1 Transient Search Data Product

Transient search data products are identical to pulsar search data products in logical structure. The only difference is in the parameters which determine data product size, defined in Table 4. We do not re-describe the data models here for brevity.

5 Review Questions

We seek your help in i) reviewing the existing models, and ii) adding new metadata information which could help improve the science data processing. When reviewing the models it may be helpful to consider,

1. Do they permit the analysis you *currently* do?
2. Do they permit the analysis you *would like* to do?
3. What additional informational do the models need?

4. What metadata is useful for enabling the reproducibility of a science analysis? For example TOA metadata must contain a reference to the standard profile and fitting method. An initial metadata list is provided in Table 6 in Appendix A for review.

5. What metadata would help you query the information, assuming it is stored in a regional centre data warehouse?

6. Are there any useful metadata items that are trivial to add during processing, but difficult to compute off-line?

7. Are their any use cases for which these data models do not work?

It would be especially beneficial to suggest metadata items that may not be present in the existing models. As a starting point, consider the pulsar catalog. This describes known sources according to many metrics, some of which are derived from the catalogue itself. Are there such values that should/could be added? Note that we present a preliminary list of metadata items in the Appendix.

5.1 Summary

We have summarised the current design of the time domain data models for the SDP. Whilst the design is mature, it is crucial to incorporate community to ensure they meet our needs. It is not too late to incorporate refinements, so long as these align with SKA science goals requirement constraints.

6 Acknowledgements

We thank Andrea Possenti, Ingrid Stairs & Willem van Straten for their reviews and helpful comments.

References

Clements P. et. al., 2010, “Documenting Software Architectures Views and Beyond”, 2nd Edition, Addison-Wesley.

Hotan A. W., Van Straten W., Manchester R. N., 2004, “PSRCHIVE and PSRFITS: an open approach to radio pulsar data storage and analysis”, Publications of the Astronomical Society of Australia, vol.21(3), p.302–309. doi:10.1071/AS04022.

Lyon R. J., 2017, “CSP to SDP NIP Data Rates & Data Models (version 1.1)”. doi:10.5281/zenodo.836715.

Merson P., 2009, “Data Models as an Architectural View”, Software Engineering Institute, `www.sei.cmu.edu/reports/09tn024.pdf`.

van Straten W., Demorest P., Oslowski S., 2012, “Pulsar Data Analysis with PSRCHIVE”, astro-ph.IM, arxiv:1205.6276. `https://arxiv.org/abs/1205.6276`.

A Metadata

Here we present a first attempt at defining possible metadata for our data products. Our metadata models are not mature, thus input is welcomed.

Table 5: The provenance specific metadata under consideration. We took inspiration from the PSRFITS format (Hotan et. al., 2004), which defines header attributes that handle data provenance.

Item	Description	Type
Scheduling Block ID	the scheduling block the data was created in.	textual
Program Block ID	the program block the data was created in.	textual
Scan ID	the scan the data was created during.	textual
Subarray ID	the sub-array configuration, if applicable.	textual
Product ID	the data product identifier.	textual
CBF Pipeline Version	pipeline processing the data.	numerical
CSP Pipeline Version	pipeline processing the data.	numerical
SDP Pipeline Version	pipeline processing the data.	numerical
History	description of the steps applied to the data.	textual
Size	Size of the product in bits.	integer
Intermediate Products	flag indicating if there are child products.	textual

Table 6: The data product specific metadata, currently planned for. We took inspiration from the PSRFITS format (Hotan et. al., 2004).

Item	Description	Type
Pulsar ID	unique identifier of the observed pulsar.	textual
Candidate count	candidates in the data product.	numerical
RAJ	right ascension (J2000) of the source.	textual or numerical
DECJ	declination (J2000) of the source.	textual or numerical
Start time	time stamp (MJD) of first sample.	numerical
End time	time stamp (MJD) of the last sample.	numerical
Sampling interval	the scan sampling interval.	numerical
Bits	bits per time sample.	numerical
Samples	total time samples.	numerical
Centre Freq.	centre frequency (MHz) of the first channel.	numerical
Channel bandwidth	filterbank channel bandwidth (MHz).	numerical
Bandwidth	total bandwidth (MHz).	numerical
channels	number of filterbank channels.	numerical
DM	optimised dispersion measure.	numerical
known period	known folding period.	numerical
search period	period found during the search.	numerical
Pulsar ephemeris	ephemeris used	textual
Polyco	polyco file/s used to predict the period	textual
Predictor	predictor file used to predict the period	textual