*Astronomical Data Analysis Software and Systems: XXIV*
*ASP Conference Series, Vol. 495*
*A. R. Taylor and E. Rosolowsky, eds.*


# Organization of the Euclid Data Processing: Dealing with Complexity

Fabio Pasian,[1] Christophe Dabin,[2] Marc Sauvage,[3] Oriana Mansutti,[1] Claudio Vuerli,[1] and Anna Gregorio[4,1]

[1]*INAF-OATs, Via Tiepolo 11, I-34143 Trieste, Italy*

[2]*CNES, 18 Av. Edouard Belin, F-31401 Toulouse Cedex 9, France*

[3]*CEA, Orme des Merisiers, F-91191 Gif-sur-Yvette, France*

[4]*University of Trieste, P.le Europa 1, I-34128 Trieste, Italy*

**Abstract.** The data processing development and operations for the Euclid mission (part of the ESA Cosmic Vision 2015-2025 Plan) is distributed within a Consortium composed of 14 countries and 1300+ persons: this imposes a high degree of complexity to the design and implementation of the data processing facilities. The focus of this paper is on the efforts to define an organisational structure capable of handling in manageable terms such a complexity.

## 1. The Euclid Mission

Euclid is the second medium-sized (M2) mission of the ESA Cosmic Vision 2015-2025 Plan, aimed at understanding the nature of dark energy and dark matter by accurately measuring the accelerated expansion of the Universe. The launch is planned in Q1 of 2020. The payload is constituted by a 1.2 m telescope and two instruments: a photometer in the visible domain (VIS), and a photometer/spectrometer in near infrared (NISP). The spacecraft operates in L2, the second Sun-Earth Lagrange point. The Euclid survey will nominally last 6 years: the extragalactic survey will cover 15,000 square degrees and around $1.5.10^{10}$ galaxies, the deep survey will cover 40 square degrees and about 10,000 galaxies. It is to be noted that the broad-band Euclid imaging data alone are not sufficient to achieve the required photometric redshift accuracy and precision, which means that additional ground-based data are required.

Details on Euclid, its instruments (imaging and spectral) and the survey are available in the Euclid Definition Study Report and in Laureijs et al. (2013).

## 2. The Euclid Ground Segment – Complexity

The spacecraft will be connected during periods of 4 hours each to one or two Ground Stations. The **MOC** *(Mission Operation Center)* monitors the spacecraft health and safety and the instrument safety, controls the spacecraft attitude, and handles telemetry and telecommands for spacecraft and instruments. MOC and Ground Station form the Mission Operations Ground Segment (MOGS) which is completely under ESA control.

The Science Ground Segment (SGS), as already described in Pasian et al. (2012) and Pasian et al. (2013), is a federation of the **SOC** (*Science Operation Center*), run by ESA and acting as the single interface to MOC, and a set of national **SDCs** (*Science Data Centers*), nine of which are currently established (Finland, France, Germany, Italy, Netherlands, Spain, Switzerland, UK, USA), with more expected to join in the future. The SDCs are part of the *Euclid Mission Consortium* (**EC**); due to the heavy processing necessary for Euclid they are often located for operations in general-purpose data centres featuring inhomogeneous hardware and software environments.

From the point of view of the organisation of work, an *ECSGS Project Office* (**PO**) has been created to coordinate SGS activity within the Euclid Consortium. The PO and the SOC have developed, and are committed to maintain, a tight collaboration to design and develop a single, truly integrated SGS. The EC organization is based on the decomposition in trans-national *Organization Units* (**OUs**), covering most of the science-related processing. Each OU produces algorithms which are integrated and executed in the national SDCs. SDCs perform both software development (*SDC-DEV*) and an operational data production task (*SDC-PROD*). The *SGS System Team* (**ST**) provides support and tools for the whole of the SGS (SOC and ECSGS). The *Euclid Archive System (EAS)* is built jointly by EC and SOC, and is managed by SOC. There are internal and public EAS functions: the latter allow access to a subset of the EA, corresponding to the data that will be accessible to the scientific community. *Science Working Groups* (**SWGs**) are external to the SGS: they turn science objectives into requirements placed on the pipeline products and performances, and verify that the requirements are met (e.g., define validation procedures).

## 3. Dealing with Complexity

### 3.1. Processing Functions

Dealing with this complexity requires the need to concentrate on the *products* Euclid needs to provide. *Processing Functions* (**PFs**) are the main product of the Euclid SGS, to be delivered to ESA at the end of the mission. They can be summarised as follows: **LE1** provides telemetry unpacking and decompression (edited telemetry), plus Level 1 (raw) VIS and NISP images; **VIS** is in charge of processing the Visible imaging data from Level 1 to Level 2, i.e., it produces fully calibrated images; **NIR** is in charge of processing the Near-Infrared imaging data from Level 1 to Level 2, i.e., it produces fully calibrated images; **SIR** is in charge of processing the Near-Infrared imaging data from Level 1 to Level 2, i.e., it produces fully calibrated spectral images and extracts the spectra in the slitless spectroscopic frames taken by NISP; **EXT** is in charge of entering in the Euclid Archive all of the external data that will be needed to achieve the Euclid science results, this is essentially multi-wavelength data for photo-z estimation, but also spectroscopic data to validate the spectrometric redshift measurement tools; **MER** realises the merging of all the Level 2 information; it is in charge of providing stacked images, source catalogues and calibrated photo-z's where all the multi-wavelength data (photometric and spectroscopic) are aggregated; **SPE** extracts spectroscopic redshifts from the Level 2 spectra; **PHZ** computes photometric redshifts from the multi-wavelength imaging data; **SHE** computes shape measurements on the visible imaging data; **LE3** is in charge of computing all the high-level science data products from the fully processed shape and redshift measurements (and any other possibly needed Euclid data). **SIM** is in charge of producing all the simulations needed.

The Processing Functions correspond to the processing steps, which are performed within an "Euclid pipeline," are algorithmically devised by the relevant OU and engineered by software development teams (SDC-DEV) and can, in principle, be run yielding the same results on any SDC site of the SGS (SDC-PROD, having different hardware environments). Since in most cases Processing Functions are developed jointly by OU members and their local SDC-DEV teams, formal OU-SDC interfaces are not needed in most cases, and it is easier to develop directly pipeline-quality code. The SGS System Team provides tools, standards and support to the code development. It is important to note that SDC Leads are members of the ST, and this simplifies the flow of information on the SGS architecture to the code developers.

### 3.2. Development, Verification and Validation

Science Working Groups (SWGs), Organisation Units (OUs) and Science Data Centres (SDCs) all have a role to play in the implementation of the Euclid SGS. But this distinction is not to be interpreted strictly. As a matter of fact, it is to be noted that individual Euclid scientists may belong to more than one of the above groups. This has in important consequence in the arrangements made to avoid over-formalisation of SWG-OU-SDC interfaces, as shown and explained in Fig. 1.

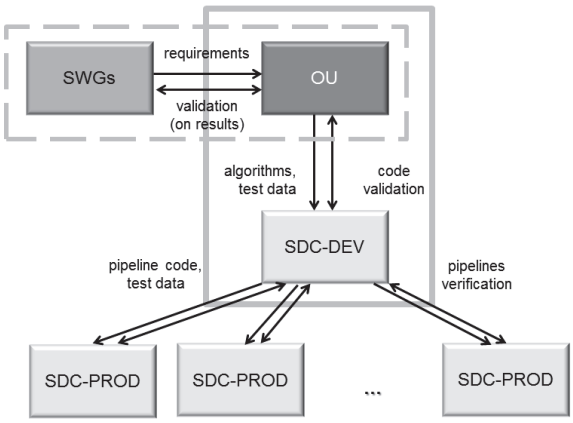


Figure 1. Relationships between SWGs, OUs and SDCs, related to each individual Processing Function. In most cases, no interfaces will exist between OU and SDC-DEV, but rather a joint development will take place (solid box). On the other hand, interactions between OU and SWG will occur only for validation of results against high-level requirements (dashed box).

A set of documents is being prepared jointly between OUs and SDCs (by product – Processing Function – and not by organisation): a Requirements Specification Document and a Validation Plan for every PF, plus a set of Development Plans organised by SDC. The validation by SWGs of the high-level data processing requirements works as follows: the various requirements contained in the high-level Data Processing Requirements Document (GDPRD) are attributed to the PFs. At this stage, the SGS will be considered as validated if every GDPRD requirement is validated. The SGS is including in the top-level IV&V Plan the inputs provided by the SWG coordinators regarding

the principles of validation, as well as the recommendations and typologies of Validation test – this top-level document will be co-signed by SGS and SWG coordinators.

### 3.3. Standards and Guidelines

Standards and guidelines help developers in taking the right decisions: e.g., by showing how/where to improve the code to meet the demanding requirements of the Euclid data processing, by encouraging the use of best practices and by providing tools to help developers improving their code.

The SGS uses a single development platform specifying operating system, programming language and support libraries. CODEEN is the Euclid collaborative development and continuous integration platform. It is important to define this environment early, since the cost of fixing bugs increases as the system integration approaches completion. Its usage is therefore mandatory for the main processing software.

Python and C++ have been adopted as the allowed languages for pipeline development, the drivers being an increased flexibility about who can contribute to development, and the long-term direction of astronomical programming.

An explicit Data Model (DM) is being built by the OUs to describe the output of their processing functions (therefore input to other PFs in most cases). DM Workshops have been held with wide participation from OUs and System Team. The first iterations of the DM seem very promising, since real data products are starting to be defined. The challenge now is to increase the coverage to all products and maintain a flexible process to allow the DM to evolve in a controlled way along with the PFs.

Thorough testing of the various pipeline "models" is made by means of "challenges" to verify if the planned architecture can be practically deployed.

## 4. Conclusions

The planning, development and operations of the Euclid Science Ground Segment are a big challenge. From the organisational point of view, the ESA-led SOC and the ECSGS Project Office provide management and control, acting in full coordination. The Euclid SGS System Team (composed of ICT experts from both SOC and the EC, and of the SDC Managers), through the work of several active working groups, deals mainly with architecture principles, logical architecture and technology watch. This activity allows to prepare standards, guidelines and tools for the code developers.

**Acknowledgments.** The authors are indebted to G.Racca, R.Laureijs, Y.Mellier, who lead different aspects of the project, to J.Hoar and G.Buenadicha of SOC, and to all the individuals participating in the SGS development within ESA and the EC, too many to be listed here. The participation in the Euclid phases B2/C is being supported by the National Space Agencies, in particular in Italy by the ASI contract I/023/12/0.

### References

Laureijs, R., et al. 2013, in ADASS XXIII, edited by N. Manset, & P. Forshay (San Francisco: ASP), vol. 485 of ASP Conf. Ser., 495

Pasian, F., et al. 2012, in Software and Cyberinfrastructure for Astronomy II (SPIE), vol. 8451 of Proceedings of the SPIE, 195

— 2013, in ADASS XXIII, edited by N. Manset, & P. Forshay (San Francisco: ASP), vol. 485 of ASP Conf. Ser., 505