Science Data Processor anticipated data products:

A quick guide for SWG members

Revision: 1.0 Author: R. Bolton
rosie@mrao.cam.ac.uk
Release Date: 2015-08-19

Introduction and Scope

The SDP consortium has been asked by the SKAO science team to provide a brief reference document for the SKA's Science Working Group members to refer to at the time of the Stockholm meeting happening in August 2015.

Fuller descriptions of the pipelines are available in the SDP Preliminary Design Review documentation, which is available on request from Rosie Bolton.
This document is intended to promote discussion and does not constitute any formal set of actual deliverable products or promises to deliver. Ultimately, all data products need to be incorporated into the SDP consortium's and the SKAO's system engineering description of the system by being written into requirements.

The SDP consortium and SKAO would like feedback from the SWGs on the data products required from the Science Data Processor, for delivery into the archive and thence to the end user. Feedback can be via the new science use cases, to Jeff Wagg at SKAO (J.Wagg@skatelescope.org) or directly by email to the author (rosie@mrao.cam.ac.uk).

Data products:

The science use cases (version C, dated 2014-07-18) mention several data products. These do not always distinguish between final science data products and the data products that the SKA (and specifically the science data processor itself) needs to deliver, and those which might be needed as intermediate data products but discarded once the experiment is complete and a final data product (e.g. 1000 hrs integrated image cube) is produced.

The SDP consortium needs to develop these into a standardised glossary of data products that can be optionally selected for each observation. This should seek to minimise data volumes where further products can be generated locally from the standard set, and to minimise the number of distinct products.

In no particular order (but grouped by science type somewhat), here are the distinct products mentioned in the science use cases to date and which are therefore known to the SDP consortium:

1. **Pipeline log**
    a. Quality Assurance Log (not mentioned in use cases but included here to complete the list)
2. **Global Sky Model**

3. **Restored Continuum image** cubes (with spectral and spatial resolution specified by the user)
4. **Residual continuum image** cubes
5. **Model continuum image** cubes

6. **Spectral Index map** (can user reconstruct from Taylor term images instead to avoid this extra product?)
7. **Spectral curvature map** (can user reconstruct from Taylor term images instead to avoid this extra product?)
8. **Source Catalogue for continuum**, to **include polarisation data** and intrinsic polarisation angle

9. **Spectral Line Image Cube** (with spectral and spatial resolution specified by the user, continuum subtracted)
10. Hi (or other **spectral line) candidate source catalogue**
11. Hi (or other **spectral line) spectra** for detected sources (format required)
12. **Postage stamps** for spectral line candidates
13. **Stacked spectra** based on input source list (details and format required)
14. **Stacked postage stamps** based on input source list
15. **Position-velocity maps**
16. **Spectral line moment maps**
17. **Total spectral line emission maps** (e.g. total Hi content)

18. **RM synthesis output cubes** (Q?, U?, P) [further clarification needed]
19. **Faraday dispersion cubes** from RM synthesis
20. **Catalogue of intrinsic polarisation angle** (will likely be part of continuum source catalogue)
21. **Circularly polarised continuum image cubes**?? (not clear if this is required to support VLBI observations)

22. **Time series catalogues for variable sources**
23. **"Full band dynamic spectra"** (requested in Exoplanets use case, definition required)
24. **High time resolution visibility data around transient events** (as detected by fast imaging pipeline, but if post-calibration data is required - with a latency of up to 12 hours - see item 29 below)

25. **Pulsar Candidate list** + metadata
26. **Pulsar timing pulse profiles**
27. **Pulsar metadata**
28. **Time-frequency power measurements** surrounding candidate single-pulse event detections

29. **Calibrated visibility data** (spectral and temporal averaging and selection ranges set by user or algorithm)
30. Calibrated, **gridded visibilities** (i.e. FFT of dirty map, to enable morphological work without artefacts of deconvolution). An Engineering Change Proposal (ECP150007) has been submitted requesting access to gridded visibility data as an output (as

opposed to cleaned images). We note that this ECP only relates to SKA1 MID, but we would hope that it could be implemented for both instruments.

How the Science Data Processor will run

For every imaging observation it is anticipated that the SDP will run a continuum imaging pipeline on the full single-track data (e.g. ~6 hours' worth). Polarisation, spectral line or other additional output, if needed, will be additionally generated after the continuum processing has been done and the best possible (continuum) sky model and calibration parameters have been found.

Additionally, we expect to have the capability to run a "fast imaging" pipeline on visibility data searching for transient sources in the image plane every 1s or so, with a latency of around 100s (TBD) – i.e the fast imaging can be done in near real time, with results available very shortly after the photons are collected. The useful output of this will be time-series catalogues for variable sources.

By contrast synthesis imaging data reduction will not begin until each observation is complete (e.g. 6 hours after the start of the observation) and would take the same length of time as the observation to process – so that the final data products from a particular 6 hour observation should be ready 12 hours after the observation begins.

Non-Imaging Processing is anticipated to represent a much smaller compute load. NIP data will be processed according to latency requirements, essentially decoupled from the imaging work.

Final data products are delivered into the long-term archive (and the out to the users), but our design also includes a mid-term archive (the "Medium Performance Buffer"), which has much faster data transfer and read/write capabilities than the long-term archive. We assume that this mid-term archive is large enough to enable some smoothing of data flow into the long-term archive, and to store image cubes or other **intermediate data products** from on-going observing programmes prior to combining these into the final data products.

Other useful concepts or considerations

The SDP consortium has developed some concepts and benchmark figures which may help place SWG's data needs into a bit of context – and especially highlight potential issues with data volume.

- ### "Discovery cube" size

It is useful to estimate the maximum possible archive size for each image cube (we call this a "Discovery Cube"). We can calculate this assuming that each observation requires the maximum possible spatial and frequency resolution. The numbers per (single pointing centre) observation are given by:

SKA1-LOW: Discovery cube size

$$= 1.1\ PetaBytes \times \left(\frac{\frac{\lambda_{max}}{\lambda_{min}}}{6}\right)^2 \times\left(\frac{N_f}{65{,}536}\right)\times\left(\frac{B_{max}}{70km}\right)^2$$

Equation 1

SKA1-MID, Band 2: Discovery cube size

$$= 2.6\ PetaBytes \times \left(\frac{\frac{\lambda_{max}}{\lambda_{min}}}{1.85}\right)^2 \times\left(\frac{N_f}{65{,}536}\right)\times\left(\frac{B_{max}}{150km}\right)^2$$

Equation 2

This can, of course, be used to estimate the maximum possible archive growth rates, though the results quickly become terrifying.

- ## Data Transport

As far as we (the SDP consortium) are aware, the data link capacities from the long term SKA archives and out into the rest of the world have not been specified. It may however be useful for SWG teams to note that these data links are unlikely (unofficially) to be more than a few x100Gbits/s where N is a small number.

1. 100Gbits/s gives 1 PByte/**day** of transfer.
2. It would therefore take over 2 days to export a **full** spatial and spectral resolution SKA MID image cube (though continuum data products or spectral postage stamp cubes would be (e.g. 3-4) orders of magnitude smaller than the discovery cube.
3. A single 6-hour visibility data set is around 10PBytes in size, for both LOW and MID.
4. After appropriate averaging the EoR visibility data set per 6 hours would be approximately 1-2 PBytes.

Appendix: Data products mentioned in Level 1 (L1) requirements

Some *imaging* products are listed in the L1 requirements. None are apparent for the NIP experiments:

Requirement number SKA1-SYS_REQ-XXXX	**Product name**	**Queries for SWG members to consider**
2336	Pipeline processing log	
2336	QA log	
2340 ("Continuum Data Products")	First 'n' moment image for Multi-frequency synthesis (with 'n' determined from signal to noise ratio)	…from which a continuum image and a spectral index map at any frequency can be reconstructed
2340 ("Continuum Data Products")	Residual images (if deconvolution has taken place)	
2340 ("Continuum Data Products")	Continuum Sensitivity image	How many to cover frequency space?
2340 ("Continuum Data Products")	Representative PSF image	How many to cover frequency space?
2342 ("Spectral Line emission data products")	Spectral line cube image (with or without continuum subtracted)	
2342 ("Spectral Line emission data products")	Continuum model image cube	
2342 ("Spectral Line emission data products")	Spectral line sensitivity image	at full frequency resolution presumably?
2342 ("Spectral Line emission data products")	Representative PSF image	How many to cover frequency space?

2346 ("Slow transients data products")	Time ordered catalogue	..of flux densities of variable/ transient objects
2346 ("Slow transients data products")	Sensitivity image	FLAG FOR REVIEW: not one image per dump time we hope? Presumably this is only needed for purpose of knowing the detection threshold appropriate for a particular time slot. Need to reduce the spatial and frequency resolution of this to avoid having too much data?? More detail in requirements please.
2346 ("Slow transients data products")	PSF (point spread function) image	FLAG FOR REVIEW: not one image per dump time we hope?