Data discovery and (interactive) analysis work in SRCNet and at ASTRON

Yan Grange

ASTRON
Netherlands Institute for Radio Astronomy

What is a science platform?

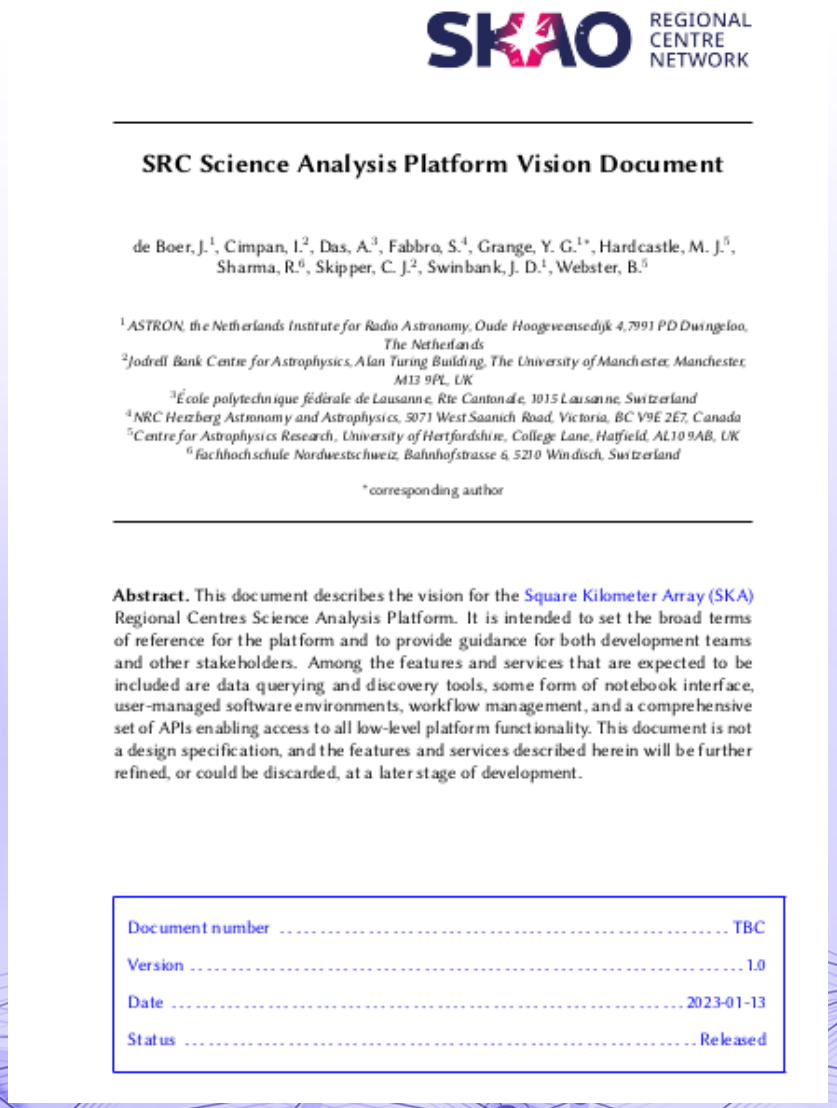
Team Tangerine (SRCNet) wrote a vision document on this!

Attributes

- consistency
- scalability
- reproducibility
- usability
- Reliability

Guiding principles:

- Highly collaborative
- end-to-end
- accessible (with focus on inclusion



SKAO REGIONAL CENTRE NETWORK

SRC Science Analysis Platform Vision Document

de Boer, J.[1], Cimpan, I.[2], Das, A.[3], Fabbro, S.[4], Grange, Y. G.[1*], Hardcastle, M. J.[5], Sharma, R.[6], Skipper, C. J.[2], Swinbank, J. D.[1], Webster, B.[5]

[1] ASTRON, the Netherlands Institute for Radio Astronomy, Oude Hoogeveensedijk 4,7991 PD Dwingeloo, The Netherlands
[2] Jodrell Bank Centre for Astrophysics, Alan Turing Building, The University of Manchester, Manchester, M13 9PL, UK
[3] École polytechnique fédérale de Lausanne, Rte Cantonale, 1015 Lausanne, Switzerland
[4] NRC Herzberg Astronomy and Astrophysics, 5071 West Saanich Road, Victoria, BC V9E 2E7, Canada
[5] Centre for Astrophysics Research, University of Hertfordshire, College Lane, Hatfield, AL10 9AB, UK
[6] Fachhochschule Nordwestschweiz, Bahnhofstrasse 6, 5210 Windisch, Switzerland

*corresponding author

Abstract. This document describes the vision for the Square Kilometer Array (SKA) Regional Centres Science Analysis Platform. It is intended to set the broad terms of reference for the platform and to provide guidance for both development teams and other stakeholders. Among the features and services that are expected to be included are data querying and discovery tools, some form of notebook interface, user-managed software environments, workflow management, and a comprehensive set of APIs enabling access to all low-level platform functionality. This document is not a design specification, and the features and services described herein will be further refined, or could be discarded, at a later stage of development.

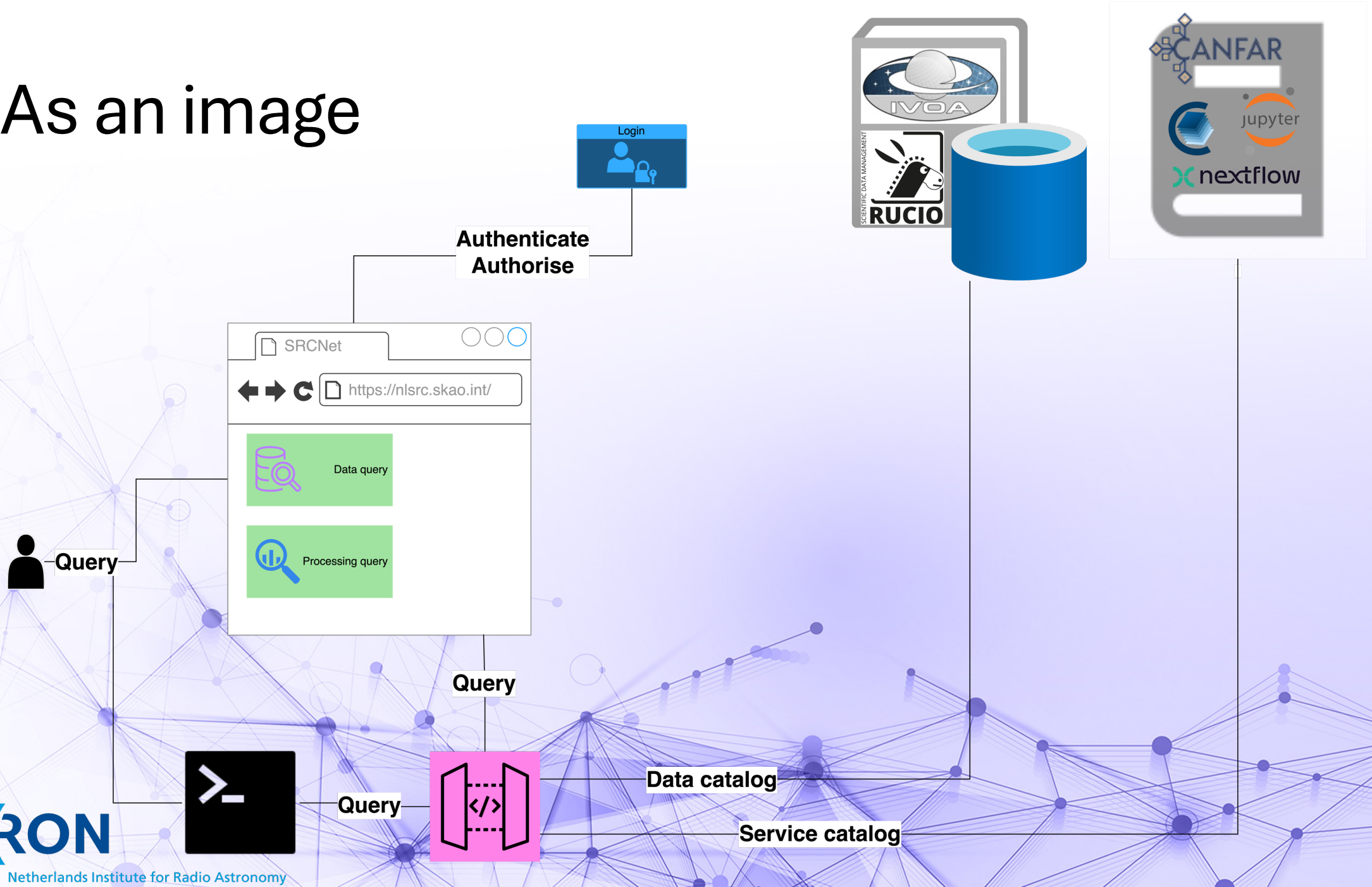
Document number ..... TBC
Version 1.0
Date 2023-01-13
Status ..... Released

ASTRON
Netherlands Institute for Radio Astronomy

In concreto

- The science platform is aimed at making the data and processing available to users with different levels of expertise
  - In the SKA/LOFAR case teams are an important concept.
- Making large data sets accessible
  - Using APIs for programmatic access
  - And a user interface for interactive interaction
- With science-ready data products, the focus is mainly at analysis work rather than large-scale pipeline processing
- **NB:** Within the scope of this talk “science platform” refers to the whole system while “science gateway” refers to the interface to the user (i.e. frontend + API layer)

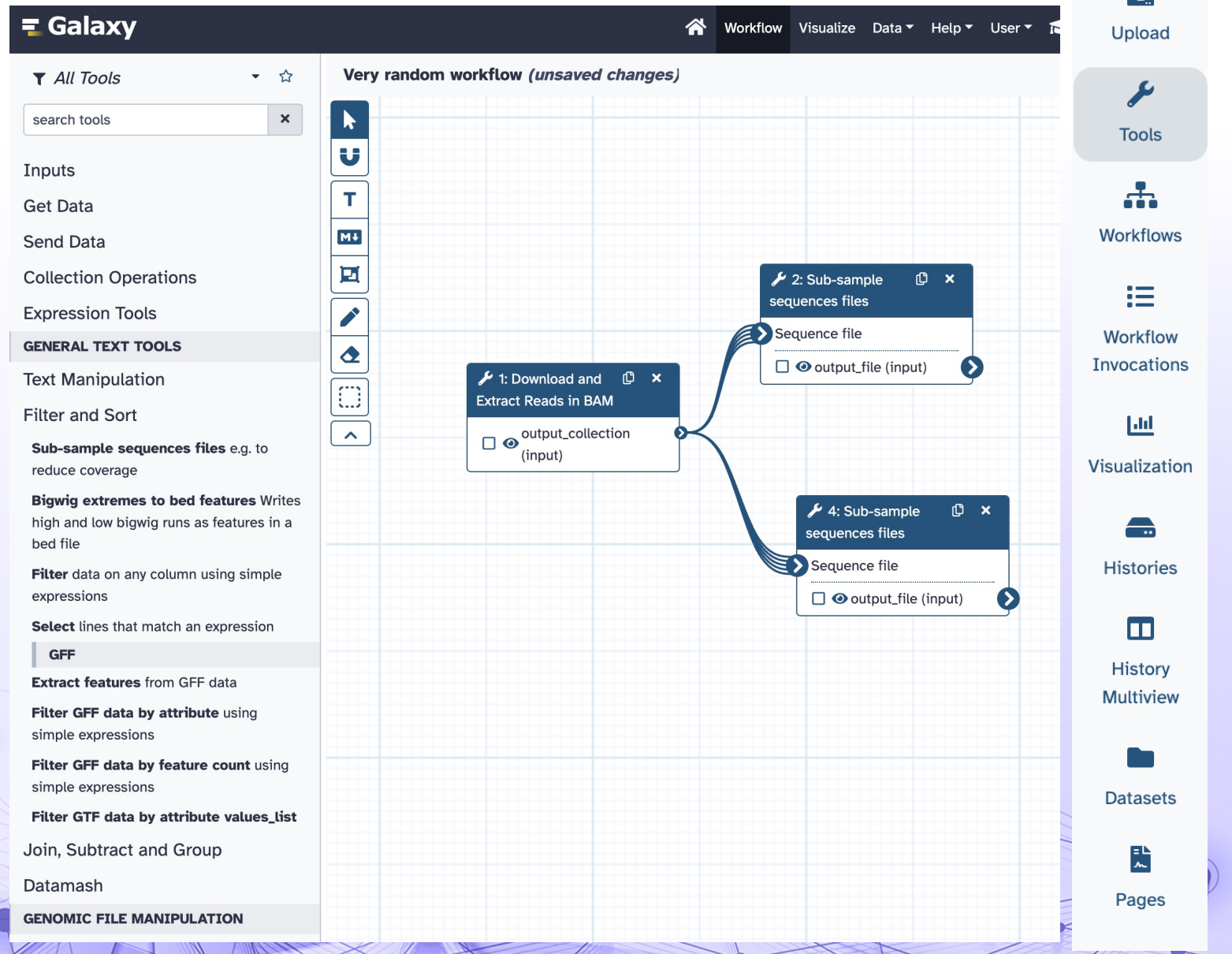
As an image



ASTRON
Netherlands Institute for Radio Astronomy

Galaxy

- Science gateways are being used in other fields. For instance Galaxy, which is focused on life science.



ASTRON
Netherlands Institute for Radio Astronomy

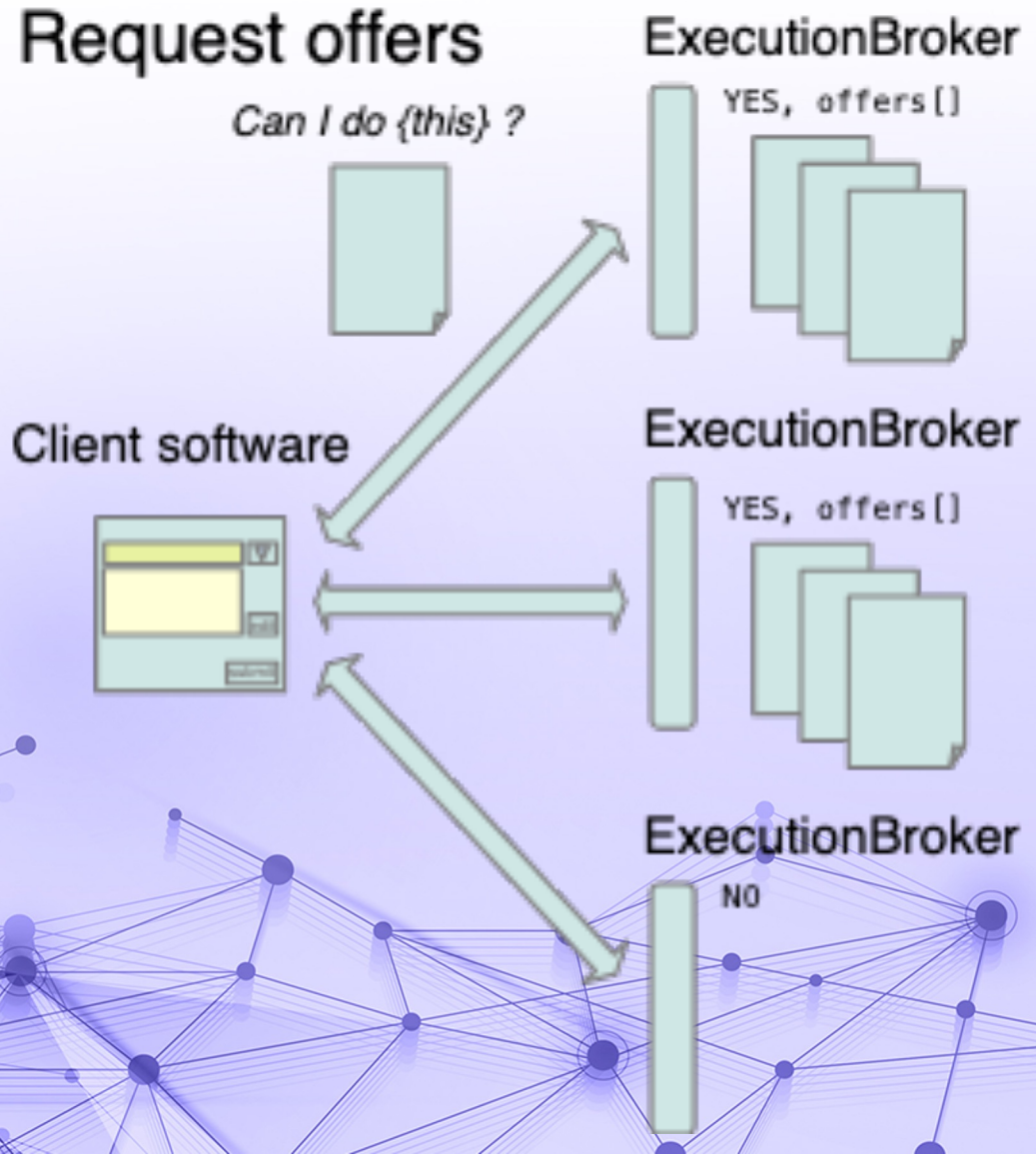
VO standards (small intermezzo)

The Virtual Observatory (VO) standards have been designed to "astronomers to interrogate multiple data centers in a seamless and transparent way (…) and gives data centers a standard framework for publishing and delivering services using their data." (from ivoa.net)

- Apart from standards for defining and sharing tables, and observational (meta) data some relevant standards that interact with the lower-level parts of an archive are:
  - Server-side Operations for Data Access (SODA): a low-level data access capability or server side data processing that can act upon the data files, performing various kinds of operations: filtering/subsection, transformations, pixel operations, and applying functions to the data.
  - The Universal Worker Service pattern (UWS) defines how to manage asynchronous execution of jobs on a service.

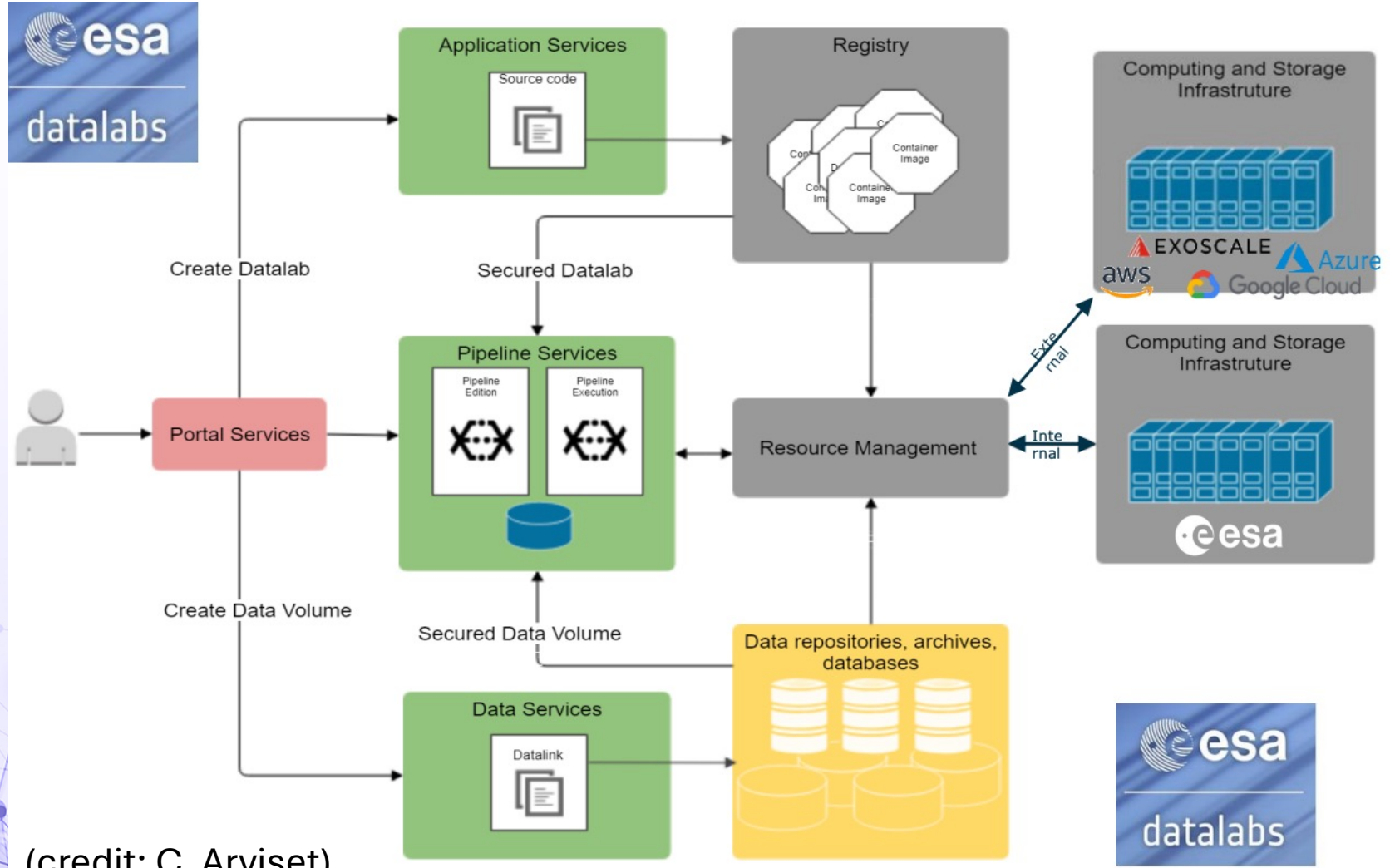
The Execution Broker standard

- Currently in DRAFT (i.e. not endorsed as an official standard yet), and being developed in the SRCNet as a way too provision resources.
- The goal is to abstract away as many details away from the user
- This feels like quite a bit of scope…



ESA Datalabs

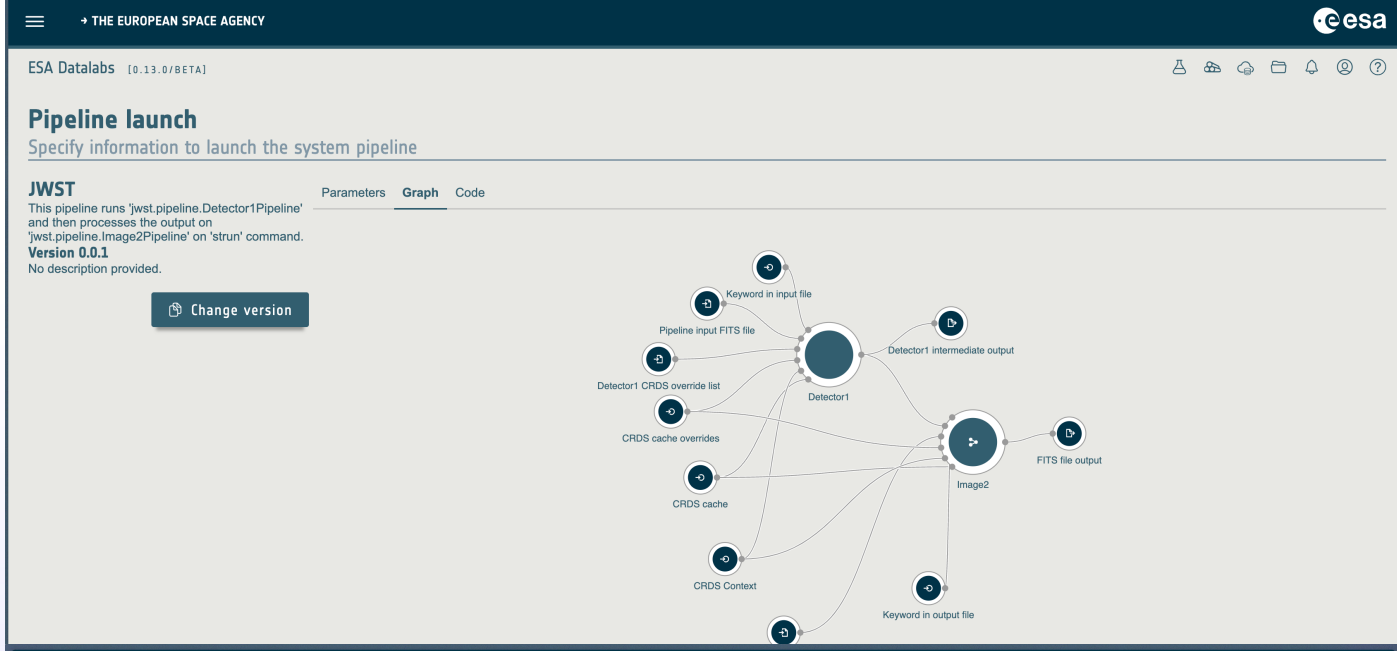
- Platform for access to ESA data

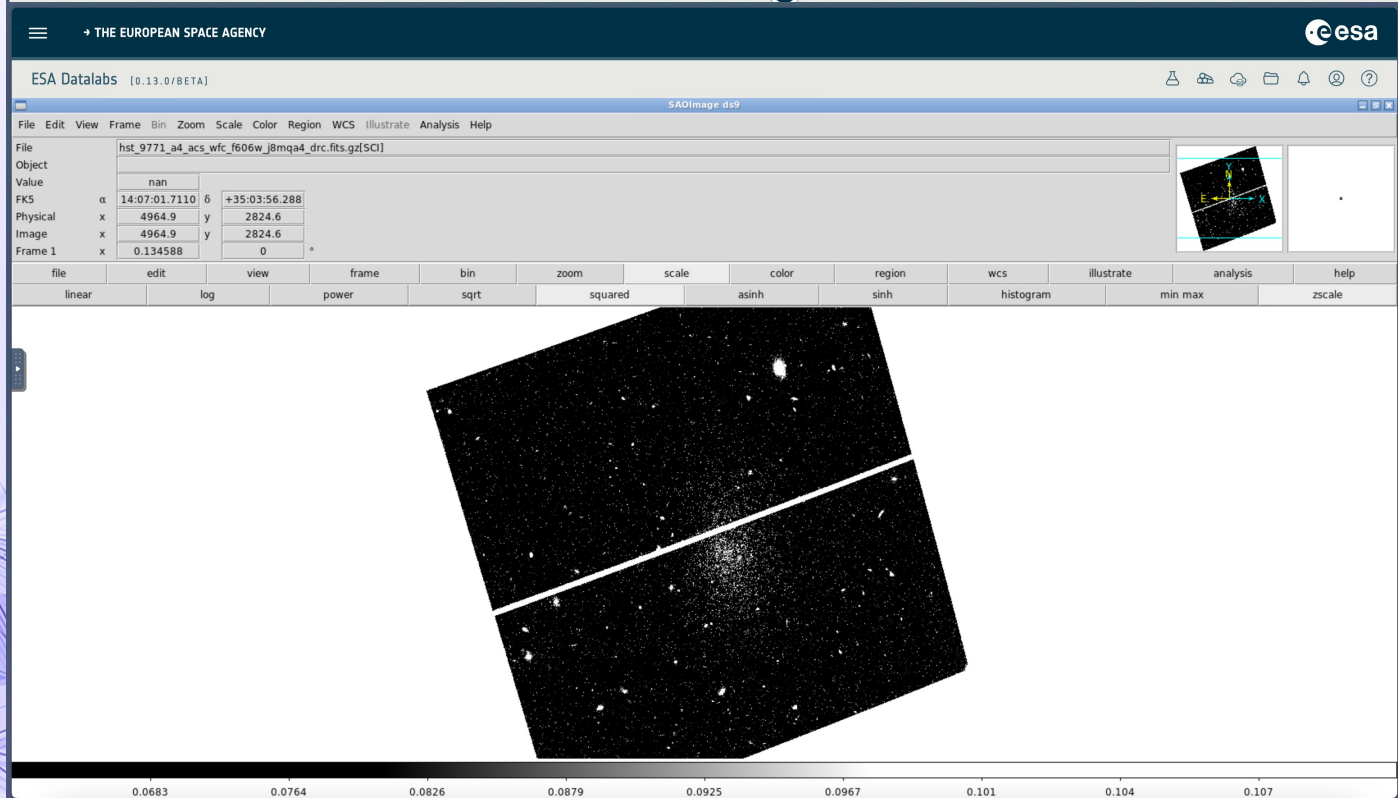


(credit: C. Arviset)

ESA Datalabs

- Notebooks
- VNC
- Pipelines (using CWL)
- Data collections
  - Directly linked to applications as “virtual directory”
- Group/team management

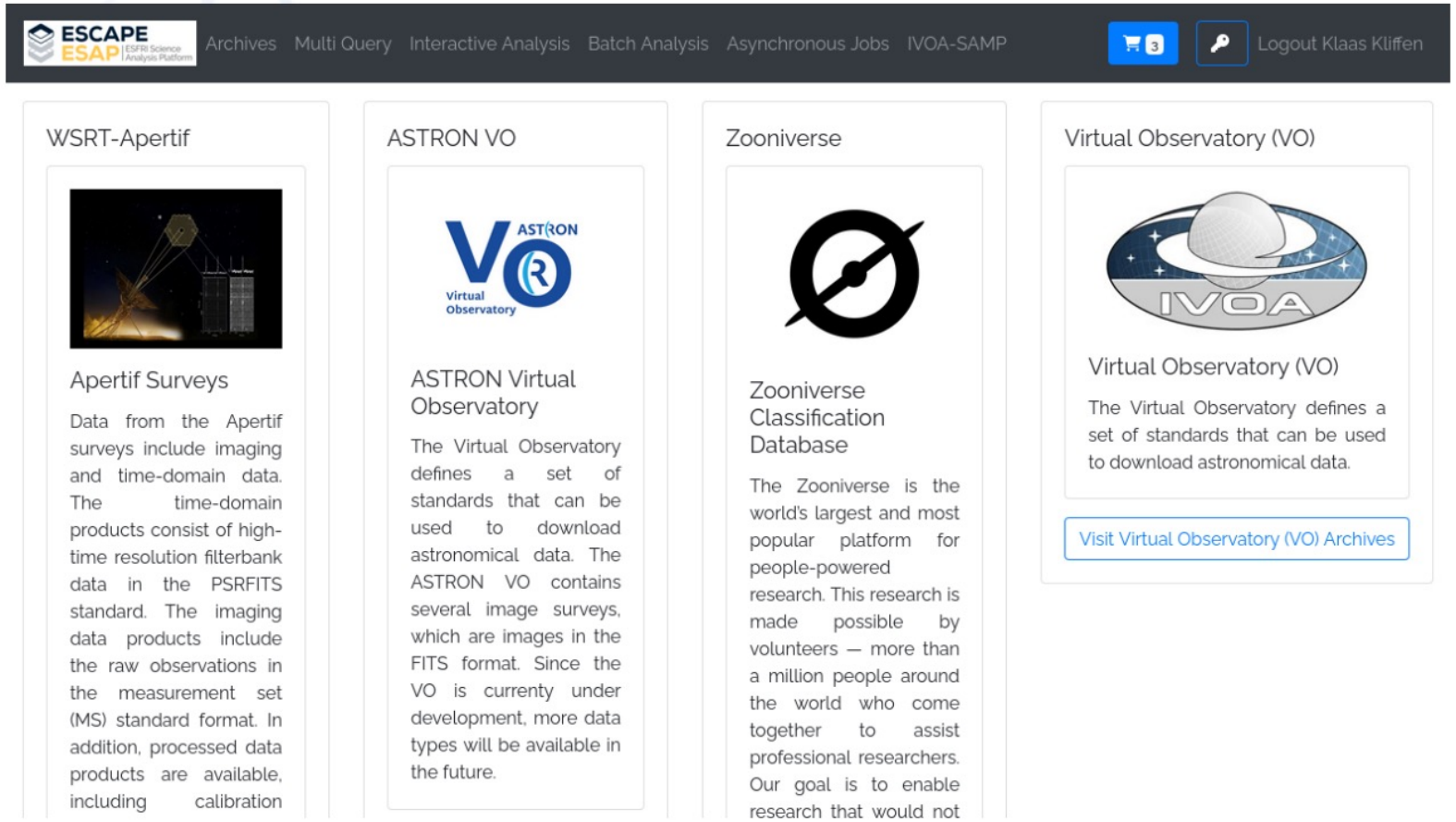




ASTRON
Netherlands Institute for Radio Astronomy

ESAP

- Developed in the ESCAPE project
- The development inspired the work on ADEX and SRCNet, but also other ESCAPE partners (e.g. CTA) are planning to reuse the concept.



(credit: K. Kliffen)

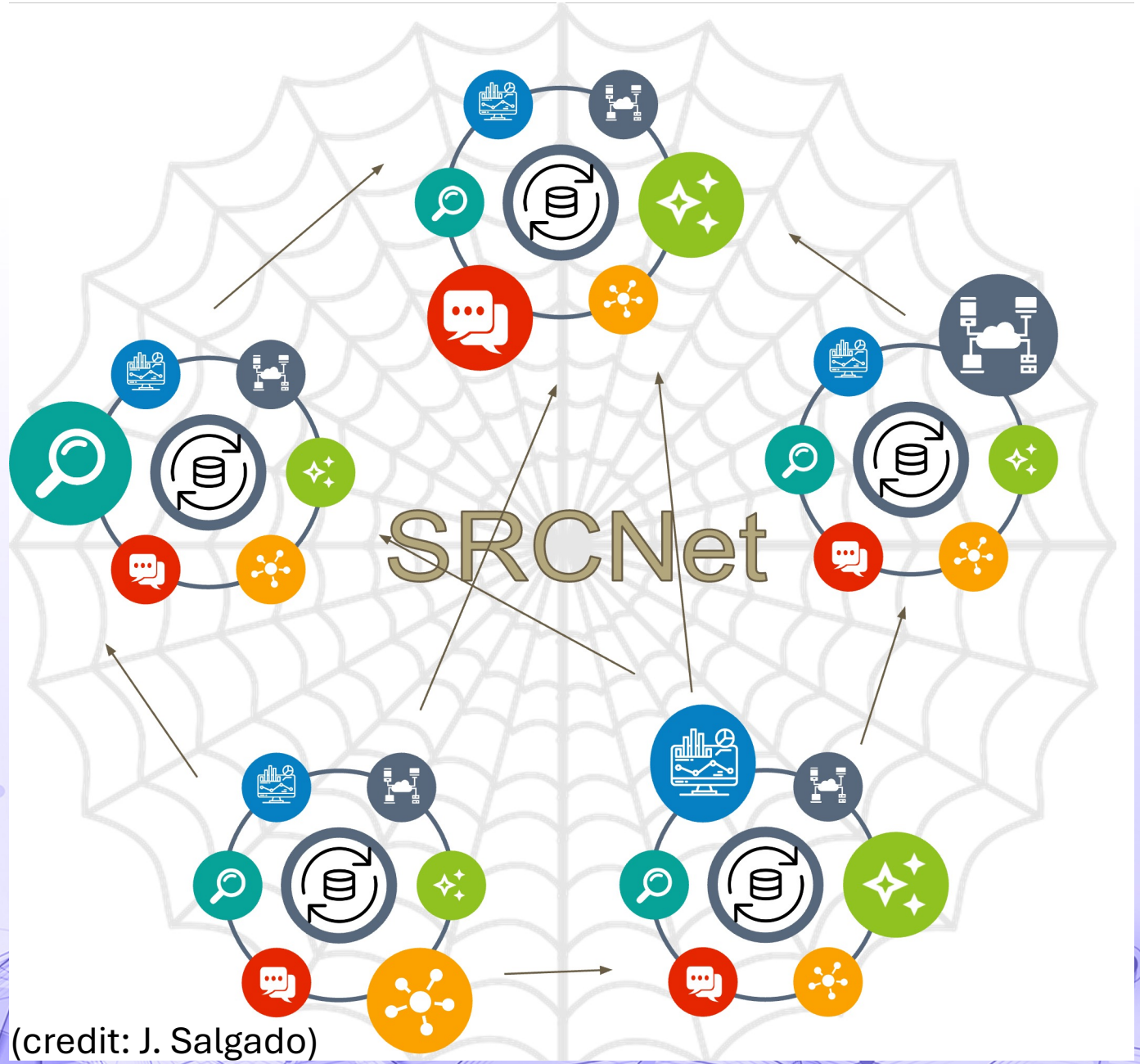
<u>For acronym soup lovers:</u>

"ESCAPE ESAP" is the European Science Cluster of Astronomy and Particle physics European Strategy Forum on Research Infrastructures research infrastructures European Strategy Forum on Research Infrastructures Science Analysis Platform. (thanks, John, for working this out!)

ASTRON
Netherlands Institute for Radio Astronomy

The SRCNet

- The numbers for SKA may be bigger, but the image looks very much like the LOFAR setup.

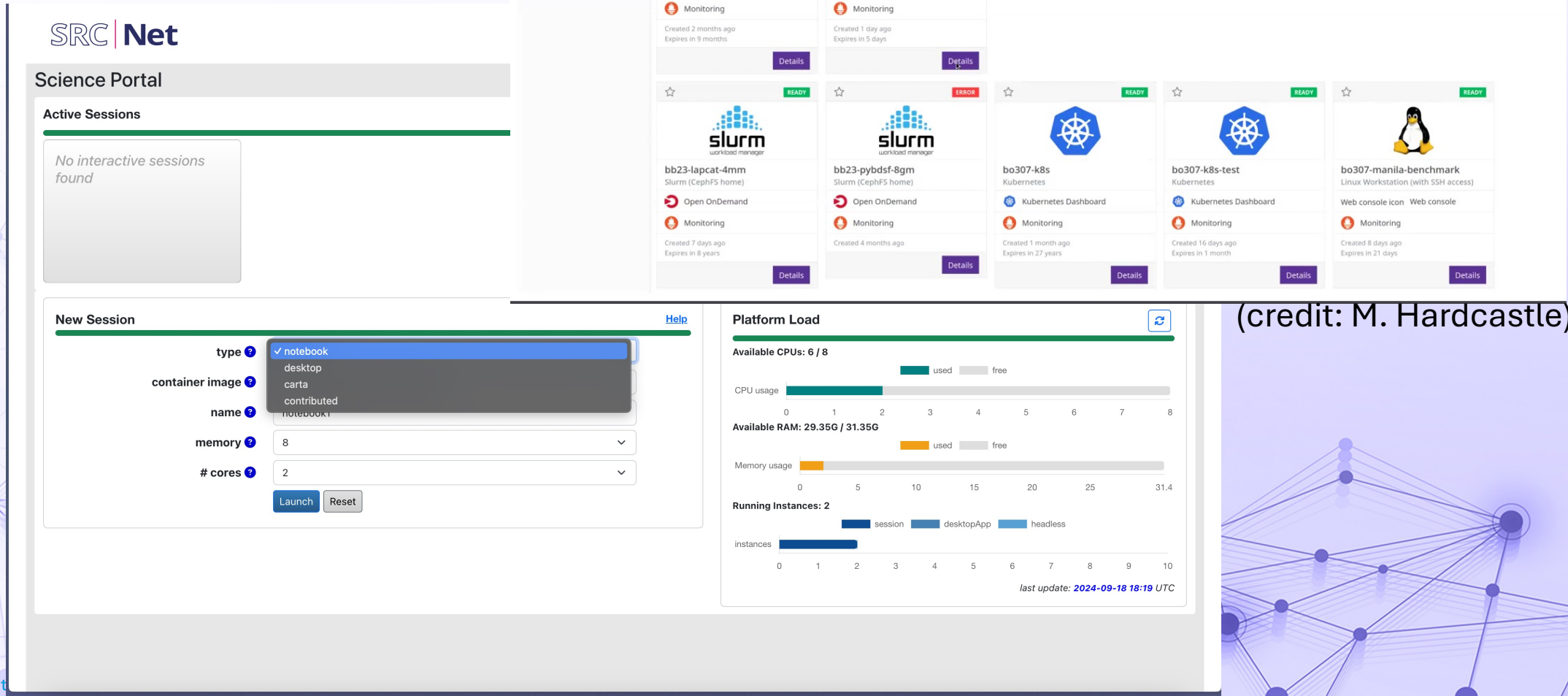


(credit: J. Salgado)

SRCNet components

- AAI through IAM
- Data management through Rucio
- Visualisation (CARTA, VISIVO, etc.)
- Several internally developed APIs (e.g. a service catalogue)
- CANFAR/Azimuth (see next slide) for provisioning of resources
  - Or providing direct access to e.g. a Jupyter lab instance
- Software repository for processing tools
- …

CANFAR, Azimuth



(credit: M. Hardcastle)

ASTRON
Netherlands Insti

ADEX

- https://sdc.astron.nl/adex-next/ (better to live-demo than to talk too much about it)

SRCNet science gateway

- https://gateway.srcdev.skao.int/ (better to live-demo than to talk too much about it)

What does this mean for the people here?

- Many of the systems, tools, software used expect the data to be accessible like it is on disk, but generally it is not.
  - Remote/network mounts can get (very!) laggy.
- Data products themselves may be very large. In the VO world, this is generally solved by offering a cutout service (only deliver the pixels the user wants rather than downloading a full image).
  - If the data is remotely stored, are there smart ways of doing this?
- Using a platform for interfacing offers flexibility of backend implementations. At what level is it needed to be homogeneous?
- How much information can/should we hide from the user?
  - Staging, data locality have effects on this.
- How is this mapped to software access, and users writing their own code?
  - In SRCNet we are looking into EESSI for this
- … and probably much more I did not think of