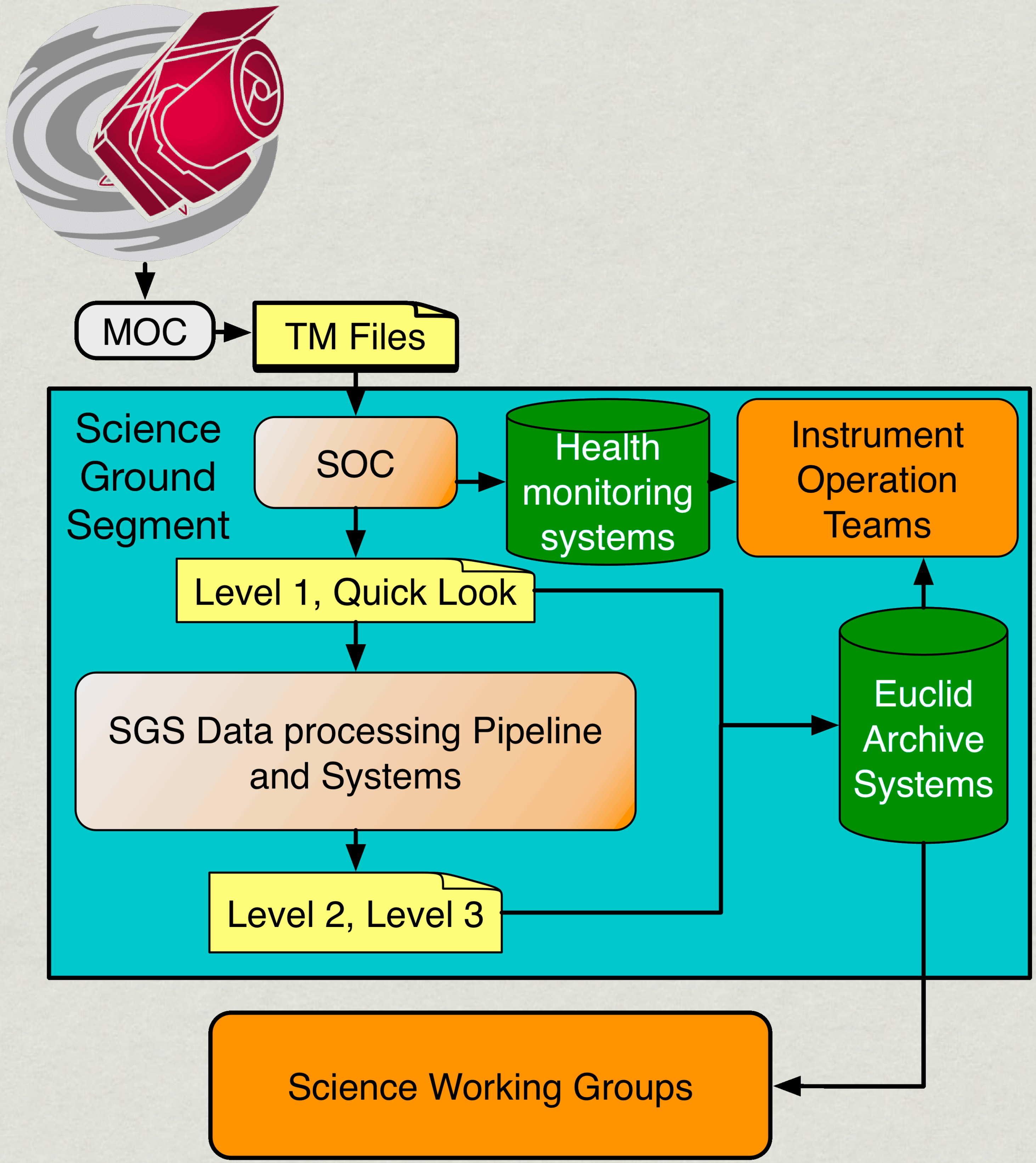
SGS Data Processing & Calibration

M. Sauvage, G. Buenadicha, Ch. Dabin, J. Hoar, O. Mansutti, C. Vuerli, A. Zacchei

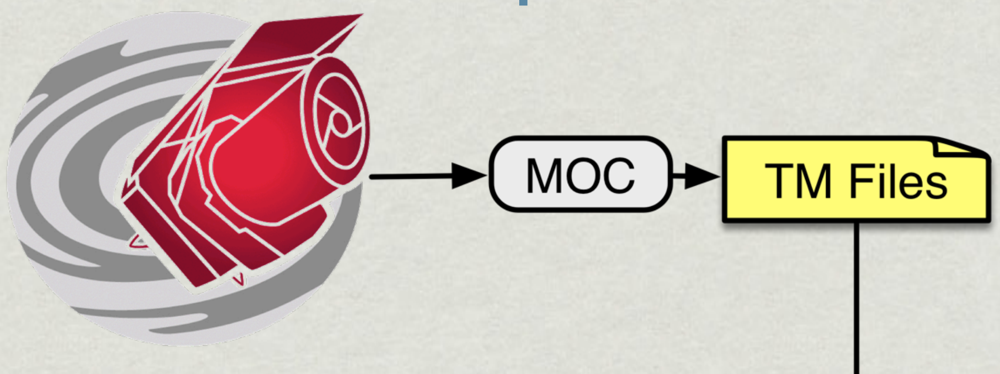
The SGS perimeter and operational role

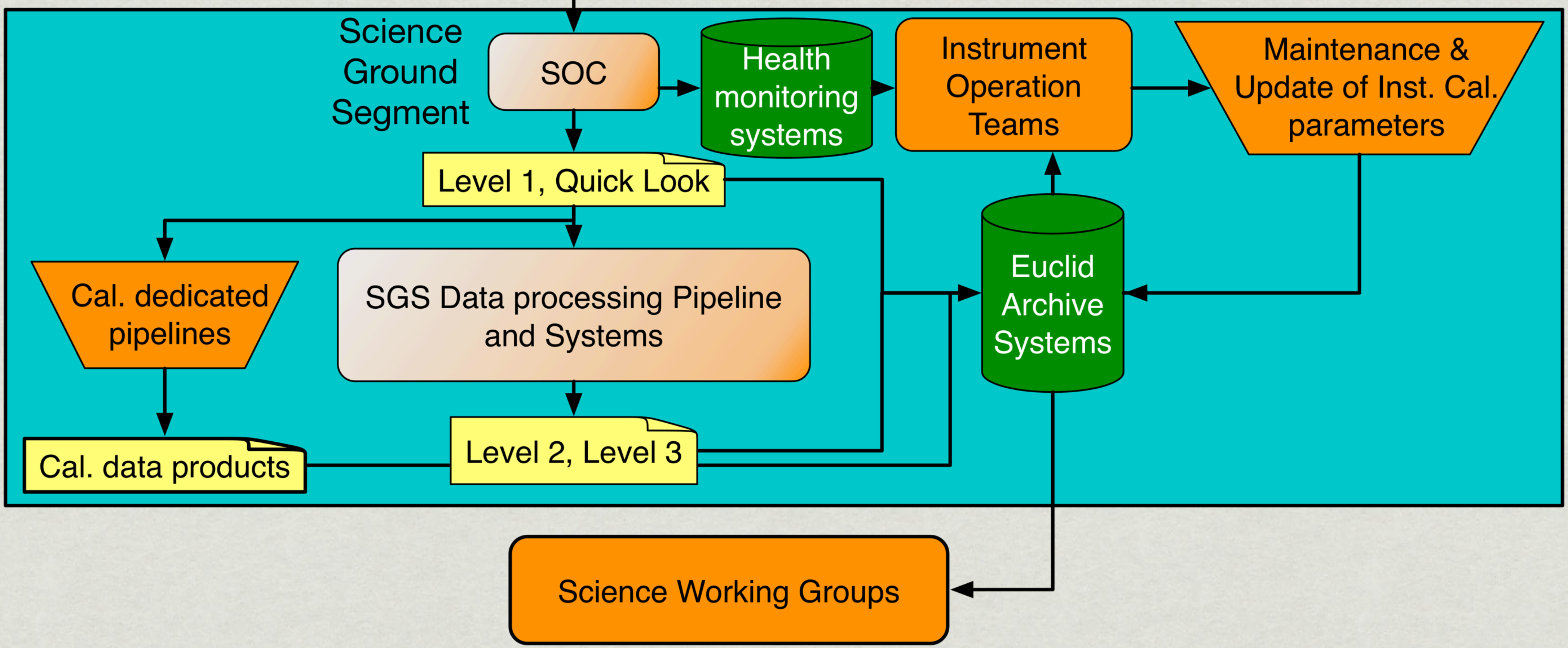


- This is a picture of what the systems we are working on are aiming at for **Euclid operations**.
- In **grey**: systems performing operations, in **orange** teams performing actions and taking decisions.
  - SGS and SOC are both!
- Not represented here are feedback actions:
  - IOTs feed the SOC team with information on instrument health to take decisions on survey execution.
  - SWGs feed the SGS with diagnostics on data science quality for pipeline improvement.
  - SOC feeds MOC with survey planning request (including re-scheduling of observation).
- **EAS is in fact an active system**, it provides data management and a transfer system for the SGS.

SGS in operation with calibration activities

The IOT's mandate does not strictly include calibration, but instrument monitoring may lead to calibration updates.





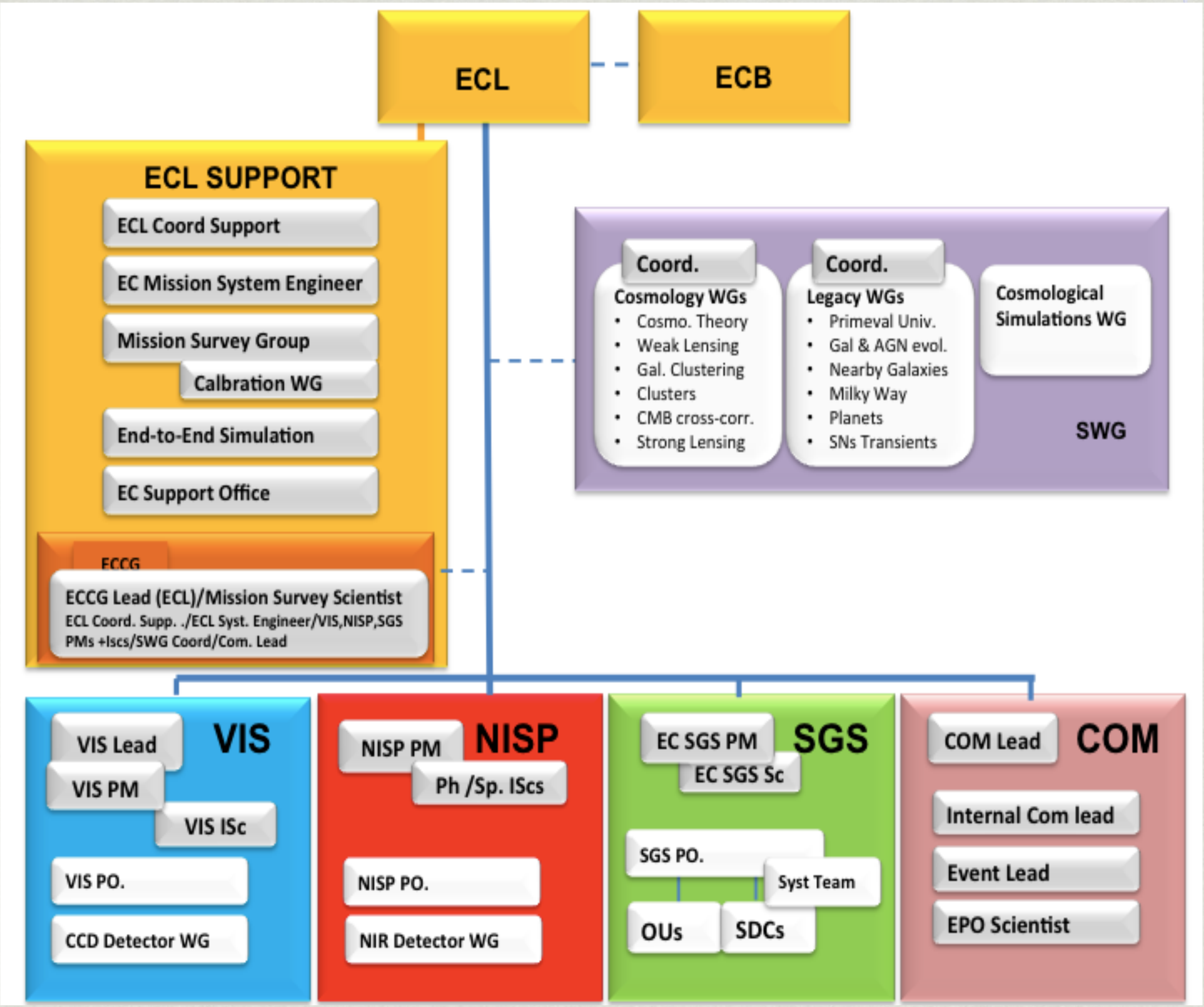
"Calibration dedicated pipelines" can be:

- specific processing functions running on dedicated calibration observations,
- Specific processing function running on standard science data,
- Standard processing functions running on dedicated calibration observations.

Automated production of calibration data products is a goal, automated usage of updated calibration data products is not!

Development organization for the SGS

The management structure: Euclid Consortium



Working structures within the EC-SGS

- For the data processing implementation phase, the EC-SGS consists mainly of two series of entities:
  - The Organization Units (OUs)
  - The Science Data Centers (SDCs)

The SDCs are built around existing national computing facilities, they gather IT support as well as developper expertise.
SDCs are in charge of the pipeline development (software & support architecture).
Production software will run in the SDCs.
The SDCs and SOC are the operational sites of the SGS.

The OUs group EC members according to their data processing expertise.
OUs are in charge of analyzing the science data processing requirements, and of producing the pre-integration version of the pipeline modules.
OUs will be in charge of future evolution of the pipeline during operations.

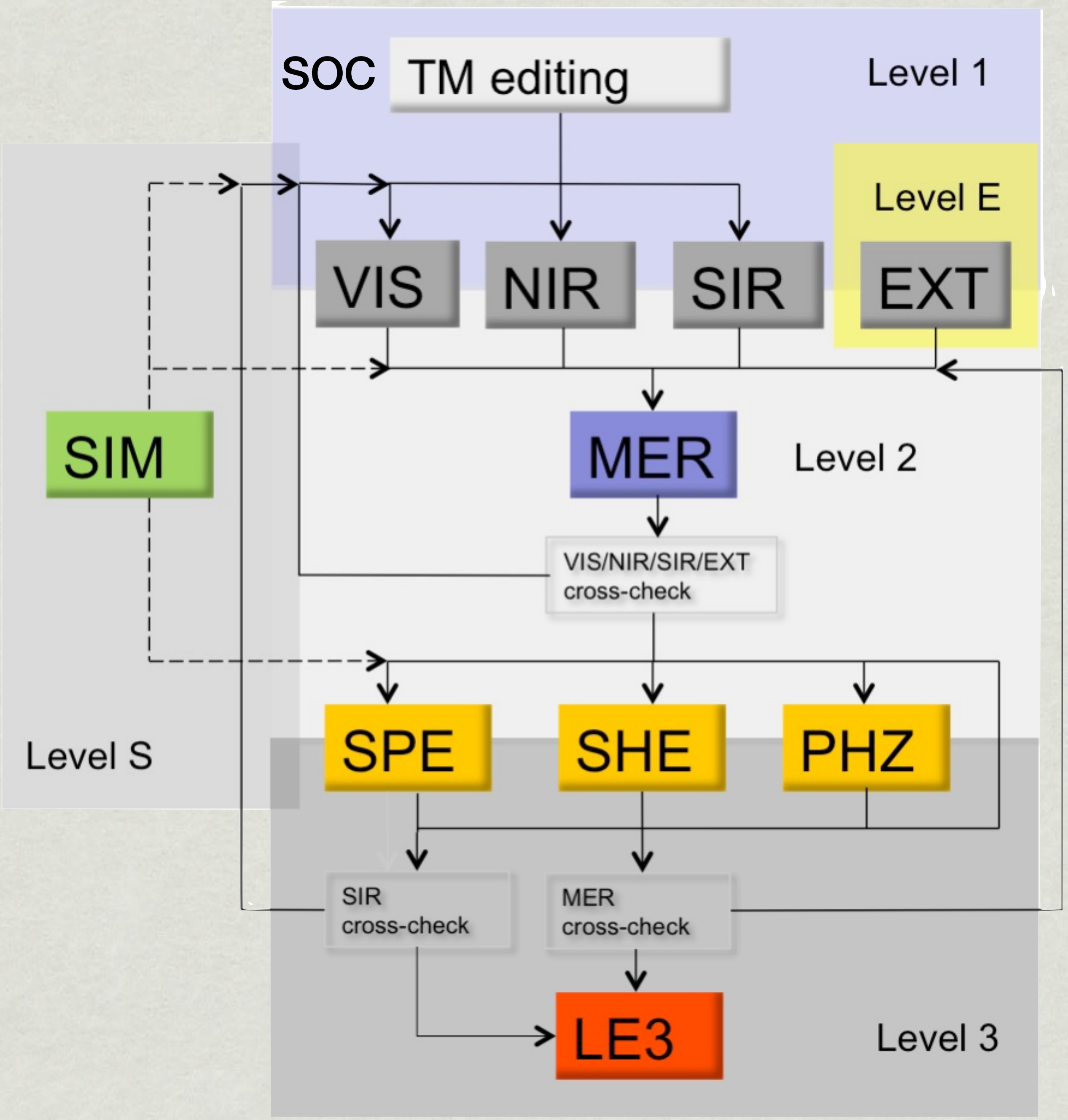
- **Production of the software to create calibration data products is also in the OUs+SDCs perimeter.**
- **OUs have been created following a plausible but a priori division of the pipeline in logical blocks.**

From OUs to Processing Functions (PF)

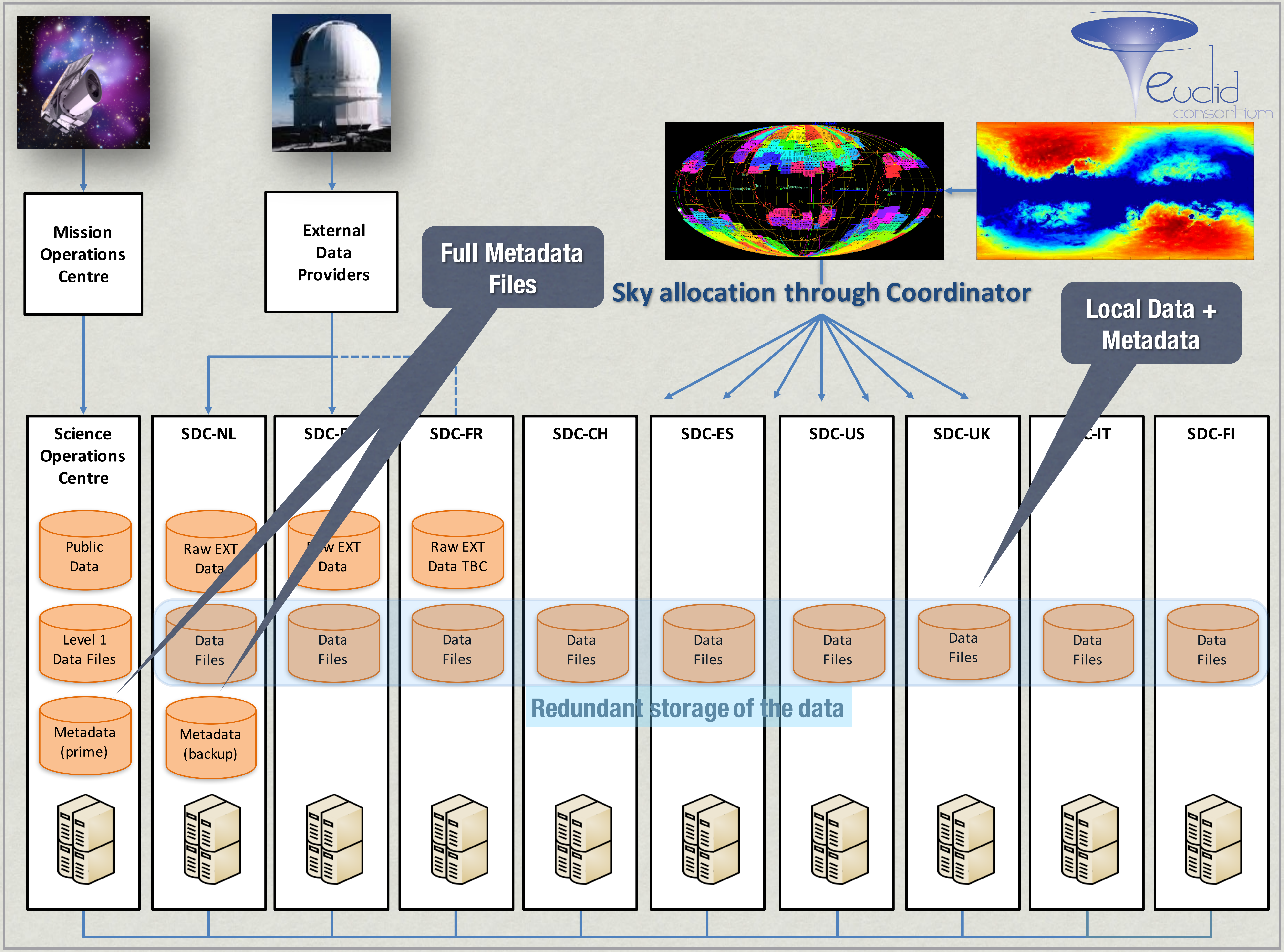
- Yet another acronym… but at attempt at clarification:
  - OUs and SDCs are groups and structures (sometimes physical structure). They collaborate to design and build a system, which is going to be the Euclid data processing pipeline and its infrastructure.
  - The pipeline elements, which are the result of the joint OU and SDC efforts are called the Processing Functions.
    - As we cannot fully redeem ourselves, the OU and the processing function it creates have the same name…

The Euclid data processing pipeline will be:
a series of Processing Functions,
designed by the OUs,
developed in collaboration between the OUs and SDC developers, integrated by the SDCs,
and running on the SDCs infrastructure.

The mapping of OUs on the pipeline



- VIS, NIR, EXT: production of fully calibrated photometric exposures from Euclid and ground-based surveys
- SIR: production of fully calibrated 1D spectra extracted from the NISP spectroscopic exposures.
- MER: production of a source catalog containing consistent photometric and spectroscopic measurements.
- PHZ: production of the photometric redshift for all catalogued sources.
- SPE: production of spectroscopic redshifts for all sources with spectra.
- SHE: measurements of galaxy shapes.
- LE3: production of all high-level science products.
- SIM: production of all the simulated data necessary to validate the data processing stages, and to calibrate observational or method biases.


Mission Operations Centre
External Data Providers
Full Metadata Files
Sky allocation through Coordinator
Local Data + Metadata
euclid consortium
Science Operations Centre
SDC-NL
SDC-
SDC-FR
SDC-CH
SDC-ES
SDC-US
SDC-UK
C-IT
SDC-FI
Public Data
Level 1 Data Files
Metadata (prime)
Raw EXT Data
Data Files
Metadata (backup)
w EXT Data
Raw EXT Data TBC
Data Files
Redundant storage of the data

The OU and SDC's work, now and in the future

- The **Organization Units** interact with the **Science Working Groups** to:
  - understand accept and possibly update the Data Processing Requirements.
  - Define the validation tests that will demonstrate that requirements are met.
- The **Organisation Units** and the **Science Data Centers** team up to produce the **processing functions** that will run on the SDC architecture.
  - This follows a development method and plan agreed between SDCs and OUs and validated at SGS level.
- The **OUs** interact with the **Calibration Working Group** and the **Instrument Development Teams** to integrate instrument characteristics appropriately in the data processing pipeline.

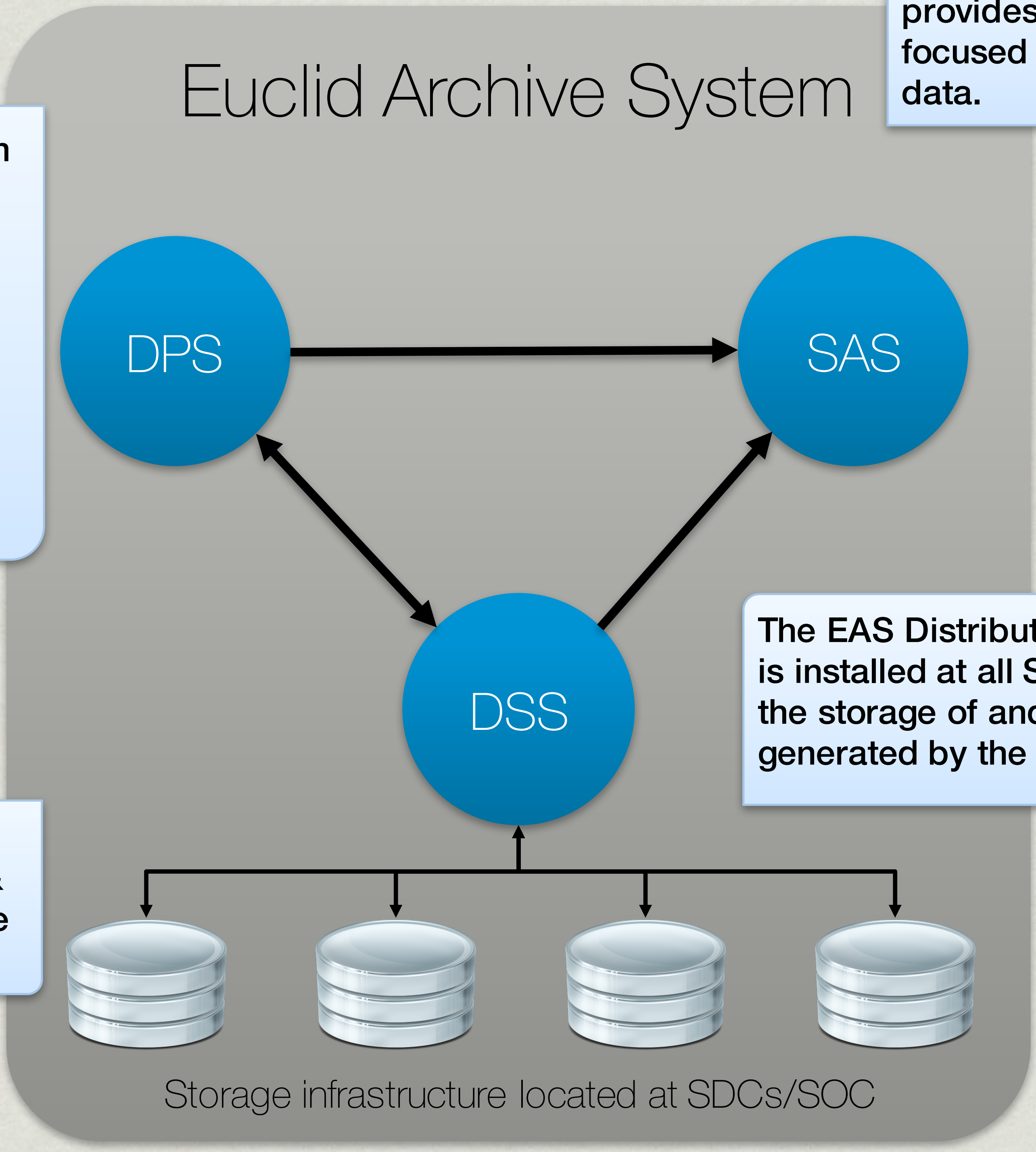
- **The SGS development plan leads to a release of a first operational pipeline in 2019.**
- **Following that release, the OUs and SDCs will continue working on the processing functions to fix, maintain, upgrade and expand them.**
- **The SGS development schedule is maintained by the Project Office on:**
  http://wiki.cosmos.esa.int/euclid/index.php/Euclid_SGS_Schedule

Euclid Archive System Overview

The EAS Data Processing System (DPS) stores the data products metadata including the locations of the data files.

It provides access to the data products to the EC members, including processing coordination, quality control and processing history tools.

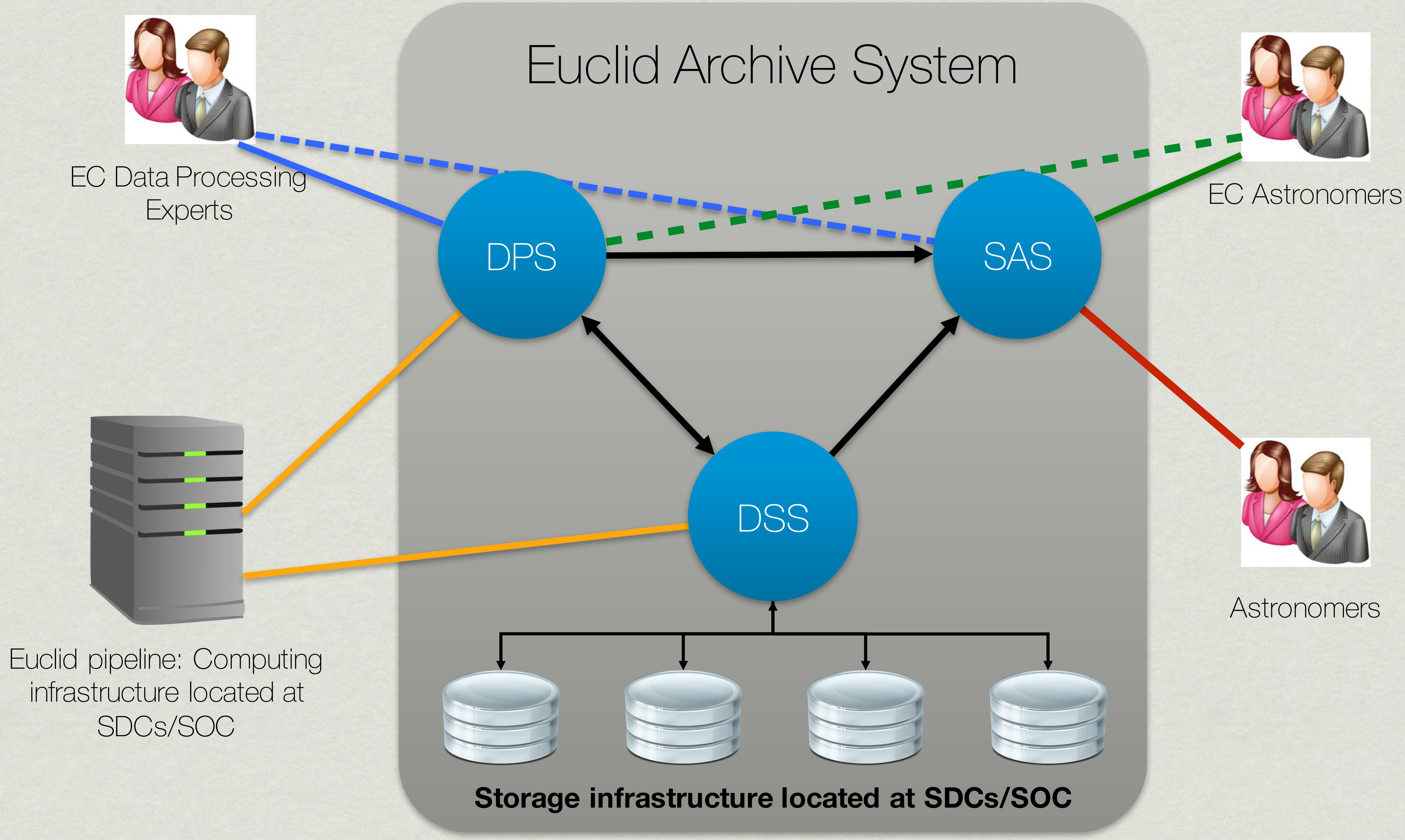
The EAS Science Archive System (SAS) provides access to the Euclid data focused on the scientific use of the data.



The EAS Distributed Storage System (DSS) is installed at all SDCs & SOC and manages the storage of and access to the data files generated by the pipelines.

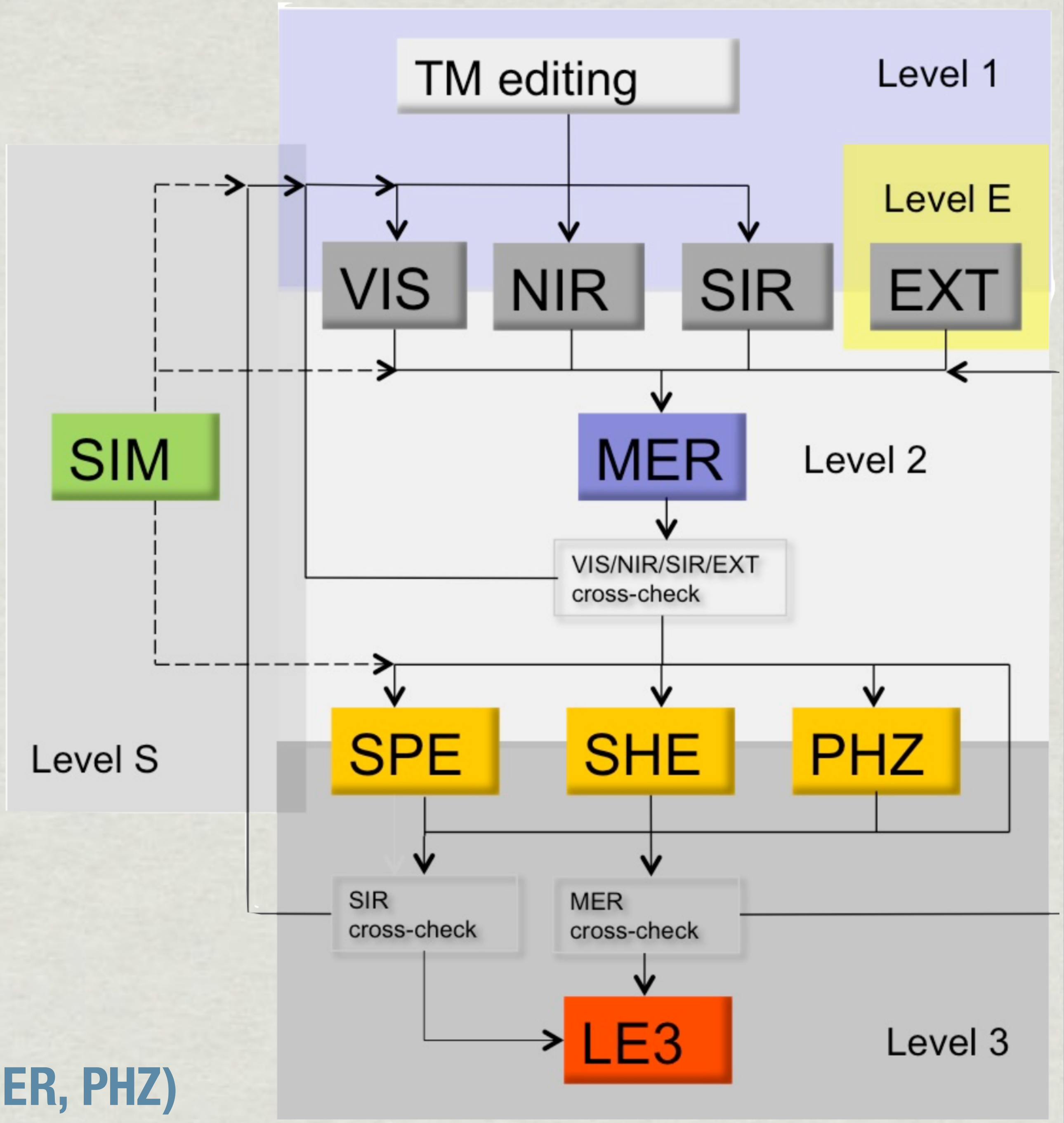
Storage systems are provided by all the SDCs & SOC. All public data will be located at ESAC.

Archive users



Purpose of this section

- Give an overview of what the science data processing pipeline will be.
- Provide the scope of each Processing Function within the global DP pipeline.
- Emphasize relevant calibration activities that we want to adress in this meeting.



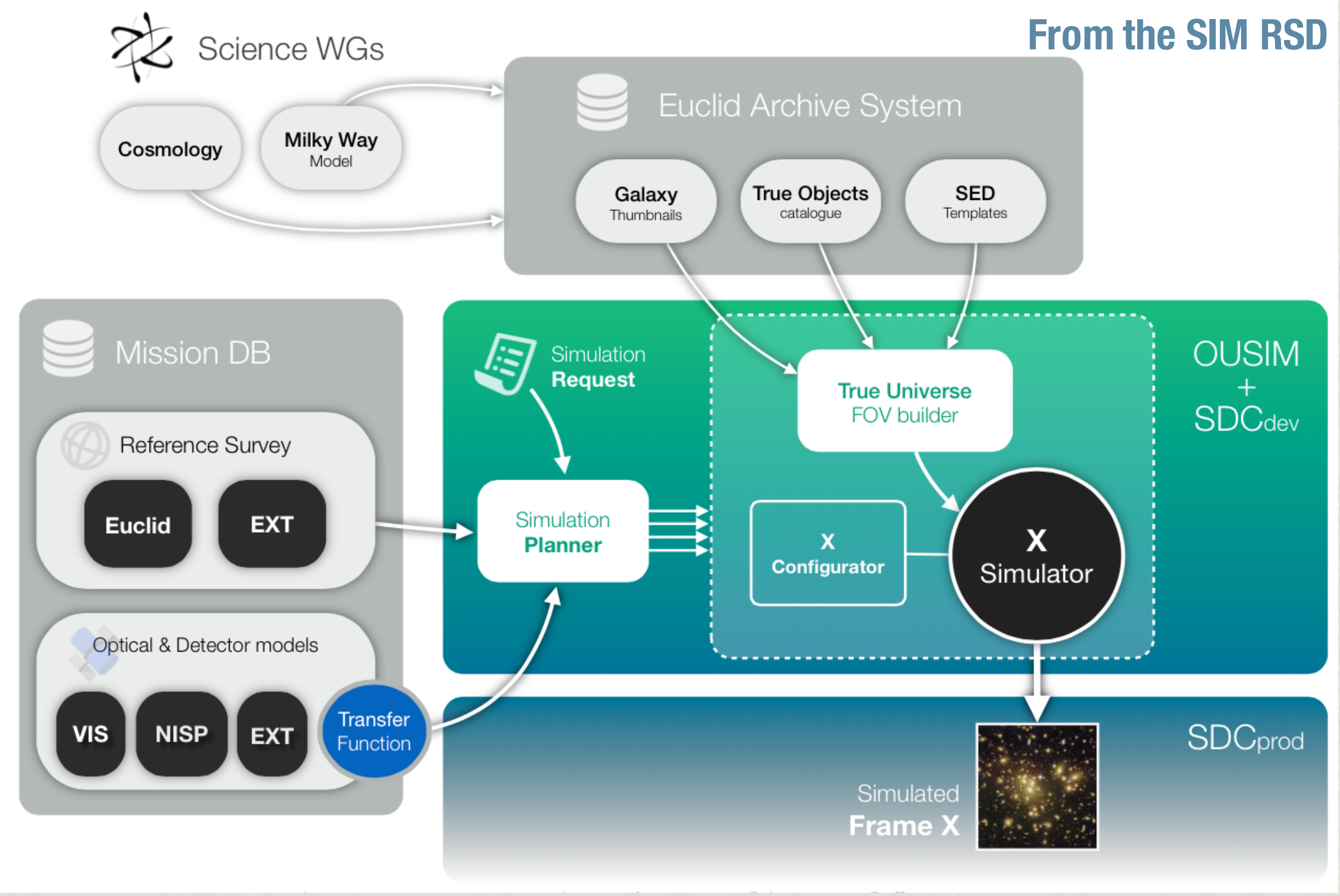
Structure for the following slides:

- **SIM for its pervasive role**
- **Imaging OUs (VIS, NIR, EXT)**
- **Spectroscopic OUs (SIR, SPE)**
- **Object photometric catalog and redshift (MER, PHZ)**
- **Shear Measurement (SHE)**
- **High level Science products (LE3)**

SIM - Provider of simulated data

- The role of SIM is:
  - To develop the facilities needed to simulate survey data (space and ground-based) at various stages of processing.
  - To deliver simulated data sets needed for the validation tests of most PFs.
  - To deliver simulated data to calibrate biases and artefacts at critical steps of the pipeline (e.g. Shear measurements).

SIM does not develop cosmology simulations, but has to pass content request from PF validation needs to the CSWG.



Imaging calibration processing functions

- We have three main branches of imaging data:
    - VIS for the calibration of images obtained with the VIS instrument (single wide filter).
    - NIR for the calibration of images obtained with the photometric filters (3) of the NISP.
    - EXT for the integration of visible ground based-surveys covering the VIS band in ~4 narrower bands.
- We are will use GAIA as the reference for astrometry, possibly enhanced if needed by VIS.
    - This may create a dependency, or a need for re-calibration before the catalog stage, between VIS and NIR+EXT.

Imaging calibration processing functions

VIS

- VIS images are used for shear measurements.This creates very strict requirements on the characterization of imaging quality.
- The VIS PSF is crucial for shear measurement, it likely has temporal and spatial dependencies. Modeling this PSF is an intrinsically complex activity and is a vast interface between VIS and SHE (joint activity).
- The PSF (possibly with relaxed constraints) also plays a part in astrometry and photometry (MER).

NIR

- The principal use of NIR images is to provide object magnitudes for PHZ.
- The same detector is at work for photometric and spectroscopic measurements with NISP, this creates a significant common branch between NIR and SIR.

EXT

- EXT supplies the spectral energy distribution for all objects detected by VIS (including stars).
- Challenges include: achieving photometric accuracies sometimes better than the original survey, masking the diversity of survey sources.
- EXT is also in charge of including in the Euclid system other external data (spectroscopic or photometric) for PHZ mostly.

Calibration issues for the Euclid imaging stage

- High dependency on detector and telescope specifics, need of strong interaction/preparation with the instrument development and operations teams.
- Definition of common standards (astrometric and photometric). This is simplified by the exact simultaneity of the two surveys (VIS and NISP).
- PSF modeling, which is a calibration activity for later stages of the pipeline, is a highly convoluted activity.

Calibration issues for the external imaging data

- External imaging dataset is highly inhomogeneous:
  - Different telescopes and observing set-ups
  - Ground-based observing
  - Possibly very long time span of observations
  - Very high requirement on photometric homogeneity

- Definition of the EXT processing function is complex:
  - We are working with a variety of organizations (from full-blown collaborations to actual partnerships in data acquisition).
  - Euclid photometric calibration requirements can be different and are probably expressed using different metrics.
  - Distribution of responsibilities between the external collaboration and the Euclid EXT OU needs to be defined on a case by case basis.

Current status of the EXT photometric data

- KIDS+VIKING
- DES

Initial southern hemisphere coverage, possibly not deep enough but providing key test data for EXT organization.
Euclid has integrated KIDS, VIKING and DES expertise as well as all data (including calibration data).

- CIFS (r, u-light, 5000 sq. degree) ⟵ **CFIS has integrated all Euclid calibration requirements. We need to build the right interface for coordination.**
- CEFCA – JT250 (g, 5000 sq. degree) ⟵ **Euclid will collaborate with the local team to build the instrument, execute the survey, process the data.**
- *Subaru/HSC (r, i, z, 2000 sq. degree) proposal submitted* ⟵ **Japanese-led proposal with a few EC member, possibly part of a long term agreement with the Japanese Cosmo community for collaboration on Euclid.**
- *LSST – under negotiations* ⟵ **Discussions re-started on practical basis this summer. Made complex by the facts that LSST has similar science objectives and a similar timeline**

Spectroscopic processing functions

- The spectroscopic analysis flow is performed by the SIR and SPE processing functions.
  - The interface between the two functions is placed at the position where 1D spectra have been extracted, calibrated (wavelength and flux) and optimally combined for each object detected on each NISP grism exposure.

SIR

- Because of the grisms, SIR processing must take into account objects located outside of the instantaneous field of view. This is an interface with the imaging side.
- Optimal extraction of the 1D spectra depends on the object shape and colors, introducing a dependency on later stages of the pipeline.

SPE

- SPE has to detect and identify lines in spectra that have a narrow range (in general only one line will be there), and where contamination by neighboring source may be high (3 orientations of the grism now).
- Objects' color or photometric redshift will be needed to ascertain the line identification. This is a dependency and an interface with MER and PHZ.

Calibration issues with Euclid spectroscopy

- Astrometric and focal plane geometry are key to locate the source spectra (based on photometric position).
- Wavelength calibration require spectroscopic standards (dedicated observations), and the very large field of NISP means this calibration procedure can be long or complex.
- Photometric calibration of the spectra is required as sources are selected on the basis of their Hα flux (calibration accuracy impact the knowledge of the selection function for GC studies).

Photometric catalog and redshifts (MER, PHZ)

- The Euclid Photometric Catalog is the main science product of Euclid.
    - It contains all the information for each object detected in any band of the survey.
    - While it is built from the imaging surveys (VIS, NIR, EXT), it will contain the spectra from SIR/SPE, redshifts from by SPE and PHZ, ellipticities by SHE, and physical parameters derived during the PHZ PF execution.
    - It is the main legacy product and may contain much more than one type of magnitude measurement per object and band (as prescribed in the Legacy Requirements Doc.).
    - A strong interface exists with PHZ as the type of magnitude impacts the performance of photometric redshift algorithms.
- Building the photometric catalog is the task of MER. It faces a number of specific challenges:
    - Dealing with surveys with significantly different PSFs. Though the constraints on PSF modeling are likely weaker when dealing with detection and photometry than when measuring shape, MER needs to handle the diversity and variability.
    - Defining the detection strategy and making sure it satisfies the constraints set by PFs operating on the catalog (i.e. object density, object deblending).
    - The MER PF execution may be split into different passes, e.g. if shape-based photometry is needed by PHZ

Photometric catalog and redshifts (MER, PHZ)

- Photometric redshifts are needed in the context of weak-lensing tomographic studies.
- As a legacy product, PHZ will produce physical parameters for all the sources in the catalog (as a result of template-fitting algorithms).
- The challenge of PHZ is not to produce photometric redshift codes (a number already exists) but to achieve the extremely strict requirements on:
  - errors and catastrophic failure rate.
  - bias on the mean redshift for color-selected samples.
- How to achieve the required level on bias is the subject of significant research and testing within PHZ in collaboration with SWG members.
  - External spectroscopic data sets are going to be needed.

Calibration issues and external data

- In principle, the cataloguing stage (MER) works with photometrically and astrometrically calibrated data from the “3” imaging sources.
- However it is the hub where all calibration efforts have to reveal their consistency, the MER is likely the best OU to coordinate the photometric calibration strategy.
- At PHZ level, calibration is more an activity applied to the algorithms, in the form of training leaning-based method with the help of a truth data set, hence the need for external spectroscopic data to provide measured redshifts to objects observed in photometry.
- External data is also required by PHZ to fully map the (colors, mean photo-z) space, in order to comply with one of its requirement.

Shear measurements

- The two main tasks of SHE are two challenges:
  - Model the VIS PSF, including “all” its spectral, spatial and temporal dependencies, so that these can be decoupled from the ellipticity of the source.
  - Select/Design a shape measurement algorithm whose biases are known and can be characterized with a minimum set of data.

PSF model:

- **Integrate as much knowledge on the instrument as built and characterized on the ground.**
- **Relies heavily on simulations produced in the SIM framework.**
- **Defines a strong interface with VIS and EXT/MER if only to identify the stars from which the model will be derived**
- **Possibly complemented with dedicated on-board calibration measurement.**
- **Using dedicated pointings as well as the science data.**

Shear measurement algorithms:

- **Identify the different branches of methods that can be implemented.**
- **Quantify the amount of calibration data that needs to be available for each method to be “bias-corrected”.**
- **Quantify the robustness of each method (i.e. its reliability when far from its calibration base).**
- **This implies significant testing on simulated/emulated data, and careful validation/test on the real data during the survey.**
- **External (HST) data to calibrate method biases.**

High-level science products from LE3

- LE3 is by far the largest Processing Function of the Euclid pipeline, and is in fact not well described by the term processing function:
- LE3 is a collection of semi-independent processing functions that each create one of the science products (SP) that is required to achieve Euclid's science goals.

6 PFs

S-Redshift catalog SP

- Photometric visibility mask (for WL)
- Spectroscopic visibility mask (for GC)
- Purity and completeness

4 PFs

Masks & Quality SP

8 PFs

Shear measurements SP

- Real-space tomographic s-s, s-p, p-p 2-point correlation function
- Fourier and real-space tomographic s-s, s-p, p-p 2nd order statistics
- Covariance matrices
- 2 and 3D mass-maps

Legacy SP

13 PFs

15 PFs

Galaxy clusters SP

- cluster catalogs
- Covariance matrices

A field for calibration – the Deep Fields

- This goes beyond calibration as foreseen in this meeting, and could be seen more as validation of the science results.
- The deep field observations are motivated essentially by the need to understand how the sensitivity limits of the wide survey impact the science results:
    - Calibration of noise bias, undetected sources in the noise affecting the shape measurement of brighter sources.
    - Computation of the actual completeness and purity of the spectroscopic samples.

Thank you!