Astronomical Data Analysis Software and Systems XXVI
ASP Conference Series, Vol. 521
Marco Molinaro, Keith Shortridge, and Fabio Pasian, eds.

The Design Strategy of Scientific Data Quality Control Software for Euclid Mission.

Massimo Brescia,[1] Stefano Cavuoti,[1] Terje Fredvik,[2]
Stein Vidar Hagfors Haugan,[2] Ghassem Gozaliasl,[3,6] Charles Kirkpatrick,[3]
Hannu Kurki-Suonio,[3] Giuseppe Longo,[4] Kari Nilsson,[5] Martin Wiesmann[2]

[1]*INAF OACN - Astronomical Observatory of Capodimonte, Naples, Italy;*
brescia@na.astro.it

[2]*Institute of Theoretical Astrophysics, University of Oslo, Norway*

[3] *Dept. of Physics, University of Helsinki, Finland*

[4] *Dept. of Physics University of Naples Federico II, Naples, Italy*

[5] *Finnish Centre for Astronomy with ESO (FINCA), University of Turku, Finland*

[6] *Helsinki Institute of Physics, University of Helsinki, Finland*

Abstract. The most valuable asset of a space mission like Euclid are the data. Due to their huge volume, the automatic quality control becomes a crucial aspect over the entire lifetime of the experiment. Here we focus on the design strategy for the Science Ground Segment (SGS) Data Quality Common Tools (DQCT), which has the main role to provide software solutions to gather, evaluate, and record quality information about the raw and derived data products from a primarily scientific perspective. The stakeholders for this system include Consortium scientists, users of the science data, and the ground segment data management system itself. The SGS DQCT will provide a quantitative basis for evaluating the application of reduction and calibration reference data (flat-fields, linearity correction, reference catalogs, etc.), as well as diagnostic tools for quality parameters, flags, trend analysis diagrams and any other metadata parameter produced by the pipeline, collected in incremental quality reports specific to each data level and stored on the Euclid Archive during pipeline processing. In a large programme like Euclid, it is prohibitively expensive to process large amount of data at the pixel level just for the purpose of quality evaluation. Thus, all measures of quality at the pixel level are implemented in the individual pipeline stages, and passed along as metadata in the production. In this sense most of the tasks related to science data quality are delegated to the pipeline stages, even though the responsibility for science data quality is managed at a higher level. The DQCT subsystem of the SGS is currently under development, but its path to full realization will likely be different than that of other subsystem; primarily because, due to a high level of parallelism and to the wide pipeline processing redundancy (for instance the mechanism of double Science Data Center for each processing function) the data quality tools have not only to be widely spread over all pipeline segments and data levels, but also to minimize the occurrences of potential diversity of solutions implemented for similar functions, ensuring the maximum of coherency and standardization for quality evaluation and reporting in the SGS.

1. Introduction

The Euclid Science Ground Segment (SGS, Laureijs et al. 2014) has the main role to provide the all the resources required to analyze the Euclid Data and to derive science data products. The Euclid SGS is in charge of the production of the science ready calibrated images and source catalogues, and all relevant quality control and meta-data required for the scientific exploitation of the Euclid mission. This, of course, includes the data from the two instrument channels of Euclid (VIS and NISP), as well as all complementary external data from other surveys.

From the data flow point of view, the SGS data processing is composed by four levels:

- Level 1: Unpacked and checked telemetry and housekeeping data;
- Level 2: Calibrated and intermediate data produced during the calibrations;
- Level 3: Final catalogues and pre-defined science data products (3D galaxy power spectra, dark matter power spectra, tomography, high order statistics, mass reconstruction map, photometric and spectroscopic redshift, etc.) but does not include data analysis beyond the production of catalogues and basic 2-point statistics or cosmological interpretation of data (joint analyses of data, dark energy studies, cosmological parameters, growth and growth rate of structures, galaxy biasing, test gravity, neutrino mass constraints, galaxy clustering, etc...).
- External Data: Euclid compliant quality-controlled data from existing surveys used for calibrations, photometric redshift, and simulation validations during all the operational phases.

The most valuable asset of the mission are the data, and due to their huge volume, quality control becomes a crucial aspect of all above items, not only for scientific data produced by the pipeline, but also for telemetry, diagnostics/monitoring/control, and calibration information. Data Quality (DQ) refers to the state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use. This means that all kinds of products (images, tables, text, etc.) have to be quality controlled by checking the right syntax of metadata and columns, detecting missing or out-of-bounds values and detecting any other inconsistencies. DQ common tools are foreseen to be available for all pipeline modules at all processing levels. They should be as general as possible, in order to be shared between different pipeline modules, thus ensuring a full homogeneity.

The DQCT is a system devoted to provide common solutions for the quality controls to be integrated within all the project pipelines. DQCT deals with information concerning quality flags, error estimates, statistics like mean, standard deviation, RMS, S/N, visual/graphical inspection products, such as thumbnails, trend analysis diagrams, histograms, scatter plots, etc., as well as any other metadata parameter produced by the pipeline, collected in quality reports specific to each data level and stored on Euclid archive during pipeline processing.

All the quality tools are based on specifications expected from the scientific organization units, and they would fulfill all common needs at all pipeline flow stages, according to standards defined among all the science teams.

DQCT APIs will also provide tools to collect quality reports at each transition between two data levels, taking also into account the traceability of the previous quality checks along the pipeline into account.

The DQCT team is composed by different Institutes from Finland, Italy and Norway, under Italian responsibility. During the prototyping and development phases, the produced packages will be tested and validated in collaboration with the Finnish Science Data Center. The DQCT will be developed by adopting the common development directives defined by the System Team in the areas Architecture & Design, Software Development Rules and Processes, Euclid Archive System, Monitoring & Control.

2. Design and Features

DQCT will provide a quantitative basis for evaluating the application of calibration reference data (flat-fields, linearity correction, reference catalogs, focal plane illumination model, photometric scale and zero-point, etc.).

DQCT will provide diagnostic tools that, among other things, will facilitate the diagnosis of problems with scientific performance and delivered image quality. Analysis of pipeline stage quality problems with DQCT will inform decisions taken by downstream pipelines and processes of whether to abort or otherwise alter their processing.

The DQCT subsystem is currently under development, but its path to full realization will likely be different in flavor than that of other subsystems. First, because there is a great deal to learn about Euclid, by reflecting historical patterns of prior challenging missions, the best strategies for data processing, new science demands and opportunities that will undoubtedly emerge during operations. Second, because the pipeline processing entails a wide redundancy (the mechanism of double Science Data Center for each processing function) and a high level of parallelism; for this reason, the data quality tools has not only to be widely spread over all pipeline segments and data levels, but also to minimize the occurrences of potential diversity of solutions implemented for same functions, ensuring the maximum of coherency and standardization for quality evaluation and reporting in the SGS.

There are many aspects of science data quality that are common among astronomical imaging surveys. Most of these tests can be automated, although in practice most prior surveys depended upon human visual inspection to confirm the computed quality metrics.

For instance, common elements of Science Data Quality Assessment for Surveys are: Image artifact flagging: Static bad pixels, cosmic rays, saturation, satellite trails, electronic cross-talk; Image background: Ghost images, scattered light, detector health, moonlight, fringing, sky glow; Delivered image quality: Size of the PSF, PSF shape (e.g., ellipticity), variation of the PSF with position in the focal plane; Astrometric fidelity and stability; Photometric fidelity and stability: Uniformity of photometric depth; dispersion about expected stellar locus in color magnitude diagrams and color-color plots.

During the instrument commissioning, followed by full-up science operations, the DQCTs will contribute to calibrate and assess both the stability and the repeatability of the science data within a variety of operating conditions; in particular, for DQCT functions there will be the possibility to correctly set thresholds for quality parameters and flags.

The DQCT system is based on three main tasks: (i) the automated computation and flagging of off-nominal conditions and quality attributes, by supporting the final assessment of science data quality; (ii) long-term trend analyses, enabling the optimization of data calibration, pipeline processing functions and investigation of quality anomalies; (iii) to avoid redundancy and inhomogeneous calculation of the same quantities.

The Euclid pipelines are required to report problems occurring during processing. At the end of SGS data flow the result of DQCTs consist of a quality report including all statistics, trend analysis plots, flags and parameters incrementally evaluated from progenitors to final science data.

Furthermore DQCT is responsible to produce all the data that has to be visualized in a tool that runs as part of the Euclid Archive System. This tool aims to provide a user-friendly method of quality information inspections for products and their progenitors. The tool is supposed to be a tailored version of QualityWise (McFarland et al. 2013).

3. Conclusions

Over time, the DQCT team will provide tools for advanced search, access, and analysis of archived data products.

For each pipeline of the Euclid SGS DQCT team is supposed to design a specific data model dedicated to list and describe all the parameters and quality flags required to perform the quality controls, as well as DQ functions and analyses, particularly suitable to monitor the status of observations. The final role of the DQCT team will be to provide an incremental quality report tools for advanced search, access, and analysis of archived data products.

Acknowledgments. MB and SC acknowledge financial contribution from the agreement ASI/INAF I/023/12/1. CK and GG were supported by the Academy of Finland grant 257989. KN was supported by Academy of Finland Grant 295114.

References

Laureijs, R., et al. 2014, in Space Telescopes and Instrumentation 2014: Optical, Infrared, and Millimeter Wave, vol. 9143 of Proceedings of the SPIE, 91430H

McFarland, J. P., Helmich, E. M., & Valentijn, E. A. 2013, Experimental Astronomy, 35, 79. URL http://dx.doi.org/10.1007/s10686-012-9296-z