DataProducts Data Model Specification

Doc. no.: CTA–SPE–COM–000000–0009-1a_d1

2023-11-06

	First/Last Name, Organization, Role	Digital signature
Prepared by	K. Kosack CTAO, CEA Paris-Saclay DPPS Coordinator, Data Model Group	
Approved by	Matthias Füßling CTAO SUSS Coordinator	
Approved by	Igor Oya CTAO ACADA Coordinator	
Released by	Stefan Schlenstedt CTAO Computing Coordinator	

Revision History				
Issue	Rev.	Created	Reasons / Remarks /Section	Author
1	a	2023-03-01	Adapted from v1b of old Top-level Document	K. Kosack

Author	
First/Last Name, Organization	Contribution Subject/Chapter
Karl Kosack, CTAO / CEA Saclay	

Table of Contents

1 Introduction **5**
1.1 Purpose . . . 5
1.2 Scope . . . 5
1.3 Applicable and Reference Documents . . . 6
1.3.1 Applicable Documents . . . 6
1.3.2 Reference Documents . . . 6

2 DataProducts Data Model **8**
2.1 Data . . . 8
2.2 Metadata . . . 8
2.2.1 Reference Metadata . . . 9
2.2.2 Context Metadata . . . 13
2.2.3 Fixity Metadata . . . 15

3 Metadata Serialization **16**

4 Data Product Identification Scheme **17**
4.1 Data Categories . . . 18
4.2 Data Product Classification . . . 18
4.3 Data Association . . . 18
4.4 Examples . . . 19
4.5 Data Models for Monte-Carlo Simulated Data . . . 19
4.5.1 Simulated Data Levels below R1 . . . 20
4.5.2 Simulated R1-DL5 products . . . 20
4.6 Data Models for Calibration Data . . . 20

A Mapping to IVOA Metadata **21**
A.1 ObsCore . . . 21

B Reference Metadata Example Representations **22**
B.1 Reference metadata in flat FITS header card representation . . . 22
B.2 Reference metadata in flat ECSV header representation . . . 22
B.3 Reference metadata in hierarchical YAML representation . . . 23
B.4 Reference metadata in hierarchical JSON representation . . . 24

Glossary **26**

Acronyms **30**

1 Introduction

1.1 Purpose

This document describes the data model for data products, i.e. serialized data and associated metadata. The metadata are intended to be standardized and sufficient, after being ingested into a file catalog, to support the querying and retrieval of all CTAO data products from the bulk archive or science archive. It is part of the overall top-level data model specified in [AD-1].

1.2 Scope

The *`DataProducts`* model only covers high-level commonalities between all CTAO data products. The model is intended to be followed by *all* data products generated by CTAO, including those resulting from acquisition, data processing, calibration, or from human-made measurements. Different types of data products will have specific models and formats, which can extend this model to include more information, however these models are outside the scope of this document. The format of the serialized data is not specified here, however guidelines for serializing metadata in Flexible Image Transport System (FITS), Hierarchical Data Format v5 (HDF5), and Enhanced Character Separated Value (ECSV) formats are provided.

In this document, the names of existing data models are written as in fixed-width italics with periods separating levels of hierarchy, e.g. *`CTAO.ScienceOperations`*.

1.3 Applicable and Reference Documents

1.3.1 Applicable Documents

[AD-1] *Top-level Data Model Specification*. CTA-SPE-COM-000000-0001. Version 2a (in preparation). CTAO. 2023.
`https://redmine.cta-observatory.org/projects/computing-info/dmsf?folder_id=1946`.

1.3.2 Reference Documents

[RD-1] *CTAO R1/Event Data Model Specification*. CTA-SPE-COM-000000-0002. Version 1f. CTAO. 2021-12-03.
`https://redmine.cta-observatory.org/dmsf/files/8627`.

[RD-2] *DL3 Data Model Specification*. CTA-SPE-COM-000000-0006. Version 1a. CTAO. 2023-06-27.
`https://redmine.cta-observatory.org/dmsf/files/17685`.

[RD-3] *Reference Model for an Open Archival Information System (OAIS)*. 650.0-M-2. CCSDS. 2012-06.

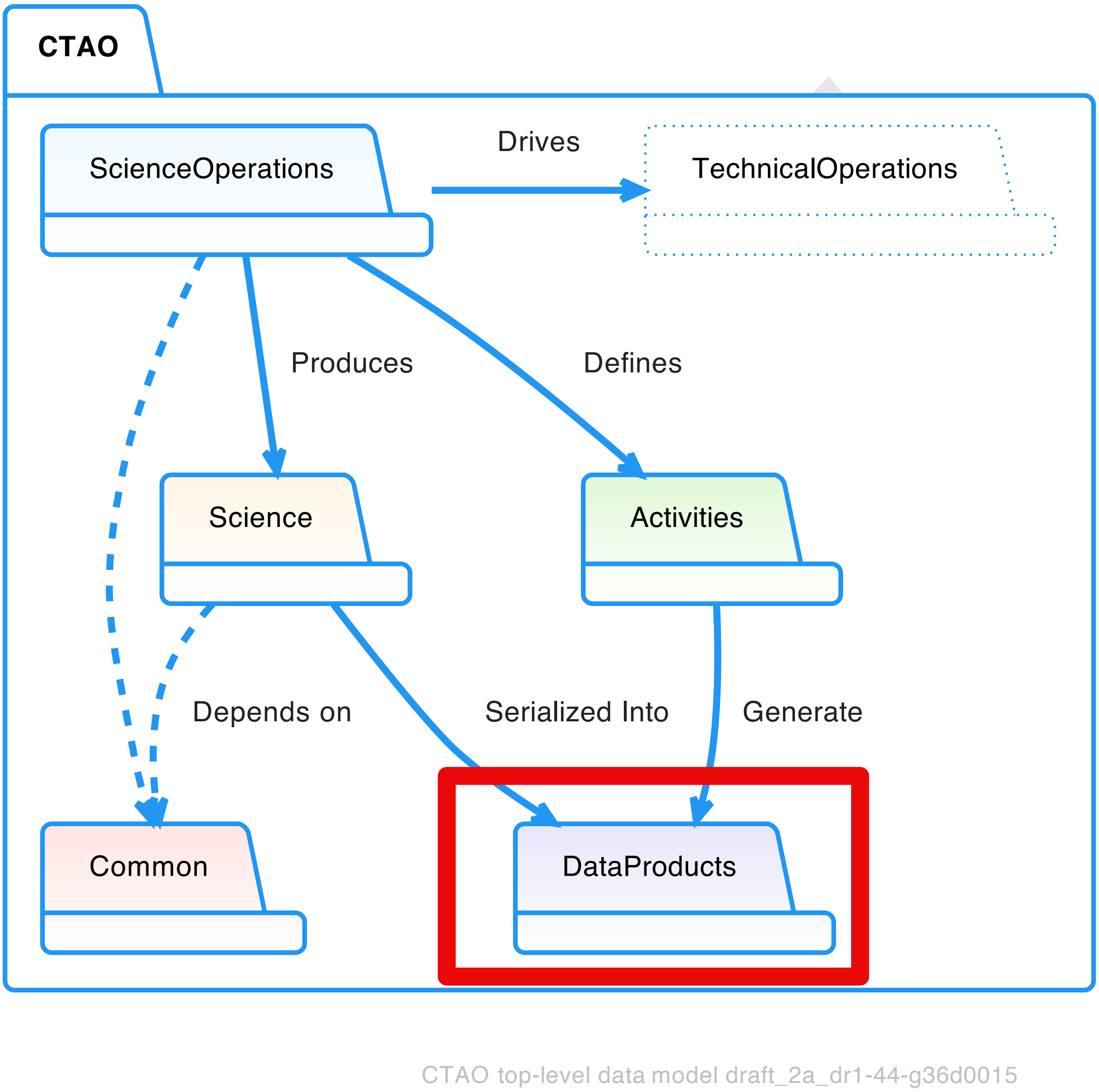

Figure 1.1 – Location of the DataProducts model within the top-level data model[AD-1]

2 DataProducts Data Model

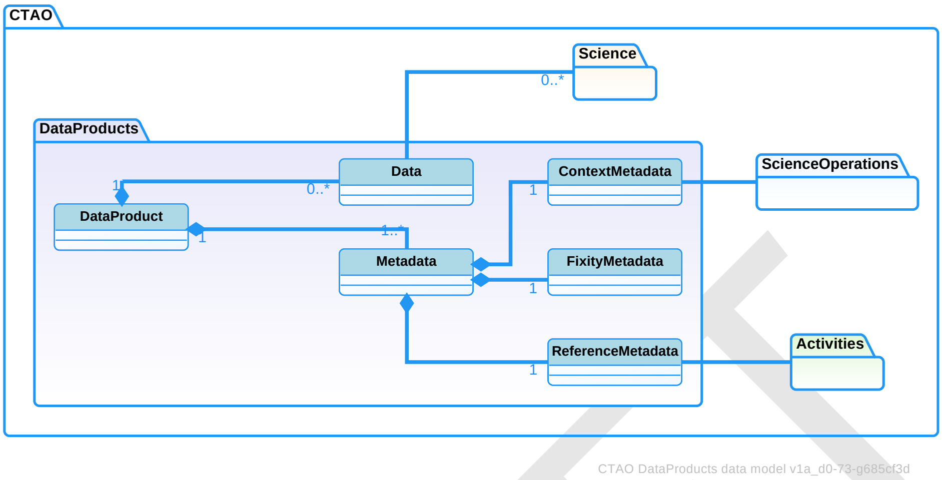

Figure 2.1 – *DataProducts* Data Model overview. This model describes serialized data and associated metadata.

This model, summarized in Figure 2.1, defines the formal concept of a CTAO data product, which is an object that contains *Data* and *Metadata*. *Data* comprises one or more elements of *Science* data that have been transformed into a form that can be written to a file (via serialization). *Metadata* associated both with that data and with *ScienceOperations* processes that created it. For definitions of *Science* and *ScienceOperations* models, see the *Top-level Data Model Specification* [AD-1].

2.1 Data

The primary contents of a data product are serialized *Data*, specifically *Science* data generated by the observatory or its software. The contents and serialization details are out of scope of this document, but are described in the *CTAO.Science* data models, which are broken into multiple data levels [e.g. RD-1, RD-2]. In certain cases, a data product may contain data from more than one *Science* data level.

2.2 Metadata

Metadata is information that describes or add context to the *Data* within a data product. The following types of metadata are defined for CTAO, following from the Open Archive Information System (OAIS) archive model [RD-3].

Reference Metadata a minimal set of headers included in any data product associated with CTAO that include the data *category*, *level*, *type*, *association* as described later in this document, along with associated provenance information. The goal is to provide enough information to understand both the content of the data, it's context within CTA, and its immediate provenance, while remaining in a standard format that can be used as keys in a file catalog.

Context Metadata Contains additional information related to the science operations process or processes associated with a data product

Fixity Metadata Contains necessary information to verify the integrity of a data product, e.g. via a checksum process.

2.2.1 Reference Metadata

This meta-data is organized in a hierarchical structure (following the FITS HIERARCH standard, but also applicable to any other file format that supports headers), where the top-level of the hierarchy is "CTA", meaning this is CTA-specific metadata, followed by 2 or more sub-levels. The meta-data is divided into the following categories:

CONTACT Describes the person or institution that is responsible for this data product (*Agent* in International Virtual Observatory Alliance (IVOA) Provenance terminology)

PRODUCT Describes the details of the data product, including its type and links to the data model definition (*Entity* in IVOA Provenance terminology)

PROCESS The top-level activity to which the activity thatq generated this product belongs, for example an Observation Block (OB), and it's associated ID (e.g. OBS_ID) . Note that a process will have many activities that produce many data products, so this only needs to be an identifier to link to a top-level description of the process, linking multiple activities and therefore data products.

ACTIVITY The specific software or task that generated this particular data product. Note that an activity may produce many data products, so this block may be repeated in many products, and again is essentially a key that links multiple data products (Same terminology used in the IVOA Provenance)

INSTRUMENT The description of the subset of CTA Instrument Description to which this data product is associated, which could be e.g. a subarray, or a small part such as a photosensor. This is a more detailed or lower-level description of the Data Association in the data model identifier. (*Context* in IVOA Provenance terminology)

2.2.1.1 Reference Metadata Use Cases

The following use cases were considered when defining this `Reference` metadata:

- store metadata for a CTA Science Observation from a subarray
- store metadata for a calibration coefficient activity from Cherenkov telescope data made on-site
- store metadata from a hand-measured calibration coefficient (like a quantum-efficiency curve) made in a lab
- store metadata for a product generated for any part of the CTA Observatory (from photosensor up to array site)
- store metadata for input products for a Monte-Carlo simulation configuration
- store metadata for a Monte-Carlo simulation of an array find out who produced a data product and when

- store metadata describing a Science-ready Data (DL3) science data product file containing multiple data sets (event list, GTI, tech time-series)
- store metadata describing a Binned Science Data (DL4) product produced by a science user or a Science User Support System (SUSS) pipeline

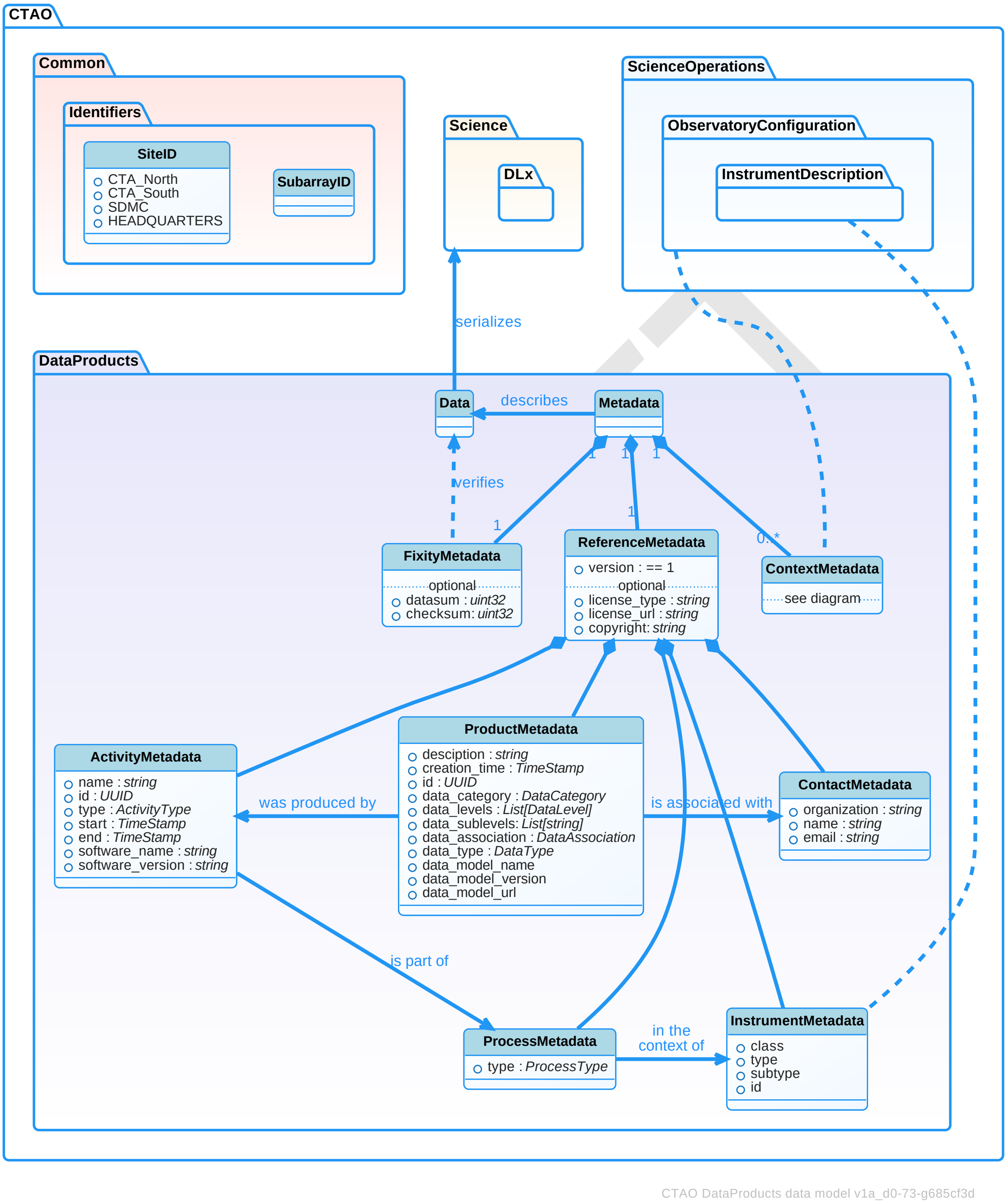

Figure 2.2 – `DataProduct` model in detail. Data Products have Data and Metadata. Metadata are what are ingested into either the bulk archive or science archive to be used to identify data products within a file catalog as well as to describe their immediate provenance. For data types, see Figure 2.3.

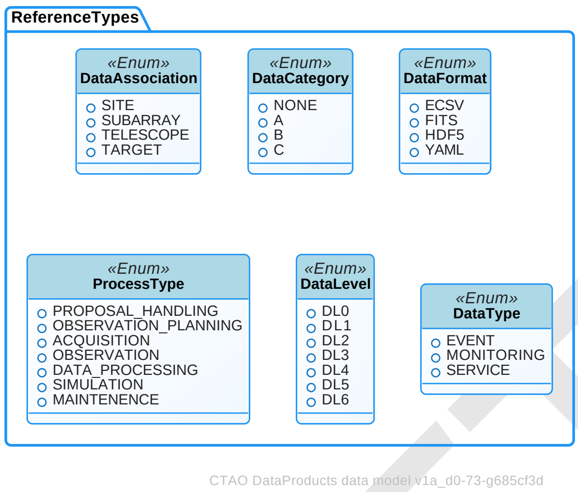

Figure 2.3 – Data Types used in the DataProducts model

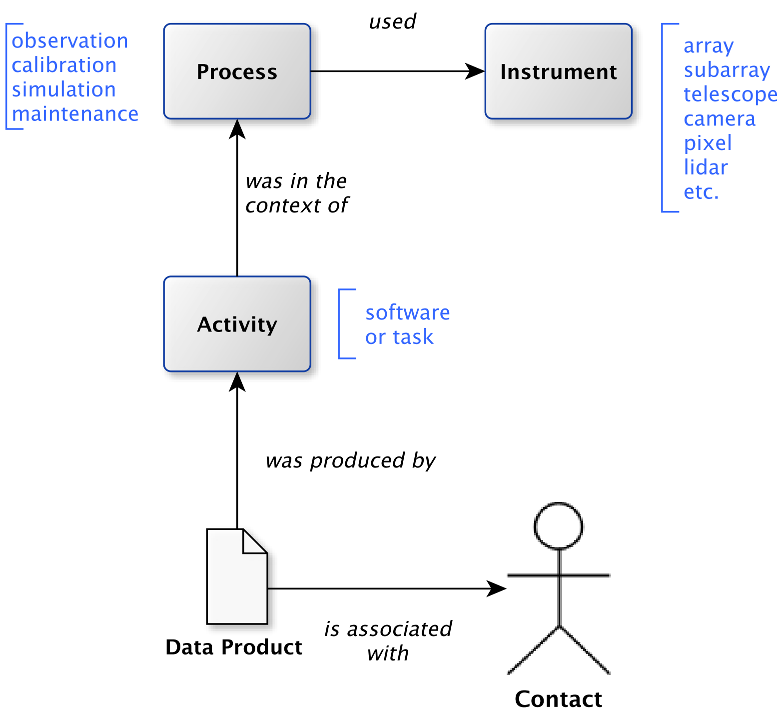

Figure 2.4 – ReferenceMetadata Relationships. This follows essentially the same scheme as the provenance data model.

2.2.2 Context Metadata

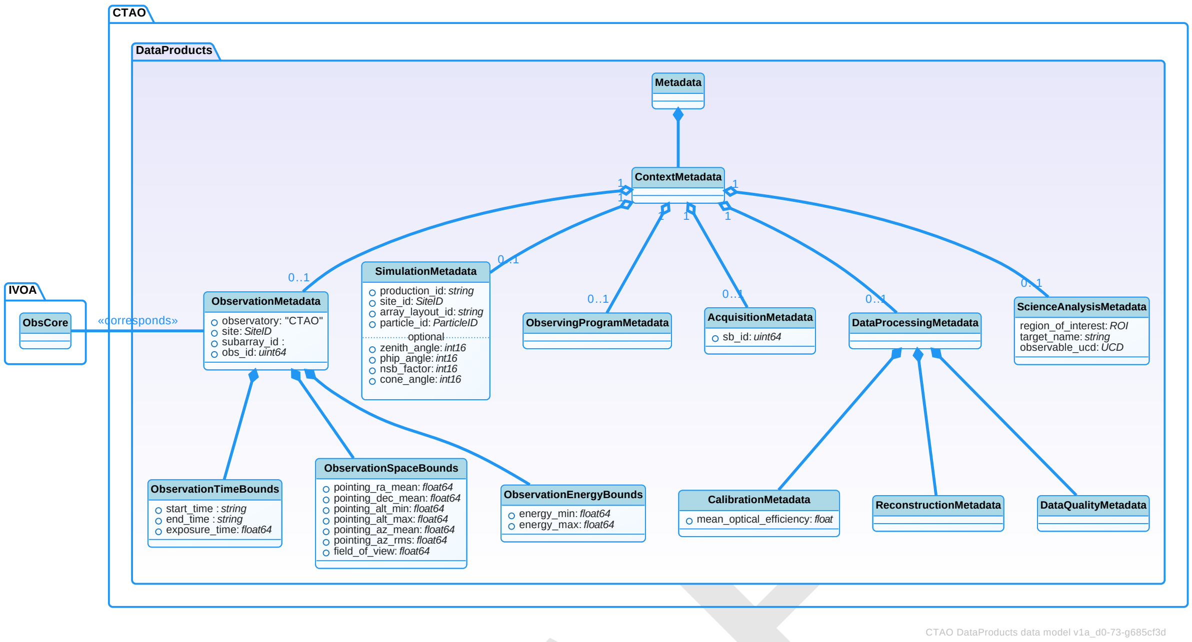

Figure 2.5 – *ContextMetadata*, a subset of the general Metadata model shown in Figure 2.2. Context metadata provides detailed information or links to one or more science operations processes.

Context metadata provides additional information related to the specific context in which the data product was produced, which for CTAO relates to the specific science operations processes that created them. The intention of this metadata is for two use cases:

1. Link the data product to more detailed *ScienceOperations* or *Science* data [AD-1] that are stored externally; i.e. provide sufficient identifiers that this information is retrievable when needed.
2. Define the information that will, along with the *ReferenceMetadata*, form a *file catalog*, allowing all data products to be queried and retrieved.

For the latter, there may be a wide variety of use cases for how data products are queried: the most basic would be a user looking for observations taken in the science archive where the spatial or temporal parameters of the observation or the proposal/program are important, but during acquisition, data processing, or maintenance, queries on more low-level parameters may be necessary. Therefore it is important that the *ContextMetadata* is sufficient to cover all common queries, or at least provide sufficient identifiers to retrieve additional information required by more complex queries.

The use cases for context metadata include:

- Identify an observed data product by location in space or time
- Identify a simulated (Monte-Carlo) data product by simulated observing conditions
- Identify an observed or simulated data product by data processing or acquisition configuration information

- Identify Science Data Challenge (SDC) data product, i.e. a toy simulated generated from existing Instrument Response Functions (IRFs).

Figure 2.5 shows the range of possible context metadata that may be present in a given data product. For example, if the data product is `Event` data corresponding to an observation, it should include `ObservationMetadata`. If it was generated from real observations, then it should additionally include `AcquisitionMetadata`; if generated from a simulation, it would include `SimulationMetadata`. In either case, if `SimulationMetadata`, and in either case if data processing has been performed, it would include `DataProcessingMetadata`.

2.2.2.1 Observation Context Metadata

`ObservationMetadata` (Class Definition 2.1, Class Definition 2.2, Class Definition 2.3, Class Definition 2.4) is intended for the discovery of observations in a catalog and to link to more detailed `ScienceOperations` and `Science` data. The elements are intended to be compatible with the IVOA Observation Data Model Core Components (ObsCore) data model as well as standard FITS conventions when appropriate. The information within describes the type of observation and rough coverage limits in time, space, and energy. This is not necessarily detailed or precise enough to use directly for science analysis use cases, where for example the full Scheduling Block (SB) and OB information, pointing history monitoring tables, live-time computed from stable observation intervals (SOIs), and energy-dependent gamma-ray Field of View (gFoV) provided in IRFs provide better detail.

Class Definition 2.1 – ObservationContextMetadata

Name	FITS	IVOA	Unit	Description
observatory	TELESCOP	facility_name		Always “CTAO”
instrument	INSTRUME	instrument_name	SiteID	The site ID (CTA-North, CTA-South)
subarray_id	*SUBARRAY*		SubarrayID	Which subarray this observation was taken with.
obs_id	OBS_ID	obs_id		Observation identifier

Class Definition 2.2 – ObservationTimeBounds

Name	FITS	IVOA	Unit	Description
reference_time	MJDREF MJDREFI		d	CTAO Reference time in MJD format
start_time	T_START	t_min	s	Starting time of observation
end_time	T_STOP	t_max	s	Ending time of observation
duration	ONTIME		s	total duration of all SOIs, without deadtime correction
live_time	LIVETIME	t_exptime	s	Total exposure time (end_time-start_time) corrected for average instrumental dead-time and any gaps between SOIs within that interval.
num_soi	*N_SOI*			Number of SOIs in the observation

Class Definition 2.3 – ObservationSpaceBounds

Name	FITS	IVOA	Unit	Description
pointing_ra_mean	RA_PNT	s_ra	deg	Mean right ascension of pointing
pointing_dec_mean	DEC_PNT	s_dec	deg	Mean declination of pointing
pointing_alt_min	*ALT_MIN*		deg	Minimum altitude
pointing_alt_max	*ALT_MAX*		deg	Maximum altitude
pointing_az_mean	*AZ_MEAN*		deg	Mean azimuth
pointing_az_rms	*AZ_RMS*		deg	RMS of azimuth within observation
reference_frame	RADESYS			Stellar reference frame used for the celestial coordinate system ("ICRS")
field_of_view	*FOV*	s_fov	deg	Approximate field-of-view radius (assuming a circular FOV) of the observation at 10 TeV energy, used to identify observations covering a point on the sky. Note that a MOC may also be available for higher-accuracy coverage.

Class Definition 2.4 – ObservationEnergyBounds

Name	FITS Keyword	Type[unit]	Description
energy_min	*E_MIN*	float[TeV]	lower-bound of safe energy interval or spectral bound
energy_max	*E_MAX*	float[TeV]	upper-bound of safe energy interval or spectral bound

Class Definition 2.5 – AcquisitionContextMetadata

Name	FITS	IVOA	Unit	Description
sb_id	*SB_ID*			Scheduling Block identifier

2.2.2.2 Acquisition Context Metadata

Sync this with GADF/VODF FITS header definitions

2.2.3 Fixity Metadata

3 Metadata Serialization

describe how the Reference, context, and fixity metadata should be serialized as headers in various file types, e.g. how the keywords are named, and what they look like. Expand the current examples to show also CONTEXT and FIXITY info, the latter for FITS can follow the FITS standard.

4 Data Product Identification Scheme

This really mixes data models and data products. Should be more specific. Categories are for data products only, the data model is the same for all of them.

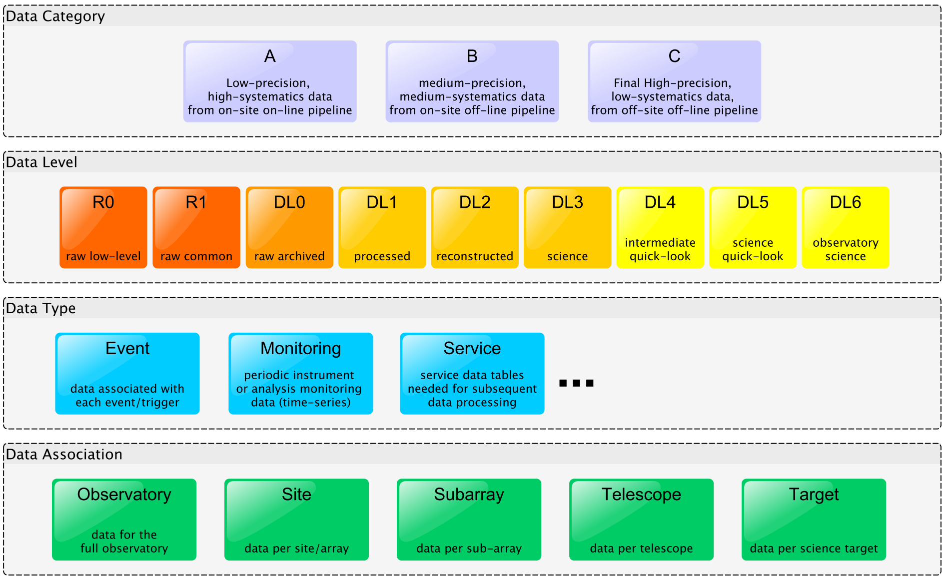

Figure 4.1 – CTA Observatory Data Products and Data Model identification scheme.

Figure 4.1 describes a dictionary for naming and identifying a CTA Observatory *Data Model* using a Category, Data Level, Data Type, and Data Association.

A data model *identifier* can be generated from this information using the following naming scheme:

```
[Category-<data product category>] [Sim-]<level>/<type>/<association>
```

Where parts in square brackets are optional (depending on context) and the *data product category* refers to the quality with respect to precision and systematics of the data, *Sim* denotes if the data products stems from Monte-Carlo simulations, the data *level* refers to the sequence in which the data item is produced (see **??** below), the *type* refers to the categorization of the data model, the *association* refers to whether the data applies to all of the CTA Observatory or a specific sub-set. The contents of each data model are not covered here. A data product may contain multiple data models, however to describe a particular data set associated with a single data model, the same naming scheme shall be used, but with the addition of the *category* to show it's origin.

Example: *`DL3/Event/Subarray`* (the data model describing a list of events from a subarray after a Data Processing Pipeline has run). To describe the file itself that contains a serialization of this and other related data models from a Category-C analysis, it is sufficient to describe it as a *`Category-C DL3/Event`* file. Note that this nomenclature also lends itself to the naming of hierarchical data-sets within a file (though the order in which the hierarchy is represented is not specified here).

4.1 Data Categories

The same data product can be produced in more than one pipeline workflow (a sequence of steps that move between data processing levels, described in the next section). For example, one might have an event-list produced by an on-site on-line pipeline in real-time, by an on-site off-line pipeline the next day, or by an off-site off-line weeks later, each with varying levels of systematics and precision from calibration and analysis techniques. To this respect, we label each data product with a predefined *data product category* describing its origin. Generally the data model for each category should be identical other than perhaps small differences, except in the case of simulated data, which may contain substantial additional information.

The following *data categories* are defined:

A (real-time) Data products produced and distributed rapidly for science alert generation and rapid data quality evaluation. These generally have the lowest precision and highest systematic uncertainties arising from basic calibration and simplified analysis techniques.

B (next-day) Data products produced and distributed after some off-line processing on-site, by the next observation day. These have somewhat better precision, lower systematics, and better data quality measurements than category-A products, but still may use simplified analysis or calibration techniques appropriate for relatively fast science alerts and proposal monitoring.

C (full)[1] Data products produced by the full high-quality data processing chain, off-site in CTA Observatory data centers, with a delay of up to a week from data taking. These use the best calibration and algorithms, providing precision and systematics meeting or exceeding CTA Observatory requirements, have the most accurate data quality assessment, and thus are the products intended for final analysis and publication of results.

4.2 Data Product Classification

This is all data product metadata, not data model

this is all now in the Top-level model document

4.3 Data Association

The *data association* allows one to see roughly to which part of the CTA instrument a data product is associated. Details about the exact association will be given in the INSTRUMENT context metadata described in subsection 2.2.1.

For identification purposes, we only allow the associations listed here, and anything at a lower or more detailed level should simply use the closest defined level (e.g. a pixel measurement would be associated with TEL, and in the detailed metadata, the fact that it is a pixel measurement would be described in the metadata).

The data associations are:

[1] We note that periodical reprocessing of all data with improved simulation and calibration methods is planned. Thus, several versions of the category C data products may exist, whereas always the latest version is distributed to the Science Users.

CTA Data that pertains to the CTA Observatory in general, independent of site or array.

Site Data that pertains to an entire array site or location (e.g. CTA-North, CTA-South, headquarters, Science Data Management Centre (SDMC), or an off-site data center). Examples include data from a common weather station, databases related to the overall array layout, or calibrations that affect the full array or array site.

Subarray Data that may be repeated for each subarray. Examples include the subarray trigger[2] data, reconstructed shower parameters, instrumental response functions, etc.

Telescope Data that may be repeated for each Cherenkov telescope. Examples include Cherenkov images, parameters, etc. These are generally indexed by a telescope id number or telescope type grouping.

Target Data associated with a science target, e.g. a specific astronomical object or region-of-interest.

4.4 Examples

Some examples of data models, named using the scheme described above.

Identifier	**Description**
DL0/Event/Telescope	Telescope-wise event data (e.g. waveform readout)
DL0/Event/Subarray	Subarray-wise event data (e.g. subarray trigger tables)
DL3/Event/Subarray	Gamma-like event data
DL3/IRF/Subarray	effective area response matrix for a subarray
DL4/DataCube/Target	Binned counts *Data Cubes* for a target region-of-interest
DL4/SkyModel/Target	physics model for a target region of interest
DL0/Monitoring/Telescope	data models pertaining to monitored telescope parameters (e.g. from Cameras, Drives, Astrometrics, etc).
DL1/Monitoring/Telescope	category for monitoring of telescope-wise quantities produced during the first stage of the data processing pipeline (average Hillas parameters, per-second computed pedestals, etc).

4.5 Data Models for Monte-Carlo Simulated Data

Thee data models for these simulated sample data products will largely be equivalent to the data products from measurements, they may contain additional information describing e.g. the applied simulation configuration and simulated "true" parameters.

The naming scheme of these simulated data models follows that of the top-level data model with an additional “Sim-” prefix to distinguish simulated samples from data.

[2] also known as SWAT (SoftWare Array Trigger)

4.5.1 Simulated Data Levels below R1

Monte-Carlo simulations require the definition of some specialized data levels below R1, that are not needed for real data, including the output of particle-tracking data (i.e. Corsika), Cherenkov-light propagation data, Ray Tracing data, and low-level instrument simulation data. For these, we will define *Sim-R0.0*, *Sim-R0.1*, and so forth. The detailed definitions will appear in a future document or future revision of this document.

4.5.2 Simulated R1-DL5 products

For data levels (*R1-DL6*), corresponding data products generated from Simulations simulations may exist. In addition, the different algorithms used in category A/B/C analyses may also be applied to simulated data to correctly describe the detector and analysis response. An example for such a Simulated data model is *Category-C Sim-DL0/Event/Telescope*.

4.6 Data Models for Calibration Data

Calibration data intentionally does not get a special category in this naming hierarchy for two reasons:

1. The concept of calibration occurs in all data processing levels and across all associations and types. For example it is possible to have calibration data that occurs for each event (e.g a measurement of the baseline of the trace readout), or one that is a *Monitoring* quantity (e.g. telescope pointing), or even a *Service* quantity (e.g. a table of ring-sampler corrections).
2. It is difficult to separate which data items are purely needed for calibration versus other uses. For example, camera images, which are normally considered “data”, may also be used to calculate calibration parameters that are applied later. Indeed the storage of specialized camera images use for calibration, e.g. those collected from flat-fielding measurements, will be included in the *DL0/Event* data model, and may even be the same as for normal observations except a few additional monitoring and configuration parameters.

Therefore all calibration data should fit into the previously defined scheme, and where each item should appear depends on the validity range of that quantity.

A Mapping to IVOA Metadata

A.1 ObsCore

Include table of mapping of IVOA ObsCore to Observation and Reference metadata

B Reference Metadata Example Representations

The following shows a few representations of the CTA Observatory `Reference` metadata as it might be represented in different formats:

B.1 Reference metadata in flat FITS header card representation

This uses the FITS ESO HIERARCH long-keyword standard (https://fits.gsfc.nasa.gov/registry/hierarch_keyword.html)

```
HIERARCH CTA METADATA VERSION = 1.0.0
HIERARCH CTA CONTACT ORGANIZATION = CTAO
HIERARCH CTA CONTACT NAME = CTAO Support
HIERARCH CTA CONTACT EMAIL = support@cta-observatory.org
HIERARCH CTA PRODUCT DESCRIPTION = Cat C DL3 event list
HIERARCH CTA PRODUCT CREATION_TIME = 2018-11-10 15:30:00
HIERARCH CTA PRODUCT ID = TBD
HIERARCH CTA PRODUCT DATA CATEGORY = C
HIERARCH CTA PRODUCT DATA LEVEL = DL3
HIERARCH CTA PRODUCT DATA TYPE = Event
HIERARCH CTA PRODUCT DATA ASSOCIATION = Subarray
HIERARCH CTA PRODUCT DATA MODEL NAME = open-gamma-astro
HIERARCH CTA PRODUCT DATA MODEL VERSION = v0.2
HIERARCH CTA PRODUCT DATA MODEL URL = https://github.com/open-gamma-ray-astro/ga
HIERARCH CTA PRODUCT FORMAT = fits
HIERARCH CTA PROCESS TYPE = observation
HIERARCH CTA PROCESS SUBTYPE = standard
HIERARCH CTA PROCESS ID = 12345
HIERARCH CTA ACTIVITY NAME = pipeline-stage-3
HIERARCH CTA ACTIVITY TYPE = software
HIERARCH CTA ACTIVITY ID = 5367fcf8-e75f-11e8-9692-3c15c2d6877e
HIERARCH CTA ACTIVITY START = 2018-11-10 15:24:11
HIERARCH CTA ACTIVITY END = 2018-11-10 15:29:00
HIERARCH CTA ACTIVITY SOFTWARE NAME = ctapipe
HIERARCH CTA ACTIVITY SOFTWARE VERSION = 2.2.0
HIERARCH CTA INSTRUMENT SITE = CTA-South
HIERARCH CTA INSTRUMENT CLASS = subarray
HIERARCH CTA INSTRUMENT TYPE = standard
HIERARCH CTA INSTRUMENT SUBTYPE =
HIERARCH CTA INSTRUMENT ID = MST20
```

B.2 Reference metadata in flat ECSV header representation

```
# %ECSV 0.9
# ---
# datatype:
# - {name: A, unit: m, datatype: float64}
# - {name: B, unit: s, datatype: float64}
```

```
# meta:
#   CTA:
#     ACTIVITY:
#       END: '2018-11-10 15:29:00'
#       ID: 5367fcf8-e75f-11e8-9692-3c15c2d6877e
#       NAME: pipeline-stage-3
#       SOFTWARE: {NAME: ctapipe, VERSION: 2.2.0}
#       START: '2018-11-10 15:24:11'
#       TYPE: software
#     CONTACT: {EMAIL: support@cta-observatory.org, NAME: CTAO Support, ORGANIZA
#     INSTRUMENT: {CLASS: subarray, ID: MST20, SITE: CTA-South, SUBTYPE: nan, TY
#     METADATA:
#       VERSION: {'1.0.0'}
#     PROCESS: {ID: '12345', SUBTYPE: standard, TYPE: observation}
#     PRODUCT:
#       CREATION_TIME: '2018-11-10 15:30:00'
#       DATA:
#         ASSOCIATION: Subarray
#         CATEGORY: C
#         LEVEL: DL3
#         MODEL: {NAME: open-gamma-astro, URL: 'https://github.com/open-gamma-ra
#         TYPE: Event
#       DESCRIPTION: Cat C DL3 event list
#       FORMAT: fits
#       ID: TBD
```

B.3 Reference metadata in hierarchical YAML representation

```yaml
CTA:
  METADATA:
    VERSION: 1.0.0
  CONTACT:
    ORGANIZATION: CTAO
    NAME: CTAO Support
    EMAIL: support@cta-observatory.org
  PRODUCT:
    DESCRIPTION: Cat C DL3 event list
    CREATION_TIME: '2018-11-10 15:30:00'
    ID: TBD
    DATA:
      CATEGORY: C
      LEVEL: DL3
      ASSOCIATION: Subarray
      TYPE: Event
      MODEL:
        NAME: open-gamma-astro
        VERSION: v0.2
        URL: https://github.com/open-gamma-ray-astro/gamma-astro-data-formats/tr
    FORMAT: fits
```

```yaml
PROCESS:
  TYPE: observation
  SUBTYPE: standard
  ID: '12345'
ACTIVITY:
  NAME: pipeline-stage-3
  TYPE: software
  ID: 5367fcf8-e75f-11e8-9692-3c15c2d6877e
  START: '2018-11-10 15:24:11'
  END: '2018-11-10 15:29:00'
  SOFTWARE:
    NAME: ctapipe
    VERSION: 2.2.0
INSTRUMENT:
  SITE: CTA-South
  CLASS: subarray
  TYPE: standard
  SUBTYPE: nan
  ID: MST20
```

B.4 Reference metadata in hierarchical JSON representation

```json
{
    "CTA": {
        "METADATA": {
            "VERSION":  "1.0.0"
        },
        "CONTACT": {
            "ORGANIZATION": "CTAO",
            "NAME": "CTAO Support",
            "EMAIL": "support@cta-observatory.org"
        },
        "PRODUCT": {
            "DESCRIPTION": "Cat C DL3 event list",
            "CREATION_TIME": "2018-11-10 15:30:00",
            "ID": "TBD",
            "DATA": {
                "CATEGORY": "C",
                "LEVEL": "DL3",
                "ASSOCIATION": "Subarray",
                "TYPE": "Event",
                "MODEL": {
                    "NAME": "open-gamma-astro",
                    "VERSION": "v0.2",
                    "URL": "https://github.com/open-gamma-ray-astro/gamma-astro-
                }
            },
            "FORMAT": "fits"
        },
        "PROCESS": {
            "TYPE": "observation",
```

```
            "SUBTYPE": "standard",
            "ID": "12345"
        },
        "ACTIVITY": {
            "NAME": "pipeline-stage-3",
            "TYPE": "software",
            "ID": "5367fcf8-e75f-11e8-9692-3c15c2d6877e",
            "START": "2018-11-10 15:24:11",
            "END": "2018-11-10 15:29:00",
            "SOFTWARE": {
                "NAME": "ctapipe",
                "VERSION": "2.2.0"
            }
        },
        "INSTRUMENT": {
            "SITE": "CTA-South",
            "CLASS": "subarray",
            "TYPE": "standard",
            "SUBTYPE": "",
            "ID": "MST20"
        }
    }
}
```

Glossary

A | B | C | D | E | F | G | H | I | M | O | P | S | W

A

array All of the Cherenkov Telescopes and other Array Elements at one of the Observatory Sites [defined by Requirement Specification for ACADA , release 2i] 9, 19, *See Also* array element

B

Binned Science Data (DL4) Data produced by binning the spatial, temporal, and/or spectral components of the Science-ready Data (DL3) over a target-specific time interval, energy interval, or region-of-interest on the celestial sphere. 10

bulk archive The archive of all high-data-volume ("bulk") DL0 to DL2 data products produced by ACADA and the DPPS. This is a distributed system managed by DPPS and comprises storage elements at all CTAO data centres. It is accessible only by internal CTAO staff and software systems. 5, 11, *See Also* archive, science archive, Archival Raw Data, Reconstructed Air Shower Data, data product, Array Control and Data Acquisition System, Data Processing and Preservation System & data centre

C

Cherenkov telescope A system composed of a Cherenkov camera and telescope structure which is used to collect and image Cherenkov light from air showers. [defined by Requirement Specification for ACADA , release 2i] 9, 19

Cherenkov Telescope Array Observatory (CTAO) An open gamma ray observatory and international user facility distributed over four primary sites: a Headquarters, Science Data Management Centre, and two array-sites in the Northern and Southern hemispheres: CTA-N and CTA-S. 1, 5, 8

D

data center 19, *See Also* data centre

data centre The physical space in which computing and storage resources and their infrastructure are housed and operated. 19

data level CTAO classifies science data into multiple levels R0, R1, and DL0-DL6 representing their level of processing. In a standard data processing workflow, each level is derived from the previous level in the sequence. 8, 20, *See Also* Device-specific Raw Data, Standardized Raw Data, Archival Raw Data, Calibrated Instrument Data, Reconstructed Air Shower Data, Science-ready Data, Binned Science Data, Advanced Science Data & Catalogues

data processing is the act of taking input and convert it to something through a set of prescribed procedures. It includes the conversion of raw data to machine-readable form, flow of data through the CPU and memory to output devices, and formatting or transformation of output. 5, *See Also* data product & Data Processing and Preservation System

data product Data and metadata that has been serialized into a file in a particular Data Format. A CTAO Data Product serializes a subset of the CTAO Data Model. 5, 8–10, *See Also* data, data format, product metadata, data model & serialization

data set pre-defined collection of Data Products, including the query that defines their relationship 10, *See Also* data product

E

effective area (A_{eff}) The energy-dependent effective collection area of an array or sub-array is defined as the number of selected gamma-rays in a given observation time divided by the incident flux and observation time. Unless otherwise specified, the Effective Area should be assumed to include all quality and background rejection cuts, including angular cuts associated with a point-source analysis. Effective Area may be given as a function of either true or reconstructed primary energy. [defined by cta-jama] 19

Enhanced Character Separated Value (ECSV) A text-only human readable data file that can be written and read by the astropy library without loss of information. The format stores column specifications like unit and data type along with table metadata by using a YAML header data structure. The actual tabular data are stored in a standard character separated values (CSV) format, giving compatibility with a wide variety of non-specialized CSV table readers. [defined by astropy] 5, *See Also* Yet Another Markup Language

F

Flexible Image Transport System (FITS) Flexible Image Transport System is an open standard defining a digital file format useful for storage, transmission and processing of data: formatted as multi-dimensional arrays (for example a 2D image), or tables. [defined by wikipedia] 5, 14

G

gamma-ray Field of View (FOV_{γ}) (gFoV) of a CTAO array or sub-array is defined as twice the angular offset from the array or sub-array pointing direction at which the differential point-source sensitivity (for a 50 hour exposure) is degraded by a factor of two. Note that the gamma-ray FoV is an energy- and analysis-dependent quantity. [defined by cta-jama] 14

H

Hierarchical Data Format v5 (HDF5) HDF5 is a data model, library, and file format for storing and managing data. It supports an unlimited variety of datatypes, and is designed for flexible and efficient I/O and for high volume and complex data. HDF5 is portable and is extensible, allowing applications to evolve in their use of HDF5. [defined by hdf group] 5

I

Instrument Response Function (IRF) a mathematical description that links the reconstructed photon arrival direction, energy and trigger time of an event to the true incident direction, energy E and arrival time t of a photon. Specifically, it allows the computation of the event density for an incoming photon intensity distribution 14

International Virtual Observatory Alliance (IVOA) The International Virtual Observatory Alliance (IVOA) was formed in June 2002 with a mission to facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and inter-operating virtual observatory [defined by IVOA website] 9, 14, 21

M

metadata data that provides information about other data 5

O

Observation Block (OB) The smallest scheduling unit, a continuous observation with a sub-array during which science data is collected on the Merged Target of the parent Scheduling

Block. During Observation Blocks, the Camera configuration and Telescope Target remain fixed. [defined by Requirement Specification for ACADA , release 2i] 9, 14, *See Also* merged target, Scheduling Block & telescope target

Observation Data Model Core Components (ObsCore) Core components of the IVOA Observation data model that are necessary to perform data discovery when querying data centers for astronomical observations of interest [defined by IVOA] 14, 21, *See Also* International Virtual Observatory Alliance

Open Archive Information System (OAIS) an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community. The term OAIS also refers, by extension, to the ISO OAIS Reference Model for an OAIS. [defined by wikipedia] 8, *See Also* designated community

P

proposal A scientific project proposed to the Observatory and assigned a unique identifier. A Proposal contains the list of Scientific Targets to be observed and associated constraints in terms of configurations, environmental conditions, observation times and external preconditions. [defined by Requirement Specification for ACADA , release 2i] 18, *See Also* science target

provenance Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance. [defined by w3c] 8, 12, *See Also* metadata

S

Scheduling Block (SB) A set Observation Blocks for the same sub-array scheduled contiguously with a Target (Single Target, Merged Target, or Split Target). Telescope targets for Observation Blocks are generated from Proposals based on the specified observing strategy. 14, 15, *See Also* Observation Block, proposal & merged target

science alert Filtered Candidate Science Alerts that have passed the associated trigger criteria and are suitable either for CTAO Observations or issuing to external observatories. [defined by Requirement Specification for ACADA , release 2i] 18

science archive The archive of DL3-DL6 data products generated by CTAO pipelines, along with proposals, schedules and other observatory metadata, and of user-contributed data. Managed by SUSS, it is exposed to external users of CTAO with restricted access for data products under embargo. 5, 11, 13, *See Also* archive, bulk archive, data product, Science-ready Data & Science User Support System

Science Data Challenge (SDC) A project to create a simulation of observed data at the DL3 level starting from pre-defined physics models that are folded through existing IRFs, following a realistic observation schedule. The goals of the project include the testing and development of the science tools, testing and development of data access methods, training of new users, and the exploration of potential science cases before CTAO starts operating. 14, *See Also* Instrument Response Function & Science-ready Data

Science Data Management Centre (SDMC) The primary centre for all CTAO science operations workflows, the management of CTAO data products, and the science user support. The selected site for the CTAO Science Data Management Centre (SDMC) is Zeuthen, Germany. 19

science target A celestial object or region of interest that is to be observed to address the scientific case of a Proposal. [defined by Requirement Specification for ACADA , release 2i] 19, *See Also* proposal

Science User Support System (SUSS) A software system responsible for the high-level science operations workflows and to provide the Science users with access to the CTAO data and software products and services through a uniform portal. Provides means for the proposal preparation and evaluation, the long-term and mid-term scheduling, the preparation and long-term preservation of science data products and tools to analyse them, MM/MWL support tools, and user support. 10

Science-ready Data (DL3) Data ready for delivery to CTAO Users that contain sets of selected Extended Air Shower events with a single final set of reconstruction and discrimination parameters per event, along with associated Instrument Response Functions and data describing the astronomical, environmental, and instrumental conditions required for science analysis. 10

serialization the process of translating a data structure or object state into a format that can be stored (e.g. files in secondary storage devices, data buffers in primary storage devices) or transmitted (e.g. data streams over computer networks) and reconstructed later (possibly in a different computer environment). [defined by wikipedia] 8, 17

simulation configuration A set of parameters selected by the user (e.g., source model, simulation software versions), provided through an interface (e.g., to obtain the observation metadata for which simulated events are generated), or determined by aspects of the IRFs (e.g., energy range or required number of simulated events). 19

stable observation interval (SOI) a time interval of observations with stable instrumental and environmental conditions (e.g., 5 min, a full night, or several nights of observations). Stable means in this context that the use of a single production configuration is sufficient to meet requirements on systematic uncertainties. 14

W

workflow a sequence, or more generally a DAG, of steps necessary to complete an automated task 18, *See Also* Directed Acyclic Graph

Acronyms

C | D | E | F | G | H | I | O | S

C

CTAO Cherenkov Telescope Array Observatory 1, 5, 8, *Glossary:* Cherenkov Telescope Array Observatory

D

DL3 Science-ready Data 10, *Glossary:* Science-ready Data
DL4 Binned Science Data 10, *Glossary:* Binned Science Data

E

ECSV Enhanced Character Separated Value 5, *Glossary:* Enhanced Character Separated Value

F

FITS Flexible Image Transport System 5, 14, *Glossary:* Flexible Image Transport System

G

gFoV gamma-ray Field of View 14, *Glossary:* gamma-ray Field of View

H

HDF5 Hierarchical Data Format v5 5, *Glossary:* Hierarchical Data Format v5

I

IRF Instrument Response Function 14, *Glossary:* Instrument Response Function
IVOA International Virtual Observatory Alliance 9, 14, 21, 31, *Glossary:* International Virtual Observatory Alliance

O

OAIS Open Archive Information System 8, *Glossary:* Open Archive Information System
OB Observation Block 9, 14, *Glossary:* Observation Block
ObsCore Observation Data Model Core Components 14, 21, 31, *Glossary:* Observation Data Model Core Components

S

SB Scheduling Block 14, 15, *Glossary:* Scheduling Block
SDC Science Data Challenge 14, *Glossary:* Science Data Challenge
SDMC Science Data Management Centre 19, *Glossary:* Science Data Management Centre
SOI stable observation interval 14, *Glossary:* stable observation interval
SUSS Science User Support System 10, *Glossary:* Science User Support System

Todo list

Sync this with GADF/VODF FITS header definitions 15

describe how the Reference, context, and fixity metadata should be serialized as headers in various file types, e.g. how the keywords are named, and what they look like. Expand the current examples to show also CONTEXT and FIXITY info, the latter for FITS can follow the FITS standard. 16

This really mixes data models and data products. Should be more specific. Categories are for data products only, the data model is the same for all of them. 17

This is all data product metadata, not data model 18

this is all now in the Top-level model document 18

Include table of mapping of IVOA ObsCore to Observation and Reference metadata . . . 21