The Role of the Euclid Archive System in the Processing of Euclid and External Data

Williams, O. R.; Begeman, K.; Boxhoorn, D.; Droge, B.; Tsyganov, A.; McFarland, J. P.; Valentijn, E. A.; Vriend, W. J.; Dabin, C.

Published in:
Astronomical Data Analysis Software and Systems XXVI

Publication date:
2019

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

The Role of the Euclid Archive System in the Processing of Euclid and External Data

O. R. Williams,[1] K. Begeman,[2] A. N. Belikov,[2] D. Boxhoorn,[2] B. Droge,[2] A. Tsyganov,[1] J. P. McFarland,[2] E. A. Valentijn,[2] W.-J. Vriend,[2] and C. Dabin[3]

[1]*Centre for Information technology, University of Groningen, Groningen;* o.r.williams@rug.nl

[2]*Kapteyn Institute, University of Groningen, The Netherlands*

[3]*CNES, Toulouse, France*

Abstract.

Euclid is an ESA M2 mission which will create a 15,000 square degrees space-based survey: the Euclid Archive System (EAS) is a core element of the Science Ground Segment (SGS) of Euclid. The EAS follows a data-centric approach to data processing, whereby the Data Processing System (DPS) is responsible for the centralized metadata storage and the Distributed Storage System (DSS) supports the distributed storage of data files. The EAS-DPS implements the Euclid Common Data model and along with the EAS-DSS provides numerous services for Euclid Consortium users and SGS subsystems. In addition, the EAS-DPS assists in the preparation of Euclid data releases which are copied to the third EAS subsystem, the ESA developed Science Archive System (SAS) where they become available to the wider astronomical community.

The EAS-DPS implements the object-oriented Euclid Common Data Model using a relational DBMS for the storage. The EAS-DPS supports the tracing of the lineage of any data item in the system, provides services for the data quality assessment and the data processing orchestration.

The EAS-DSS is a distributed storage system which is based on a set of storage nodes located in each of the ten Science Data Centers of the Euclid SGS. The storage nodes supports a wide range of solutions from local disk, using a unix filesystem, to iRODS nodes or Grid storage elements. In this paper the architectural design of EAS-DPS and EAS-DSS are reviewed: the interaction between them and tests of the already implemented components are described.

1. Introduction

The Euclid mission will be a milestone in the understanding of the geometry of the Universe (Laureijs et al. 2011). The Euclid SGS and EAS have two main challenges during the data processing: firstly, the unprecedented accuracy which must be achieved in order to meet the scientific goals; secondly, the heavy dependence on the processing and reprocessing of ground-based data which will form the bulk of the stored data volume (Pasian et al. 2014). In total Euclid may produce up to 26 PB per year of data (Williams et al. 2014).

The EAS must provide to the Euclid SGS a distributed scientific information system, able to handle hundreds PBs of the data,together with tools to help in the assess-

ment of the data quality for each produced item. It must be possible to build the lineage of data products from the raw data to the science-ready images, spectra and catalogs.

2. EAS and SGS

The EAS is responsible both for the delivery of the science ready data to the astronomical community and the support of data processing during the mission. The responsibility for delivery of science ready data lies with the EAS Science Archive System and is outside the scope of this paper. The responsibility for data processing support lies within two components of the EAS: the EAS EAS-DPS and the EAS-DSS.

The EAS-DPS stores metadata related to the data being processed, the orchestration of the processing, the quality assessment of the data products and the preparation of data releases. The EAS-DSS stores the data files themselves, from raw frames to processed and calibrated images and spectra. Figures 1 shows an overview of the three components of the EAS and their principle interactions.

To implement the scientific requirement for traceability of the data and to enforce data lineage, all operations during the data processing must be reflected in the metadata and all data products necessary for the next step in the processing must be ingested into EAS. The EAS is thus not merely an archive for the storage of data, but rather an information system which can at any moment give to a scientist a detailed overview of the status of the data which has been processed or is being processed. This information includes full backward and forward lineage for each data item. Such lineage is crucial for the quality assessment of the data and also to prevent unnecessary reprocessing.

Euclid processes data in a distributed environment which consists of nine national Science Data Centers and the Science Operation Center. The EAS-DPS and EAS-DSS must allow access data from this distributed environment and guarantee data distribution according to the needs of the Euclid processing plan.

The design of the EAS-DPS and EAS-DSS draw on lessons learned from earlier archives for OmegaCAM (Begeman et al. 2013) and the LOFAR Long Term Archive (Begeman et al. 2011).

3. EAS-DPS Design

The binding between the different processing steps in the Euclid SGS is defined by the Euclid Common Data Model (ECDM). The ECDM describes not only input and output of each pipeline but also contains the processing and orchestration information for the Euclid SGS. The ECDM is based on the XML Schema Definition Language and forms an object-oriented data model. The EAS-DPS takes each stable release of ECDM and implements it by creating first Python stubs from each definition in ECDM and then generating a DDL schema to be created in the metadata database of the EAS-DPS.

The Python stubs, interfaces to the metadata database and the XML library together forms the Metadata Access Layer (MAL) of EAS-DPS. The task of the MAL is to hide the complexity of the metadata database implementation from the user and allow the user to interact with objects formed according to ECDM instead of the selection of rows in tables.

The EAS-DPS provides a number of services for other SGS components and users: the Consortium User Service allows users to browse the content of the meta-

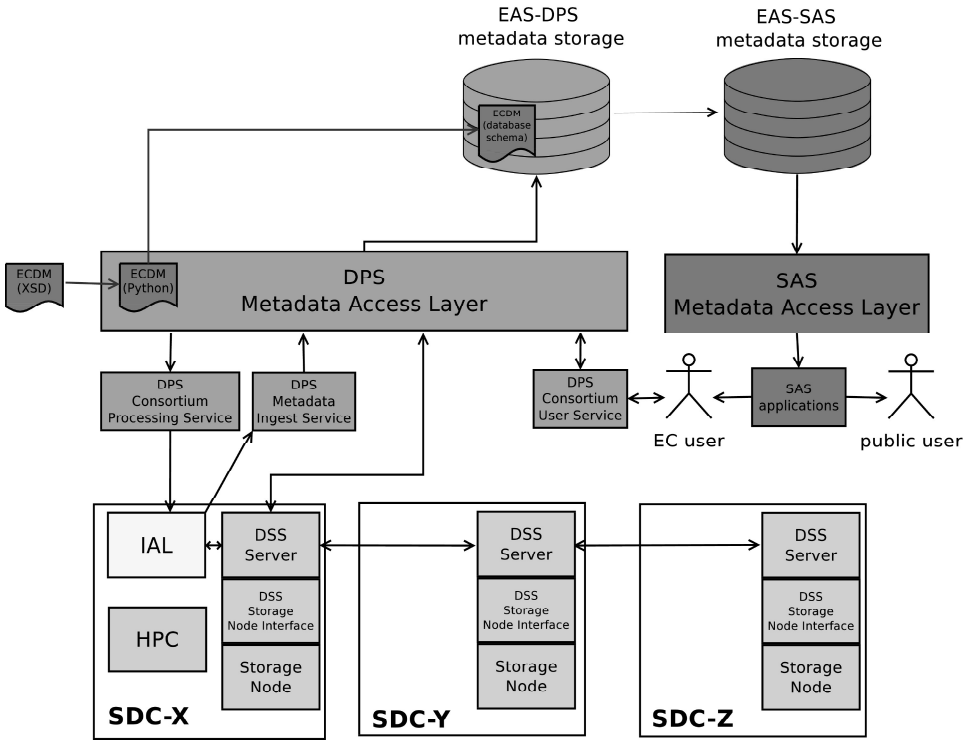

Figure 1. An Overview of the three components of the EAS. The design of the EAS-DPS and EAS-DSS are described in this paper.

data database in the web browser; the Consortium Processing Service is used by Infrastructure Abstraction Layer (IAL) to retrieve metadata from the EAS-DPS; finally, the Metadata Ingest Service which allows to transform XML to Python objects and commit them to the metadata database. All EAS-DPS services are based on the MAL.

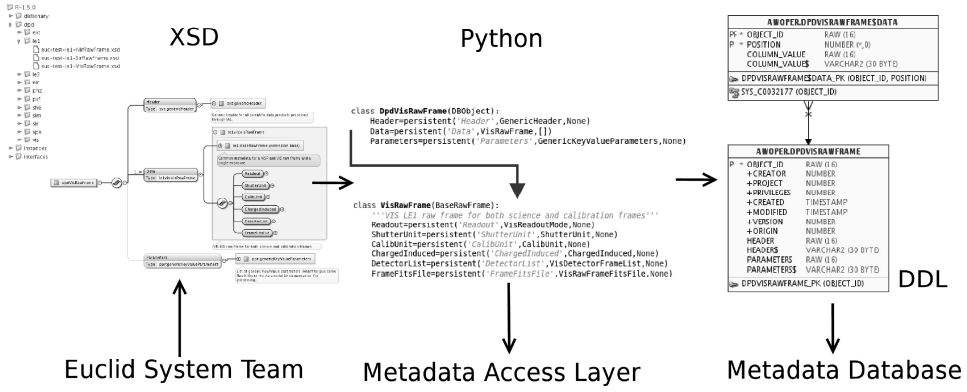

Figure 2. The Euclid Common Data Model. The model is initially defined by the Euclid System team in XSD. This is then used to generate the Python Code in the DPS Metadata Access layer. This Python Layer then generates the Data Defininion Language, used to define the contents of the Metadata Storage

4. EAS-DSS Design

The EAS-DSS serves both the EAS-DPS and the EAS-SAS as a common distributed file storage solution.

The EAS-DSS is a data storage grid with a single https-based user interface. The DSS servers act as an interface to the non-homogeneous data storage solutions deployed in the SDCs to the SGS and its users. Currently a DSS server can be deployed on top of a local POSIX filesystem, an iRODS server, a sftp server, a Grid storage element or an Astro-WISE dataserver. At least one DSS server is installed at each SDC and stores the data files processed or created during the running of a pipeline at this SDC.

To cope with the data volume of Euclid mission the data processing orchestration minimizes data file transfer between SDCs by assigning sky patch to each SDC. To ensure zero loss of the data at least 2 copies of each data file is created in the DSS. Each copy is registered in the metadata storage.

5. EAS Status and Future Development

The EAS Prototype (composed of the EAS-DSS and and the EAS-DPS) was developed and tested in 2013 and 2014. In 2015 the prototype formed the basis of the first version of the EAS itself. Initial interfaces for the EAS-DPS and the EAS-DSS were released and tested during several IT challenges organized by the Euclid Consortium in which simulated Euclid data was produced and stored. We have successfully tested massive metadata ingestion for the data objects with extensive data lineage (KIDS Data Release 1) (de Jong et al. 2013).

In 2015 the EAS team tested a master-slave configuration of the EAS-DPS metadata storage. Metadata was transferred between the EAS-DPS metadata storage mirror in ESAC and the current master site Groningen.

We are currently undertaking a systematic study of RDBMS systems for the EAS-DPS, so we can make a final selection in 2017. We plan as well to create a system which will support a dynamic data model: accommodating changes in ECDM without re-creation of the metadata storage scheme and migrating the data.

References

Begeman, K., et al. 2011, Future Generation Computer Systems, 27, 319
— 2013, Experimental Astronomy, 35, 1. 1208.0447
de Jong, J. T. A., et al. 2013, Experimental Astronomy, 35, 25. 1206.1254
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
Williams, O., Belikov, A., & Koppenhoefer, J. 2014, in Proc. of NETSPACE Workshop, edited by O. Sykioti, & I. Daglis (11), 491