SRC | Net

SKAO Regional Centre Network

SRCNet Science Analysis Platform Vision

SRC-0000003	Revision 01B
Classification:	UNRESTRICTED
Document type:	PLN
Date:	2023-08-25
Status:	DRAFT
Authors:	Skipper, Chris; Das, Arpan; de Boer, Janneke; Cimpan, Iulia; Fabbro, Sébastien; Grange, Yan; Hardcastle, Martin; Sharma, Rohit; Swinbank, John; Webster, Brendan

Role	*Name*	*Designation*	*Affiliation*	*Signature*	*Date*
Author	Chris Skipper	Product Owner, Tangerine Team	JBCA		
Owner	Shari Breen	Head of Science Operations	SKAO		
Approver	Michiel van Haarlem	On behalf of the SRCSC	ASTRON		
Released by	Lewis Ball	On behalf of the SKAO	SKAO		

Disclaimer:

The scope of this document is to set out a vision for a science analysis platform to serve the SKA Regional Centres. Its goal is to define what the concept of a Science Analysis Platform means within the SRC Network. It is not intended to be an implementation plan or to make concrete technology choices. Any technology mentioned in the document should be interpreted as an example rather than an implementation decision.

At the time of writing, the architecture of the SRC Network contains many unknowns. Therefore, any reference to external components that are to be provided by partners in the SRC Network can be changed in a future version.

This document is a living document, and will therefore continue to evolve as the project develops. It is not the role of the authors to define how or when decisions will be made about the implementation of different technologies – this will be agreed through further work conducted as part of the SKA Regional Centre (SRC) ART.

TABLE OF CONTENTS

LIST OF FIGURES

No table of figures entries found.

1 Introduction and Context

1.1 Purpose of the Document

This document describes the SRC vision for the first iteration of the Science Analysis Platform intended to be delivered across the SRCNet. This platform will be consistent across the nodes and provide an interface between Square Kilometre Array (SKA) users and their data, and a powerful and intuitive environment for carrying out reproducible data analysis, collaboration, and other essential research tasks, while offering enough flexibility for individuals to customise the platform to meet their own needs.

Within this vision document we describe a science platform appropriate to the needs of the SRCNet in broad terms. It is not a set of detailed requirements or specifications, nor is it a design document. Rather, it sketches a high-level view of the necessary platform functionality and identifies key interfaces with other SRCNet technologies and areas of development. The vision outlined here is sufficiently broad as to allow the proposed platform to be moulded to the use-case requirements at a later stage of development. It is intended to serve as a reference for both development teams and other stakeholders.

The document has its origin in discussions undertaken within the responsible SRCNet prototyping team ("Team Tangerine") during the period of March to September 2022. It draws on the expertise of team members, inputs from other members of the SRC development community, existing requirements derived in the SKA Regional Centre Steering Committee (SRCSC) working-group phase and a wide-ranging literature review covering the philosophies and practicalities of many existing science analysis platforms, (e.g., CANFAR, ESCAPE ESAP, CERN REANA, GALAXY).

This is expected to be a *living document*: it will be updated over time to reflect lessons learned during development and refined understanding of user needs. Just as this document is intended to be *living* so is the platform itself, and future upgrades will be necessary to meet the evolving needs of the SKA user.

1.2 What is a science analysis platform?

A science analysis platform provides scientists with a computing environment that permits the collaborative handling of large and diverse datasets and allows them to access large-scale computing resources that they may not have locally available. Science analysis platforms are often linked to a specific project, instrument or method of analysis (for example, CERN SWAN[1], or the Rubin Science

[1] https://indico.cern.ch/event/679940/attachments/1570151/2476533/SWAN-Service_Interactive_Analysis-Cloud-2018.pdf

Platform[2]) and are typically designed to provide consistency for all users while providing access to appropriate tools and data. Thus, a science analysis platform must provide features that are relevant to the user base and avoid any pitfalls that will make it difficult for scientists to complete their work.

1.3 Platform aims and objectives

The SKAO will generate around 700 PB/year of science-ready Observatory Data Products (ODPs), each with appropriate provenance and quality assessment information.

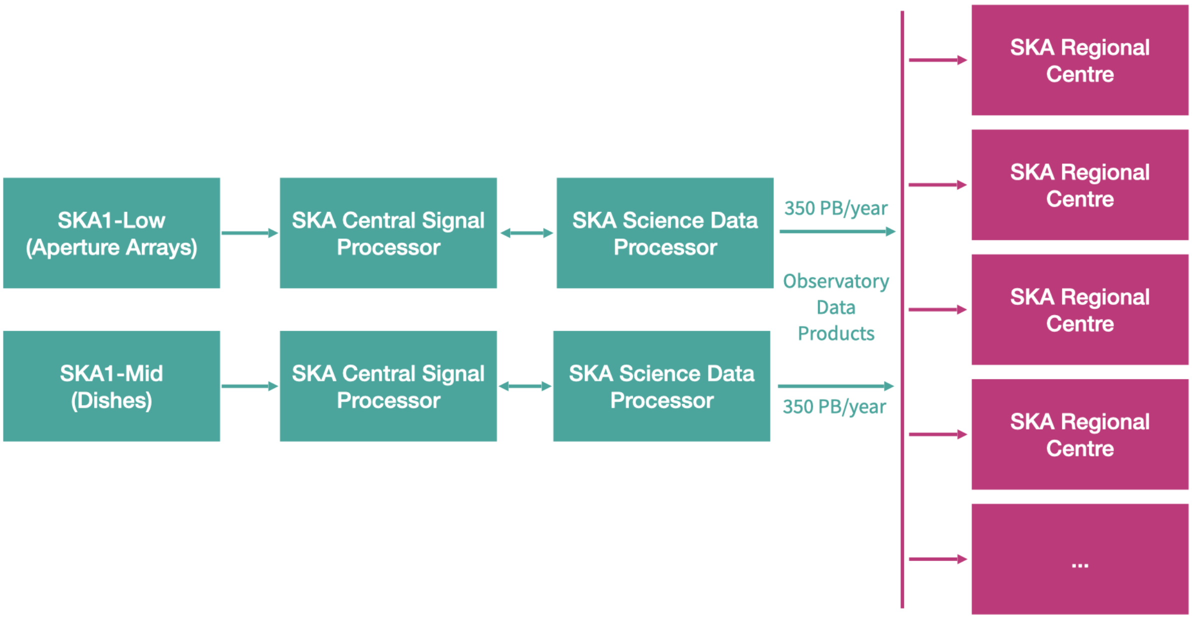
SKA1-Low (Aperture Arrays)
SKA Central Signal Processor
SKA Science Data Processor
SKA1-Mid (Dishes)
SKA Central Signal Processor
SKA Science Data Processor
350 PB/year
Observatory Data Products
350 PB/year
SKA Regional Centre
SKA Regional Centre
SKA Regional Centre
SKA Regional Centre
...

Figure 1: The relationship between SRCNet and the SKAO.

These data products will include — amongst others — image cubes, *uv*-grids, calibrated visibilities, and pulsar timing solutions [AD1], as well as non-science ODPs such as calibration data and observatory metadata. However, these products will not directly be made available to the scientific community. Instead, as illustrated in Figure 1, ODPs will be provided to a worldwide network of SRCs [AD2]. The SRCNet then assumes responsibility for the following user-facing functionality [AD3]:

- Data logistics, including making data available to end users;
- Data processing, including providing computational and storage resources and appropriate software environments to enable end users to interact with SKA data products produced by

[2] https://data.lsst.cloud/

the Science Data Processing & Handling (SDH&P) and produce additional, higher-level data products, some of which will be approved and accepted into the archive as Advanced Data Products (ADPs);

- Data archiving and curation, including data discovery and re-use;
- User support, provided through a suite of documentation and a single helpdesk staffed by both SKAO and SRCNet staff to service all user enquiries.

The platform will be the primary mechanism by which end users interact with the SRCNet. As such, it will provide users with a unified place where they can come to search for and discover SKA data (using the functionality described in §3.2), to perform interactive data analysis tasks (§§3.3, 3.4 & 3.5), and to schedule and manage long-running or complex workflows (§3.7). In principle, batch workflows could also be quick and simple, but highly numerous. The platform will also provide capabilities for users to collaborate, sharing and publishing both their analysis workflows and their results in the form of ADPs. All of this functionality will be available both through a modern and accessible User Interface (UI) (§3.1) and through an Application Programming Interface (API) which will facilitate access from user scripts and other automated tooling (§3.6).

The platform will be highly collaborative, allowing the sharing of workflows, data, and code in order to bring teams together, and end-to-end, such that scientists will be able to perform analysis, and publish their data products and analysis code in a transparent way (e.g. through a connection with GitHub/GitLab and through Digital Object Identifiers (DOIs)) that follows Findability, Accessibility, Interoperability, and Reusability (FAIR) [RD1] principles, thus being supported through all stages of the lifecycle of a scientific project.

1.4 The design of the platform

At the time of writing, we expect that the set of interfaces constituting the SRC science analysis platform will be primarily web-based, with both interactive and programmatic access, but the design should also be capable of responding to new standards as they evolve. However, it will be possible to get command line/terminal access as well for experienced programmers/analysts. The design and operation of the SRC science analysis platform will provide the following attributes:

- *consistency*: the user experience will not be dependent on location or the available infrastructure;
- *scalability*: the platform can be expanded to accommodate increased numbers of users and larger volumes of data;
- *reproducibility*: analysis produced on the platform can be reproduced at a later time;
- *usability*: interfaces must be usable by a user base with a wide range of skills across a wide variety of local compute resources and accepted standards;

- *reliability*: interfaces must be fully functioning at all times and must provide transparency on a user's resource usage.
- *accessibility*: The platform's interfaces will be configurable to provide maximum accessibility to the diverse scientific user base (see §1.5).
- *Transparency*: clear visual reporting of the system resource usage to users (e.g. compute resource utilisation, queue wait times, etc.)

Reliability and scalability are dependent on the underlying components of the SRCNet and compute infrastructure. The SRC science analysis platform will pass on information about the state of this infrastructure to its users. Implementing all of these attributes *simultaneously* will be challenging but is necessary for the efficient and thorough exploitation of SKA data.

The Platform will be designed and built to comply with the high-level SRC system requirements[3], and to follow the SRCNet Architectural Principles[4]. In particular, we highlight the following:

- The platform will provide accessible and straightforward interfaces, where possible adopting paradigms which are already well established within the astronomical community, to ensure that SKA data is available to the widest possible user base;
- The platform will provide users with maximal flexibility in deploying sophisticated custom tools, and personalising the UI to suit their needs;
- The platform will provide appropriate graphical and API interfaces to the lower-level services and tools deployed across the SRCNet;
- The platform will provide seamless access to available services regardless of the physical location of the user and/or of the SRC node providing a particular service;
- The platform will follow FAIR principles by ensuring that all ADPs have appropriate metadata describing their provenance so that they are reusable. Providing public access (following the conclusion of the proprietary period) to the workflows that produced the ADPs will be strongly preferred;
- The platform will provide abstractions so that users are appropriately insulated from the details of the underlying SRCNet infrastructure;
- The platform will provide users with access to services and/or data products to which they are entitled, and will include public access to data products that have exceeded their proprietary period.

1.5 Accessibility and inclusion

Proper consideration of accessibility issues are crucial to the design of the platform and its UI. The platform will be available to as many users as possible, and, for example, will avoid excluding users

[3] https://confluence.skatelescope.org/display/SRCSC/High+Level+Requirement

[4] https://confluence.skatelescope.org/display/SRCSC/SRCNet+Architecture+Principles

on the grounds of disability, language, cultural background, domain specialism or experience. This requirement has two aspects:

1. The platform must facilitate individual accessibility requirements by making sure everyone can use, adjust and configure the UI to suit their own needs. Examples may include choice of interaction languages, colour filters, screen readers, text scaling among others. Since our expectation is that scientific users will be interacting with SKA data almost exclusively through this platform (see §1.3), we need, to the extent possible, to allow the user to have as much control over the web interface presented to them as they would expect to have over a well written local application. As legislation and best practice will vary throughout the network, it will be essential to keep up-to-date with evolving web accessibility guidelines[5], relevant laws and community feedback.

2. The platform must be usable to people from a wide variety of backgrounds, specialisms, and career stages, and must therefore present an interface that is accessible to all of those groups, again potentially through the use of configurations that allow the system to be customised to meet the needs of both novice and highly experienced users.

In order to meet these needs the users will need to feel supported, directed towards the right resources, and helped, inspired and motivated by collaborating closely with them through user support from a single helpdesk, workshops, tutorials, cookbooks, clear documentation and user-friendly interface. These resources should be suitable for the full range of user expertise.

1.6 Data product type definitions

Throughout the document several types of data are referenced:

- ODPs: Divided into Observation-Level Data Products (OLDPs), which are calibrated data products generated by SDH&P workflows, and based upon data obtained from a single execution of a scheduling block, and Project-Level Data Products (PLDPs), which are calibrated data products generated by combining several, related OLDPs.
- ADPs: User-generated products, generated by workflows from ODPs and data from other SKA observations or other facilities.

The SRC science analysis platform will mostly be used to generate ADPs from ODPs obtained from the archive, but will also be used to run SDH&P pipelines to

- create PLDPs,
- test and fine tune parameters (e.g. for calibration or imaging) on small volumes of data prior to a large observing run.

[5] e.g. https://www.w3.org/TR/wcag-3.0/

For further details on SKA science data products please refer to the SKAO Science Data Products Summary Report [AD1].

2 Back-end features and services

In this section we briefly describe the back-end features required for an SRC node in order to provide the context for the user-facing aspects that will then be discussed in §3.

2.1 Compute services

Analysis through the SRCs will be carried out over different regional, national or supranational compute infrastructures. Computing requirements are varied and include processing of very large datasets for generation of ADPs or running of the SDH&P pipelines (to create PLDPs or test appropriate workflows and parameters of ODPs), as well as user-driven batch processing or interactive processing via, for example, a notebook. It will also be possible for users to test new workflows on small datasets and receive immediate results, thus enabling workflows to be refined prior to submitting large batch jobs.

Resource management will be carried out by SRCNet to make the most efficient use of compute resources, and resource allocation will start from the project approval stage so that appropriate staging of the data will be done. Resource allocations may be made to users wishing to exploit exclusively public data (i.e. without an approved SKA project), following an approved SRCNet processing proposal.

An important feature of the distributed nature of the SKA archive is that it will in many cases be more efficient to generate an ADP on the SRC node that is local to the data, rather than the one that is local to the user. Similarly, when a workflow requires that an ADP be present on a specific SRC node it will in many cases be more efficient to regenerate that ADP rather than transfer it from another node. In the latter case the SRCNet will automatically determine the most efficient way to provide the required data product based upon previous run time and resource requirements held within the metadata of the ADP. Storing this information in the metadata also allows the user to view these details.

2.2 Archive and distributed data

The SRCNet will be responsible for providing an archive for the ADPs (as well as the ODPs), together with a software repository for the code used to generate the ADPs (see Section 3.9 for more details). The archive will be needed to allow users to query the data sets within it, and retrieve data for further processing. There will be a process for users to upload ADPs created outside the SRCNet, and their associated metadata, to the archive. Once ADPs are submitted to the archive, it will not be possible for users to alter their contents, although it will be possible to both upload a new version and mark older versions as obsolete.

The users will be able to run queries on the archive using appropriate query languages, e.g. Table Access Protocol (TAP), TAP+ [RD2], Astronomical Data Query Language (ADQL), etc., providing both synchronous and asynchronous query modes depending on datasets. The TAP-like queries will support uploading the table to user space from multiple inputs like URLs, files, tables from a job, and tables from an astropy table. These are implemented in astroquery[6] and CADC user database[7], and can be built on.

2.3 User data storage

Users will be provided with a persistent, personal file system (for example, Portable Operating System Interface (POSIX)), to which they can upload and download files at will, including from the archive, to within the per-project resource allocation. These files are available for further processing and visualisation, as well as to hold code and additional uploaded data for analysis. Processing logs will be stored along with the data for all processing operations and will record information on software and resources used and any other parameters to ensure reproducibility of new data product generation.

The file system will allow fine-grained sharing of files or directories for collaboration purposes, making use of the Authentication and Authorisation Infrastructure (AAI) service (§2.4).

In addition to the file system, users can generate their own databases which can be created, accessed, updated and queried using tools such as Structured Query Language (SQL), in order to store structured data relevant to their science cases. Database rights will be shareable between groups of users.

2.4 Authentication and authorisation

The SRCNet will provide AAI services which will provide mechanisms for verifying user identity ("authentication") and establishing the services and data that each user has the right to access ("authorisation")[8].

The platform will integrate with this AAI system. In particular:

[6] https://astroquery.readthedocs.io/en/latest/utils/tap.html
[7] https://ws.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/youcat/
[8] https://jira.skatelescope.org/browse/SRC-147

- Access to the platform itself will be subject to appropriate authorisation;
- All services and data that are offered to or accessed by users through the platform will be subject to appropriate authorisation;
- Platform-provided graphical UI (§2.4) and APIs (§3.6) will provide seamless interoperability with the AAI system.

In general, the platform will not offer users access to services or data to which they do not have rights; where appropriate, opportunities for acquiring elevated privileges may be indicated.

Ultimate responsibility for controlling the access to data or services is the responsibility of the system which hosts the data or service. For example, the underlying storage system, not the platform, is the final arbiter of whether a given user can access a particular data element. However, the platform assumes responsibility for forwarding user credentials to appropriate systems and presenting the results of data or service-access operations to platform users.

3 User-facing features and services

3.1 Main user interface

The main UI for the platform will be a web page (the ‘portal’, which will be hosted by the SRCNet) providing access to the functionality of the SRC node through a range of services. The user can sign on to the portal on the front page using single sign-on criteria (§2.4) and will then be given access to the full UI; without signing on there will only be limited public data access to, for example, image previews and catalogs. The UI will be consistent across different SRC nodes.

Many users will simply want to access either their own data or data from the archive, e.g. by obtaining previews of images or catalogs or downloading reducedvolume datasets directly to their own computer; due to the amount of data expected, and the nature of a server-side oriented science platform for analysis, we do not expect that users will be routinely downloading large volumes of data to external compute resources. Simple data-querying and discovery (§3.2) will be the default panel, allowing users who do not wish to carry out more sophisticated analysis to go straight to the data. Once a dataset has been selected, users will need to be able to visualise the data (images, spectra, cubes, catalogs, light curves etc) either by running tools built in to the platform, or by spawning a tool that runs within a software environment or notebook. The visualisation tools provided by the platform will include tools suitable for large datasets. Other panels will provide access to a notebook for interactive analysis (§3.3) and the ability to run containers (§3.4), Virtual Machines (VMs) (§3.5), or distributed jobs, as well as constructing more complex workflows (§3.7). The user will also be able to perform simultaneous tasks, e.g. download some data while running a complex workflow and viewing an image, and will have access to information covering their resource usage, either as a single user or a member of one or more groups (§3.8).

Many elements of the UI of the portal will be configurable by the user to support (1) the user's particular accessibility needs (§1.5) and (2) the user's ability to simplify access to the tools that they are most likely to need. Customisation of the UI will be on a per-project basis, and the platform will provide a quick and simple means for users to switch between projects.

In order to satisfy evolving user requirements, an SRCNet administrator will be able to add or modify functionality available through the UI without the need to recompile or redeploy the software. Wherever possible, all functionality will be implemented through workflows, written in the language of a supported workflow-execution system (§3.7), and the UI will be configurable by an administrator such that this functionality is accessible from an appropriate place within the portal. For example, if new functionality is required that converts visibility data to a new file format then an administrator would define a new workflow to perform the task, and - via the portal - add a new button/menu option/etc to an appropriate location, with a suitable description and icon.

3.2 Data querying and discovery

In order to maximise the science potential of the data held within the SRCNet archives, users will be provided with a variety of tools for data querying and discovery. Any queries entered by the user will call an underlying API that will process the data and provide the results. The API will be publicly available so that users can execute archive searches and examine the results within other environments such as a notebook.

In order for data discovery to work correctly, each archived data product will require associated metadata. Therefore, whenever a data product has been approved as an ADP it will be stored in the archive along with the associated metadata, which will include links to the code used to generate it (§3.9).

In the case of image data, users will be able to access either the whole file or make a cutout. Cutouts can be made from any slices of data, such as spatial and/or frequency axes, and also time (if applicable), polarisation product, etc.

Among the mechanisms that will be available for data discovery are

- Web-based search engine: The archive search will allow users to search both the ODPs and publicly available ADPs as well as compatible non-SKA archives. Users will be able to create simple searches using web forms, enter more advanced ADQL-type queries and use TAP for database cross-matching. Search results will be presented to users in tabular form (which can also be downloaded/exported to a Comma-Separated Value (CSV) file) with the ability to access the associated data product.

- Quick look images: The user can choose to view a quick look image that is identified by their search parameters.
- Drill-down browsing: The user will be presented with a list of data collections, and can explore these collections by clicking through to a lower level.
- Graphical selection: The user will be shown a cutout of a region of sky, and can select data by clicking on regions of this cutout. This process for data discovery is similar to that used by the Virtual Observatory.

In order to facilitate multi-wavelength investigations, the SKA archive should be linked to other standard astronomy archives via Virtual Observatory (VO) protocols to allow users to discover and conduct combined analysis of available data products.

3.3 Notebook interface

The notebook interface enables the user to do science through their own web browser by creating and running notebooks, defined broadly as structured interactive environments combining executable code, text and visualisation; at the time of writing the predominant notebook type in the astronomy community is the Jupyter notebook, but the platform will need to be capable of evolving to meet new community standards.

Notebooks will run ‘next to the data’ [RD3] within the SRC science analysis platform and will offer user environments with preinstalled packages that provide the functionality required by a standard scientific user, including the generation of new data products from ODPs and ADPs as well as the ability to customise the environment by installing new packages. The results of data queries or discovery (§3.2) will be available within the notebook environment, and notebooks will be persistent for each user. Thus, the user experience will be similar to that of running a notebook on their own local compute resources, but they will have transparent access to the much larger compute resources provided by the SRC.

Sharing of notebooks will be a convenient way for users to collaborate on analysis tasks in a reproducible way and this will be linked to a software repository (§3.9) to allow for well-documented and clear software sharing. When notebooks are shared beyond the project team, measures will need to be taken to ensure that proprietary data access is respected.

3.4 Command-line applications

The platform will provide, through its web interface, the possibility for researchers to launch, provision, manage, build, and share customised environments that include complete software dependencies for running complex applications [RD4]. Technologies for providing these at present include VMs and containers in systems like Docker or Singularity, which will run on top of the computing resources provided to the SRC node (§2.1). Researchers will be able to run code in a specific environment, extend it and resubmit it as a new environment and also create snapshots of a running instance of an environment for reproducibility. Software environments used as part of a scientific analysis will be stored in the software repository (§3.9) and will be linked to the metadata of datasets that they generate. More generic environments will be accessible through web-based graphical UIs, and the option of opening a remote terminal connection (e.g. Secure Shell (SSH)) will be provided.

3.5 Graphical applications

Many astronomers will be familiar with using graphical applications, or applications requiring an X server, for data reduction and analysis (e.g. DS9, HEASoft, CASA, etc), and one way to provide access to these applications is through VMs. We envisage that the platform will support users in launching instances of VMs preconfigured with a specific operating system and software. To support this feature the platform will provide the users with resources to develop VM images, and allow the users to have their own individual space for testing and development. VM images will also be provided preconfigured with software suites, work environments, and community-contributed software, whether designed by the user themselves or taken from the pre-existing software repository.

3.6 Web APIs

It is important that APIs are publicly available to interact with the system. These APIs will serve two purposes:

- They enable remote discovery of, and access to, all data products, including access to a user’s proprietary and private data. This will allow the results to be combined with data from other facilities.
- They enable users to remotely access all low-level services and functionality provided by the SRC. This will allow users to create new tools built on top of the SRCNet.

Machine-accessible web APIs will handle the access to databases, images, and other files which will be exposed to the public internet. These will make remote data access easy. However, we will go beyond this and follow the LSST approach [RD5] of making all low-level platform functionality available through these APIs with suitable authorisation. The interfaces provided by e.g. notebooks (§3.3) will then make use of middleware wrapper libraries around APIs. All APIs will be thoroughly documented to make them accessible.

The archive will allow users to access non-proprietary SKA data through VO interfaces, which will increase the impact/legacy of SKA data and make it accessible to non-radio astronomers or non-SKA users.

3.7 Workflow management

The SRC science analysis platform will facilitate workflow management with a tool that will allow users to specify individual workflow steps and combine them to form a larger workflow which can be stored in the software repository (§3.9) and re-used by others. It will be possible to combine workflow steps sequentially as well as using simple programming constructs (and, or, if), and each workflow step will draw on tools from the software repository, such as code within a notebook, a call to one of the pre-defined APIs, a call to a piece of software that is pre-installed within an existing defined software environment, or a call to another workflow.

After a workflow is defined, there will be options to run the workflow either in real-time or as a background process that can be scheduled to maximise efficiency and prioritisation, with the user being notified upon completion/failure.

3.8 User resource management

Users will necessarily have allocations of compute (both CPU and Graphical-Processing Unit (GPU)) and storage resources on the computing infrastructure underpinning the SRC nodes, as discussed in §2.1 and §2.3. These allocations will be set by SRCNet global policies, so the resources allocated and used are visible in a consistent way across SRC nodes. Users may be granted allocations across multiple SRC nodes, and although the distribution of those allocations shouldn't matter to them, it would be of interest to some users to see this information.

Resource allocations will be assigned to user groups, which will typically be at project level, but smaller allocations of storage will likely be granted to individual users in the form of single-user groups. In addition, there may potentially be institutional resource allocations, but this type of allocation will likely be a matter for individual SRC nodes. In order to manage resource usage within their project, principle investigators will be able to create sub groups of users and divide their resource allocations amongst these groups.

Resource utilisation information will make the user's current consumption of the different types of resources clear to them, and allow them to take steps to reduce their usage if necessary. For example, the user will be able to see currently-staged ODPs and take the decision to remove some of them to free up capacity for another task. They will be able to see and if necessary terminate background operations that they have initiated, such as staging ODPs or processing them into new data products. If they have VMs running, they will be able to see their overall resource allocation, as well as the CPU and GPU consumption on each VM.

3.9 Software repository

Code written for scientific analysis and the 'software environment' used to generate it (§3.4 and §3.5) must be preserved. Therefore, all such code and environment descriptions (which may internally be VM or container build instructions) must be held in a software repository which is integrated with the platform and transparently keeps track of user changes. The SKA archive will reference the software repository so that ADPs and the code and environments used to generate them are linked. The repository will be searchable and queryable in the same way as the SKA archive, and will be shared across SRC nodes.

4 Architectural outline

The SRC science analysis platform component will be the primary interaction point between the SRC system and its users, and will therefore have connections with or dependencies on all other components being developed in the SRC context. An overview of the relations between the various components of the SRCNet is visualised in Fig. 2. In the figure, the boxes delineated by dashed lines represent the different components under current development by the SRCNet team (as of start 2023; teams and responsibilities are expected to evolve), and the arrows represent the flow of requests.

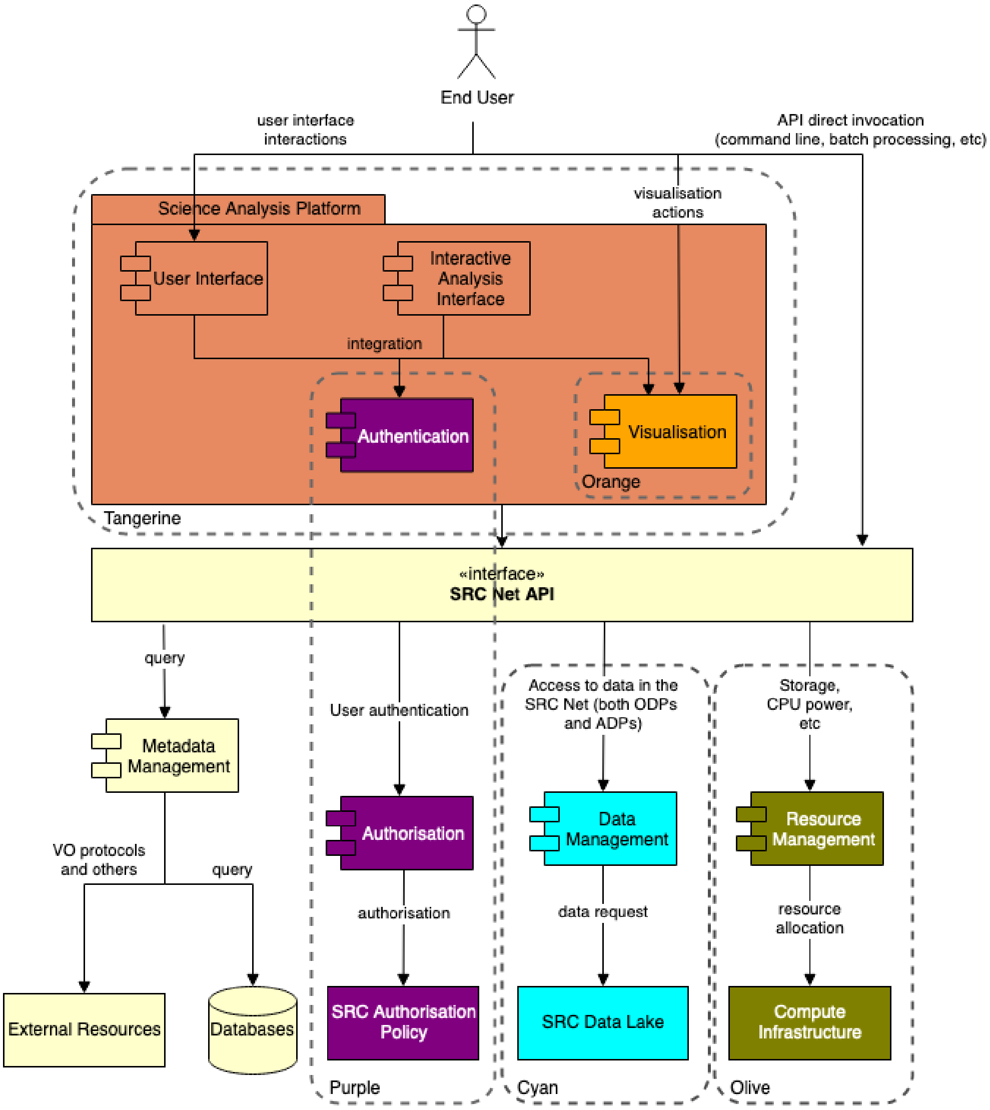

Figure 2: Dependencies between the SRC science analysis platform and other components of the SRCNet.

At the top of the figure, the end user is shown. As discussed in §3.1 and §3.6, users will be able to access the system both through a graphical UI and a user-facing API.

The functionality that is offered within the UI is shown inside the tangerine-coloured block. The UI will offer access to interfaces to support interactive analysis (§3.3) as well as the SRCNet visualisation environment (represented in orange). Also, the UI needs to implement methods to authenticate and authorise against the SRCNet AAI system (which is represented by the blocks in purple), and will need to offer access to the SRC functionality that is handled by the API.

The API will be acting as the broker between user requests on one side and backend services on the other. The first set of blocks (in purple) represents the AAI infrastructure, which will take care of authorisation and authentication throughout the platform, as well as the underlying services. The second set of boxes (coloured in cyan) corresponds to the data management subsystem that will be responsible for making sure the data in the SRCNet can be accessed from the processing environment (both interactive and batch). The third set of boxes (in olive) is part of the component that manages available compute resources and makes those accessible to the processing environments. To the far left of the figure, the three yellow boxes are parts of the system that make it possible for users to query data, both within and outside the scope of the SKA archive, based on metadata.

5 Summary

The volume of the data collected by SRCs will be in exabyte scale. Hence, the science analysis platform used by the SRCs needs to be developed based on advanced technologies for large data handling with the support and active involvement of the astronomy community.

The computing services of the SRCs will include regional, national or supra-national compute infrastructures. SRCs will need to provide an archival data-product storage, which will allow users to query the archive for ODPs and ADPs, and retrieve them for further processing, and store approved generated data products as new ADPs. Users will be provided with a personal POSIX-like file system, to upload and download files and to store intermediate data products for further processing and visualisation as well as to hold code and additional uploaded data for analysis. Centrally managed databases will be created and made available for the released data products, detected sources etc. Users will also be able to generate their own databases through similar TAP, ADQL-like queries, and can store them. The SRCNet will provide AAI services for authentication and authorisation. However, the platform will not offer users access to services or data to which they do not have rights.

The main UI for the platform will be provided by a web page which will allow access to the functionality of the SRC node in a set of panels. Users will be able to sign on to the portal on the front page to access the full UI. Many elements of the UI of the portal will be configurable by the user. Users will be provided with a web-based search engine to query data using an underlying API that will process the data and provide the results. The archive search will allow users to search both the ODPs and publicly available ADPs as well as compatible non-SKA archives. Users will be provided a notebook interface which will allow them to do science in their own web browser by creating and running the notebooks. Sharing the notebooks will be a convenient way for users to collaborate with each other and this will be linked to a software repository. The platform will allow users to define workflows that can run as a background process that will be scheduled according to efficiency and prioritisation considerations. The platform will also provide, through its web interface, the possibility for researchers to launch, provision, manage, build, and share customised environments that include complete software dependencies. These will be provided in the form of VMs and containers in systems like Docker or Singularity, which will run on top of the computing resources provided to the SRC nodes. Users can use their resource allocation either to make a high-level data discovery/processing request or to carry out further operations on the data products that are the results of the previous step. In order to collaborate with others, users must be able to set up, join and leave groups of other users dynamically. All the code written for scientific analysis and the 'software environment' used must be preserved essentially as part of the metadata.

6 Acknowledgements

This publication was partially supported by FuSE: Fundamental Sciences e-Infrastructure, Dutch Research Council (NWO) Roadmap Project 184.035.004.

A References

A.1 Applicable Documents

The following documents are applicable to the extent stated herein. In the event of conflict between the contents of the applicable documents and this document, **the applicable documents** shall take precedence.

[AD1] Shari Breen, Rosie Bolton, and Antonio Chrysostomou. SKAO Science Data Products: A Summary. Technical Report SKA-TEL-SKO-0001818, SKAO, 2021.

[AD2] A. Chrysostomou and the SRCCG. SKA Regional Centres: Background and Framework. Technical Report SKA-TEL-SKO-0000706, SKAO, 2017.

[AD3] Peter Quinn et al., SKA Regional Centres: A White Paper. https://confluence.skatelescope.org/display/SRCSC/SRCSC+ White+Paper+V1.0, May 2020.

A.2 Reference Documents

The following documents are referenced in this document. In the event of conflict between the contents of the referenced documents and this document, **this document** shall take precedence.

[RD1] Mark D. Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, March 2016.

[RD2] J. Salgado et al., The esa gaia archive: Data release 1.´ *Astronomy and Computing*, 21:22–26, 2017.

[RD3] Rubin Science Platform Notebook Aspect documentation. https://nb.lsst. io/.

[RD4] Nirav Merchant, Eric Lyons, Stephen Goff, Matthew Vaughn, Doreen Ware, David Micklos, and Parker Antin. The iplant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS biology*, 14(1):e1002342, 2016.

[RD5] M. Juric, D. Ciardi, G.P. Dubois-Felsmann, and L.P. Guy. LSST Science Platform´ Vision Document. https://lse-319.lsst.io/, 2019.

LIST OF ABBREVIATIONS

AD	Applicable Document
AAI	Authentication and Authorisation Infrastructure
ADP	Advanced Data Product
ADQL	Astronomical Data Query Language
API	Application Programming Interface
CSV	Comma-Separated Value
DOI	Digital Object Identifier
FAIR	Findability, Accessibility, Interoperability, and Reusability
GPU	Graphical-Processing Unit
ODP	Observatory Data Product
OLDP	Observation-Level Data Product
PLDP	Project-Level Data Product
POSIX	Portable Operating System Interface
RD	Reference Document
SDH&P	Science Data Processing & Handling
SKA	Square Kilometre Array
SKAO	Square Kilometre Array Observatory
SQL	Structured Query Language
SRC	SKA Regional Centre
SRCNet	SKA Regional Centre Network
SRCSC	SKA Regional Centre Steering Committee
SSH	Secure Shell
TAP	Table Access Protocol
UI	User Interface
VM	Virtual Machine
VO	Virtual Observatory

DOCUMENT HISTORY

Revision	Date Of Issue	Engineering Change Number	Comments
A	2023-01-13	N/A	Initial Release for comment
B	2023-06-23	N/A	Update to address feedback from external science experts, and WG6
1	2023-08-25	N/A	Released version following science-expert feedback

DOCUMENT SOFTWARE

	Package	Version	Filename
Word processor	MS Word	Office 365	SKAO-TEL-0000000-01B_GenDocTemp_Unclassified_EmptyTemplate.docx
Block diagrams			
Other			

ORGANISATION DETAILS

Name	SKA Observatory
Registered Address	Jodrell Bank Lower Withington Macclesfield Cheshire, SK11 9FT, UK
Fax	+44 (0)161 306 9600
Website	www.skao.int