SDP Preservation Design

Document number	SKA-TEL-SDP-0000023
Document type	DRE
Revision	03
Authors	Markus Dolensky, Chris Broekema, Patrick Dowler Iain Emsley, Julián Garrido, Kevin Vinsen, Andreas Wicenec
Release Date	2016-07-21
Document Classification	Unrestricted
Status	Released

Lead author	Designation	Affiliation
Markus Dolensky	DATA Deputy Team Lead	ICRAR
Signature & Date:		

Owned by	Designation	Affiliation
Bojan Nikolic	SDP Project Engineer	University of Cambridge
Signature & Date:		

Approved by	Designation	Affiliation
Paul Alexander	SDP Project Lead	University of Cambridge
Signature & Date:		

Released by	Designation	Affiliation
Paul Alexander	SDP Project Lead	University of Cambridge
Signature & Date:		

Version	Date of Issue	Prepared by	Comments
1	2015-02-09	M. Dolensky	Issued for PDR
2	2015-05-29	M. Dolensky	Updated with PDR close-out actions
02C	2016-03-24	M. Dolensky	updated for d-PDR; new title & scope
03	2016-07-21	V. Allan	Prepared for SDP delta-PDR sign-off

ORGANISATION DETAILS

Name	Science Data Processor Consortium
Address	Astrophysics Cavendish Laboratory JJ Thomson Avenue Cambridge CB3 0HE
Website	http://ska-sdp.org
Email	ska-sdp-pa@mrao.cam.ac.uk

1 Table of Contents

2 List of Figures

3 List of Tables

4 List of Abbreviations

MAID Massive Array of Idle Disks

5 References

5.1 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.

Reference Number	Reference
AD01	SDP Assumptions and Non-Conformance, SKA-TEL-SDP-0000014
AD02	SDP Architecture, SKA-TEL-SDP-0000013
AD03	SDP Execution Framework Design, SKA-TEL-SDP-0000015
AD04	SDP Glossary, SKA-TEL-SDP-0000056
AD05	SDP L2 Requirements Specification, SKA-TEL-SDP-0000033

5.2 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.

Reference Number	Reference
RD01	Data Challenge Supplement, SKA-TEL-SDP-0000024
RD02	SDP Archive Size Estimates from the Parametric System, SKA-TEL-SDP-0000016

RD03	SDP Data Processor PlatformDesign, SKA-TEL-SDP-0000018
RD04	SDP Delivery System Design, SKA-TEL-SDP-0000025
RD05	SDP Local Monitoring and Control Element Design, SKA-TEL-SDP-0000026
RD06	SDP Pipelines Desing, SKA-TEL-SDP-0000027
RD07	SDP iPython Parametric Model Handbook, SKA-TEL-SDP-0000041
	RD08-RD10 were obsoleted and removed
RD11	Parametric models of SDP compute requirements, SKA-TEL-SDP-0000040
RD13	Bonnarel F.. et al. IVOA Dataset Metadata Model, 2015, http://www.ivoa.net/documents/DatasetDM/20151021/WD-DatasetDM-1.0-20151021.pdf
RD14	Louys M., et al., Observation Data Model Core Components and its Implementation in the Table Access Protocol, 2011. http://www.ivoa.net/documents/ObsCore/20111028/index.html
RD15	McDowell J., et al., IVOA Spectral Data Model, 2015. http://www.ivoa.net/documents/SpectralDM/20150528/index.html
RD16	Louys M., et al., Data Model for Astronomical DataSet Characterisation, 2008. http://www.ivoa.net/documents/REC/DM/CharacterisationDM-20080325.pdf
RD17	Bonnarel F., et al., Characterisation DM: Complements and new features. Observation quality and variability - complex datasets, 2012. http://www.ivoa.net/documents/Characterisation2/20121029/WD-Characterisation2-1.0-20121029.pdf
RD18	Richards A. M.S., Radio interferometry data in the VO, 2010. http://wiki.ivoa.net/internal/IVOA/SiaInterface/Anita-InterferometryVO.pdf
RD19	Dowler P., et al., Table Access Protocol, 2010. http://www.ivoa.net/documents/TAP/20100327/

RD20	Dowler P., et al., IVOA Datalink, 2015. http://www.ivoa.net/documents/DataLink/20150617/index.html
RD21	Regional Centres, SKA-TEL-SDP-0000060
RD22	SDP Standard Data Products, SKA-TEL-SDP-0000075
RD23	Rots A. H., Space-Time Coordinate Metadata. http://www.ivoa.net/documents/latest/STC.html
RD24	Salgado, J., et al., Photometry Data Model, 2013. http://www.ivoa.net/documents/PHOTDM/
RD25	Lemson, G., et al., VO-DML: A Consistent Modeling Language for IVOA Data Models, 2015. http://www.ivoa.net/documents/VODML/index.html
RD26	Bonnarel, F., et al., Server-side Operations for Data Access, 2016. http://www.ivoa.net/documents/SODA/index.html
RD27	Osuna, P., et al., Catalogue Data Model. http://wiki.ivoa.net/internal/IVOA/IVAODMCatalogsWP/IVOACatalogueDataModel.pdf

6 Document Scope

This document covers the Preservation System. The Preservation System for final and intermediate data products and associated metadata including the Science Product Catalogue. The Preservation System consists of the Preservation Platform and the Preservation Software. The scope, size, location, lifetime and resilience of the Preservation System will largely depend on policies and operational requirements [AD01] not available at the time of writing. Therefore, the document should be considered a snapshot of work in progress.

7 Functional Description

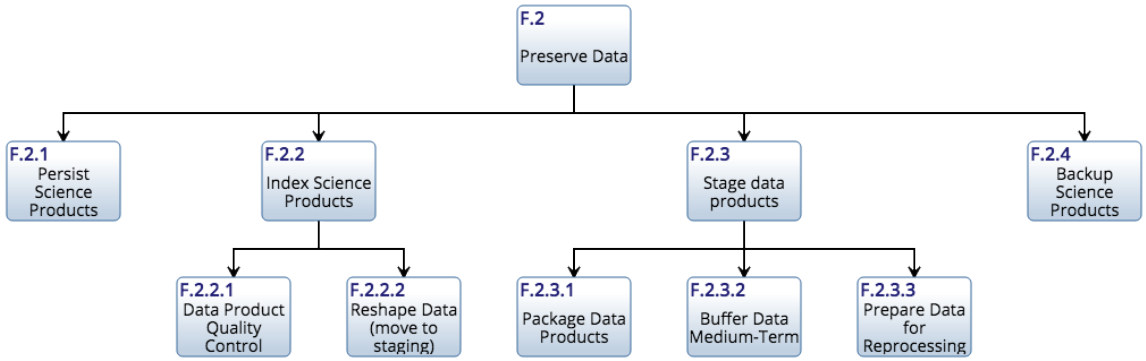

Figure 1: Functions of Preservation System

7.1 Persist Science Products

The Persist Science Products function receives Science Products from the Preservation and Index Science Products function. It maintains Science Products long term, including the generation of multiple copies and/or parity or erasure code protection, checksumming and other means to ensure the resilience of the Science Products.

7.2 Index Science Products

The Index Science Products function has two aspects:

1. It maintains the local catalogue of Science Products; i.e., the ones created at this SDP site: it therefore provides a service function interface.

2. It provides functionality to manage the preservation of data into and out of Persistence and organise and index Science Products. It associates data and metadata including provenance information in science products. It creates entries in the Science Product Catalogue for new Science Products.

The data are not moved in the process. Access references to data are managed by the Data Lifecycle DB and may take the form of URIs.

The Science Product Catalogue includes all of the science metadata that is needed to characterize and discover Science Products. The catalogue information can be queried and replicated, as needed by the Delivery System.

7.3 Stage Data Products

Per the functional breakdown and architecture [AD02] Staging is the quasi omnipotent function that packages data from Local Telescope Manager, Quality Assessment and Local Sky Model in internal format prior to Index(ing) Science Products and Persist(ing) Science Products for association with Science Products; it reorganises science data products into forms suitable for long term preservation; it acts as interface between processing functions and preservation.

7.4 Backup Science Products

This function provides for back up of persisted science data products potentially off site from the main SDP facility. Possible offsite backup locations are the alternate SDP instance, specific off-site facilities and regional centres.

7.5 Allocation of Functions to Products

Product	Function
C.4 Preservation System	F.2 Preserve Data
C.4.1 Preservation Platform	F.2.1 Persist Science Products
C.4.1.1 Long Term Storage	F.2.1 Persist Science Products

C.4.1.2 Medium Term Storage	F.2.3 Stage Data Products
C.4.2 Preservation Software	F.2 Preserve Data
C.4.2.1 [=C3.2.6.1.1] Data Lifecycle DB	F.2.2.2 Populate Product Status
C.4.2.2 Science Persistence Software	F.2.1 Persist Science Products
C.4.2.2.1 Data and Metadata Services	F.2.2 Index Science Products F.2.3.1 Package Data Products F.2.3.2 Buffer Data Medium-Term F.2.3.3 Prepare Data for Reprocessing F.2.4 Backup Science Products
C.4.2.2.2 Science Product Catalogue	F.2.2.1 Populate Science Metadata

Table 1: Allocation of Functions to Products

8 Preservation System

The Preservation System consists of the Preservation Platform and the Preservation Software. It optimises data placement through the support of policies for persisting final data products, unfinished intermediate products as well as backup copies.

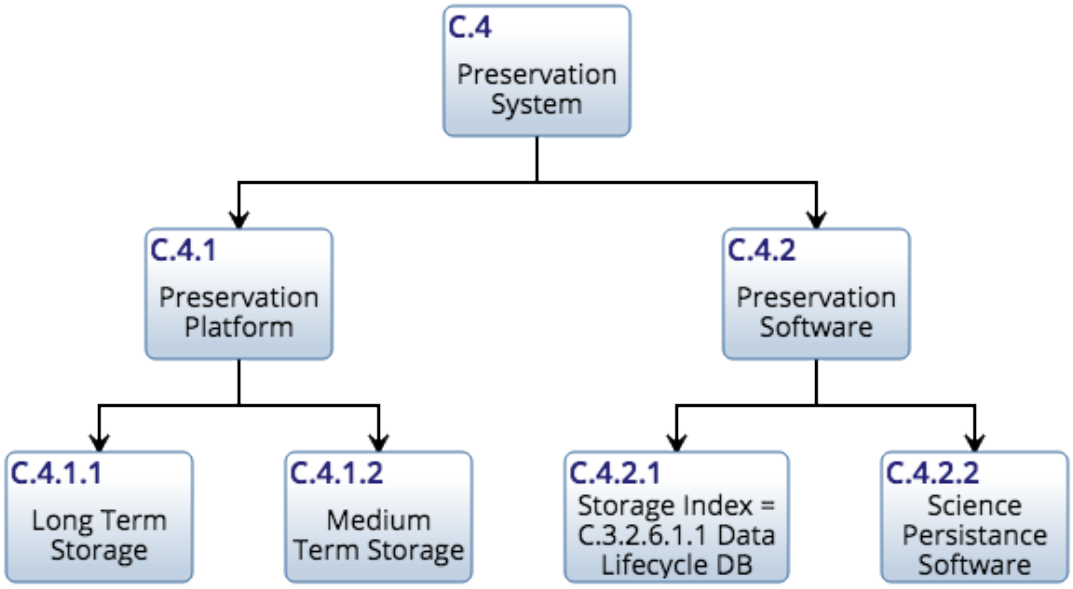

Figure 2: Products of Preservation System

8.1 Preservation Platform

The Preservation Platform has two major purposes:

1. long term storage of science data products
2. mid term storage of intermediate data products

While the two purposes can be served by a single system, one has to keep in mind that the long term storage capacity needs to keep up with the archival growth rate [RD02].

The Long Term Storage provides the long term curation *platform* for SKA data products. Data products can also be buffered on other parts of the SDP infrastructure and intermediate products of multi part observations are a typical use case for the Mid Term Storage. Once procured, the Long Term Storage will be an integral part of the SKA Science Archive and provide low cost and secure storage for the data products. User access is provided by the Delivery Platform [RD04].

The Preservation Software makes use of the Data Lifecycle Manager which maintains the binding between the physical location of stored Drops [AD02] to the Science Product Catalogue. The Science

Product Catalogue is part of the Science Persistence Software that also has Data and Metadata Services for curating and redistributing data and Science Data Products to the delivery mechanism. To make the Science Product Catalogue accessible to external users and tools it is compliant with IVOA data models. Part of the Science Product Catalogue is an Observation Database that provides context for Science Data Products (see also Observation Data Model and Metadata Tools).

Figure 3: Science Preservation Software

8.1.1 Mass Storage System

The following potential options for a mass storage system suitable for Long and Mid Term Storage were considered, whereby one or a mix of these can be employed to serve either purpose:

- **Enterprise Storage:** In this context it means a disk based system by a big vendor coming with its own ecosystem of middleware software and tools for volume management, monitoring, error recovery, etc.
- **Storage Pods:** These are given as a low cost alternative to enterprise solutions. The low cost is partly due to the more limited or missing middleware functionality and the use of commodity hardware to create a minimalistic system optimized for a good volume/price ratio.
- **MAID Pods:** A Massive Array of Idle Disks is a storage technology in which only those disk drives in active use are spinning at any given time.
- **Tape Library:** A jukebox-like mechanism with several drives for read and write operations. It holds many passive, i.e., unpowered media. A robotic device transfers media from their parking slots to the tape drives and swaps them out on demand.
- **Cloud Storage:** Since Cloud providers can internally employ any of the other solutions or a combination thereof this is about the flexibility to define and change the desired service level on short notice.
- **Offline Storage:** A physical pile of unpowered storage media as the ultimate low end backup solution.

Mass Storage System (horizontal)	Enterprise Storage	Storage Pods	MAID Pods	Tape Library	Cloud Storage	Offline Storage
Media Type	disk	disk	disk	tape	selectable	any
Latency	low	low	medium	medium	selectable	high
Bandwidth	high	high	high	medium	selectable	low
CapEx	high	medium	medium	medium	n/a	medium
OpEx incl. Licenses	high	medium	medium	high	high	low
Power Consumption	high	medium	low	low	n/a	low
SW Complexity	low	low	medium	medium	high	low

Table 2: Mass Storage Options

The basic classification in Table 1 is relative, meaning that *high* is significantly higher than that of the candidate solution scoring lowest in the given row. The absolute dynamic range between *low* and *high* arbitrarily differs between criteria (rows).

Enterprise mass storage systems are generally a combination of hardware and a middleware layer plus tools. Some support the definition of storage policies for data lifecycle management. Low latency, high throughput and high error resilience are readily achievable. Naturally that comes at a price, mostly in the form of software licenses and also in terms of power consumption. Advanced features such as lifecycle management tend to tie in closely with a proprietary system.

Storage Pods are built up using commodity hardware that vendors procure in bulk in order to bolt them together in stackable, customizable and direct-attached storage units. There is often little or no specific software coming along. So, pods provide comparable hardware performance without licensing overhead.

MAID systems are a storage technology in which only those disk drives in active use are spinning at any given time. It is an option to tape libraries. That reduces power consumption and can prolong the lives of the drives. Some of the reduction in power comes through reduced cooling. This is partially offset by the need to keep disk controllers up. The drawback is increased latency and possibly reduced disk reliability because of the power cycles. When the management software prefers powered up

devices over powering up new devices it can lead to issues with the filesystem.

Tape libraries reduce power consumption even further. A fair comparison needs take into account that it requires an additional staging buffer based on one of the above technologies and that this mix adds to the software complexity. Current storage robots typically come with a hefty licensing overhead for their proprietary management system.

The Cloud solution is somewhat orthogonal to the above. Since the underlying technology is potentially a mix of all of the above it is not meaningful to compare hardware performance. Instead it is a matter of using a commercial service and to allocate resources in a possibly fully dynamic way. Infrastructure and system support are outsourced. The exploitation of such flexibility comes with increased software complexity and requires significant network bandwidth to the provider.

The most minimalistic solution is storing data offline. This is only meaningful when data are not or very infrequently used. Provided storing such 'cold' data is considered worthwhile at all then the obvious drawback of maintaining such a heap of unpowered media is to that it requires human intervention to find and access any particular datum. Media defects only become apparent upon access. On the bright side, power consumption is minimal and offline data are naturally very secure from unauthorized access.

8.1.2 Hierarchical Storage System

Hierarchical Storage Management (HSM) is a data storage technique which automatically moves data between storage media. HSM systems address the fact that high-speed storage hardware (i.e. excluding software layers) such as disk drives are more expensive per byte stored than slower devices, such as magnetic tape drives. While it would be ideal to have all data available on high-speed devices all the time an HSM system stores the bulk of data on slower devices. It effectively turns the fast disk drives into a cache for the slower mass storage devices. The HSM system monitors the way data is used and decides which data it moves to slower devices and which data should stay on the fast devices. Some systems are implemented on a lower level, providing a file-system view across multiple storage layers. Other systems support policies and provide APIs and plug-ins for implementing higher level functionality.

As a minimum, HSMs cover the functions F.2.1 Persistence and F2.4 Backup, whereby remote replication to the Delivery Platform can be considered a form of backup. There is a large diversity in feature richness, system requirements, maturity, licensing schemes and available system support. Some HSMs support policies and provide APIs and plug-in mechanisms. Programming these interfaces potentially allow support of Data Lifecycle Management, erasure coding for enhanced resilience, format conversion, compression, subsetting, access control, checksumming, regular integrity checks

as part of long term persistence as well as migration support when swapping out whole storage volumes when next generation devices become available.

Criteria

License - Is the system open source? If not, what is the type of license that can be purchased and are there any benefits that come with the license, such as support?

HW - Is specialist hardware required, or can the system run on commodity hardware? Are there any lists of supported hardware?

OS - Will the software run on any operating system? If Linux is supported, then does the system support either the RedHat or Debian families?

FS - Does the system have any file system requirements? Does it have any limitations with a file system?

Throughput - Network: The storage manager will have to ingest and send data in at varying rates. What rates can the systems achieve and are there any limitations?

Concurrency - Can the system run on concurrent streams to aid the parallelism of the data? Are there any limits to this?

Latency - Are any latencies reported for storing or retrieving data from the storage system?

Replication - Does the system allow replication? Are there any limitations to this?

Customisation - Can the system be customised and does it have any interfaces to retrieve data from either to move data or report any issues to LMC?

Notes - Are there any other notes from the research?

Systems Considered

Tivoli is a storage manager solution from IBM. It is modular with a range of licenses.

DMF is SGI's storage virtualization manager that uses its own file system to manage the placement and storage of files.

OHSM is an open source project that appears to have stalled. It worked in the Linux user space to provide data migration.

LessFS is another open source project that provides an open source de-duplicating file system. As version 2 has been mentioned but does not appear to have made much progress.

BTier is an open source block storage device from the makers of LessFS. It is a kernel module to automate data movement between block devices.

LVMTS is an open source data mover that appears to have been designed for Small Office / Home use.

StorNext is commercial offering with various license options and the policy driven options for data movement.

DiskXtender is EMC's offering in the area but appears to have been made into a Windows only product.

Compellent is Dell's software and hardware storage solution.

HPSS is a collaboration between various large US laboratories and IBM to provide a larges scale storage solution.

ArchiveManager from QStar is another commercial offering with plugins to provide functionality.

SAM-QFS is the open source library used in Oracle's offerings. It appears to be more targeted towards Solaris but might work on other Linuxes.

Versity is a new company who have ported the SAM-QFS library to Linux. The library underlies the Cray Tiered Adaptive Service.

Moonwalk is a commercial company who do not offer much information about their service.

Lustre HSM is an integration of the Lustre file system with a policy engine that has appeared in the 2.5 release.

Results

Name	Producer	License	System requirements	Features	Notes
Tivoli	IBM	Commerical. CPU or volume. A single license for the Client Manager with Software Support is available. Bundles of 10 Processor Value Units for Storage Management and Storage Area Networks are slightly cheaper than individual licenses. There is an HSM for a Terabyte for Windows license but none listed for Linux. These licenses include software and support for one year.	Might not write to all file systems, such as Andrew FS.	Replication supported. Policies can be moved on disk space. Client API can be customised and repackaged. Supports HDD, tape; cloud; virtual environments;de-duplication; multi-site replication; analytics included;bare metal recovery; increased performance with GPFS.	

DMF	SGI	Commercial. Volume.	Uses CXFS file system.	Simple policy based on file ageing, file size, last access time. Limited customisation (no priority queue, no file content-based migration, no file event management, etc.)	
OHSM	FSCops	GPLv2	Uses Linux kernel 2.6.32	An XML file controls the policy movement based on file access time, file type, I/O activity, file access activity, or file size. A GUI may exist but it appears doubtful that it fully works.	Code has not been updated since 2010.
lessfs	Mark Ruijter	GPLv3	Linux only.	De-duplication FS; includes support for lzo, QuickLZ, bzip compression; data encryption and replication; LessFS2 appears to be in development; 130MB/s for new data and 170MB/s previously stored data on SATA	Last code is 4 years old.
btier	Mark Ruijter	GPLv2	Linux only	User scripts control data migration and	

				the retrieve API. Up to 16 devices per btier device; provides user-space API; built-in kernel space data migration; transfers up to 1100MB/s; works with any file system	
LVMTS	Hubert Kario	-	Linux only.	Uses LVM on Linux; only one volume supported; uses a config file for movement rules. Last code is two years old. Runs in user space and is designed as a SOHO solution not enterprise.	Code has not been updated recently.
StorNext	Quantum	Commercial. Varying types of license: Permanent, Temporary, Transfer, and Multi-Mount License Requests. Multi-Mount license applies to mount a single metadata controller from a SAN.	Can be installed on Quantum hardware or a list of supported hardware is available. Supports major Linux families.	Capacity 5 billion files; policy-based, automated migration for tiered storage and archive; access from SAN and LAN, supports SSD,disk,object storage, LTO/LTFS tape;optimised for StorNext products	
DiskXtender	EMC	Commercial. DiskXtender licensing is capacity-based.			Windows only. Linux product appears to be discontinue

					d.
Compellent	Dell	Commercial. Perpetual software license.	Dell hardware.	Supports SSDs, HDDS; policy based automation; virtualised storage platform; up to 2PB capacity per 48U rack.	
HPSS	IBM	Commercial. first year services include planning, second include maintenance and support. Other licenses may be purchased.	Supports HACMP hardware. Supports Aix (Power), RHEL (Power, Intel & AMD).	Supports HDD, SSD and most tape archives; single file disk data rate of > 2Gbps; interface mounted as VFS; claims up to 100 petabyte storage.	
Archive Manager	QStar	Commercial. Permanent license tied to the host id of the CPU.	Supported hardware list. Supported OS list.	Supports encryption; supports cloud, optical, object storage, tape; optional policy based management, mirroring and replication;online & offline storage management; uses CIFS/NFS as interface	Server software appears to only support Windows.
SAM-QFS	Oracle	Oracle license excepting some open source components under Apache 2,GPL2, OpenSSL license, SSLeay license libSAM - LGPL	Performance is best on Solaris. Runs on other Linuxes with best effort support in forum.	Policy based data movement; supports tape access, Distributed I/O (uses Solaris); LTFS import (from Solaris); staging priorities; media validation	SAM-QFS is the library on which StorageTek runs.

				(with Solaris).	
Versity	Versity	Commercial. Monthly, quarterly or annual. Volume license.	This is a Linux port of SAM-QFS.	Policy based data movement; writes to tar file; builds on POSIX; files and metadata stored in open format; builds on SAM-QFS; open APIS; file restoration in hardware failure.	Does not appear to have come out of private beta yet and Cray are one of the backers. Currently in private beta. Versity is part of Cray's Tiered Adaptive Storage.
Moonwalk	Moonwalk	U		Policy based data movement; supports RHEL, Windows, NFS, CIFS, cloud; retrieval at native speeds of storage and network	
Simpana	CommVault	Commercial. Capacity, managed service provision or perpetual license on server, application or platform	Supported list of hardware and software on CommVault site.	De-duplication; search; VM integration & protection; workflow automation; User defined policies for retention; Cloud, disk and tape; integrated encryption; replication across WAN & LAN	
Lustre	Open Scalable	Linux versions under GNU GPL	OEL 5, RHEL 5, SLES 10	Policies can run on time but	Lustre HSM appears in

	File Systems	but ports to other systems might require change of license or become proprietary	and 11, Scientific Linux , and Fedora 11 2.6.30 OSes supported.	ignore files owned by a group or size.	version 2.5. it appears to use an open source library to manage the policies and is used in front of other HSM systems.

Table 3: HSM Characterization

8.2 Preservation Software

8.2.1 Observation Data Model and Metadata Tools

The Common Archive Observation Model (CAOM) is a general purpose data model developed and maintained by the Canadian Astronomy Data Centre (CADC) and defines uniform metadata about astronomical observations. It serves as a design template for the SDP. The given database scheme defines an interface between the Science Product Catalogue and the Delivery System. See [RD01] for more details of the model.

Probably more importantly, the model is the basis for a common set of software tools. This set of tools includes those for harvesting metadata into the Science Product Catalogue as well as for querying and accessing the Catalogue.

8.2.2 Science Data Products and IVOA Compliance

An analysis of potential pipeline products resulted in a preliminary classification with respect to their preservation metadata which are tightly linked to requirements that the delivery system puts on the preservation system:

Data Product Type	IVOA Discovery	IVOA Access	IVOA Data Model

Image cubes (image product)	TAP, SIAv2, DataLink	SODA	ObsCore, NDimCubeDM (future)
UV-grids (image product)	TAP, SIAv2, DataLink	SODA	ObsCore, NDimCubeDM (future)
Calibrated visibilities	TAP, SIAv2?, DataLink	SODA?	ObsCore?, NDimCubeDM? (future)
LSM catalogue	TAP?, DataLink?	TAP	RegTAP?, CharDM? STC?, PhotDM?, VO-DML (future)
Pulsar timing solutions	TAP?, DataLink	?	-
Transient buffer data	-	-	-
Sieved pulsar and transient candidates	TAP?, DataLink	TAP	-
Science product catalogue	TAP?, DataLink	TAP	RegTAP?, STC, PhotDM; future: VO-DML Catalog DM, Source DM

Table 4: IVOA support per Science Pipeline Product Category

Some data products have IVOA standard support, some will potentially be covered by planned standards and pending data product definitions and policies [Table 4]. The traceability table at the bottom of [RD04] lists IVOA standards required by the delivery system.Auxiliary file types can be bundled using the Datalink mechanism. For the remaining product types the user interface will either lack certain capabilities (e.g. for searching) or requires a custom solution.

Image cubes (image product)

The IVOA N-dimensional Cube DM [RD12] is a working draft and it may still evolve. However, it is expected that it will be a recommendation for the beginning of the SDP construction. At the time of writing, it counts with several prototyping implementations and it is built considering the Dataset DM working draft [RD13] and, therefore, the ObsCore [RD14], Spectral [RD15] and Characterization [RD16] models.

UV-grids and calibrated visibilities

According to the IVOA WD-Characterization2 document [RD17], radio interferometry visibility data can be described with the Characterization data model. Basically visibility measurements are given at a given time, for a source at a given position, for several spectral channels and polarization feeds, and at several ”spatial frequencies” or equivalently ”baselines”. This document does not include a detailed description of the metadata required for this use case. However, some work has been done within the IVOA to study the characterization of radio interferometry data [RD18]

Pulsar timing solutions

There are currently some efforts within the IVOA to characterize time dependant data. The Spectral DM standard [RD15] presents a core data model describing the structure of spectrophotometric datasets with coordinate axes and associated metadata. This data model may be used to represent spectra, time series data, generic segments of a Spectral Energy Distribution (SED) and other spectral or temporal associations.

The PSRFITS format[1] defines a set of pulsar-related parameters to be included in the header. Part of that information could be included in a data model to provide a better characterization of these objects.

Catalogues

There are no IVOA recommendations for the different catalogues that will provide the SDP i.e. LSM catalogue, Science Product Catalogue, and sieved pulsar and transient candidates. However, there are two IVOA recommendations to support discovery and provide advanced query capabilities. TAP [RD19] defines a service protocol for accessing general table data, including astronomical catalogues as well as general database tables. Datalink [RD20] describes the linking of data discovery metadata to access to the data itself, further detailed metadata, related resources, and services which perform operations on the data.

8.2.3 Indexing and Metadata Services

Three potential database use cases were identified: Data Lifecycle Management, Science Product Catalogue, Observation Catalogues. Below is a generic analysis of these scenarios in terms of database requirements.

Traditional database technology that dominated the market for the past 40 years is facing severe limitations concerning scalability and flexibility of the data model. There is a raft of new database families and products on the market. Their basic properties are given in [Table 5].

[1] http://www.atnf.csiro.au/research/pulsar/index.html?n=Main.Psrfits

DB Technology	Performance	Scalability	Flexibility	Connectedness	Functionality
Key–value Stores	high	high	High	low	variable (none)
Column-families Store	high	high	moderate	low	minimal
Document Database	high	variable (high)	high	low	variable (low)
Graph Database	variable	variable	high	high	graph theory
Relational Database	variable	variable	low	moderate	relational algebra

Table 5: Rating of Database Families against Key Design Attributes

1. Data Lifecycle DB

- moderate to large number of entities
- limited number of attributes per entity; schema is mostly static
- high insert rate in append-only fashion (data is versioned)
- heterogeneous but mostly static queries
- high read and write availability

Function	**Description**
DB Technology	relational DBMS and/or Document DB; possibly key-value store
Reliability	read and write access within specified boundaries; multiple copies; transparent failover, distributed, replicated architecture to avoid single point of failure; backup procedure
Consistency	allow the application to choose the required consistency model
Provisioning	replication of entire DB to secondary locations and in parallel to bulk data mirroring
Data access	data segmentation, time-based catalog access

Table 6: Data Lifecycle Management Scenario - DB Design Criteria

2. Science Product Catalogue

[RD13-RD20] describe metadata schemes and standards and potential implementations that can be adopted to assure the adequacy of the Science Data Product.

- high read availability; consistency can be traded for availability
- secondary and possibly lagged copies can improve availability
- provide consistent view of science archive

Function	Description
DB Technology	relational DBMS and/or Document DB
Reliability	1 primary writeable DB; secondary distributed read-only copies also to improve accessibility as needed, read access within specified boundaries, distributed, replicated architecture to avoid single point of failure; backup procedure
Consistency	allow the application to choose the required consistency model
Provisioning	optional replication of entire DB to secondary locations and in parallel to bulk data mirroring
Data access	data segmentation, heterogeneous (not fixed) queries

Table 7: Science Product Catalogue Scenario - DB Design Criteria

3. Observational Catalogues

This may include source catalogues for continuum, possibly combined with a catalogue of intrinsic polarisation angles and Hi (or other spectral line) candidate source catalogues [RD22]. Managing these may overlap with the scope of the Delivery Platform [RD04], however, this is not relevant for the given analysis.

- huge number of entities
- many attributes per entity
- high insert rate in append-only fashion; multiple observations of sky patch can produce time-series
- key and range queries; possibly joins
- read availability through multiple independent copies

Function	Description
DB Technology	relational DBMS and/or Document DB; possibly key-value store
Reliability	1 primary writeable DB; secondary distributed read-only copies, read access within specified boundaries, distributed, replicated architecture to avoid single point of failure; volume may make backup from scratch impractical
Provisioning	optionally provision portions of data catalogues on demand
Data access	data segmentation and distribution to multiple, independent DB instances, efficient spatial and time-based catalogue access

Table 8: Observational Catalogue Scenario - DB Design Criteria

9 Interfaces

Delivery System

The interaction with the Delivery System is described in [RD04].

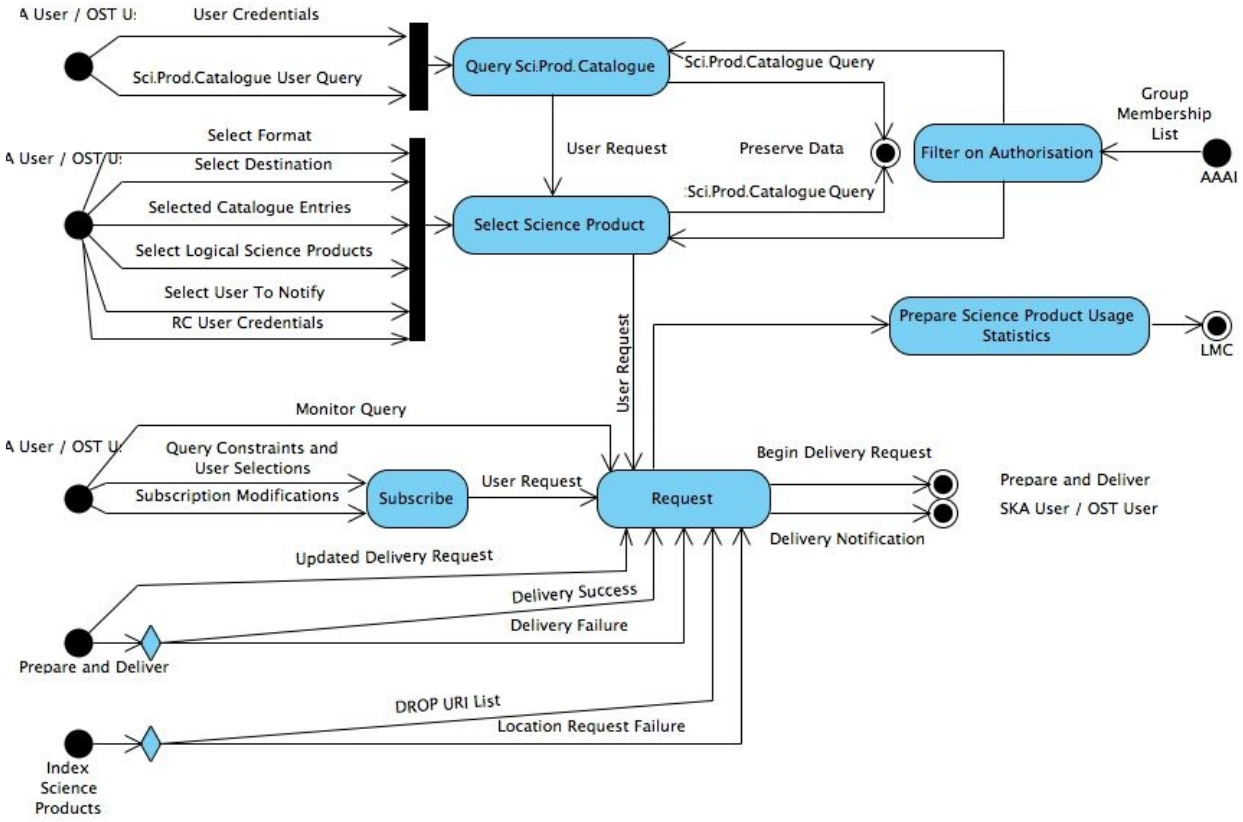

Figure 4: Delivery Query and Request Workflow.

VoEvent

One Metadata Service will be a VoEvent client. It subscribes to an LMC VoEvent broker. The client will persist VoEvent messages as part of the Science Product Catalogue database. In this way the VoEvent history is accessible to the Delivery System.

Metadata Synchronization between SDP Instances

Various L1 requirements set the scene for archiving and distributing science data products (SKA1-SYS_REQ-2366, SKA1-SYS_REQ-2355, SKA1-SYS_REQ-2353) and VO compliance. The requirements also state the need to protect telescope operations from effects of archive usage. There is less guidance on how to provide a unified view on metadata of both SDP instances. Various options for employing database replication methods between SKA-AU and SKA-SA permitting a consistent representation of data products from both SDPs are discussed below. This is an interface between the Science Product Catalogue and Science Catalogue Service API of the Delivery System.

Assumptions:

- information transfer is asynchronous
- metadata are continuously synchronized in near real-time
- R/W and R/O access happens at the closer site with the remote one as potential fallback
- free and open source tools are sufficiently up to the task

Given the physical distance between AU and SA a high network latency in the order of hundreds of milliseconds prevents synchronous replication. For completeness, one should mention that some DBMS can be extended with so-called multi-master semi-synchronous replication modules. They replicate synchronously and apply deltas optimistically, while the actual data is placed in a queue. The following certification phase detects if the propagated information can be effectively applied on the receiving end. This mode is highly latency sensitive and the behavior, when lacking a properly modeled schema, may appear erratic in case of conflict.

To avoid these problems one should consider the deployment of two fully independent database systems, one for each SDP. Not only does it preserve the symmetry and independence of each SDP. Such a dual master-slave approach (AU → SA + SA → AU) avoids above problems. Drawbacks are relatively minor:

- somewhat more resources required for two full systems compared to a single master/slave setup
- schema duplication takes in the risk of introducing unexpected inconsistencies (divergence)
- the unified metadata view is managed by a replication logic sitting between the DB and the application layer

Most DBMS efficiently and natively produce/store/manipulate data in XML or JSON format. Much of post processing and conversion tasks can be offloaded to the database engine.

Firstly, classic DBMS SQL solutions are considered followed by NoSQL solutions. In short, they are both capable whereby NoSQL is not known to have been used at large scale and for this purpose in astronomy data centres yet. Here are the details of the analysis:

- MySQL/MariaDB:
  - Native master-slave replication is statement or row based and inherently asynchronous, can be used to implement master-master using two, opposite, propagation streams
  - Conflicts are not natively addressed
  - The topology can be quite easily extended when using Global Transaction IDs
  - The Galera Cluster extension, natively part of MariaDB, is a semi-synchronous solution is use in the industry
- PostgreSQL:
  - native streaming replication is of physical type; for dual master-slave architecture but not for multi-master topologies
  - some extensions do provide multi-master capabilities (Bucardo, Slony-I,...), they are all based on triggers and demanding to implement and maintain

    - BDR (Bi-Directional Replication) is a native multi-master solution which is going to be available in version 9.5; its maturity is questionable
- Replicators:
    - are third-party components sitting on top of database engine
    - extract information from write-ahead load stream; assign Global Transaction ID to deal with complex topologies
    - support filtering and conversions
    - useful for heterogeneous data deployments

In the free and open source domain MySQL and PostgreSQL have certain multi master capabilities.The majority of the commercial products include a component to deal with multi-master asynchronous replication. They allow filtering at any stage and to some extent automated conflict management. The ALMA project uses Oracle Streams/GoldenGate.

NoSQL DBMS: The majority of NoSQL databases natively implement an asynchronous master-slave replication protocol. This choice has been made to reduce the complexity of the engine and avoiding to have to deal with conflicts.

Many products relaxed consistency, meaning that the database, as seen by the application, can be in a state of transition.

Pros of NoSQL products:

- the data model is semi-structured and often JSON or JSON-like
- it scales by adding shards and auto-rebalancing data
- R/W connections can be directed to an installation site by tagging the shards
- Tagging also allows to locate newer and hotter data on faster storage
- R/O sessions transparently connect to the nearest cluster member with others as potential fallbacks

Drawbacks:

- heterogeneous query languages
- in some cases consistency can be compromised

10 Risk Analysis

The technological risks of hardware and software for large scale storage are considered low. It really boils down to affordability and in that context power consumption as well as licensing:

CapEx (moderate):

The biggest risk is affordability vis a vis the archival growth rate [RD02] and deviation in terms of excessive use of high data production observing modes. Mitigation options are:

- gradual expansion of capacity

- migration to higher density media over time
- agreements with Regional Data Centres to host data [RD21]
- policy adjustment: e.g. allow for excision of unused science data products
- fit high data rate observing mode usage to available storage budget

Licensing (moderate):

Licensing models are dynamic and can move, for instance, from a per core to a capacity basis.

Energy consumption (moderate):

To state the obvious, energy consumption roughly scales with storage volume.

Mitigation: The Cloud solution trades on-site energy consumption for the need of sufficient network connectivity to the service provider. The reduced infrastructure is traded for higher OpEx.

Changing Access Patterns (low):

A given storage system configuration is optimal for certain access characteristics. Typical access patterns (junk size, cadence, randomness) may change over the observatory lifetime.

Technology (low):

- Hardware: Because of the need to leverage on the latest and therefore most power efficient systems this means that development takes place on previous generations of hardware. This specifically applies to storage solutions that do not come with their own middleware.
- Software: Middleware software layers can be complex and may require special skills to maintain and troubleshoot.
- Database Maturity: NonSQL database families outperform classic RDBMS systems in certain areas, however, they are relatively new and products are appearing and disappearing on the market frequently, making liftetime the dominating risk factor [Figure 5].

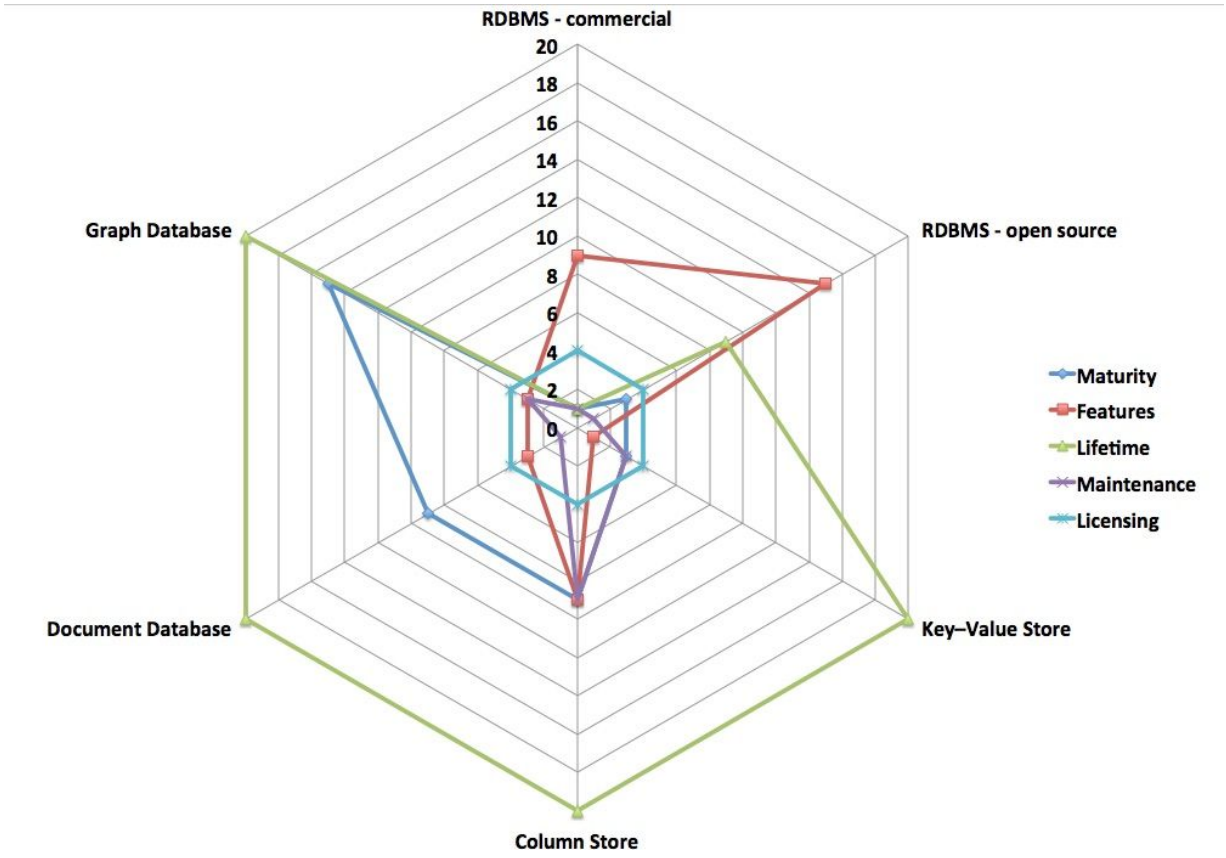

Figure 5: Risk Radar for Database Families

11 Function To Requirements Traceability

Function	L2 Requirement	Name	Description
F.2.1 Persist Science Products	SDP_REQ-289	Maximum science product preservation lifetime	The SDP shall preserve science data products for not less than 50 years from the start of science operations.
	SDP_REQ-705	Archive growth rate	The SDP shall support an archive growth rate per year covering at least the science data

			products of the High Priority Science Objectives plus 500 channel continuum observations and Transient Buffer Data (TBC).
	SDP_REQ-706	Delivery latency	The SDP shall start delivering any science data product
	SDP_REQ-707	Data product retrieval rate	The SDP shall support the retrieval of no less than 1 PB (TBC) per day of preserved data.
	SDP_REQ-708	Partial data product retrieval	The SDP shall support retrieval of partial data products. The minimum size of a delivered part may be bigger than the actual request to coincide with the storage partitioning of the preserved products.
F.2.2 Index Science Products	SDP_REQ-276	Data Product Provenance	The SDP shall create and maintain provenance links between science data products and observing projects and proposals.
	SDP_REQ-709	Science Catalogue	The SDP shall generate and maintain a catalogue of science oriented meta data for all preserved science data products.
	SDP_REQ-710	Science Catalogue	The SDP science

		access	catalogue shall contain sufficient meta data to support standard IVOA queries as a minimum.
	SDP_REQ-711	Science Product Indexing performance	The SDP shall complete indexing of a science data product within 10 minutes (TBC) of the final processing step of the data product being completed to allow queries of the science product catalogue.
	SDP_REQ-712	Science product catalogues	The SDP shall generate and maintain a catalogue of product-oriented meta data for all preserved science data products. This includes
F.2.3 Stage data products	SDP_REQ-617	Restore images from archive to buffer	The SDP shall be able to carry out processing and updating of preserved science data products.
	SDP_REQ-713	Staging performance	The SDP shall allow access to processed data at 200 GB/s (random access) (TBC) for 2PB (TBC) (twice the daily data product retrieval rate) of data for the purpose staging processed data products before preservation and staging preserved data products for further processing.

F.2.4 Backup Science Products	SDP_REQ-260	Restoration of data products	The SDP shall be able to restore a corrupted or lost data product within 24 hours.
	SDP_REQ-281	Protection against data loss	The SDP shall protect the preserved science data products against data loss and malicious or incidental modification. The level of protection shall be maximised within the budget limits of the preservation function.
	SDP_REQ-283	Disaster recovery	The SDP shall provide the capability to restore access to the entire set of preserved data products within one week following a disaster.
	SDP_REQ-714	Initial Preservation Level	The SDP shall protect new science data products locally and instantly against data loss by mechanisms such as RAID
	SDP_REQ-715	Full Preservation Latency	It shall take less than 24h (TBC) for a new Science Data Product to reach the final preservation level in the SDP. This includes the creation of e.g. additional copies

Table 9: Mapping of Functions and Requirements

12 Requirement Traceability

ID	Name	Trace incl. Changes since PDR
SDP_REQ-247	Data Layer Management	data layer concept was dropped
SDP_REQ-248	Data Life Cycle Management	description now in [AD03]
SDP_REQ-262	Data Layer Product Distribution	dropped
SDP_REQ-271	Data Products	dropped
SDP_REQ-275	Remote Processing	dropped
SDP_REQ-276	Data Product Provenance	Index Science Products
SDP_REQ-281	Protection against data loss	Backup Science Products
SDP_REQ-282	Backup Archive Retrieval	dropped
SDP_REQ-283	Restore archive from backup	Backup Science Products
SDP_REQ-287	Continuous performance monitoring.	dropped
SDP_REQ-289	Maximum science product preservation lifetime	Preservation Platform
SDP_REQ-290	Secure Archive Environment	dropped
SDP_REQ-573	[no] Third party data products	parent SKA1-SYS_REQ-2358 dropped
SDP_REQ-574	Science Data Archive	Science Archive
SDP_REQ-598	Storage Hierarchy	dropped
SDP_REQ-599	Data Service Layer	dropped

SDP_REQ-610	Consistent Release State	dropped
SDP_REQ-618	Transfer images from archive to buffer	dropped
SDP_REQ-705	Archive growth rate	Preservation Platform
SDP_REQ-706	Delivery latency	Mass Storage System
SDP_REQ-707	Data product retrieval rate	Preservation Platform
SDP_REQ-708	Partial data product retrieval	Stage Data Products
SDP_REQ-709	Science Catalogue	Indexing and Metadata Services
SDP_REQ-710	Science Catalogue access	Science Data Products and IVOA Compliance
SDP_REQ-711	Science Product Indexing performance	Indexing and Metadata Services
SDP_REQ-712	Science product catalogues	Indexing and Metadata Services
SDP_REQ-713	Staging performance	Preservation Platform
SKA1-SYS_REQ-2128	Continuum and spectral line imaging mode	dropped SKA1_Survey telescope
SKA1-SYS_REQ-2174	Combined SKA1_Mid configuration	was moved to SKA1_Mid Configuration Coordinates; TBC; drives data product size [RD02]
SKA1-SYS_REQ-2178	Combined SKA1_Mid configuration	was moved to SKA1_Mid Configuration Coordinates; TBC; drives data product size [RD02]
SKA1-SYS_REQ-2340	Continuum imaging data products	Science Data Products and IVOA Compliance
SKA1-SYS_REQ-2342	Spectral line emission data products	Science Data Products and IVOA Compliance
SKA1-SYS_REQ-2344	Spectral line absorption data products	Science Data Products and IVOA Compliance

SKA1-SYS_REQ-2346	Slow transient data products	Science Data Products and IVOA Compliance
SKA1-SYS_REQ-2350	Mirror sites	SDP_REQ-281, [AD05]
SKA1-SYS_REQ-2353	Virtual Observatory Interface	Science Data Products and IVOA Compliance
SKA1-SYS_REQ-2354	Archive API	Science Data Products and IVOA Compliance
SKA1-SYS_REQ-2355	Data product provenance	SDP_REQ-276, [AD05]
SKA1-SYS_REQ-2358	Third party products	dropped by SKAO
SKA1-SYS_REQ-2360	Science data product archive policy	now part of AAAI within Delivery System
SKA1-SYS_REQ-2361	Archive access	was moved to SKA Operational Requirements Document; TBC
SKA1-SYS_REQ-2363	Archive lifetime	dropped by SKAO
SKA1-SYS_REQ-2364	Data migration plan	was moved to SKA Operational Requirements Document; TBC
SKA1-SYS_REQ-2366	Distribution of data products	SDP_REQ-262, [AD05]
SKA1-SYS_REQ-2479	Archive security	SDP_REQ-290, [AD05]
SKA1-SYS_REQ-2616	SKA1_Mid Pulsar phase binning	SDP_REQ-252, [AD05]
SKA1-SYS_REQ-2660	Backup archive retrieval	was moved to SKA Operational Requirements Document; TBC
SKA1-SYS_REQ-2661	Backup archive user access conversion	was moved to SKA Operational Requirements Document; TBC
SKA1-SYS_REQ-2688	Commensal Observing Data access rights	dropped by SKAO
SKA1-SYS_REQ-2728	Data migration design	was moved to SKA Operational Requirements Document; TBC
SKA1-SYS_REQ-2739	Levels of access to archive	now part of AAAI within Delivery System

Table 10: Requirement Traceability Matrix