

Structuring metadata for Cherenkov Astronomy

Mathieu Servillat

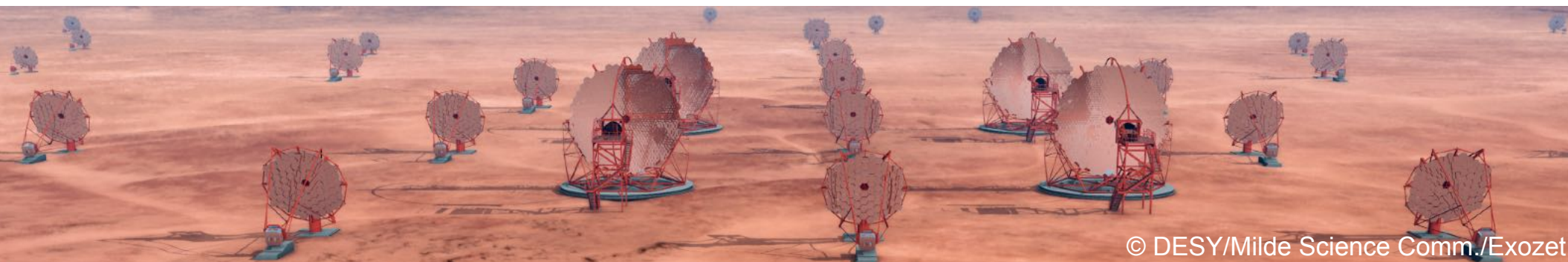
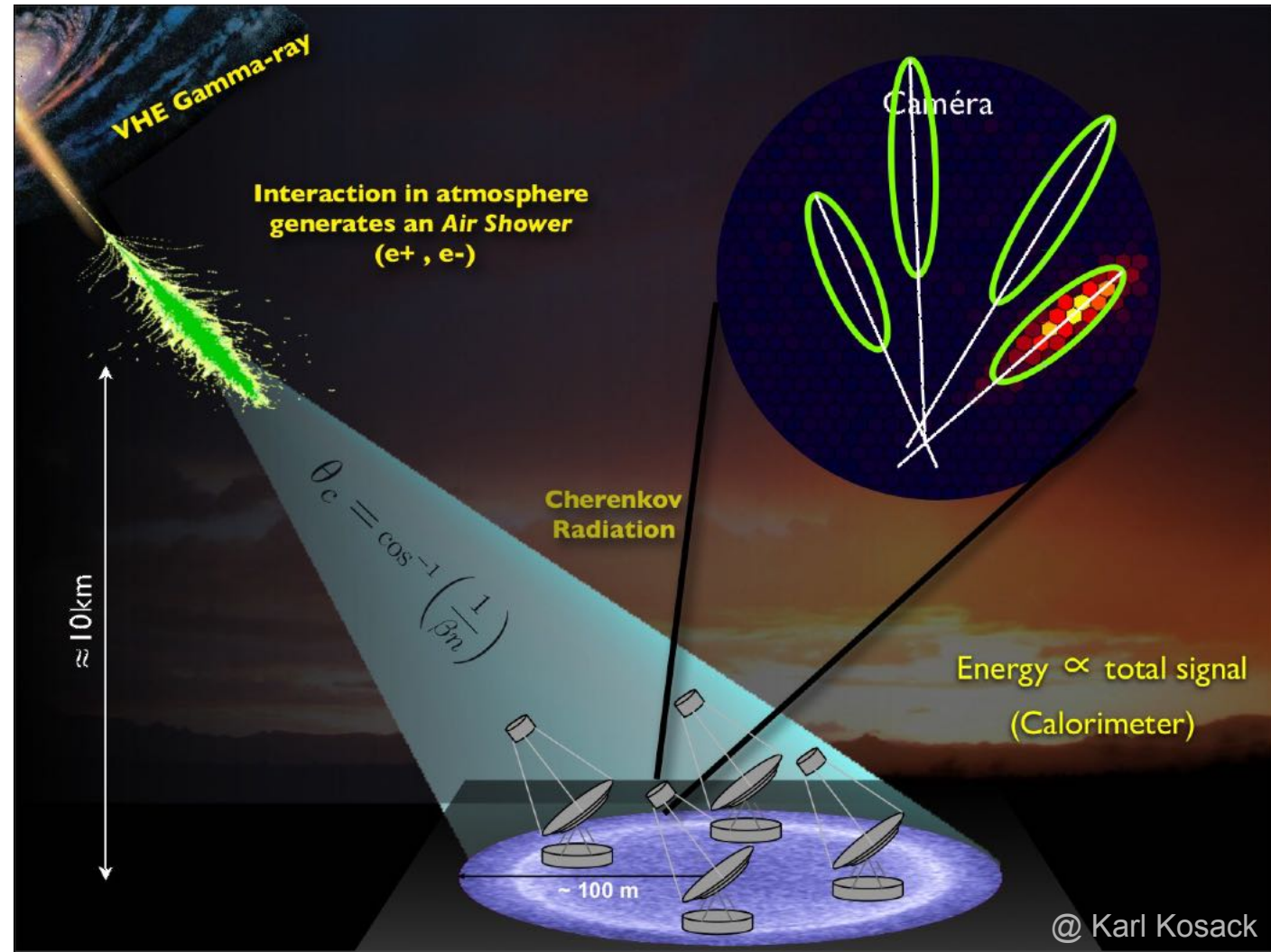
Laboratoire Univers et Théories
Observatoire de Paris
PSL Research University

ASTERICS European Data Provider Forum



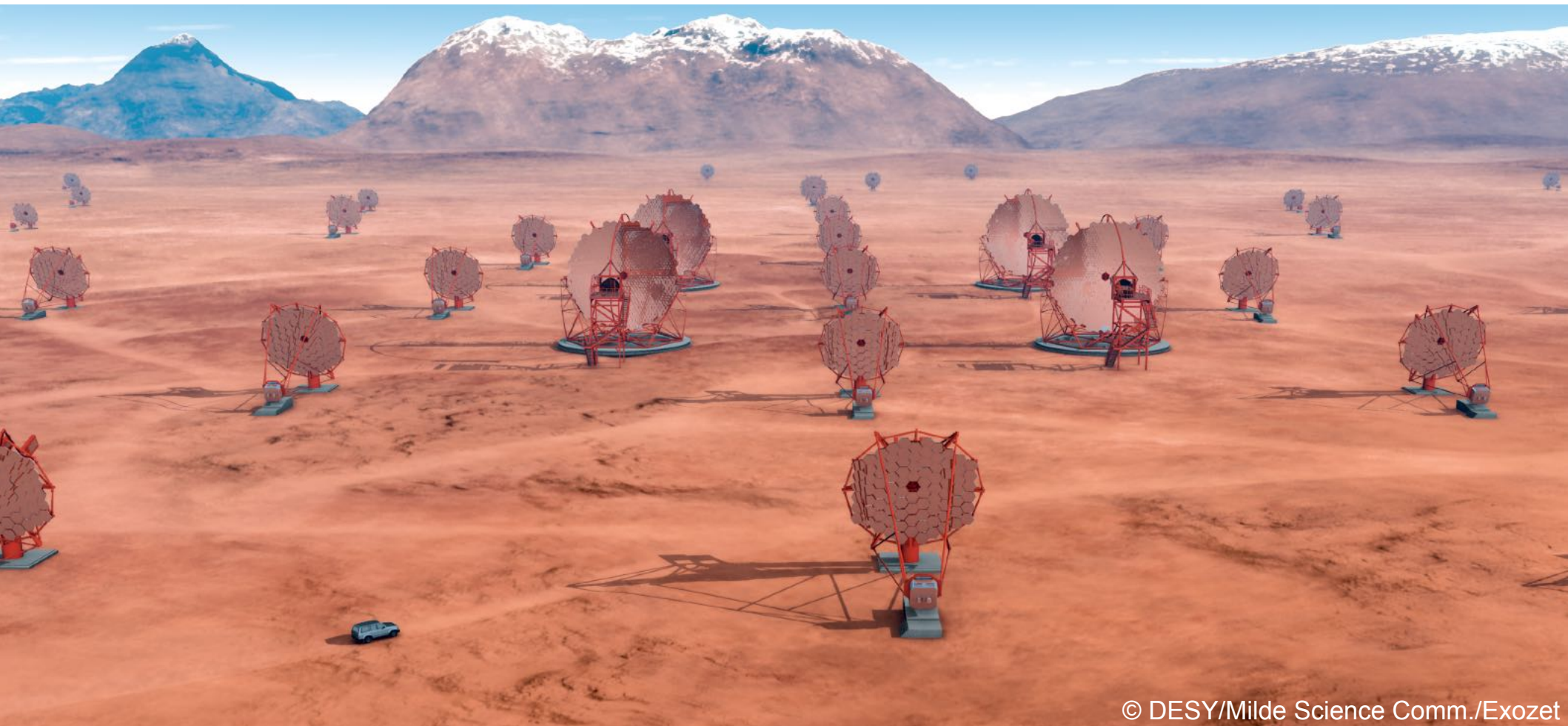
Cherenkov Imaging

- ◆ **Dark nights** (small duty cycle)
- ◆ Field of view: 5-8 degrees
- ◆ **Event Reconstruction:**
photon, particle shower,
Cherenkov light
(faint, few nanoseconds)
- ◆ **Atmosphere = calorimetre**
Simulations, assumptions
- ◆ **Complex Metadata,**
need to be structured



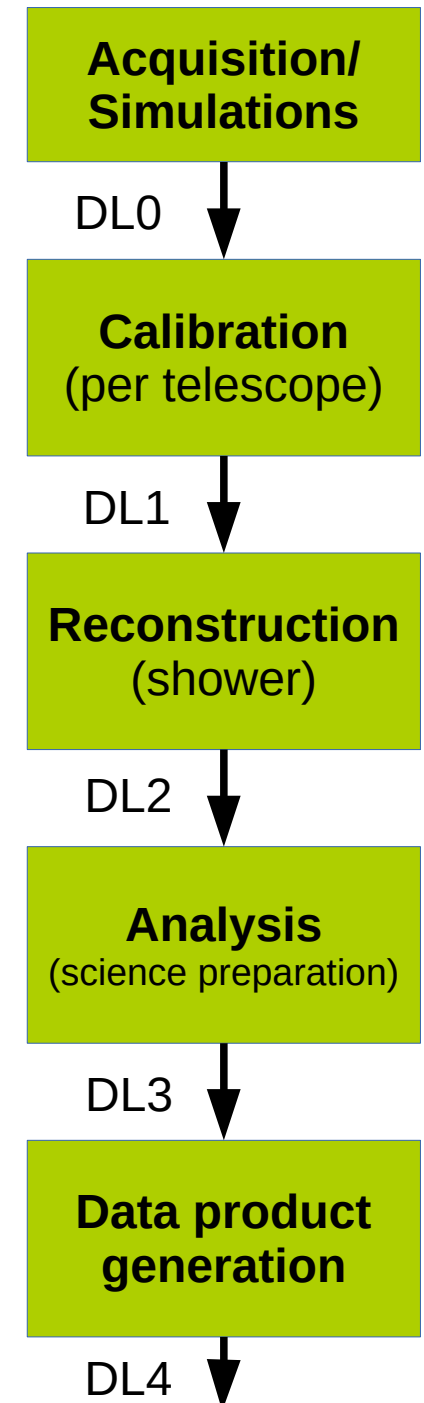


- ◆ **Two arrays** of **100 (South)** et **20 (North)** Cherenkov telescopes (4, 12 et 24 m in diametre)
- ◆ July 2015: **site selection**, Chile (ESO) and La Palma
- ◆ 2016: **pre-production** phase
- ◆ 2018-2013: **production** phase
- ◆ Observatory **open** to the Astronomy community

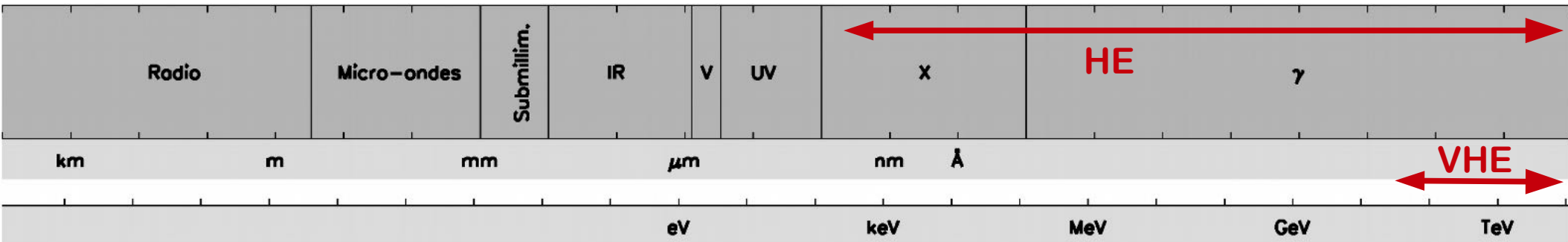


Data levels and workflow

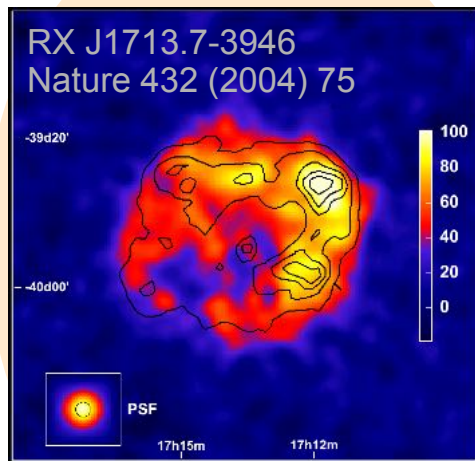
Data Level	Short Name	Description	Data reduction factor
Level 0 (DL0)	DAQ-RAW	Data from the Data Acquisition hardware/software.	
Level 1 (DL1)	CALIBRATED	Physical quantities measured in each separate camera: photons, arrival times, etc., and per-telescope parameters derived from those quantities.	1-0.2
Level 2 (DL2)	RECONSTRUCTED	Reconstructed shower parameters (per event, no longer per-telescope) such as energy, direction, particle ID, and related signal discrimination parameters.	10^{-1}
Level 3 (DL3)	REDUCED published	Sets of selected (e.g. gamma-ray-candidate) events, along with associated instrumental response characterizations and any technical data needed for science analysis.	10^{-2}
Level 4 (DL4)	SCIENCE	High Level binned data products like spectra, sky maps, or light curves.	10^{-3}
Level 5 (DL5)	OBSERVATORY	Legacy observatory data, such as CTA survey sky maps or the CTA source catalog.	$10^{-5} - 10^{-3}$



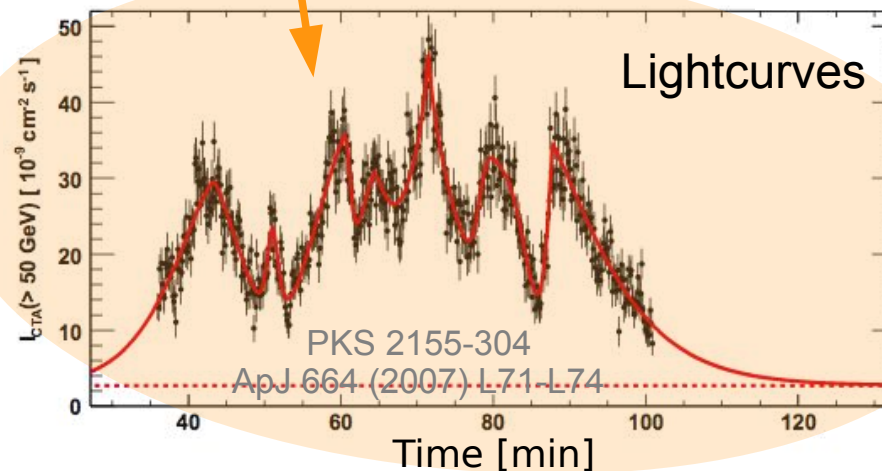
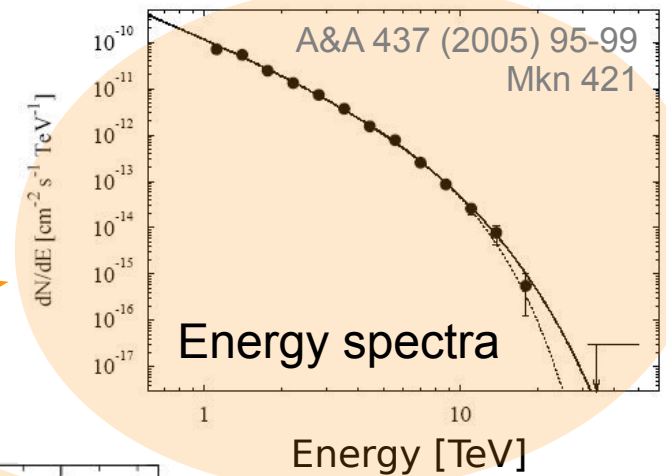
Very high energy (VHE) data



- ◆ Several orders of magnitude
- ◆ Photon counting
- ◆ Low count statistics, high background
- ◆ **Event lists**
(coordinates, time, energy)

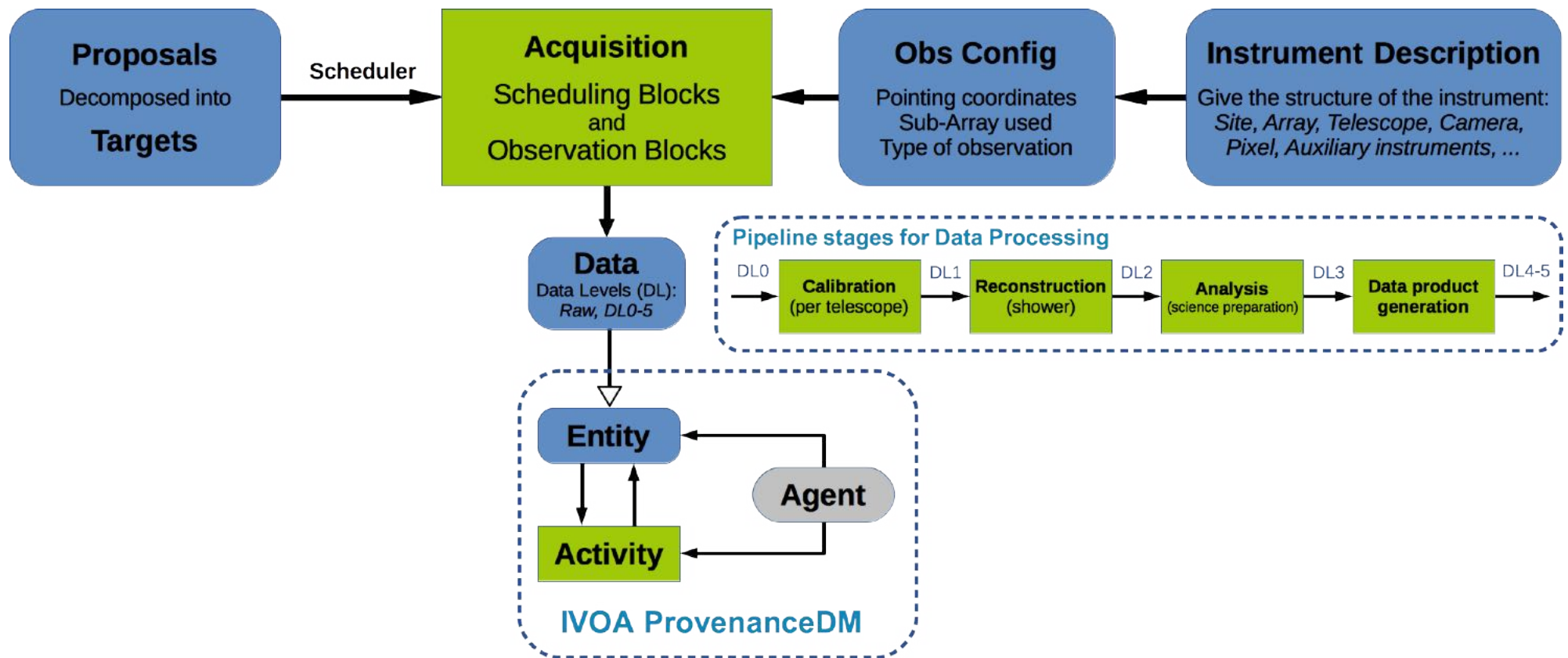


Images



High level data model

- ◆ Defines **structure** of services, content and context of data
- ◆ Can be seen as a **global interface**



High level data model

- ◆ **Proposals** → **Targets** + requirements and constraints
 - ◆ **Scheduling Blocks**
(sequence of observations planned for a given Target)
 - ◆ **Observation Blocks**
(effective start and stop times with a given configuration)
- ◆ **ObsConfig**
 - ◆ Defines coordinates, SubArray, type of Observation, strategy, pointing and trigger modes...
- ◆ **InstrumentDescription**
 - ◆ Static part of the ObsConfig → simply point to a description file
 - ◆ SubArray: fixed set of telescopes, list of active telescopes
- ◆ **Acquisition**
 - ◆ Raw Data then processed to higher Data Levels

Acquisition as a stream of data

◆ Scheduling Blocks (SB) definition

- ◆ a unit of observation that includes all necessary calibration observations/procedures for the Observatory and the Guest Observers needed for reduction and analysis. They include descriptions of configurations and calibrations.

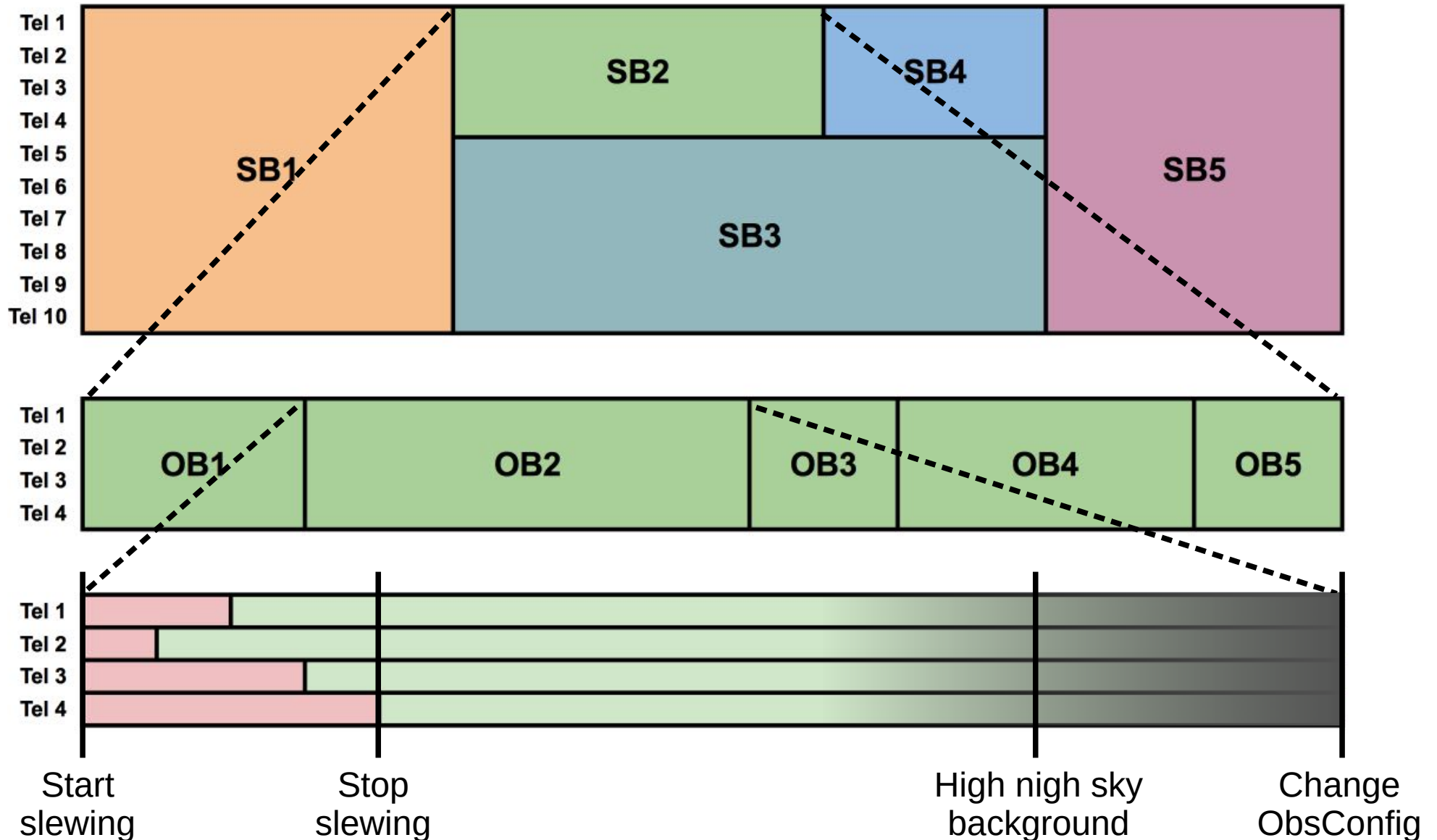
◆ Observation Block (OB) definition

- ◆ a part of the acquisition data stream with a **start** time, a **stop** time and an **ObsID**. An OB uses one and just one **ObsConfig** (SubArray, pointing, ObsType). If the **ObsConfig** changes during the acquisition, the current ObsBlock is closed and a new one is started with another **ObsID**.

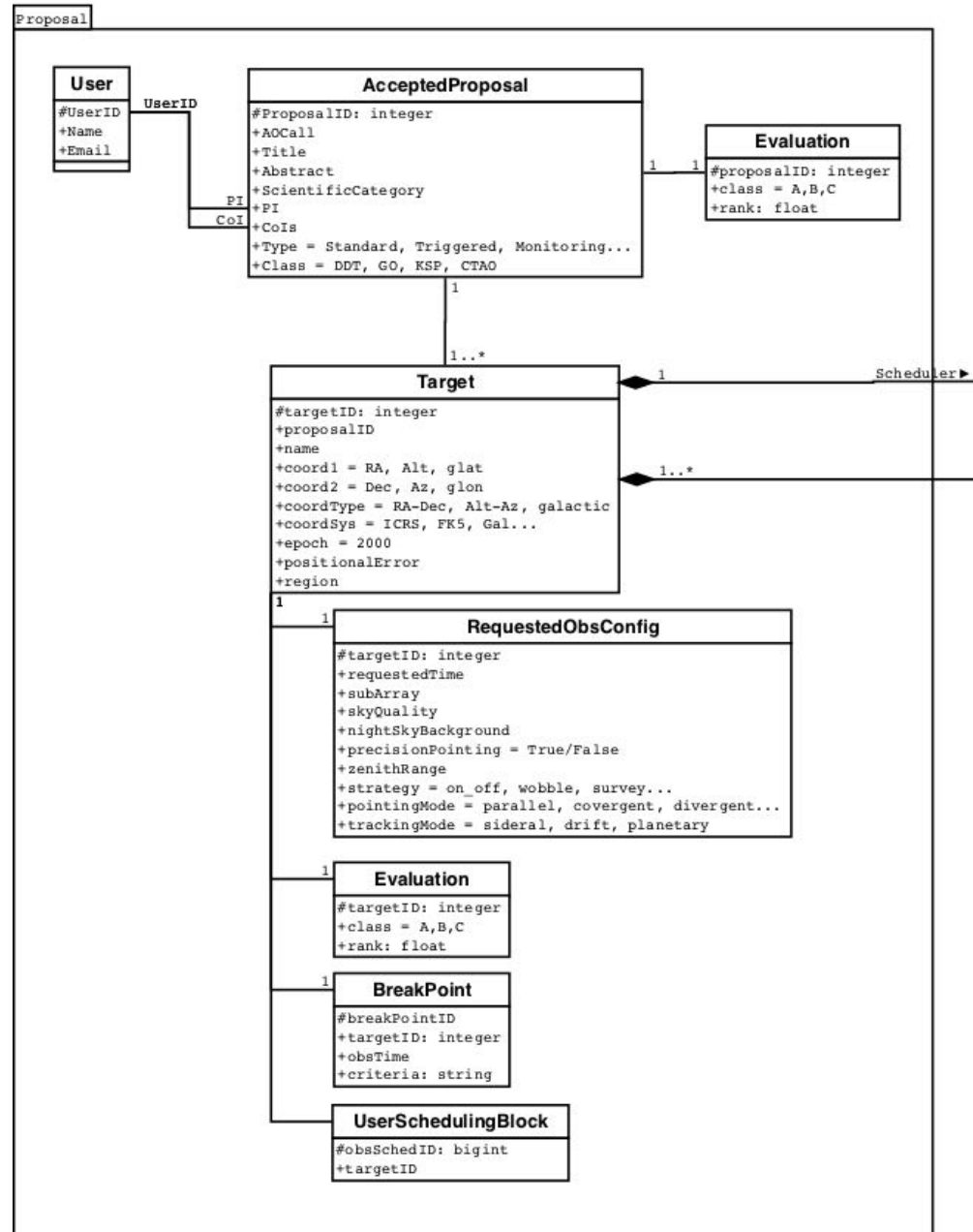
◆ Time Intervals (TI) definition

- ◆ a part of an OB with a **start** time, a **stop** time, and **common characteristics** (slewing, high NSB, calibration, ...). Time Intervals may be defined from a list of events occurring during the OB, e.g.: start slewing, stop slewing, hardware failure, high trigger rate suggesting high NSB. Some TIs could require different processing or extra MC.

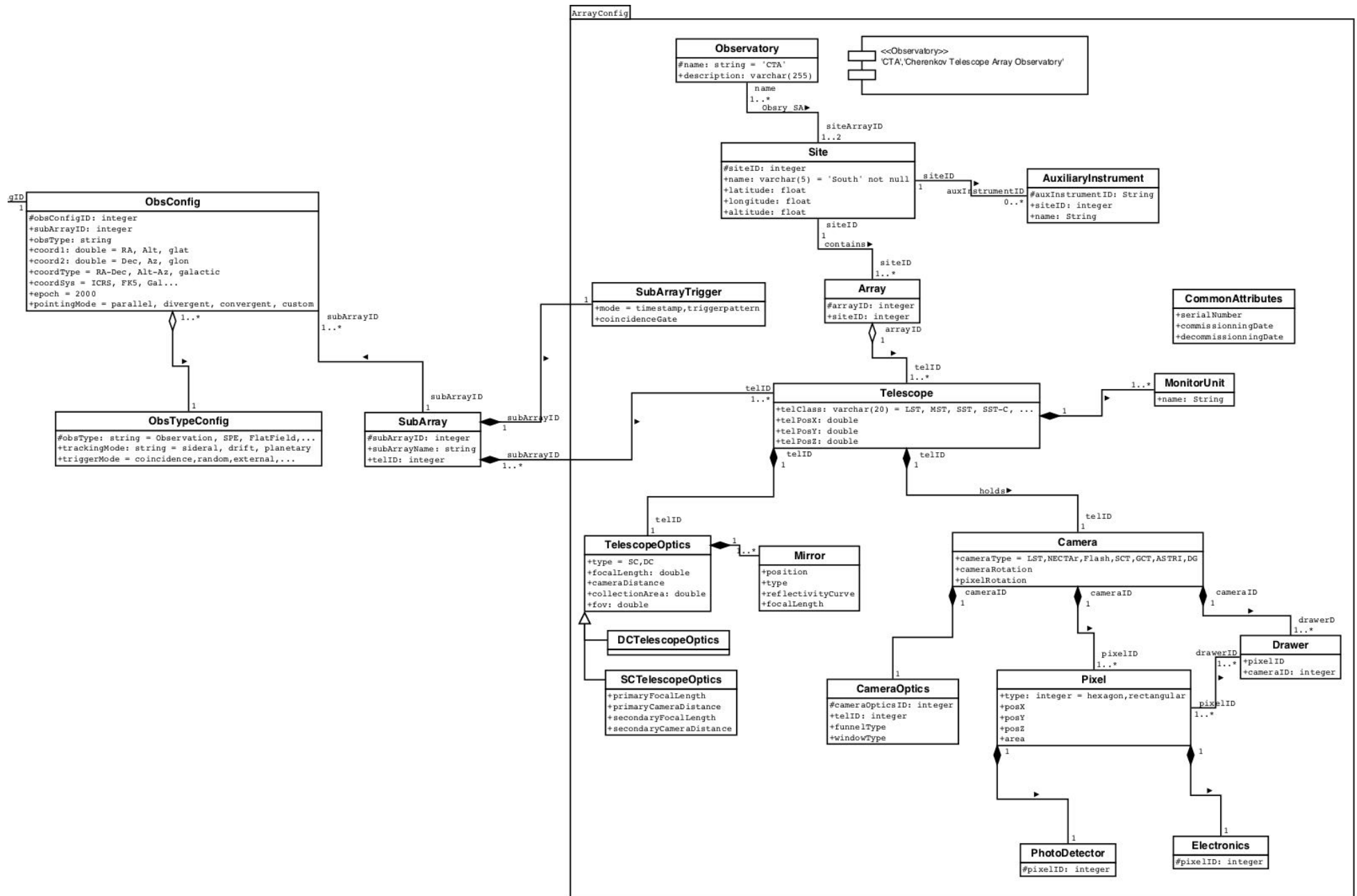
Acquisition as a stream of data



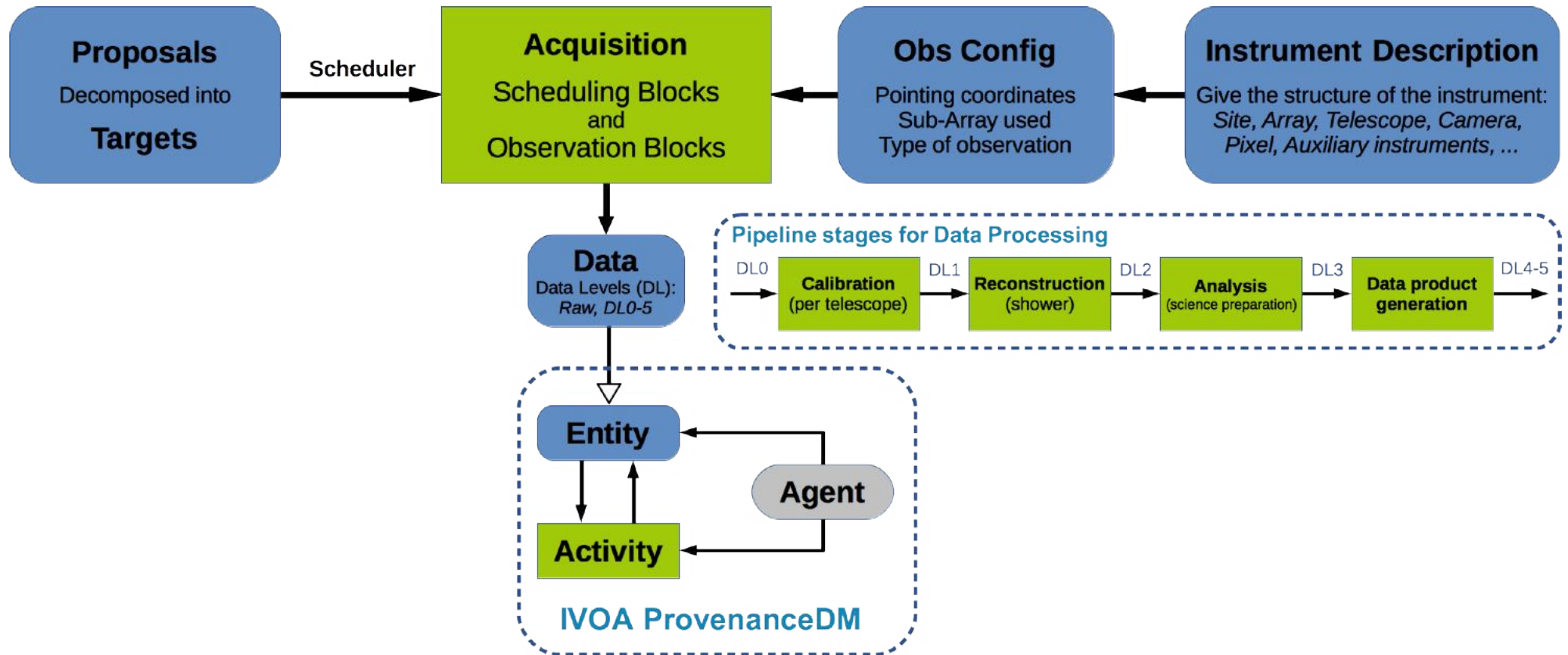
Proposal and Targets



ObsConfig and InstrumentDescription



Acquisition and Data Processing



ObsCore fields for CTA

dataproduuct_type: has to be one of the following: image, cube, spectrum, sed, timeseries, visibility, event. Set to "event" in the prototype, has it exposes the 1DC DL3 files.

calib_level: one of the following integer values: 0 (instrumental or raw data in a non-standard/proprietary format), 1 (instrumental data in a standard format, e.g. FITS), 2 (calibrated data in standard format, with instrument signature removed), and 3 (more highly processed data product). CTA defines 5 data level, for example DL3 data are calibrated data in scientific units but still include an instrument signature, hence its calib_level would be between 1 and 2.

access_url: to be defined by the Archive, however the CTA 1DC data should not be accessible to the public. We thus include simulated data hosted on <http://voplus.obspm.fr/cta/> and always point to this URL in the prototype. In the VO context, the access URL is generally a public link. To handle data rights, this may point to a retrieval system with the ID of the requested data product.

em_min, em_max: The spectral coordinates are in TeV for us and should be converted to meters to follow the ObsCore standard. This could lead to precision issues in spectral data (though it is not an issue for discovery purposes).

facility_name: we use the observatory name, e.g. "CTA".

instrument_name: As our test data comes from several experiments, we describe them here: HESS, MAGIC, VERITAS or CTOOLS (for simulated data with the ctools). This could be use to expose the CTA SubArray used to acquire the data?

Extended ObsCore fields for CTA

◆ **Optional ObsCore fields:**

- ◆ **dataprodct_subtype**: show DL0-5?
- ◆ **obs_release_date**
- ◆ **data_rights** (Public/Secure/Proprietary)
- ◆ **s_resolution min, s_resolution max** (as it is dependent on energy)
- ◆ **proposal_id**

◆ **ObsConfig:**

- ◆ **site**: North or South site.
- ◆ **sub_array_name** (or directly in **instrument_name**)
- ◆ **pointing_mode**: parallel, divergent, convergent, custom...
- ◆ **obs_mode**: wobble, scan, on, off
- ◆ **obs_type**: flatfield, science, SPE...

◆ **Provenance:**

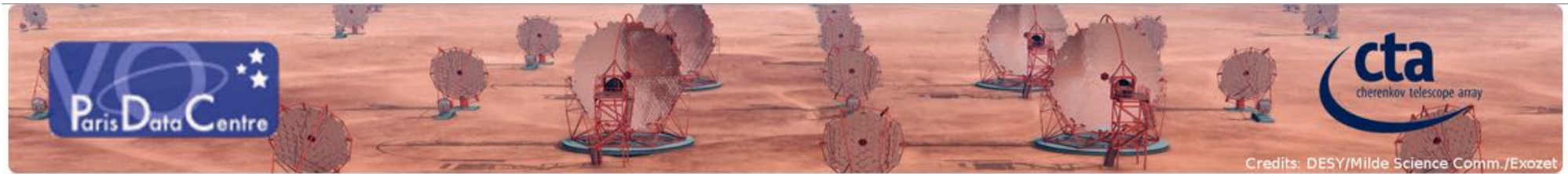
- ◆ **data_quality**: flag giving information on the data quality
- ◆ **calib_version**: version of the calibration stage of the Pipeline
- ◆ **reco_version**: version of the reconstruction stage of the Pipeline
- ◆ **reco_method**: reconstruction method used to obtain DL2 data
- ◆ **applied_cuts**: selection criteria used to obtain e.g. a DL3 photon event list
- ◆ **spectral_model**: spectral model assumed to obtain spectrum

Data mining use cases for CTA

Use case	Description
Cone Search	Search data available for a given Target
ObsCore search	Search data available corresponding to ObsCore keywords (target_name, time interval, ...), e.g.: <ul style="list-style-type: none">• search data for a given target at a given time• search data in a given region of the sky• search data that contain events at energy higher than 50 TeV
ObsCore optional search	Search data available corresponding to ObsCore optional keywords (target_class, data_rights, ...), e.g.: <ul style="list-style-type: none">• search public data for all blazars• search data for a given proposal_id
ObsConfig search	Search data available corresponding to ObsConfig keywords (sub_array_name, pointing_mode, obs_mode ...), e.g.: <ul style="list-style-type: none">• search data that include the Large Size Telescopes (LSTs)• search data for a given target, that do not include the divergent pointing mode
Provenance search	Search data available corresponding to Provenance keywords (calib_version, creation_date ...), e.g.: <ul style="list-style-type: none">• search data produced by a given version of the pipeline and for a given target• search data produced using a given reconstruction method• search data for a given target produced with loose cuts

CTA Data Distiller

<https://voparis-cta-test.obspm.fr>



CTA Data Distiller

🔍 Search Form

⚙️ Job List

✕ Sign out user

Cone Search

Target Name

Crab Nebula

Used to query Simbad with Sesame and set RA/Dec.

Source RA (deg)

83.633

Source Dec (deg)

22.514

Search radius (deg)

0.001

Submit

Reset

- ◆ Django, jQuery, Bootstrap3
- ◆ Name resolver
- ◆ Simbad through Sesame
- ◆ Builds and Sends the ADQL query

▼ ObsCore Search

proposal_id

Proposal ID

dataprodct_type

Nothing selected

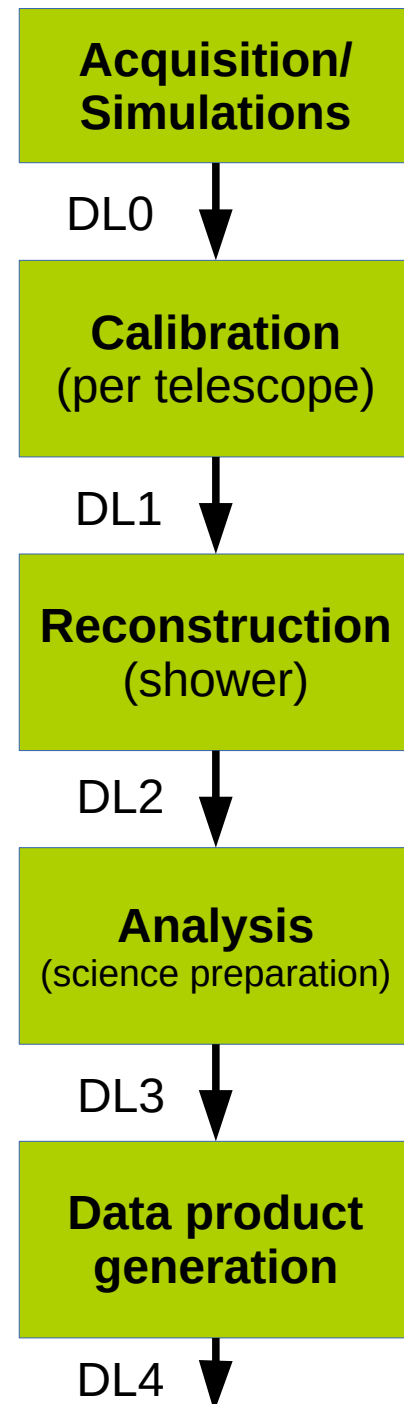
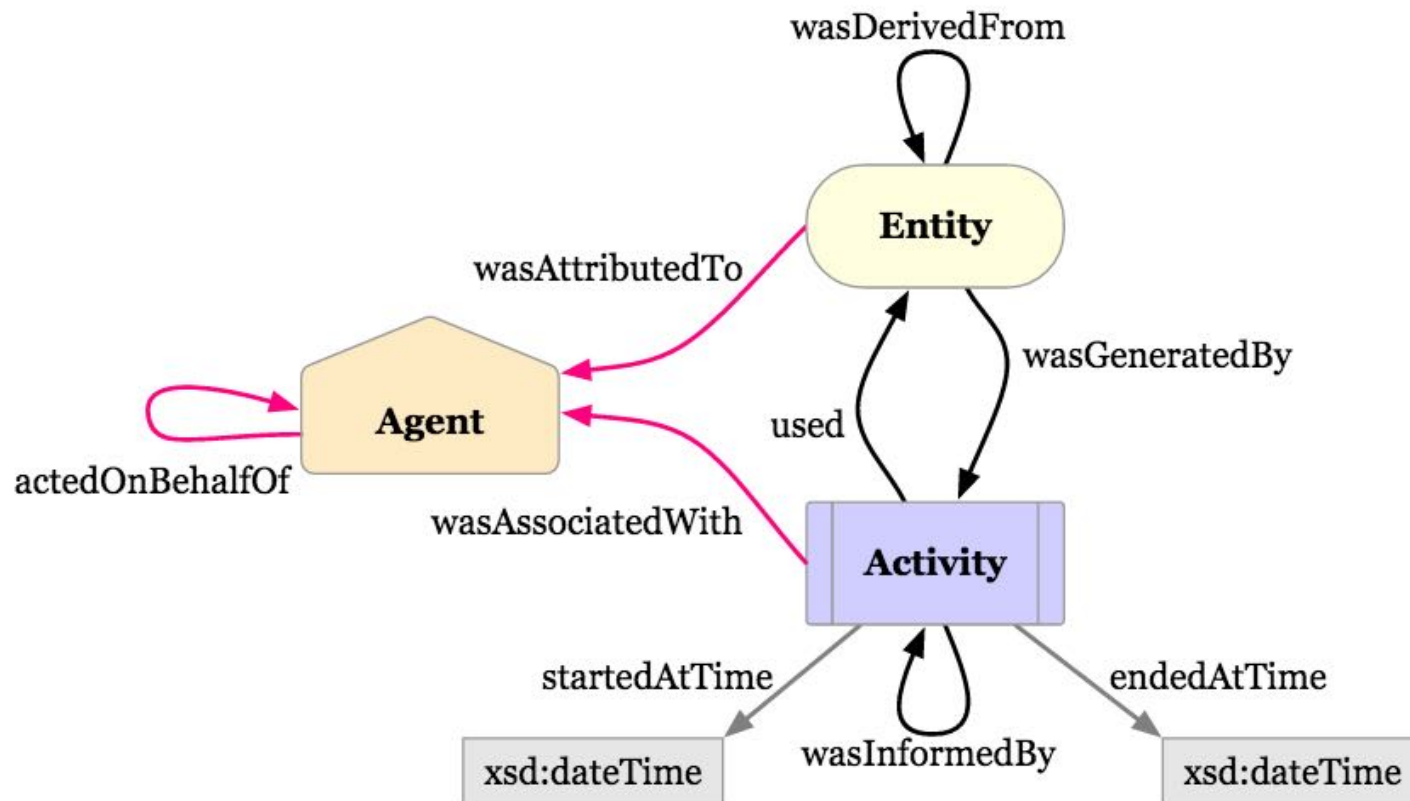
Data product (file content) primary type

dataprodct_level

Nothing selected

DL0-5

IVOA Provenance data model



IVOA ProvenanceDM working group

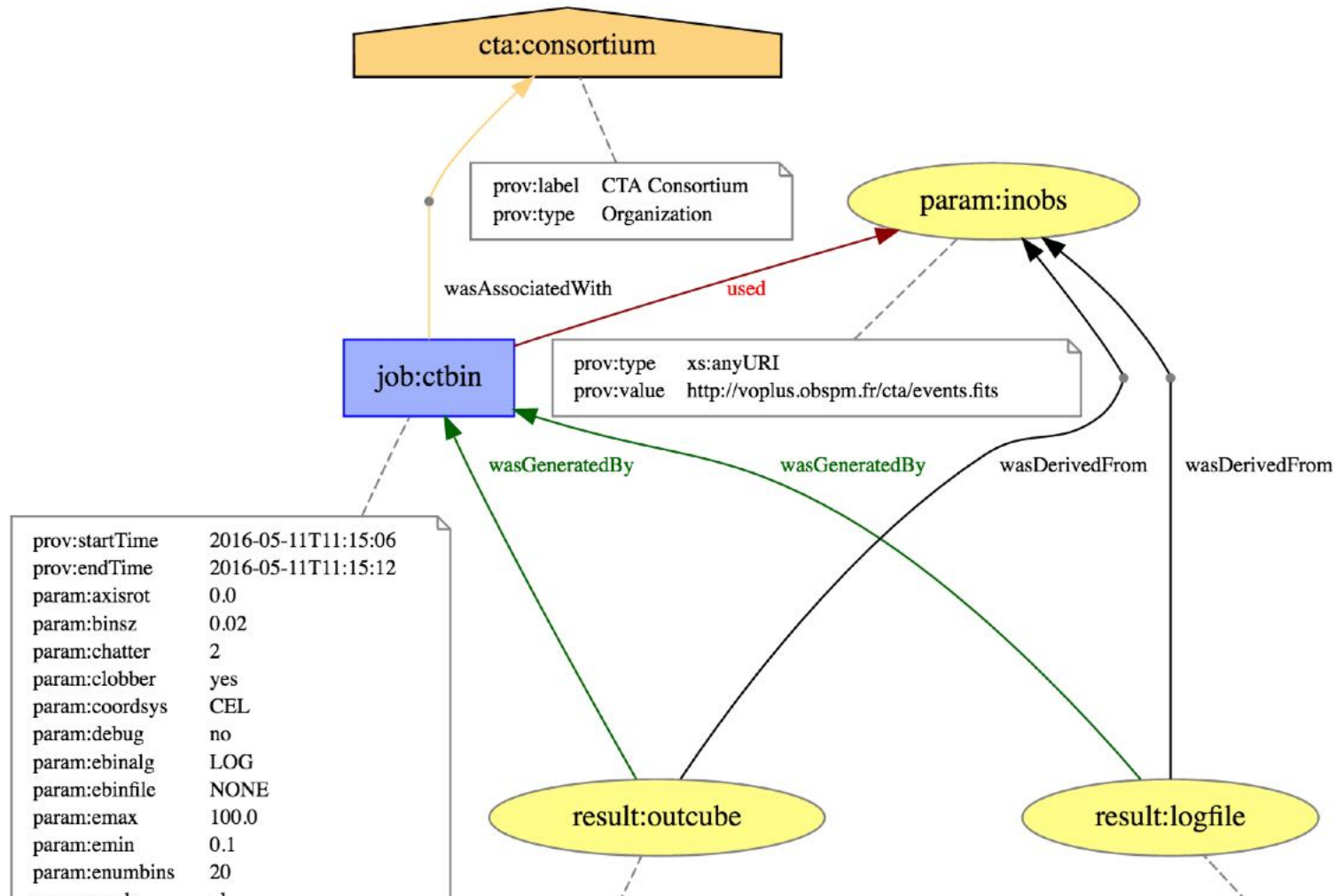
<http://wiki.ivoa.net/twiki/bin/view/IVOA/ObservationProvenanceDataModel>

W3C PROV Ontology

<https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/>

Example: analysis step with OPUS

- ◆ OPUS is a light job controller for the Paris Observatory work cluster <https://github.com/ParisAstronomicalDataCentre/OPUS>
- ◆ Provides Provenance files and visualization:



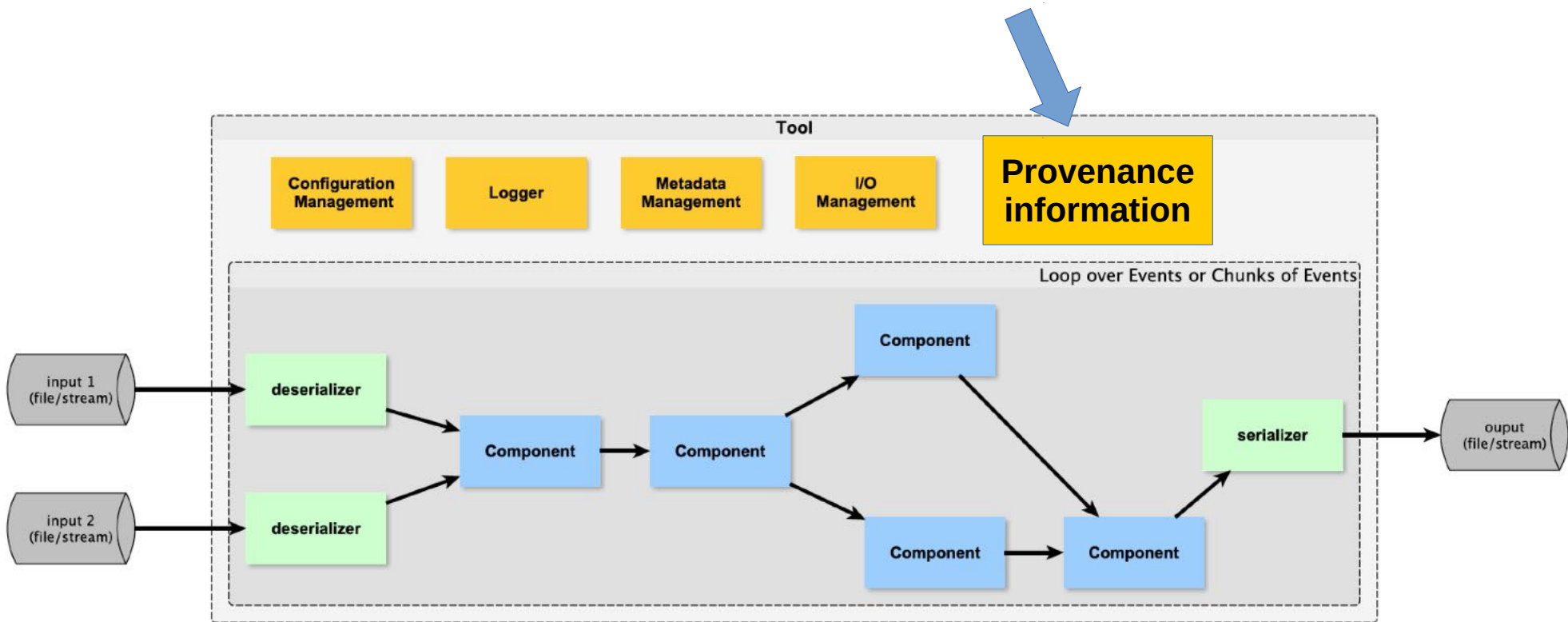
Output files (PROV-XML and PROV-JSON)

```
<prov:document xmlns:ctadata="ivo://vopdc.obspm/cta#" xmlns:ctajob
  <prov:activity prov:id="ctajobs:ctbin">
    <prov:startTime> 2016-03-13T23:44:46 </prov:startTime>
    <prov:endTime> 2016-03-13T23:44:56 </prov:endTime>
  </prov:activity>
  <prov:agent prov:id="cta:consortium">
    <prov:type xsi:type="xsd:string"> Organization </prov:type>
  </prov:agent>
  <prov:wasAssociatedWith>
    <prov:activity prov:ref="ctajobs:ctbin" />
    <prov:agent prov:ref="cta:consortium" />
  </prov:wasAssociatedWith>
  <prov:entity prov:id="uwsdata:parameters/inobs" />
  <prov:used>
    <prov:activity prov:ref="ctajobs:ctbin" />
    <prov:entity prov:ref="uwsdata:parameters/inobs" />
  </prov:used>
  <prov:entity prov:id="uwsdata:results/outcube" />
  <prov:wasGeneratedBy>
    <prov:entity prov:ref="uwsdata:results/outcube" />
    <prov:activity prov:ref="ctajobs:ctbin" />
  </prov:wasGeneratedBy>
  <prov:wasDerivedFrom>
    <prov:generatedEntity prov:ref="uwsdata:results/outcube" />
    <prov:usedEntity prov:ref="uwsdata:parameters/inobs" />
  </prov:wasDerivedFrom>
  <prov:entity prov:id="uwsdata:results/logfile" />
  <prov:wasGeneratedBy>
    <prov:entity prov:ref="uwsdata:results/logfile" />
    <prov:activity prov:ref="ctajobs:ctbin" />
  </prov:wasGeneratedBy>
  <prov:wasDerivedFrom>
    <prov:generatedEntity prov:ref="uwsdata:results/logfile" />
    <prov:usedEntity prov:ref="uwsdata:parameters/inobs" />
  </prov:wasDerivedFrom>
</prov:document>
```

```
{
  - wasAssociatedWith: {
    - _:id1: {
      prov:agent: "cta:consortium",
      prov:activity: "cta:anactools_v1.1"
    }
  },
  - agent: {
    - cta:consortium: {
      prov:type: "Organization"
    }
  },
  - entity: {
    uwsdata:results/fit_results: { },
    uwsdata:results/configfile: { },
    uwsdata:results/butterfly: { },
    uwsdata:results/spectrum_plot: { },
    uwsdata:results/spectrum: { }
  },
  - prefix: {
    uwsdata: "https://voparis-uws-test.obspm.fr/rest
    cta: "http://www.cta-observatory.org#",
    voprov: "http://www.ivoa.net/ns/voprov#"
  },
  - activity: {
    - cta:anactools_v1.1: {
      prov:startTime: "2016-04-07T00:26:00",
      prov:endTime: "2016-04-07T00:27:15"
    }
  },
  - wasGeneratedBy: {
    - _:id5: {
      prov:entity: "uwsdata:results/butterfly",
      prov:activity: "cta:anactools_v1.1"
    },
    - _:id4: {
      prov:entity: "uwsdata:results/fit_results",
      prov:activity: "cta:anactools_v1.1"
    },
    ...
  },
  ...
}
```




To be integrated in ctapipe

- ◆ **ctapipe**: code for exploring a CTA data processing framework. It is not official and not recommended for use!
<https://github.com/cta-observatory/ctapipe>
- ◆ Tool class providing configuration (set of parameters), logger, I/O management... and possibly Provenance data






Manipulating Provenance

Storing Provenance:

- ◆ Write to files 
- ◆ Store with data product (header, fits-plus...) 
- ◆ Store in a database (using data model) 

Retrieving Provenance:

- ◆ Request Provenance path
 - ◆ From files 
 - ◆ From database (API) 
- ◆ Search data products based on Provenance 
 - ◆ A given Activity was performed (with given version)
 - ◆ A given input parameter was set to...

Next steps

- ◆ **High level data model** to be completed
 - ◆ Interactions with CTA working groups
- ◆ Use ProvenanceDM to define a **database**
- ◆ **I/O package** for this database
 - ◆ Using descriptions: activity/data/parameters
 - ◆ Based on prov?
- ◆ Could be included in the CTA framework
 - ◆ **ctapipe** project in Python
 - ◆ Fill the Provenance info from DL0 to DL3
- ◆ **Query** systems
 - ◆ PROV-AQ, IVOA SSA/TAP/..., files, headers...