The Euclid Archive System

Nieto, Sara; de Teodoro, Pilar; Salgado, Jesus; Altieri, Bruno; Buenadicha, Guillermo; Belikov, Andrey; Boxhoorn, Danny; McFarland, John; Valentijn, Edwin A.; Williams, O. R.

Published in:
Astronomical Data Analysis Software and Systems XXVI

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

The Euclid Archive System: A Data-Centric Approach to Big Data

S. Nieto,[3] A. N. Belikov,[1] O. R. Williams,[2] B. Altieri,[3] D. Boxhoorn,[1] G. Buenadicha,[3] B. Droge,[2] J. P. McFarland,[1] J. Salgado,[3] P. de Teodoro,[3] A. Tsyganov,[2] and E. A. Valentijn[1]

[1]*Kapteyn Astonomical Institute, University of Groningen, Groningen, The Netherlands*

[2]*Donald Smits Centre for Information Technology, University of Groningen, Groningen, The Netherlands*

[3]*European Space Astronomy Center, European Space Agency, Spain*

Abstract. We review the architectural design and implementation of the Euclid Archive System (EAS) which is in the core of the Euclid Science Ground System (SGS) and represents a new generation of data-centric scientific information systems. It will handle up to one hundred PBs of mission data in a heterogeneous storage environment and will allow intensive access both to the data and metadata produced during the mission. This paper makes a particular emphasis on the access to science-ready products and interfaces which will be provided for the end-user.

1. Introduction

Euclid is the ESA M2 mission (Laureijs et al. 2011) which will map the sky in a single optical band and three near-infrared bands. It will measure photometric and spectroscopic redshifts of galaxies to understand the properties and nature of dark matter and dark energy. Euclid will be launched at the end of 2020 and once at its nominal L2 orbit will start a 5 and a half year observing program to complete a wide survey (covering 15000 deg^2) and a deep survey (covering 40 deg^2 and 2 magnitudes deeper than the wide survey).

To achieve its scientific objectives the Euclid mission will combine space surveys with ground-based surveys which will boost the data volume produced by Euclid SGS up to 26PB per year and a catalogue of up to 10 billion objects (Pasian et al. 2014). To manage such an amount of information, the Euclid Archive System (EAS), described below, will handle data and metadata according to mission premises.

2. SGS AND EAS

The Euclid Science Ground Segment is a distributed data processing and data storage system, which is responsible for the delivery of the science-ready data to ESA. The SGS is formed by 9 national Science Data Centres (SDCs) and the Euclid Science Operations Centre (SOC). The task of the SGS is to process the data from ingested raw

frames to science-ready images, spectra and catalogs and deliver them to ESA (Pasian et al. 2014).

The data in the Euclid SGS is a combination of images, spectra and catalogs in FITS files and an extensive metadata description of these data products. In addition, the EAS will make ground-based surveys available to the SGS, possibly after additional reprocessing. Experience from previous missions (e.g. Astro-WISE (Begeman et al. 2013) and the LOFAR Long-Term Archive (Begeman et al. 2011)) shows that the volume of metadata will not exceed 5% of the total data volume.

3. EAS Design

According to requirements on the SGS and the EAS, the EAS design was established as a combination of 3 independent subsystems:

- The Data Processing System (DPS) consists of metadata storage and services which support the data processing inside the SGS;
- The Science Archive System (SAS) which is a gateway for end-users to Euclid data. It supports the scientific use-cases, the release delivery of data to the wide astronomical community, and long-term data preservation.
- The Distributed Storage System (DSS) consists of data file storage for both the DPS and SAS.

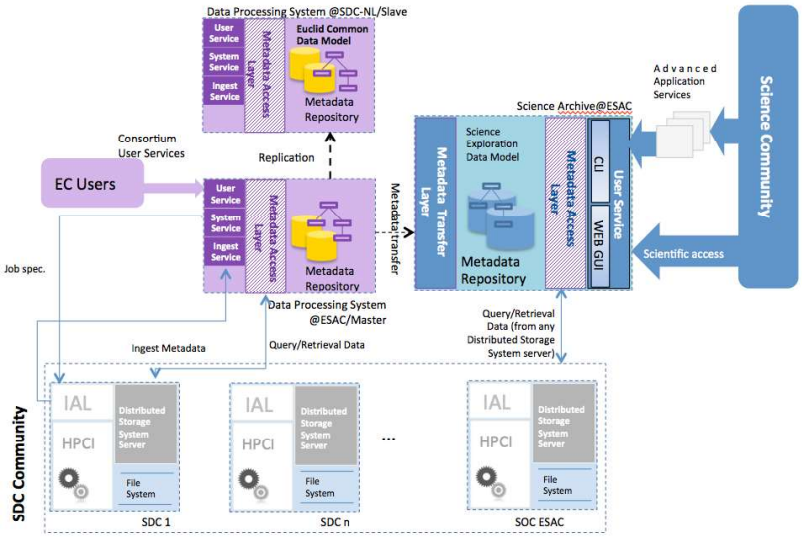

Figure 1. EAS Architecture

The DPS contains metadata for the data processed as well as operational and orchestration metadata. The main objective of the DPS is to provide data and metadata to other SGS operations and to trace any operation on the data inside the SGS. The DPS keeps the full data lineage for each data product. The DPS will assist in the preparation of Euclid data releases as a subset of the data processed by SGS. The DPS services

allow the retrieval of metadata in the form of XML files, the ingestion of data products, and the browsing of available data products through web applications.

The DSS is a distributed file storage for both the DPS and SAS and implements the distribution of data files according to the Euclid processing plan. The DSS consists of a grid of DSS servers which utilize a simple interface to the different storage solutions implemented by each SDC. Each DSS server communicates with all other DSS servers and allows a set of operations on files: ingestion, retrieval, copy to designated SDC and registration of the file in DPS metadata database.

4. Science Archive System

The SAS is part of the EAS and it aims to support the needs for scientific data exploration for the Euclid Consortium and the wider astronomical community. It will face the challenge of Big Data, as it will store a huge and increasing amount of scientific metadata and catalogues up to 10 billion galaxies. This will provide the worldwide astronomical community with an extremely large source of targets for future missions. Under these premises, the SAS will face the challenge of guaranteeing the long-term preservation of Euclid data while providing the scientific community with access to this data.

The SAS is being built at the ESAC Science Data Centre (ESDC), which is responsible for the development and maintenance of the scientific archives for the Astronomy, Planetary and Heliophysics missions of ESA. The SAS is focused on the needs of the scientific community. In this context, the SAS will provide access to the most valuable scientific metadata coming from the EAS-DPS through a set of public data releases. According to the policy of public releases defined by the Euclid Consortium, the plan is to deliver 3 data releases to the scientific community every 2 years after the nominal start of the mission.

The design of the SAS follows the latest generation of archives being developed by the ESDC, taking full advantage of the existing knowledge, expertise and code. The SAS will provide two ways of access through the Archive User Services (AUS): a web-based portal and a command line interface for programmable access. The archive web-portal is based on the Google Web Toolkit technology and the server side components are based on Java and integrate a set of standard VO protocols to manage requests from the users.

SAS implements the Science Exploitation Data Model (SEDM), which describes the scientific metadata. The SAS is populated through the Metadata Transfer Service (MTS) that acts as a transfer layer from the ECDM in DPS to the SEDM. This mechanism is the implementation of the public releases policy defined by the Euclid Consortium. The RDBMS supporting the SEDM is PostgreSQL, the same used for other scientific archives, (e.g. Gaia Archive).

The scientific requirements on the SAS cover three main areas: parametric search for metadata and catalogues, data retrieval and the visualization of images, spectra, etc. Regarding the visualization of maps, the current prototype is based on the technology developed for the ESA Sky (Arviset et al. 2016) that allows the exploration of the astronomical resources using a useful and intuitive web interface.

SAS will provide the tools and VO interfaces to enable the Software-to-Data Paradigm and "bring the software to the data" for the Euclid science. The distributed nature of the user community requires standard protocols oriented to access, exchange

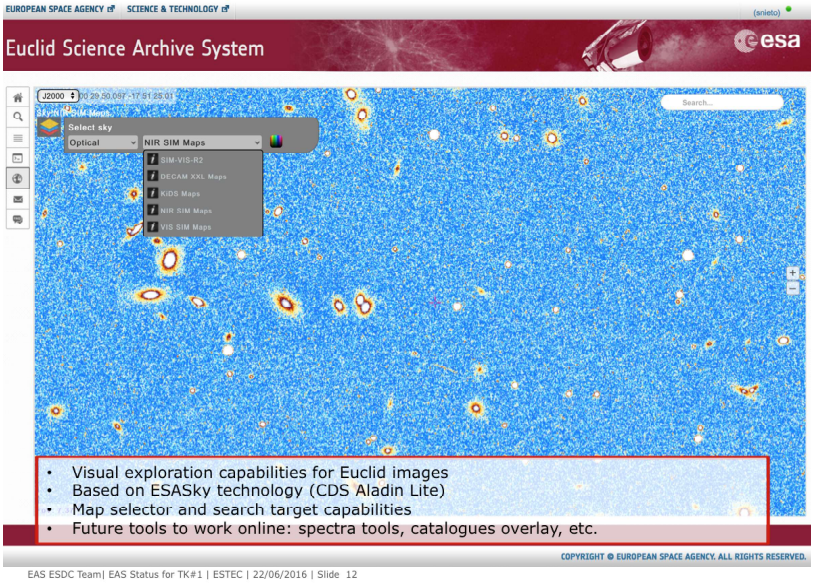

Figure 2. SAS Visualization Interface

and store data to guarantee the information accessibility (González-Núñez, J. and others 2017); Table Access Protocol (Dowler et al. 2010) to perform efficient parametric search and VOSpace (Graham et al. 2013), for distributed data storage are also part of the SAS infrastructure.

In the following years prior to the start of the mission the EAS will go through a number of crucial steps in its development including the final selection of an RDBMS and scaling up of the archive subsystems to successfully face the challenge of Big Data.

References

Arviset, C., et al. 2016, in Proceedings of the 2016 conference on Big Data from Space (BiDS 16), edited by P. Soille, & P. Marchetti, 9. 10.2788/854791

Begeman, K., et al. 2011, Future Generation Computer Systems, 27, 319

— 2013, Experimental Astronomy, 35, 1. 1208.0447

Dowler, P., Rixon, G., & Tody, D. 2010, Table Access Protocol Version 1.0, Tech. rep. URL http://www.ivoa.net/documents/TAP

González-Núñez, J. and others 2017, in ADASS XXV, edited by N. P. F. Lorente, K. Shortridge, & R. Wayth (San Francisco: ASP), vol. 512 of ASP Conf. Ser., 141

Graham, M., et al. 2013, IVOA recommendation: VOSpace specification v2.0, IVOA Recommendation 29 March 2013. 1509.06049

Laureijs, R., et al. 2011, arXiv. arxiv1110.3193

Pasian, F., et al. 2014, in ADASS XXIII, edited by N. Manset, & P. Forshay, vol. 485 of ASP Conf. Ser., 505