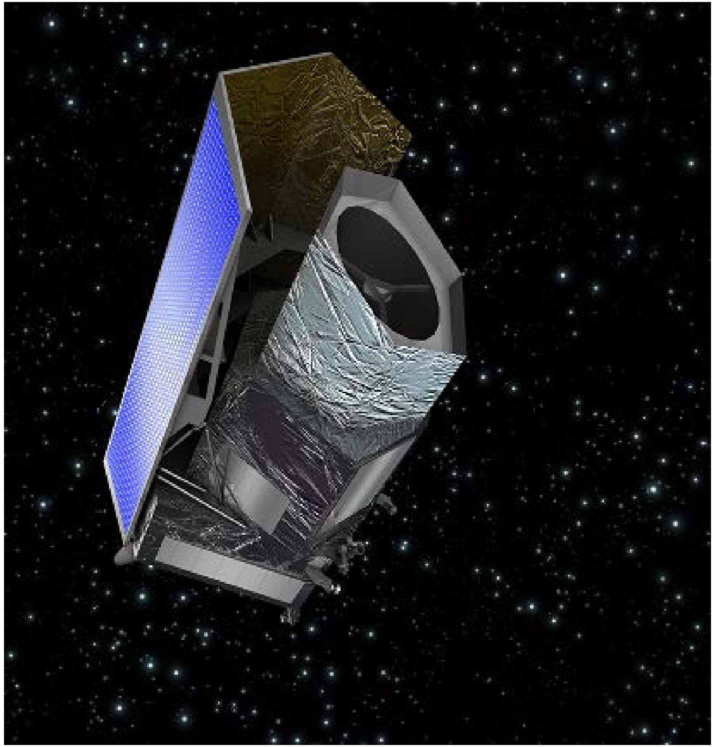
The Euclid Archive Data Processing and Storage Systems: a distributed infrastructure for Euclid

Rees Williams

University of Groningen

on behalf of the Euclid Archive Development Team

Euclid



- ESA mission to map dark matter & dark energy
- Launch on Soyuz in 2022

- Wide field survey
- 1.2m telescope
- 3 Optical & IR instruments

- 15 countries
- 200+ institutes
- 2000+ consortium members

Data Processing Challenge

Unprecedented data volumes for astronomy satellite mission

1 Peta byte images from Euclid

10 Peta byte images from earth observatories

Massive simulation efforts required

Key features

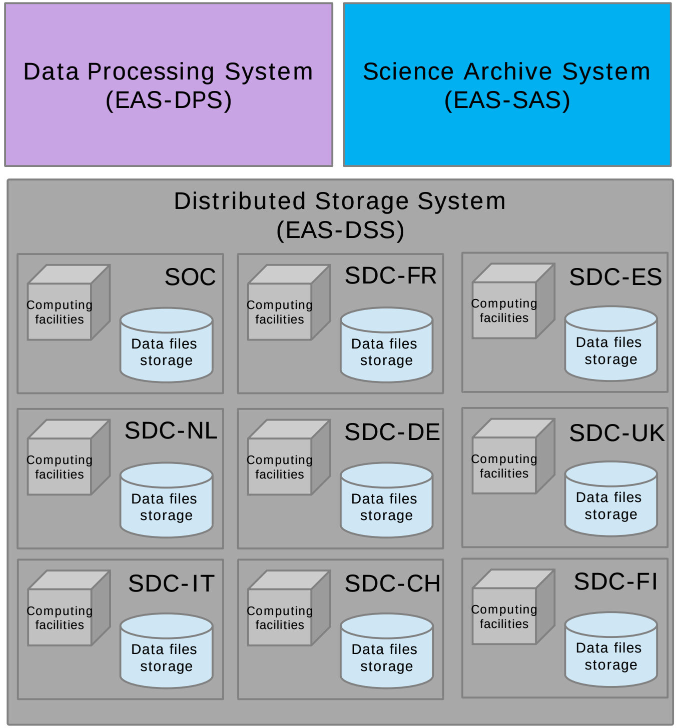
- ➢ Very strict requirements on quality and calibration
- ➢ heavy (re)processing needed from raw data to science products
- ➢ mission could generate 25 Pbytes/year
- ➢ 10 billion objects in catalogue
- ➢ 10 national science data centres (SDC)
- ➢ distributed data processing and storage

The Euclid Archive System: Data Centric Approach

- The Euclid Archive System (EAS) has a central role in Euclid data processing
- Traceability and data lineage are strong requirements
    - EAS acts as interface between all ground system components
    - EAS data distributed across the 10 national data centres
    - EAS metadata is centrally stored
    - EAS metadata contains all information except images and spectra
    - EAS stores dependencies of data products

- Builds on experience with the Astro-WISE information system used by
    - OmegaCAM (KiDS), MUSE

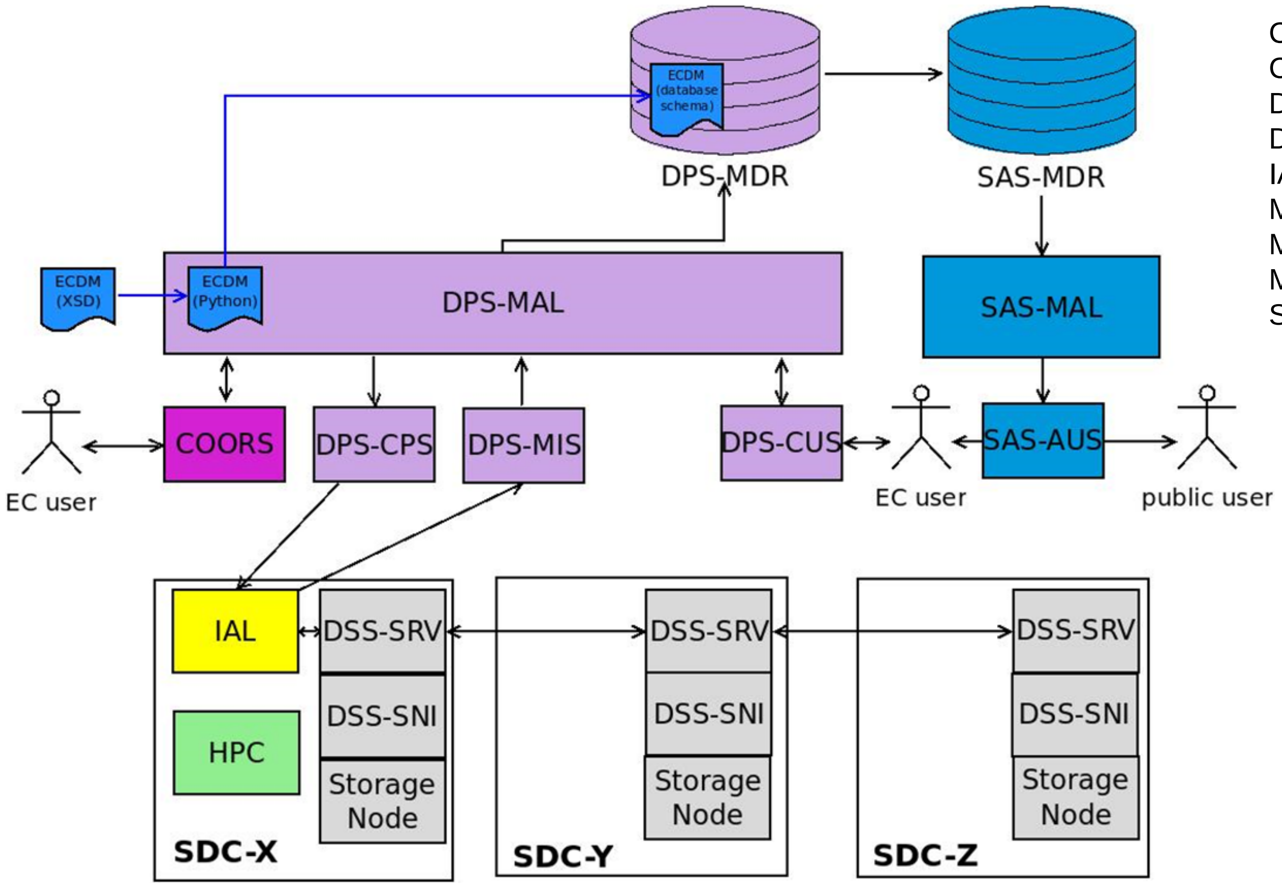
Euclid Archive System: Components



Data Processing System (EAS-DPS)
Science Archive System (EAS-SAS)
Distributed Storage System (EAS-DSS)
SOC
SDC-FR
SDC-ES
SDC-NL
SDC-DE
SDC-UK
SDC-IT
SDC-CH
SDC-FI
Computing facilities
Data files storage

- EAS-DPS: Functionality needed to support operations of the SGS, most notably Data Processing. Implements the complex Euclid Common Data Model (poster by T. Nutma).
- EAS-DSS: Functionality to manage and transfer data stored across many locations.
- EAS-SAS: Functionality needed to support the scientific community (talk by S. Nieto).

Euclid Archive System Architecture



COORS- Coordination & orchestration
CPS – Consortium processing service
DPS – Data Processing System
DSS - Distributed Storage System
IAL - Infrastructure abstraction layer
MAL- Metadata access layer
MDR - Metadata Repository
MIS – Metadata Ingestion service
SAS – Science Archive System

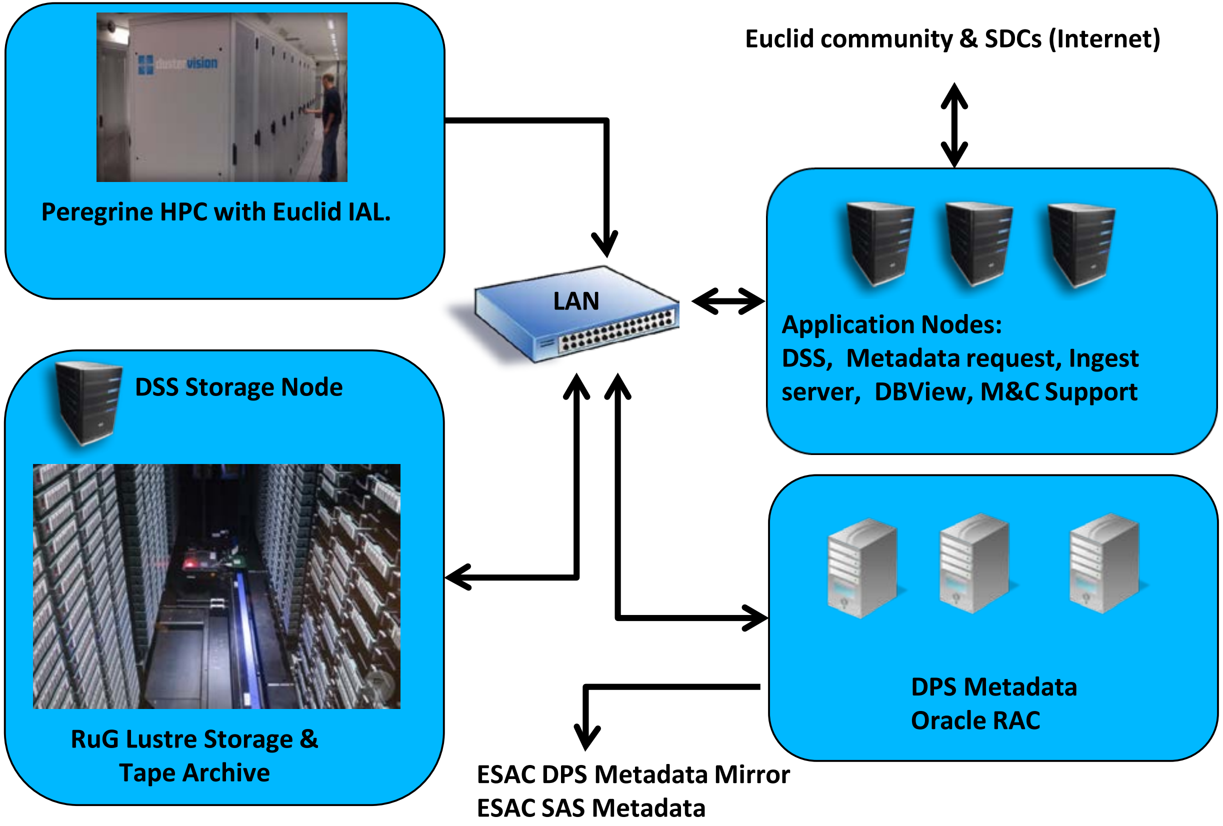
Distributed Storage System Servers: I

- File location stored in central metadata system - provides a global “file system”
  - Process coordination system can run jobs at cluster closest to data
- DSS server installed at each SDC can access data to other SDCs.
  - Users can run jobs locally without knowing the location of data
- Common code base with Astro-WISE, MUSE-WISE, NOVA-LTA
  - Decreases maintenance overhead
- Simple interface:
  - store, retrieve, copy, delete, check, make_local
- Extended interface: cut-out services

Distributed Storage System Servers : II

- No constraint on storage systems at data centres
- DSS server currently supports the access to data using the following protocols
  - Posix file system
  - sftp
  - https
  - gridftp,
  - Astro-Wise dataserver
  - iRods
  - XRootD
  - Openstack

Euclid infrastructure at SDC-NL



Euclid community & SDCs (Internet)

Peregrine HPC with Euclid IAL.

LAN

Application Nodes:
DSS, Metadata request, Ingest server, DBView, M&C Support

DSS Storage Node

RuG Lustre Storage & Tape Archive

DPS Metadata
Oracle RAC

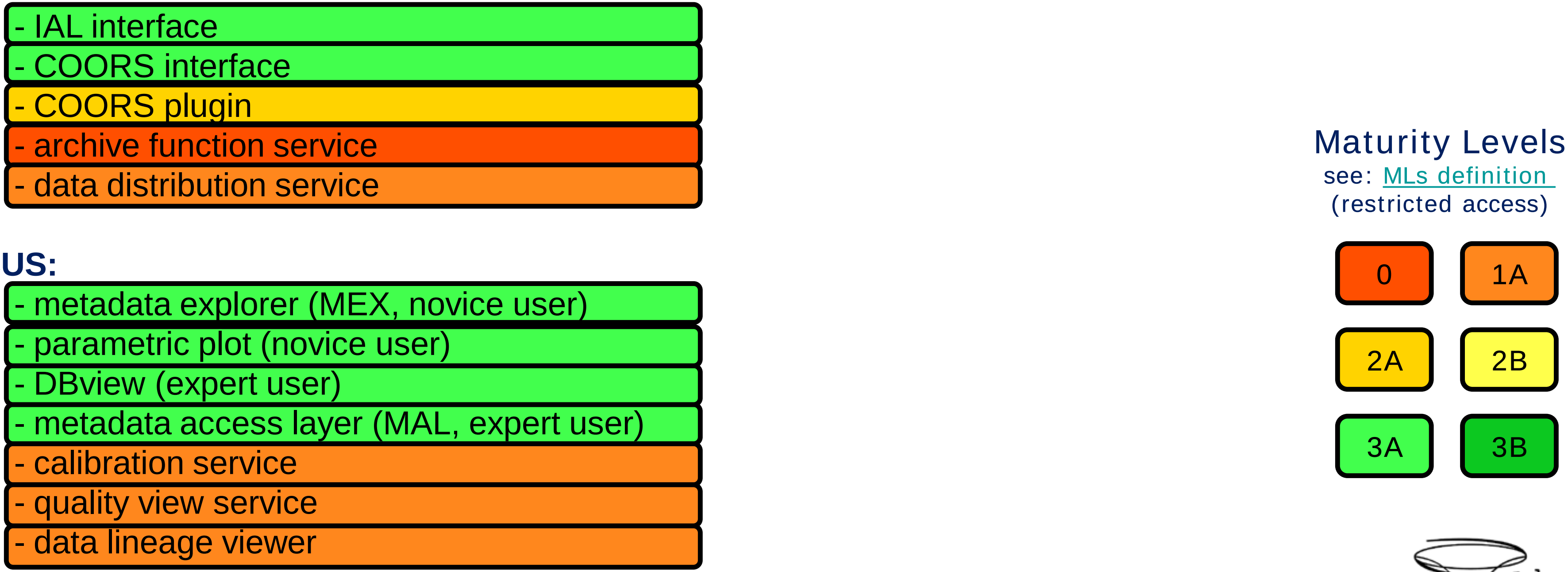
ESAC DPS Metadata Mirror
ESAC SAS Metadata

DPS – Data Processing System
DSS - Distributed Storage System
IAL - Infrastructure abstraction layer
SAS – Science Archive System

EAS Data Processing Services

EAS-DPS services:

- consortium processing service (CPS)
- consortium user service (CUS)



CPS:

- IAL interface
- COORS interface
- COORS plugin
- archive function service
- data distribution service

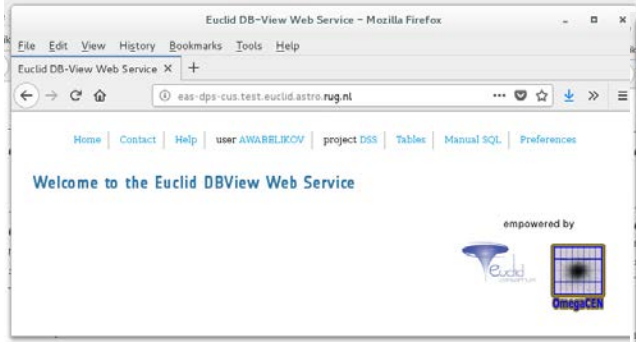
CUS:

- metadata explorer (MEX, novice user)
- parametric plot (novice user)
- DBview (expert user)
- metadata access layer (MAL, expert user)
- calibration service
- quality view service
- data lineage viewer

Maturity Levels

see: MLs definition
(restricted access)

User Interface I: DBview – data model viewer



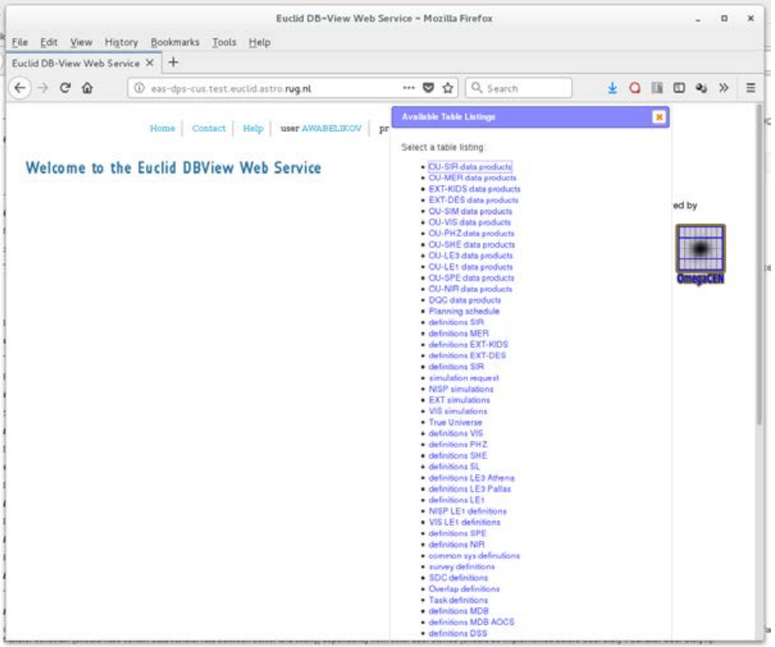
Euclid DB-View Web Service - Mozilla Firefox

File Edit View History Bookmarks Tools Help
Euclid DB-View Web Service
eas-dps-cus.test.euclid.astro.rug.nl
Home | Contact | Help | user AWABELIKOV | project DSS | Tables | Manual SQL | Preferences

Welcome to the Euclid DBView Web Service

empowered by

Euclid consortium
OmegaCEN



Euclid DB-View Web Service - Mozilla Firefox

File Edit View History Bookmarks Tools Help
Euclid DB-View Web Service
eas-dps-cus.test.euclid.astro.rug.nl
Search
Home | Contact | Help | user AWABELIKOV | pr

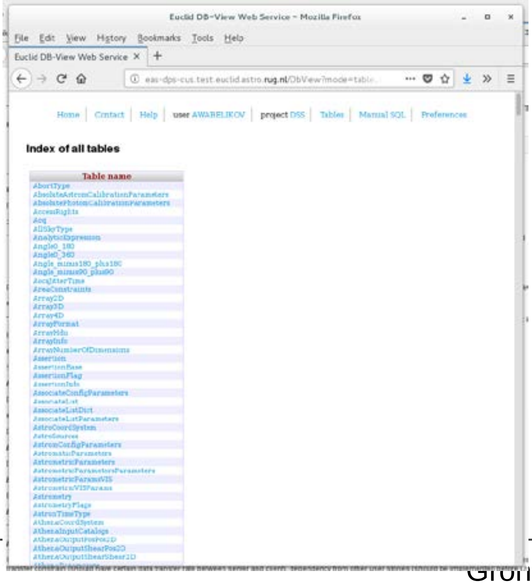
Welcome to the Euclid DBView Web Service

ed by

Available Table Listings

Select a table listing:

- OU-SIR data products
- OU-MER data products
- EXT-KIDS data products
- EXT-DES data products
- OU-SIM data products
- OU-VIS data products
- OU-PHZ data products
- OU-SHE data products
- OU-LE3 data products
- OU-LE1 data products
- OU-SPE data products
- OU-NIR data products
- DQC data products
- Planning schedule
- definitions SIR
- definitions MER
- definitions EXT-KIDS
- definitions EXT-DES
- definitions SIR
- simulation request
- NISP simulations
- EXT simulations
- VIS simulations
- True Universe
- definitions VIS
- definitions PHZ
- definitions SHE
- definitions SL
- definitions LE3 Athena
- definitions LE3 Pallas
- definitions LE1
- NISP LE1 definitions
- VIS LE1 definitions
- definitions SPE
- definitions NIR
- common sys definutions
- survey definitions
- SDC definitions
- Overlap definitions
- Task definitions
- definitions MDB
- definitions MDB AOCS
- definitions DSS



Euclid DB-View Web Service - Mozilla Firefox

File Edit View History Bookmarks Tools Help
Euclid DB-View Web Service
eas-dps-cus.test.euclid.astro.rug.nl/DbView?mode=table...
Home | Contact | Help | user AWABELIKOV | project DSS | Tables | Manual SQL | Preferences

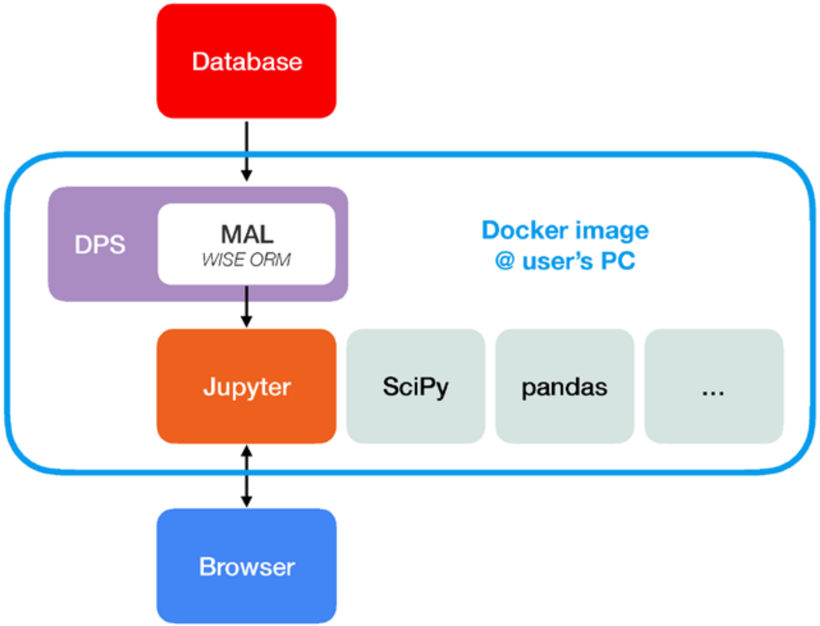
Index of all tables

Table name
AbortType
AbsoluteAstromCalibrationParameters
AbsolutePhotomCalibrationParameters
AccessRights
Acq
AllSkyType
AnalyticExpression
Angle0_180
Angle0_360
Angle_minus180_plus180
Angle_minus90_plus90
AocsJitterTime
AreaConstraints
Array2D
Array3D
Array4D
ArrayFormat
ArrayHdu
ArrayInfo
ArrayNumberOfDimensions
Assertion
AssertionBase
AssertionFlag
AssertionInfo
AssociateConfigParameters
AssociateList
AssociateListDict
AssociateListParameters
AstroCoordSystem
AstroSources
AstromConfigParameters
AstromaticParameters
AstrometricParameters
AstrometricParametersParameters
AstrometricParamsVIS
AstrometricVISParams
Astrometry
AstrometryFlags
AstromTimeType
AthenaCoordSystem
AthenaInputCatalogs
AthenaOutputPosPos2D
AthenaOutputShearPos2D
AthenaOutputShearShear2D

Full data model review
Data products/definitions grouped

User Interface II: Metadata access layer support for Jupyter

SC456 software stack



Euclid Science Challenges

- EAS is currently supporting the so called Euclid Science Challenge #4, #5 & #6.
  - Distributed processing across all 10 data centres
  - Two small wide field regions (7 observations)
  - One large wide field region (38 observations c.f. 40,000 observations in total)
  - Challenge due to finish Nov 2019
- Issues found
  - Management of data processing too manpower intensive (needs more tools)
  - Spatial queries need to be faster

Conclusions

- Basic functionality for the Euclid archive is in place.

- WISE technology shown to scale to Euclid data scales
  - datacentric distributed processing and storage

- Performance tests show minimum levels reached
  - Except for complex spatial queries

- More work on user interface required for
  - Calibration scientists
  - Data quality investigations