AENEAS European SRC Design

Science Archive +
Data Processing and Storage

WP3

Science Archive

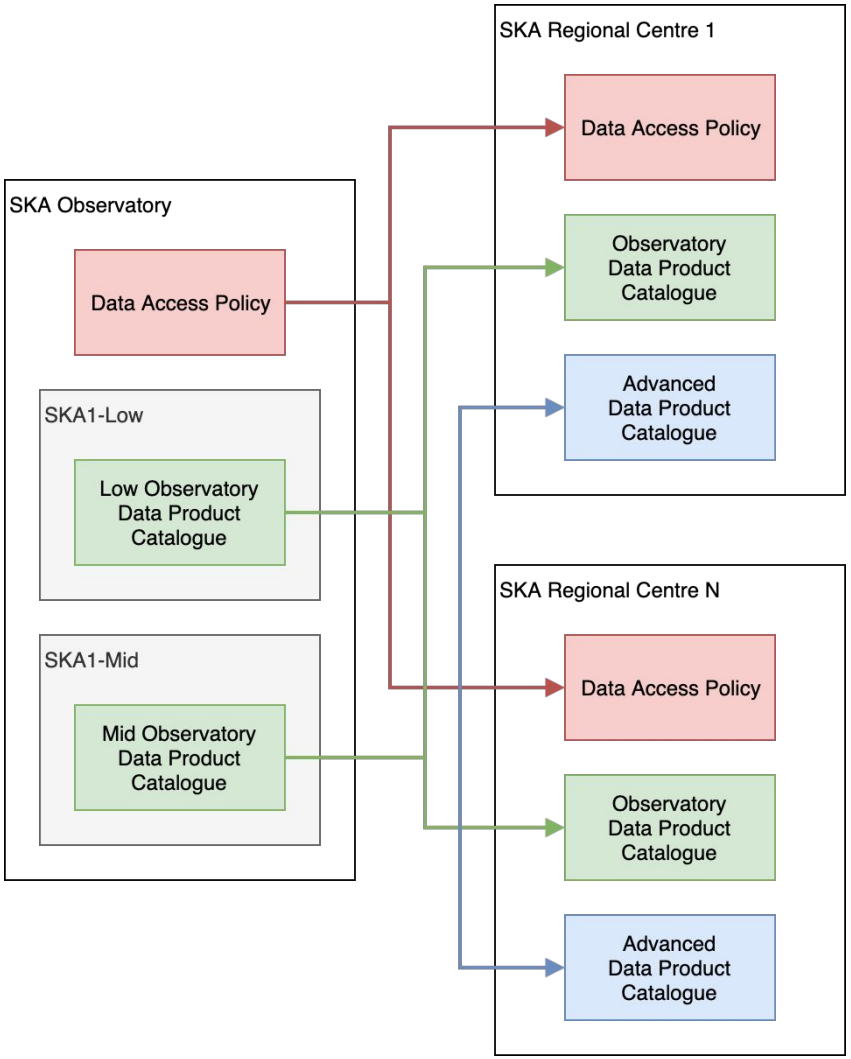
The Science Archive contains

- Observatory data products (ODPs)
  - Generated by SDPs at the two telescopes with no user access (mostly automatically)
  - Distributed from SDPs to SRCs using push model
  - Observatory sets data distribution policy (destination SRC, etc.)
- Advanced Data Products (ADPs)
  - Generated by users at the SRCs (manually or automatically)

Data products are distributed across SRC network, with one or more copies available somewhere in the network

Science Archive

- Data products are globally distributed, so a global catalogue is required to discover and locate them
- Data access policy is defined by the Observatory and the SRCs must enforce it
- Each SRC must make data products available to other SRCs when required
  - Data distribution policy should be designed to minimise transfers
- Each SRC must publish the ADPs it generates in the catalogue
- Science Archive must adhere to FAIR principles
  - Findable, Accessible, Interoperable, Reusable



Data Product Catalogue

Data Access Policy

- Defined by the Observatory
- Published to SRCs
- Enforced by the SRCs

Observatory Data Product Catalogue

- ODPs are added by the SDP that generates them
- Published to SRCs

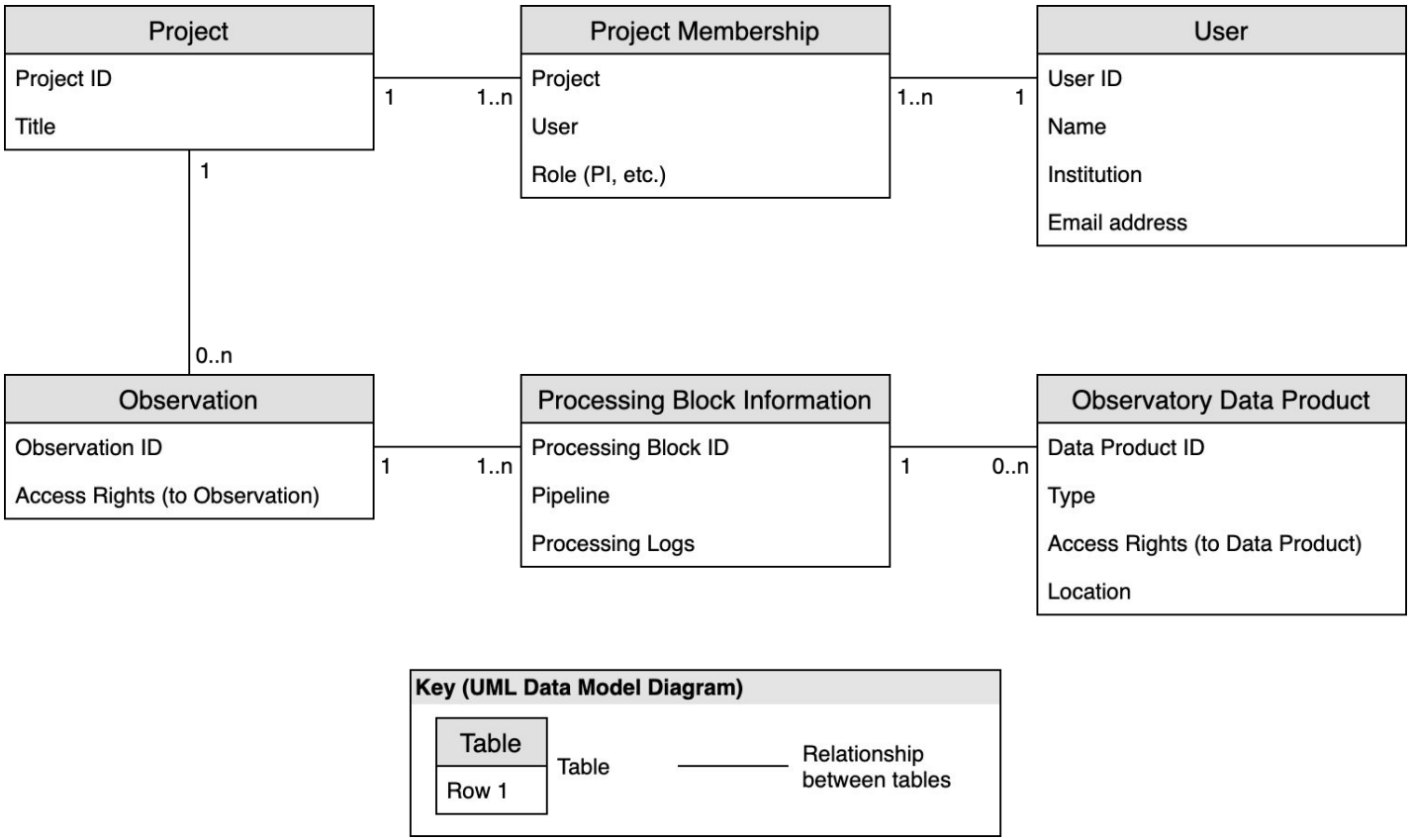
Advanced Data Product Catalogue

- ADPs are added by the SRC that generates them
- Published to other SRCs

Provenance of ODPs

- Processing of a observation by SDP is defined in a “processing block”
  - Specifies the pipelines to run and their parameters
- Each processing block produces the desired data products, plus one more called the “Science Data Model” containing full provenance information
  - Telescope configuration and state at the time of observation
  - Processing block information
  - Local sky model (subset of GSM)
  - Output of pipeline monitoring (logs, QA metrics, etc.)
- Some of this information will be included in the ODP catalogue too

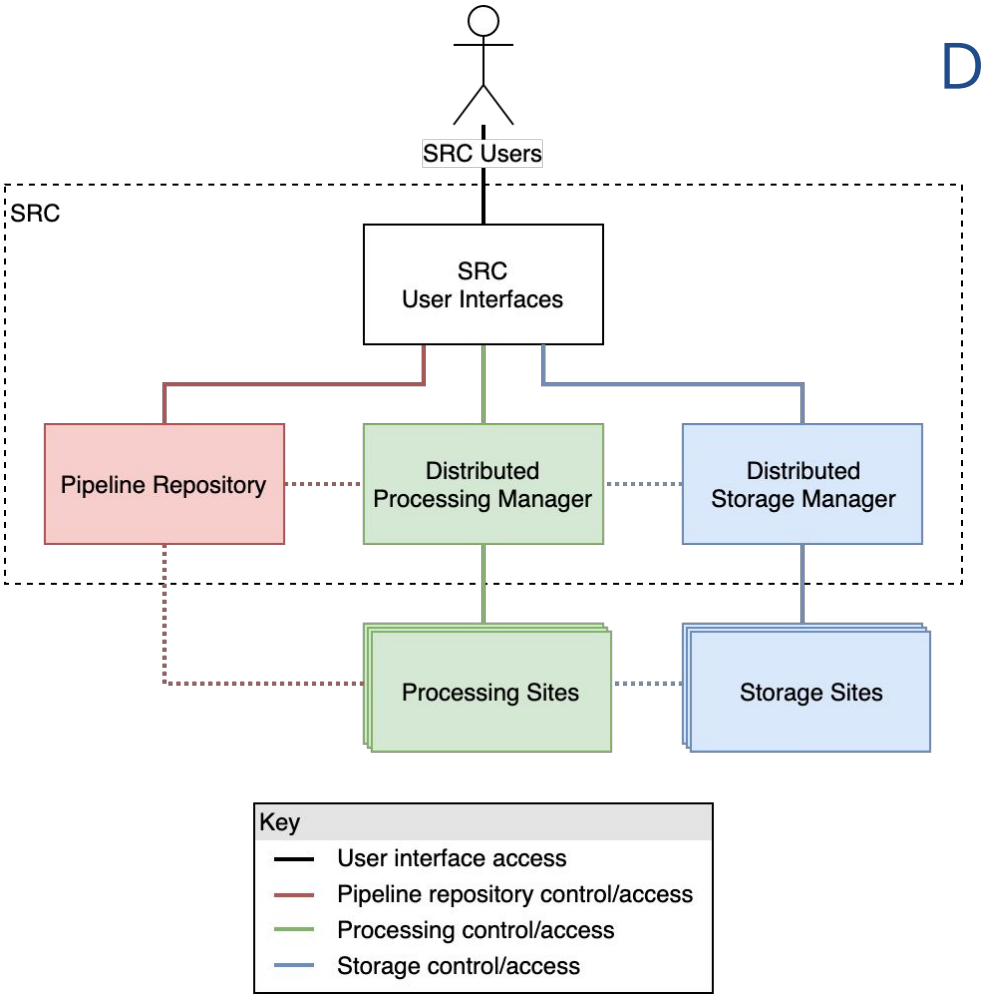
Data model for ODP catalogue



Provenance and Reproducibility of ADPs

- Advanced data products must contain full provenance information and be reproducible
- Catalogue must store provenance information to make this possible
  - Pipeline used to generate the data product (including version information)
  - Pipeline input data (ODPs, ADPs, external data)
- SRCs must store all of the necessary data and the pipeline itself
- This implies that the SRC requires a pipeline repository

Data processing and storage



- Use external providers for processing and storage
- Dashed line contains systems provided and operated by the SRC
  - User interfaces
  - Pipeline repository
  - Dist. processing manager
  - Dist. storage manager
- At minimum these systems need to be interoperable with those at other SRCs
- Ideally adopt the same systems across the SRC network

User Interfaces

User interfaces tie together the processing and storage resources and present them to the users in a unified way

Types of interface to support

- Web interface
- Application programming interface (API)
  - With command line client
- VO interface
  - Implied in requirements but not explicitly stated
  - Mostly for access to data, but some processing may be necessary to generate data products if they are not already available

Pipeline Repository

Need a way to encapsulate pipelines, to

- Make it easy to run them on available platforms
- Preserve them for provenance purposes (including versioning information)
- Make them discoverable

Containerisation seems to be a good solution to this requirement

- More than one possibility: Docker, Singularity
- Need to adopt suitable conventions e.g. Scientific Filesystem
- Must ensure source code is available, not just binaries

Data Processing and Storage

- Processing manager needs to take into account location of data products when processing
- Storage manager must support
  - Receiving data from the SDPs
  - Moving data between sites
  - Migration between different tiers / latencies of storage
- Observatory and SRCs must have a mechanism for planning and agreeing the amount and types of processing and storage needed in the network on a ongoing basis