Euclid – Feedback on IVOA data models

JC Malapert, C. Dabin

IVOA

11-13 October 2019, Groningen

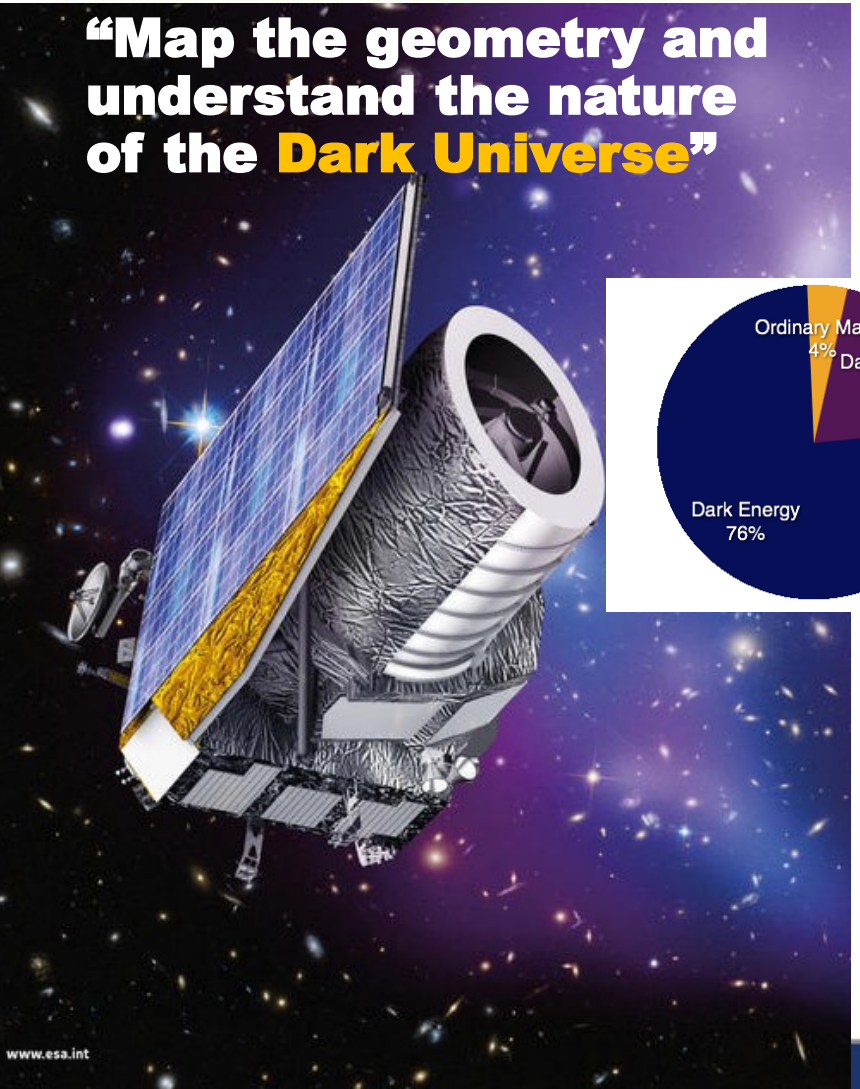
cnes
CENTRE NATIONAL D'ETUDES SPATIALES

Overview

- **Context: The Euclid project**
  - In a nutshell
  - The distributed processing of the Science Ground Segment
- **The Euclid data model**
  - Presentation
  - Differences betwee IVOA schema and Euclid schema rules
  - Decisions about IVOA schema
  - Suggested improvements

The Euclid project

cnes
CENTRE NATIONAL D'ETUDES SPATIALES



M2 mission of ESA’s Cosmic Vision Program

- Launched in 2022

1.2m space telescope + external data

- VIS instrument for visible imaging
- NISP instrument for near-IR and spectro imaging
- Ground-based surveys (KIDS, DES, LSST, J-PAS, UNIONS,…)

One of the largest organizations

- 16 countries
- 220 labs
- 1700 members

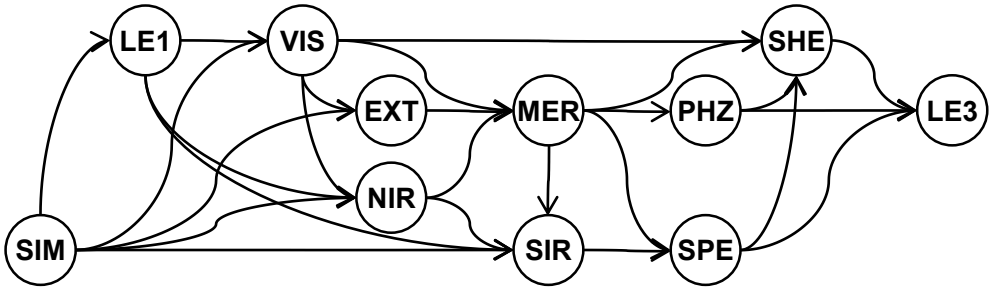
“Move the processing, not the data”

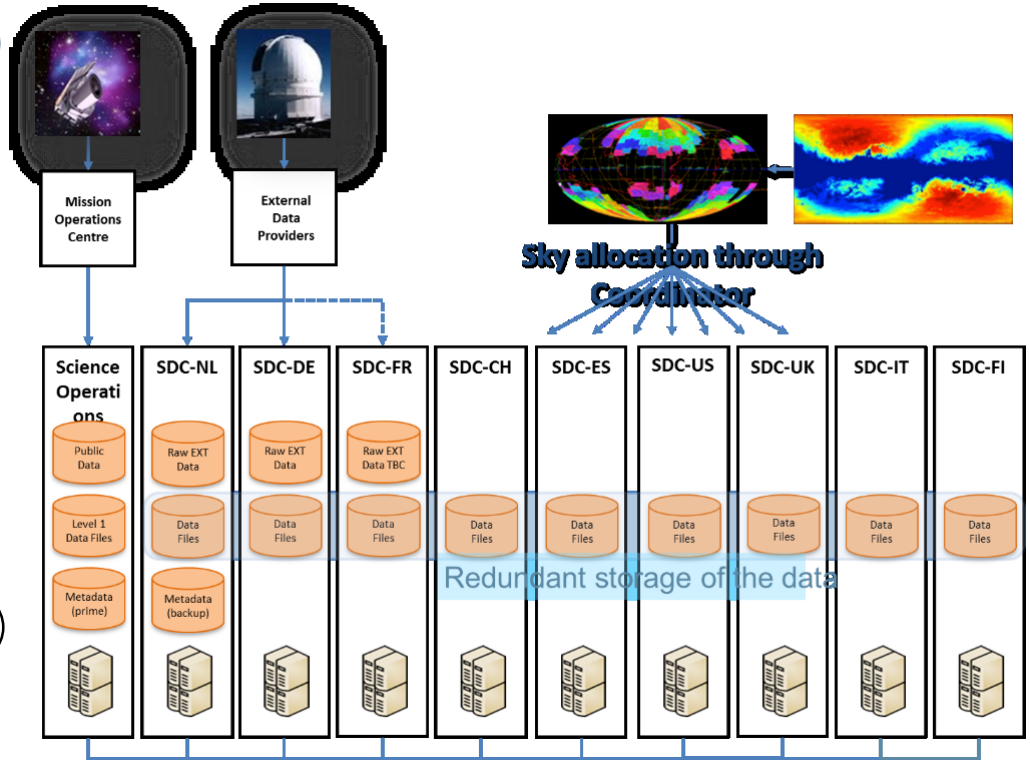
9 Science Data Centers (SDCs)

- Sky area (set of Adjacent Observations) allocation

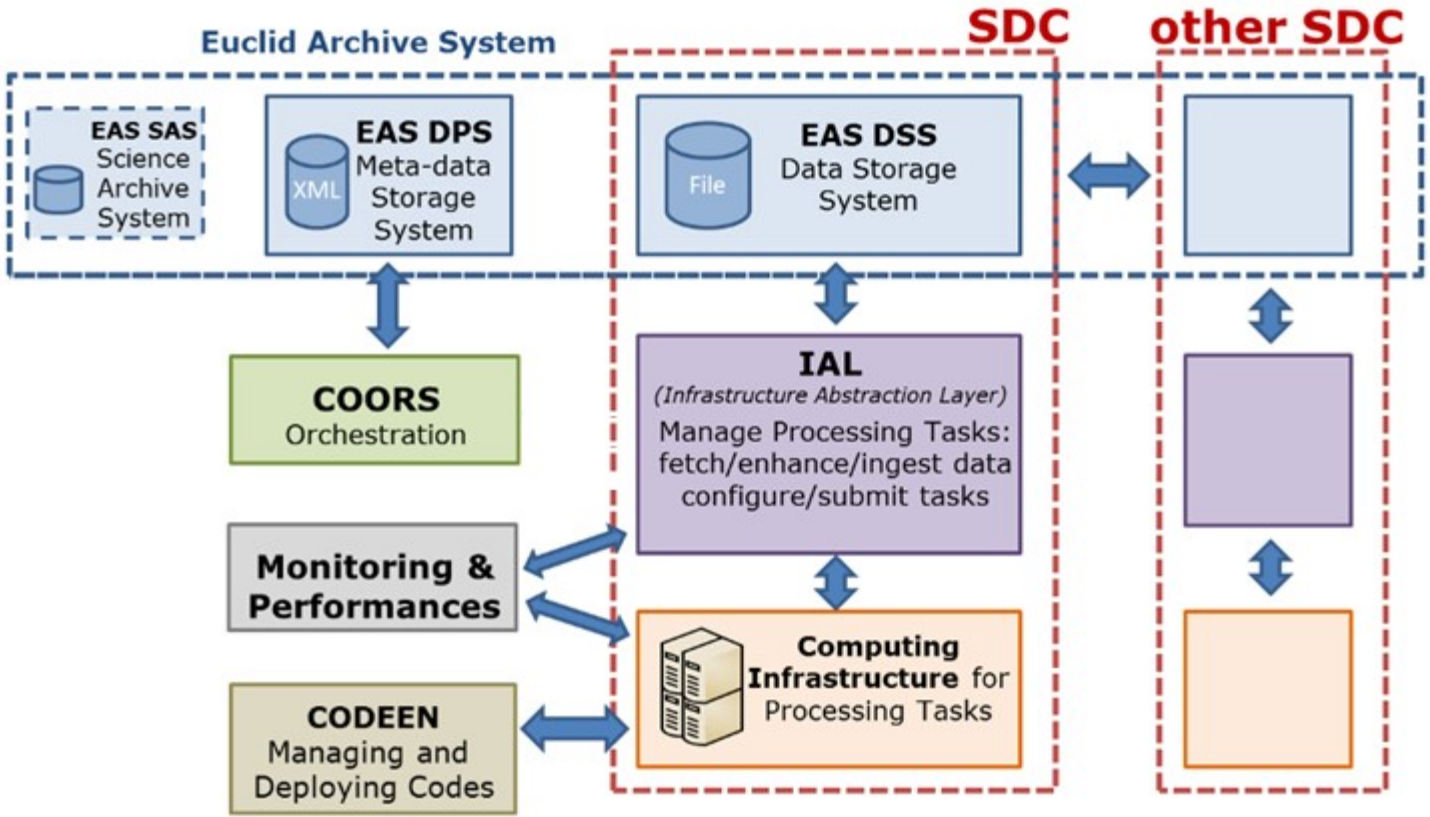
11 Processing Functions (PFs)

- Developed by the Science Ground Segment (SGS) labs
- Each PF runs on all SDCs





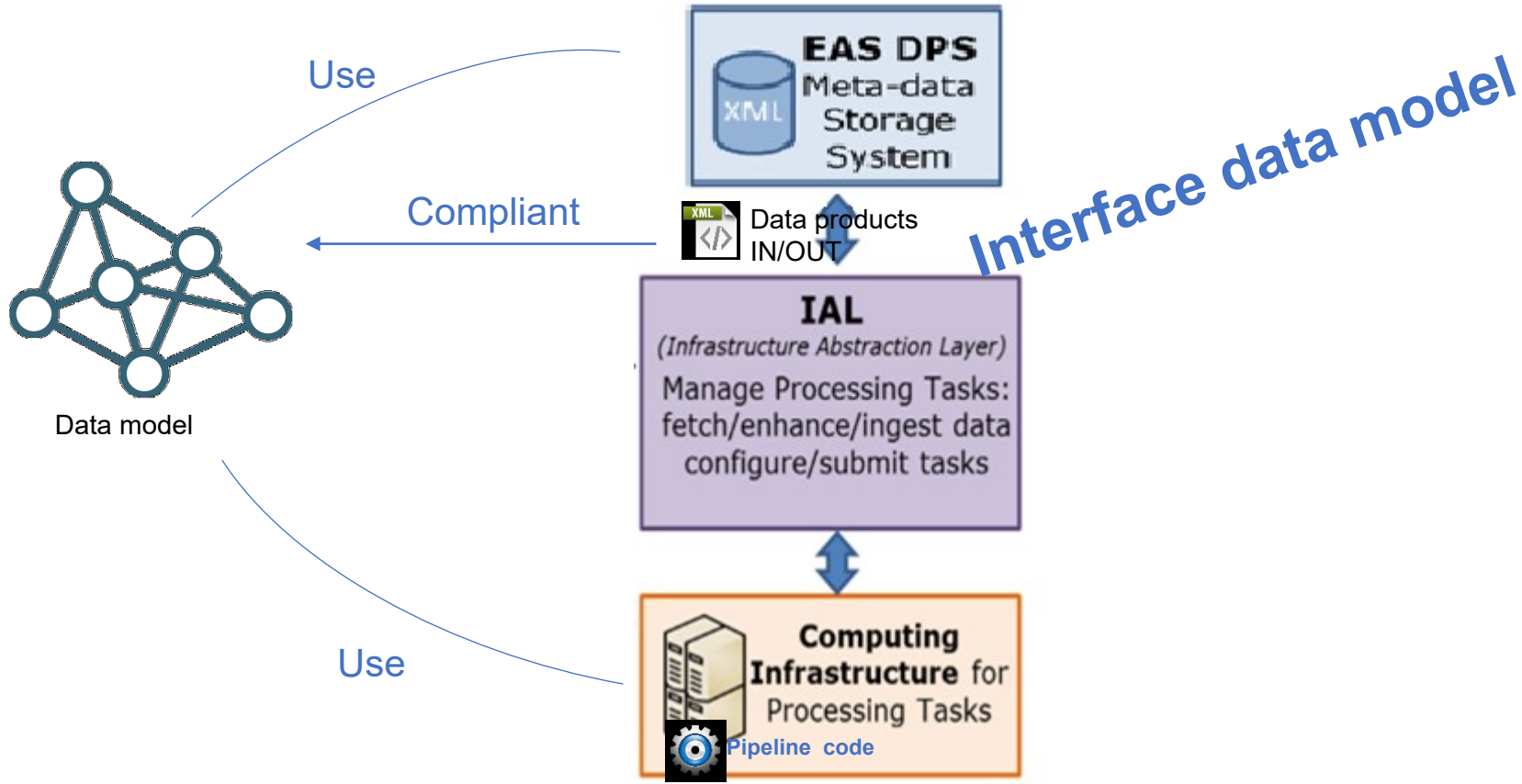
Euclid SGS components



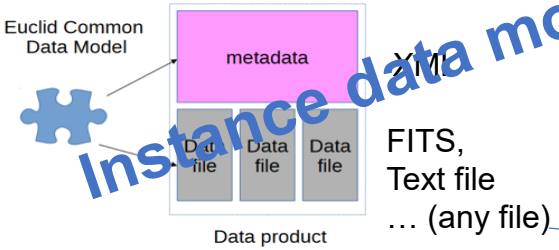
Euclid Data Model

cnes
CENTRE NATIONAL D'ETUDES SPATIALES

What is the data model used for ?



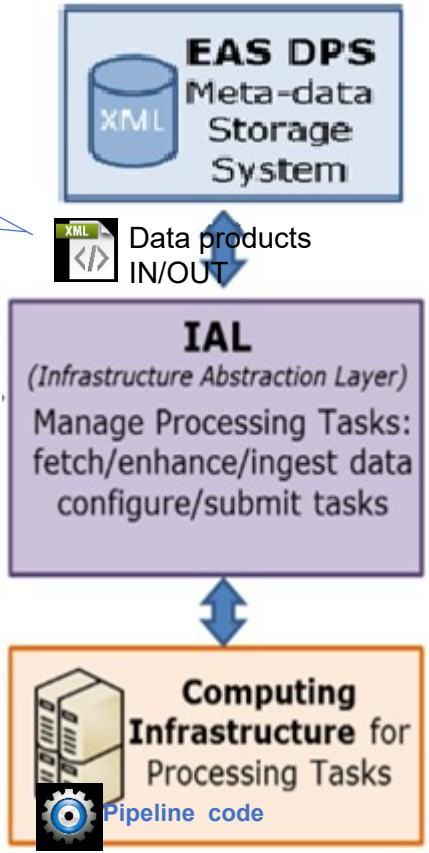
What is the data model used for ?



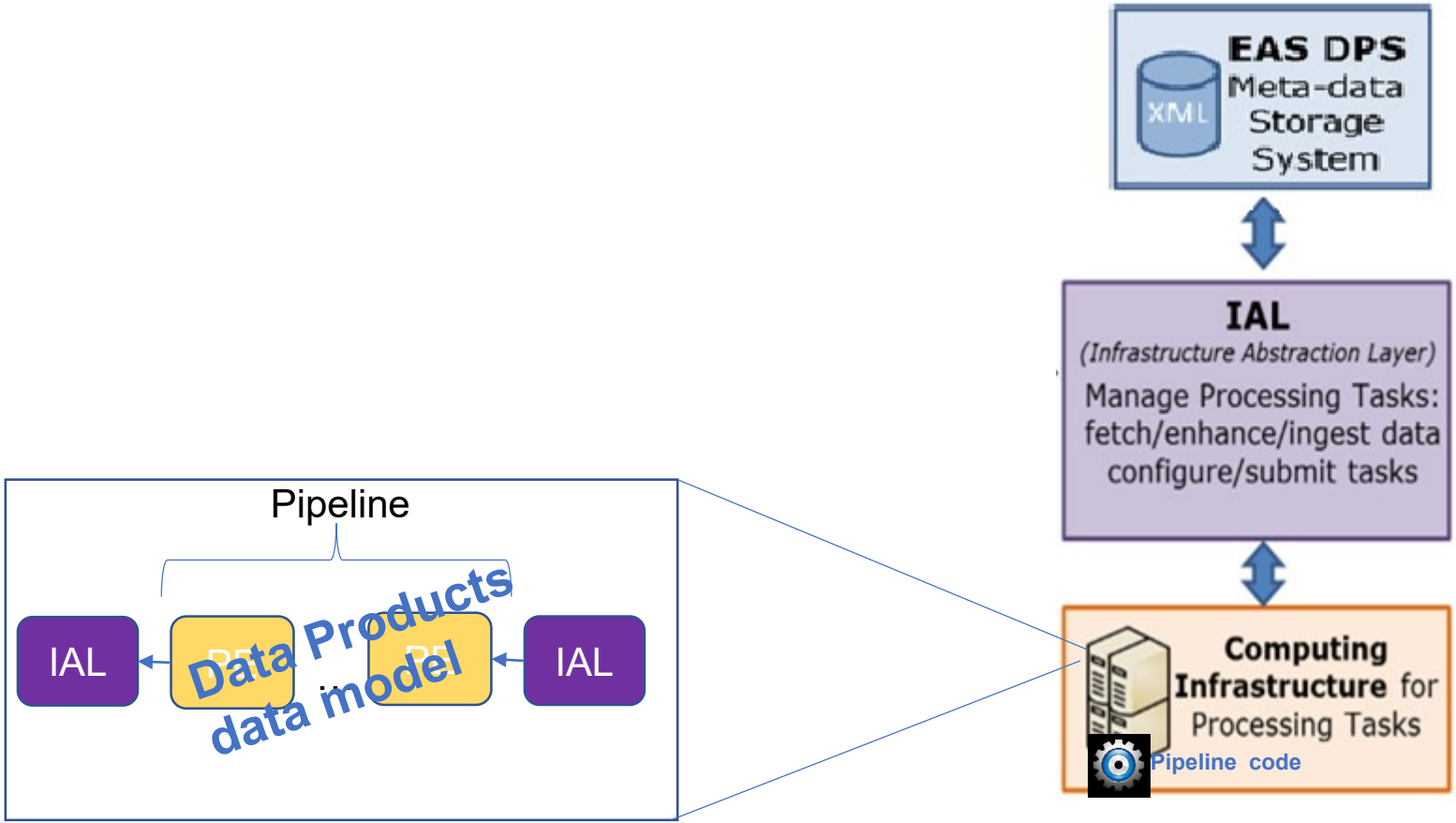
Instance data model

XML

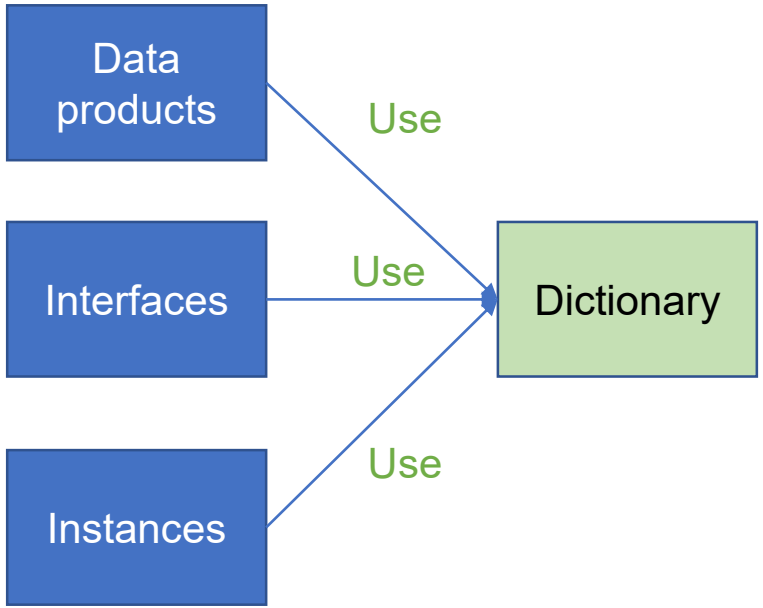
FITS,
Text file
… (any file)



What is the data model used for ?



How is the data model structured ?



4 different themes (each theme has its own namespace) :

- **bas : common definitions shared by everyone**
- **ins : instrument specific definitions**
- **pro : processing function specific definitions**
- **sys : system specific definitions (storage, processing, survey, production plan,…)**

Dictionary : ins namespace

this branch (acronym of Instruments) is related to the specific data types describing the instruments characteristics

(ex: position of the observatory, filter used)

Dictionary : pro namespace

this branch (acronym of Processing) is related to the different processing functions of the overall Euclid processing pipeline. This branch is broken down into as many processing functions and sub processing functions as implemented: le1, vis, nir, sir, ... and could be completed as much as desired.

Dictionary : sys namespace

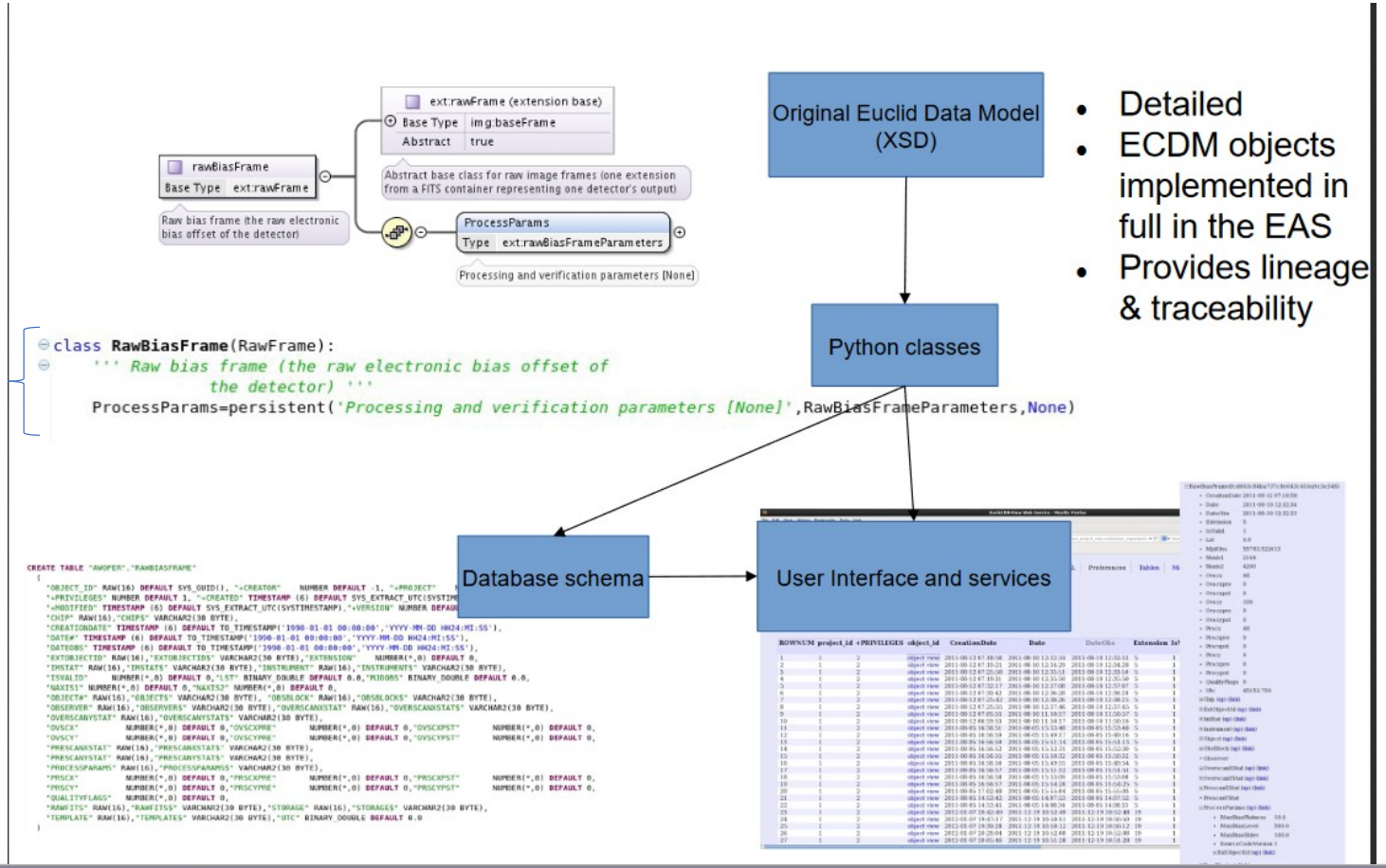
this branch (acronym of system) is the common namespace for processing and operational data types used in the Common Data Model and that are shared by transverse components and non PFs components of Euclid SGS such as Infrastructure Abstraction Layer, Orchestration component in charge of the data and processing distribution on SDCs and EC-SURV that is the component in charge of the Survey definition.

Dictionary : bas namespace

- cat: definition of catalog of astrophysical sources (transients, galaxies, stars, ...)
- cot: coordinates-time domain such as time, space, spectral etc. not covered by standard IVOA STC scheme (see: imp/stc)
- dqc: definition of Data Quality Common types shared by all Processing Functions
- dtd: data types introduced by Euclid common data model or redefined from standard XML data types: array, matrices, angles, lists, …
- fit: description of fits files format used in Euclid (each fits file used by Processing Functions shall be referenced in this section)
- img: basic definitions for Euclid (and external) images (frames)
- imp: import of external schema derived from IVOA (stc), CCSDS, ESO or FITS standards (elementary data types derived directly from the FITS specification)
- mat: mathematical definitions for polynomial computation, Gaussian,…
- msk: definitions of the binary masks related to images
- ppr: definitions related to the Parameters of some common processing steps of the overall Euclid processing pipeline
- psf: definitions related to the Point Spread Function object and its processing
- utd: Units used by Euclid, we have to mention that the units used are conform to the IVOA standards referenced here: http://www.ivoa.net/internal/IVOA/UnitsDesc/Units.html

Data model implementation

Use binding generation in pipeline code



- Detailed
- ECDM objects implemented in full in the EAS
- Provides lineage & traceability

Choosing XML schema as the formal language of data description brings numerous advantages; some of them are particularly desirable:

- XML documents conform with xsd are structured
- XML documents conform with xsd are portable and readable without proprietary tools
- XML documents conform with xsd are easily customised

A major shortcoming of this choice is the verbosity and the volume of XML documents. This drawback is addressed via specific rules on xsd.

Concerns related to IVOA schema vs Euclid rules

1/ Lack of maintenability and reusability

- Every schema is self-contained and redefine its own types with a lot of similarity with other schemas
- coordSysType is defined in spectrumDM
- Some types ‘redshiftFrameType’ are defined twice (in STC and spectrumDM) with different definitions
- No namespace, no import, no versioning of schema inside

2/ Complexity :

- STC make an intensive use of Substitution Group. Not recommended
- Type « abstract » : lack of visibility
- No « mixed content » : complexity
- No « anyType » : complexity

3/ Understandability :

- Elements and types are mixed : it may be confusing

Decisions about IVOA schema

Facing above incompatibilities between ivoa schema defined in: http://www.ivoa.net/xml/ and EUCLID rules , the following decisions were made:

- Preferred STC metadata linear string implementation,
- Huge data sets such as catalogs that describe similar objects with same attributes, and same reference frames should be packed as far as possible in a specific header (to minimize the volume of the instance of schema),
- Extract types (simple and complex) from IVOA xsd schema relevant for Euclid,
- Dictionary (thematic tree) to set up from extraction of these types: basic, units, misc, projection, timescale, coordinates, photo, spectrum, astroobjects, ….
- Keep track of the IVOA source through dedicated annotation, keep IVOA syntax for a better identification of reused IVOA types,
- Re use of ref Frame enumeration: projection, time, code projection,…
- Simplification of attribute Group

Conclusion

Suggested improvements on schema :

- **Define common types in IVOA data models**
- **Regroup types by namespace and manage versionning**
- **Reuse types in scientific data models (characterisation, spectrum, …)**
- **Use only one root element**
- **Keep it as simple as possible**
- **Have a look on ‘similar’ standards from EO : OGC**
  - Spatial objects have simpler definitions