NL Data Intensive Astronomy Centres workshop 12 March 2024

Session 2: Processing & Pipelines Challenges

Outline

- OmegaCEN Astronomical Science Data Center
- {data model + database}-centric processing & pipelines challenges
- Euclid experience and lessons learned

OmegaCEN

Astronomical Science Data Center

Gijs Verdoes Kleijn

Rees Williams

Willem-Jan Vriend

university of groningen

CIT & Kapteyn

Sterrewacht Leiden

NOVA

ESO

esa European Space Agency

ATG europe accelerating innovation

240312

Opt/IR Astro Information Systems
survey production, quality control, analysis

Opt/IR survey astro-science

OmegaCEN

Societal astronomy

Data Science Systems Education

Astronomical data-intensive science through information systems aligned with NL Opt-IR instrumentation

societal / non-astro project portfolio

DSSC Data Science & Systems Complexity

TarGet

TarGet Field Lab

AstroWISE

FOTOS OPS

Space Situational Awareness Platform

sky darkness, fireballs satellites, asteroids

information Systems

AstroWISE → MuseWISE → EuclidDPS → ELT Science Platform

OmegaCEN → Euclid Science Data Center → Astro SDC

ESO Data Processing System

Euclid Science Platform

2000 2005 2010 2015 2020 2025 2030

NL instrumentation & data

VST

VL

Euclid

EL

OmegaCAM

KiDS

MUSE

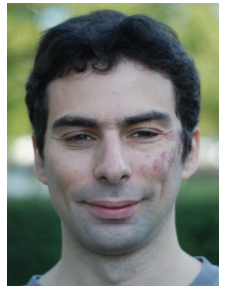
Euclid

FOTOS OPS

METIS MICADO

Kapteyn + CIT OmegaCEN & partners (Leiden Obs, ATG)

Gijs
Verdoes Kleijn
PhD, astro



Andrey
Tsyganov
PhD
database infra

Danny
Boxhoorn
MSc
system architect
database, storage

Willem-Jan
Vriend
MSc
system architect
compute, QA services

Pablo
Corcho-Caballero
PhD
pipelines, algos
science

astro
MSc PhD

Pedro
Beirao
PhD
pipelines, infra

Hugo
Buddelmeijer
PhD
pipelines, infra

Rees
Williams
PhD, IT

Bob
Dröge
MSc
compute infra

Andrey
Belikov
PhD + PhD
database system,
storage system

Jelte
de Jong
PhD
pipelines, algos, QA
services

Eduardo
Balbinot
PhD
pipelines, algos, science

Zuzanna
Kostrzewa
PhD
pipelines

Matthew
Horrobion
PhD
pipelines

Edwin
Valentijn
Prof

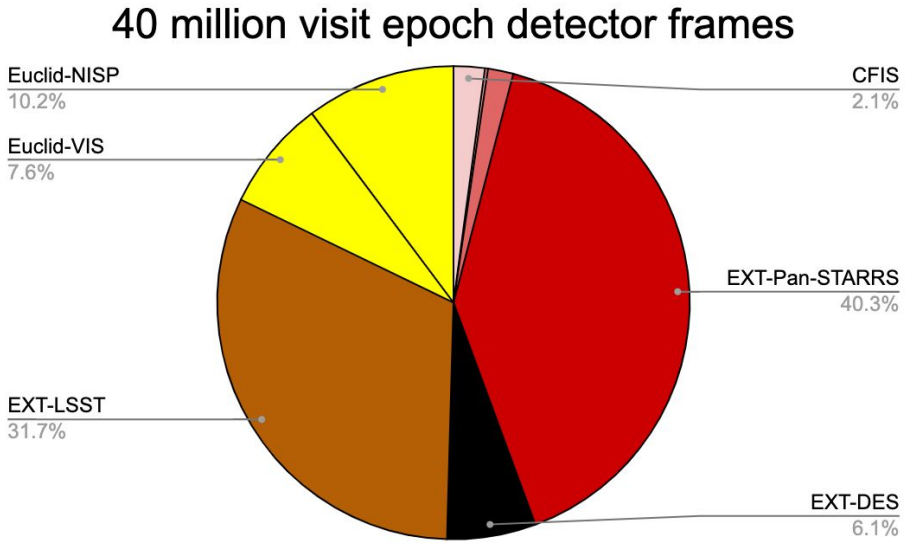
Euclid Science Platform

Space Situational Awareness Platform

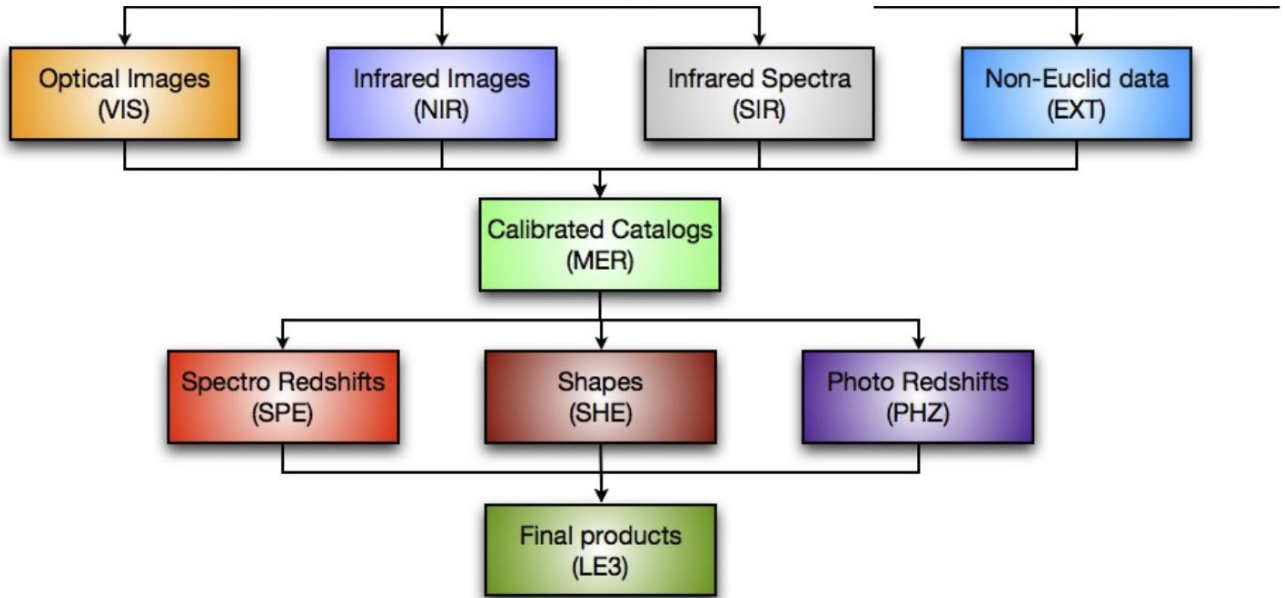
ELT Science Platform

Euclid experiment: 6 instruments and 10 data & processing centers spread over Earth and L2:

space

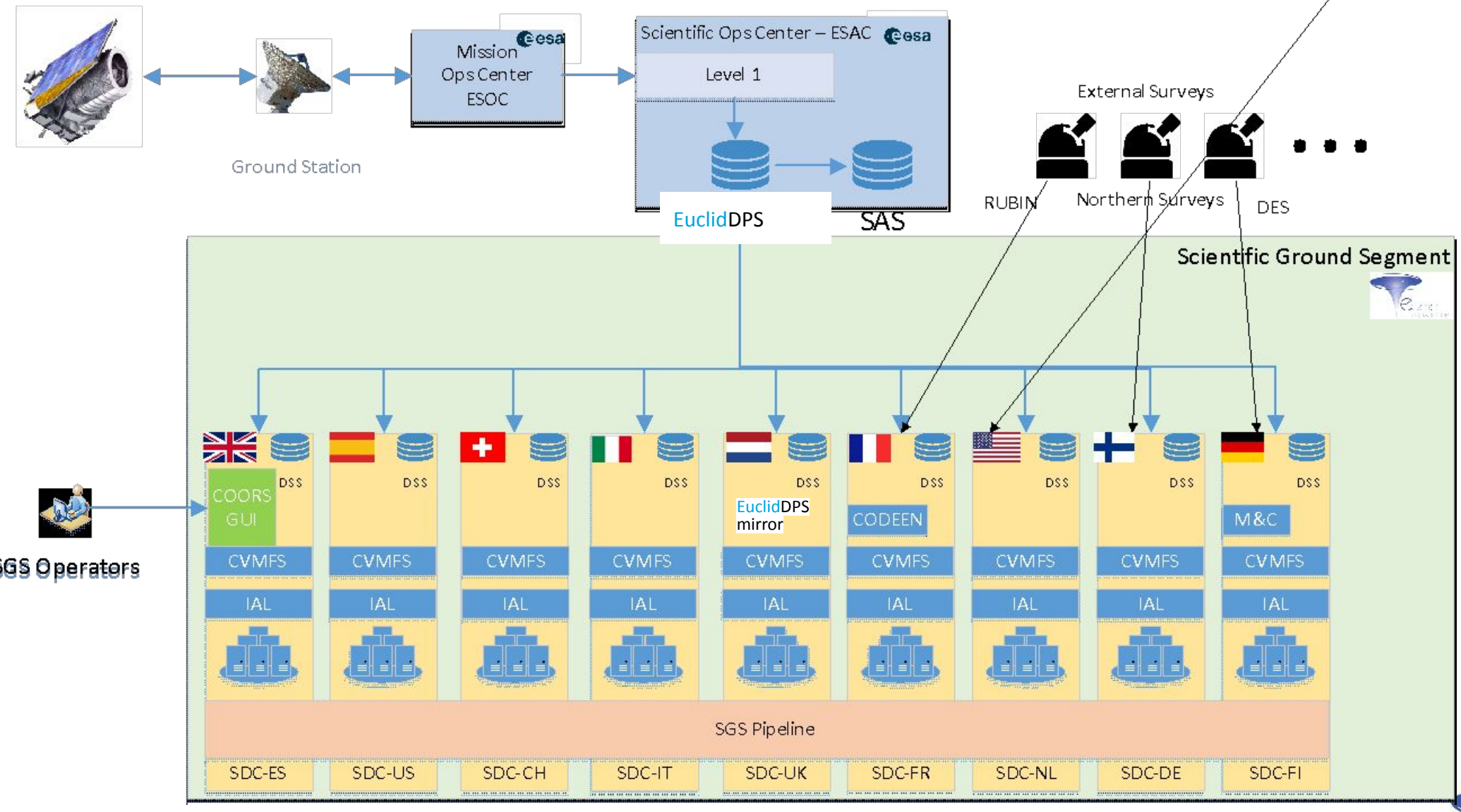


Pipelines (up to few dozen in each box):



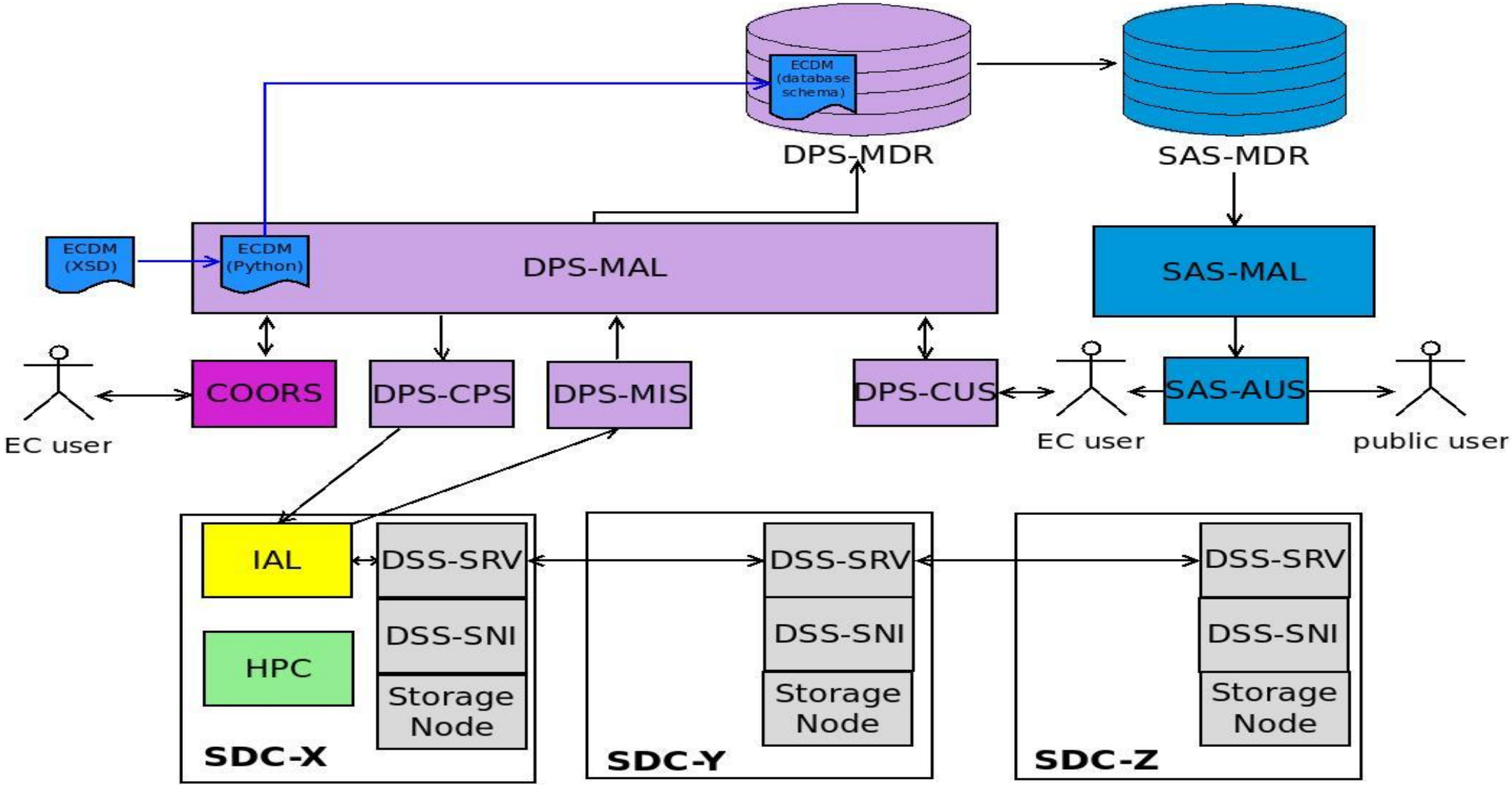
raw: order 3 Pbyte
total: order 30 Pbyte
information system users: order 2500

Euclid Science Ground Segment architecture overview



Euclid Science Ground Segment

Euclid Consortium



Organizing distributed communities – Open Science

for survey production for science analysis

Cathedral

Database

Data model

Access layers

DPS MAL

DPS x

IAL

DSS x

Containers -data Meta data

Bazar

Hardware

processors

storage

Jupyter-based Science Platform

Slide courtesy: Edwin Valentijn

Main lessons learned Euclid

Choice: database-centric: all data items specified via an XML Schema Definition: XSD Data Model

+ acts as software-readable interface control document: automates creation database schema and database interfaces
- development of human readable version was under-prioritized: readability perceived as challenging for humans. Even for some pipeline developers.
- relationship between data items was not captured: hampers automated orchestration workflow: becomes human intensive

Choice: pipelines do not interface with database: input data items specified by separate process

+ immune to Data Processing Center rules about interfaces to databases external to Data Processing Center

- metadata via XML is major overhead on system.

Choice: full freedom on types of data storage and types of compute in data processing centers

+ cheap in terms of infra-hardware & its personpower: maximizes resident expertise and

- expensive in terms of infra-software & its personspower: need to debug up to 10 times as many infrastructures. (Just posix compliancy would already have helped)

END